Mitigating Racial Biases for Machine Learning Based Skin Cancer Detection

Julian Abhari
The University of Tulsa
Tulsa, OK, USA
jsa3484@utulsa.edu

Ashwin Ashok
Georgia State University
Atlanta, USA
aashok@gsu.edu

ABSTRACT

Machine learning (ML) based skin cancer detection tools are an example of a transformative medical technology that could potentially democratize early detection for skin cancer cases for everyone. However, due to the dependency of datasets for training, ML based skin cancer detection always suffers from a systemic racial bias. Racial communities and ethnicity not well represented within the training datasets will not be able to use these tools, leading to health disparities being amplified. Based on empirical observations we posit that skin cancer training data is biased as it's dataset represents mostly communities of lighter skin tones, despite skin cancer being far more lethal for people of color. In this paper we use domain adaptation techniques by employing CycleGANs to mitigate racial biases existing within state of the art machine learning based skin cancer detection tools by adapting minority images to appear as the majority. Using our domain adaptation techniques to augment our minority datasets, we are able to improve the accuracy, precision, recall, and F1 score of typical image classification machine learning models for skin cancer classification from the biased 50% accuracy rate to a 79% accuracy rate when testing on minority skin tone images. We evaluate and demonstrate a proof-of-concept smartphone application.

CCS CONCEPTS

• **Human-centered computing → Ubiquitous and mobile computing systems and tools**.

KEYWORDS

skin cancer, machine learning, domain adaptation, racial, skin tones, artificial intelligence, bias, computer vision, app, generative adversarial networks

ACM Reference Format:
Julian Abhari and Ashwin Ashok. 2023. Mitigating Racial Biases for Machine Learning Based Skin Cancer Detection. In *The Twenty-fourth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing (MobiHoc '23), October 23–26, 2023, Washington, DC, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3565287.3617639

MobiHoc '23, October 23–26, 2023, Washington, DC, USA

ACM ISBN 978-1-4503-9926-5/23/10…$15.00
https://doi.org/10.1145/3565287.3617639

1 INTRODUCTION

Every hour two people die in the U.S. from skin cancer [2]. To make matters worse, skin cancer is far more lethal for people of color, where it was found that statistically doctors and physicians are racially biased, misdiagnosing real cases of skin cancer [6]. There is hope however, as 99% of skin cancer cases are curable if detected early [4]. Thankfully, this is the perfect problem for the revolutionary technologies from artificially intelligent (AI) assisted solutions from the field of machine learning (ML) to potentially help eliminate racial and ethnic health disparities. In the case of skin cancer, AI solutions can be particularly effective to bring early skin cancer detection tools to everyone. An op-ed article asserts further that the application of AI to improve early detection of diseases is an example of a rare but transformative advance in medicine, while stating that the limits of this technology lies within the data [1]. Furthermore, this paper goes on to show how a community's health could be severely undermined if there is not enough data representing that community.

Unfortunately, the lack of data representing diverse racial and ethnic communities is especially apparent for skin cancer datasets. A research paper found that the race and ethnicity of subjects in dermatologic ML scientific papers underrepresented darker skin types [10]. A separate research paper from machine learning experts and medical researchers present a framework to mitigate racial and ethnic biases [3]. They show that if the lack of data is a bottleneck, as is the case with diverse images for skin cancer, then we must create a model that checks and corrects differences across demographic groups, and adjust the performance of the model so that the majority group does not dominate the learning.

AI skin cancer detection tools have the power to democratize early detection of skin cancer for everyone and mitigate racial and ethnic health inequalities around the world, but as the technology stands right now, it would only worsen the health disparities due to a lack of diverse representations in the data. The question is, what innovative solution should be used to mitigate racial biases within demographics? The traditional data-hungry computer vision approach will not work in a timely manner to mitigate racial biases, as getting a strong representation of data from all racial and ethnic groups does not exist right now. Instead we need to use the innovative domain adaptation technique of style matching models to match the minority data to the majority – thereby allowing a much smaller dataset to generalize.

2 RELATED WORK

Given the difficulty of the alignment problem and mitigating racial biases, one of the first issues to help mitigate this bias is by getting diverse skin tones to be represented. A research paper showed how to improve skin color diversity by generating synthetic darker skin tone skin cancer images based off of lighter skin tone images using generative AI approaches from Deep Blending and Style Transfer [14]. Their resulting synthetic darker skin tone data was evaluated by medical professionals and used to train a convolutional neural network (CNN), resulting in 76% accuracy when tested on synthetic data. A drawback with this approach is that they simply were not able to test it on real images of darker skin tone data.

Another research paper went on to create a dataset of diverse dermatology images with diverse skin tones [8]. They used it to test current approaches, which resulted in accuracies ranging from 0.56 to 0.67. When they used their diverse dermatology dataset to fine-tune the current approaches they resulted in 0.77 and 0.78 accuracies for light and dark skin tones respectively. However, a large drawback with their solution was a great imbalance in classes based on their dataset, resulting in far more benign moles to malignant moles. In fact, when we reproduced their results using an Inception V3 CNN architecture, we were also able to achieve similar accuracies of 0.77 for overall skin tones, however, due to the unbalanced amount of benign moles to malignant moles, we found that our results had a very low precision and recall for classifying malignant moles, resulting in a low F1 score of 0.32 for the malignant mole class. After discovering the imbalance in the dataset it became clear that the accuracies reported were highly inflated due to the umbalance in the dataset. Therefore the problem of mitigating racial and ethnic bias was still not solved. Using an evaluated and highlighted overview of all publicly available skin image datasets used for cancer diagnosis from researchers [15], we were able to identify dataset features they showcased like ethnicity data and fitzpatrick skin type data. With their systematic overview, we were able to identify another dataset that contains darker skin tone fitzpatrick types for benign and malignant moles, the PAD-UFES-20 dataset.

Researchers [13] collected clinical images of skin lesions taken with smartphones which contain darker skin tone fitzpatrick types. Another group of researchers [3] construct a four-stage framework to analyze where biases can emerge and they outline that if the lack of data is a bottleneck, another mitigation strategy is to create a model that checks and corrects learned differences across demographic groups so the majority group doesn't dominate the learning.

A field of ML that specializes in this topic is the field of domain adaptation techniques, in which a solution from this field is explored to mitigate the biases present in these datasets. A research paper presents an experiment where they use ML models to stylize the minority data to look like the majority to be used in domain adaptation for breast cancer data with the goal of generalizing data between varying tissue types and scanners [5], a problem not unlike generalizing cancerous moles with varying skin tones. With their stylizing approach for domain adaptation, they achieved much more consistent accuracy than traditional ML methods.

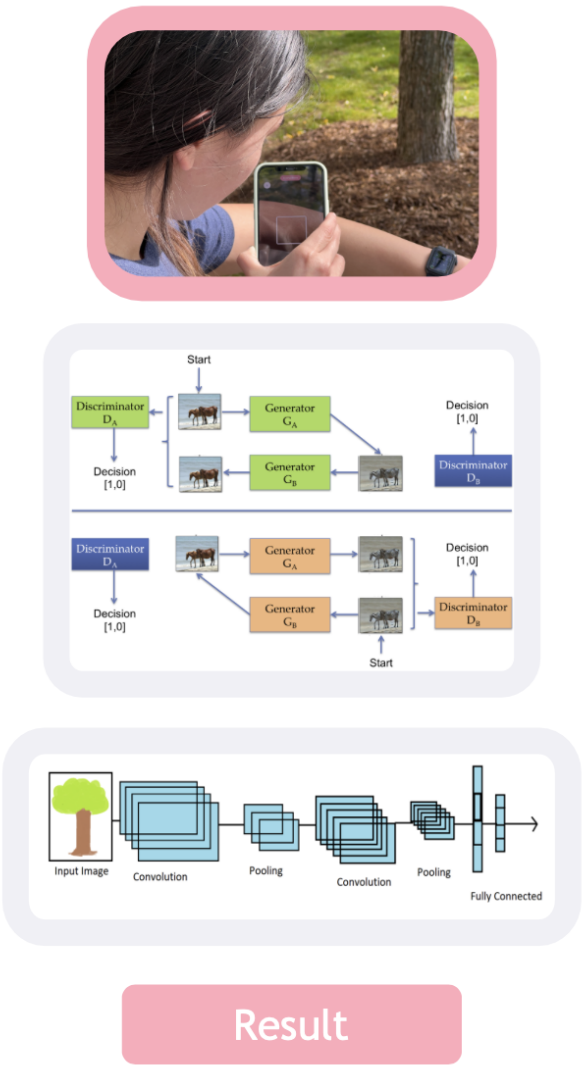

Figure 1: Processing pipeline.

3 METHODOLOGIES

3.1 System Overview

Our datasets play an important role in creating a classification model, as the balance of these datasets in their classes will play a large factor in the resulting model's precision and recall for each classification [7]. Moreover, it's clear from the related works that the datasets are racially biased, containing mostly lighter-skin tone data, though, there does exist darker skin tone data within the 2.15% of datasets that report fitzpatrick skin type [15]. The problem is simply that there is a lot less data from the datasets that report skin tone. More specifically, those researchers found that of the 106,950 skin lesion images that are readily available, only about 2,298 of those images have skin tone reported, this dataset being the PAD-UFES-20 dataset [13]. From this dataset only about 465 of those images represent people of darker minority skin tones not usually present, Fitzpatrick types 3 through 6. We can still use

this dataset to create our domain adaptation method, however. The PAD-UFES-20 dataset will help us to represent darker minority skin tones not present in the majority of the data. With our minority dataset established, our majority dataset will come from the ISIC 2020 dataset also referenced in the research overview on publicly available datasets [15], as mentioned in the related works.

For our domain adaptation method, we're trying to create an effective generator to adapt our minority images to match the majority, thereby mitigating the differences in data and hopefully resulting in a classifier that can generalize for both the majority and minority of data. However, it's important to note that with our majority and minority datasets established, there is still a discrepancy in the amount of samples we have for both datasets. More specifically, from the ISIC 2020 dataset we use 5,714 malignant samples and 5,747 benign samples, and from the PAD-UFES-20 dataset we use 1,089 malignant samples and 1,209 benign samples. There is still a difference in dataset size between these majority and minority datasets.

We then need to figure out how we're splitting our data into training and testing sets for both the domain adaptation training and the image classification training. For our domain adaptation training we used 10,869 images of both benign and malignant majority data samples, and 821 images of both benign and malignant minority data samples. Meanwhile for validating our domain adaptation we used 592 images of both benign and malignant majority data samples, and 92 images of both benign and malignant minority data samples. For the image classification training we used 2,089 benign and 2,089 malignant data samples of both the majority and minority datasets, where 1,089 images came from the adapted minority dataset, and 1,000 images came from the raw majority dataset. Meanwhile for the image classification testing we used 221 benign and 221 malignant data samples of both the majority and minority datasets, where 109 of those data samples came from the minority dataset and 112 of those data samples came from the majority dataset.

Inspired by a research paper [9] explaining the creation of a deep neural network to detect benign or malignant moles from images, we start by creating a typical machine learning based skin cancer detector by using deep convolutional neural networks and data from the International Skin Imaging Collaboration (ISIC) 2020 Vienna challenge. As outlined previously, datasets like the ones found from the ISIC don't represent racially or ethnically diverse communities. We should expect to see racial biases from a typical machine learning based deep convolutional neural network model trained on these racially biased datasets. After training an Inception V3 architecture for a deep convolutional neural network model on the ISIC 2020 Vienna challenge dataset using images of benign and malignant moles, we were able to achieve an 89.7% training accuracy with an 88.8% validation accuracy. **When we tested this model on a testing dataset consisting of 592 images we omitted from the ISIC 2020 dataset we achieved 85% accuracy. However, when tested on only the minority skin tones present in the PAD-UFES-20 dataset, the accuracy dropped to only 50%, showing the clear bias and the need for domain adaptation.**

Inspired by this research paper [5], which explains how they used domain adaptation for minority tissue types and scanners within breast cancer, we use domain adaptation for our minority

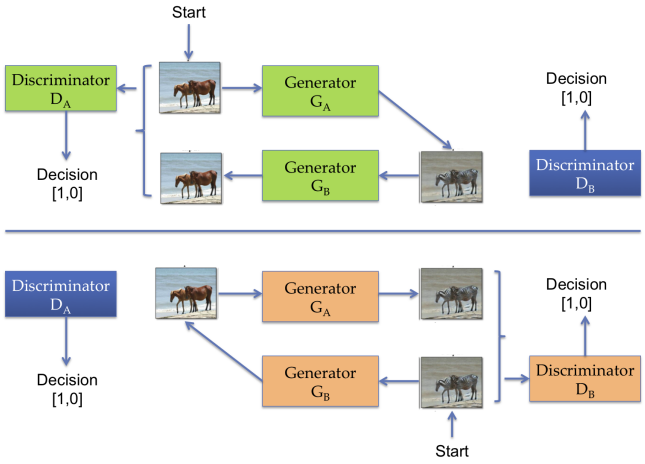

Figure 2: The Cycle-GAN architecture used in our project.

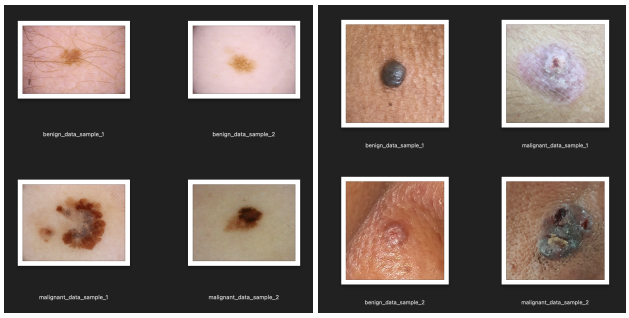

Figure 3: Skin data samples from majority and minority datasets left to right respectively

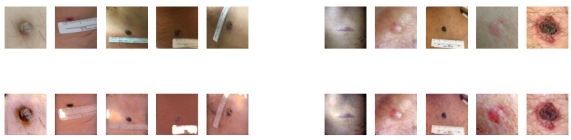

Figure 4: Generated lighter skin tone images with real samples on the top and generated samples on the bottom

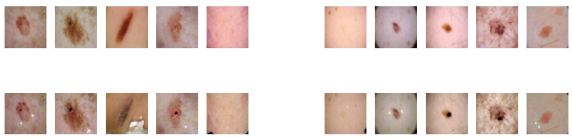

Figure 5: Generated darker skin tone images with real samples on the top and generated samples on the bottom

darker skin tone data present in the PAD-UFES-20 dataset. More specifically, these researchers use a variety of techniques and find that the CycleGAN model to stylize images of the minority dataset to look like the majority yields the best results. With this in mind, we create a CycleGAN to adapt our minority darker skin tone data to our majority lighter skin tone data.

After training our CycleGAN for 10 epochs we use it to augment the PAD-UFES-20 minority dataset to look like the majority ISIC 2020 dataset. With these augmentations we create an Inception V3 architecture for a CNN and train that on the augmented minority data with the majority data. Finally, after creating our classification model trained on the augmented data, we test it on augmented darker skin tones from the PAD-UFES-20 dataset. When used in a mobile application the data-pipeline process starts with a captured photo that is then augmented to look like the majority dataset using our CycleGAN, then that augmented photo is fed into the classifier which was trained and tested on similarly augmented minority and majority data, resulting in a classification. Figure 1 shows the processing pipeline in our system architecture.

3.2 Domain Adaptation

Generative adversarial networks (GAN) are often used in domain adaptation techniques, as they can alter the data from one domain to match that of another. In our case we use CycleGAN from the original research paper [16]. The CycleGAN is a specific type of generative adversarial network that uses multiple loss functions to ensure cycle consistency when adapting data between different domains, and ensures that the content of the data is preserved. More specifically, the CycleGAN architecture (Figure 2) [11] uses two pairs of discriminators and generators, one for domain A and the other for domain B. The two discriminators provide the first two loss functions, ensuring that the generated images could exist in the corresponding domain. Then as an image from domain A is passed to generator B, it will stylize the image to approximate domain B, then that generated image is passed to generator A to stylize it back to domain A, and the difference between the original image and the reconstructed image through the forward cycle process determines the forward cycle consistency loss. The same process can be done in reverse to determine the backward cycle consistency loss. Finally another loss function is determined by having an image from a specific domain be passed into the generator of that specific domain which should approximate it to belong in the domain, however, since the image is initially of that generator's domain, there should be no change when the image is passed to that generator. The evaluation of the original image from the specific domain and the augmented version when passed through that domain's generator determines the identity loss.

What this looks like for our majority and minority skin tone stylizing application is that we have a majority dataset of lighter skin-tone images, we'll call that domain A, and a minority dataset of darker skin-tone images, we'll call that domain B. When given a benign or malignant mole of a darker skin tone, domain B, the generator for domain A should be able to make it look like it was of a lighter skin tone. Likewise, when given a benign or malignant mole of a lighter skin tone, domain A, the generator for domain B should be able to make it look like it was of a darker skin tone.

Our hyperparameters for training this CycleGAN model are as recommended from the original paper [16], in which we use a batch size of 1 and instance normalization. In addition, we also update discriminators using a history of generated images, as opposed to using one image from the generator. This was recommended from the original paper as a way to help prevent model oscillation. More specifically, it's the history of the previous 50 generated images for the discriminator models to evaluate. Due to the architecture of this technique, it requires a lot of computing power to train, which can serve as a limitation for training this model. However, as soon as the model is finished training, the generator of a specific domain can be exported as a file to fit a mobile application. We trained our CycleGAN using the majority and minority skin cancer datasets (Figure 3) as previously mentioned, and trained it for 10 training epochs, while saving an instance of both generator models after each epoch. Ideally, the model should improve at generating images after each training epoch.

The results of training for the minority to majority domain generator for epoch four and six respectively, are shown in Figure 4, where training epoch 4 is shown first, then training epoch 6 is shown second. In these figures the real data samples are shown on top, with the generator's adaptation shown beneath. Quickly the generator learns to adapt the skin tone difference, however, there are artifacts that are created by the model through the process. In epoch four there are clear gaps in the image, shown as empty spots of white or black pixels (Figure 4). By epoch six these artifacts appear to be a lot less apparent, however. Similarly, the results of training for the majority to minority domain generator for epoch nine and ten respectively, are shown in Figure 5. In this case the goal for the generator is to create visibly darker skin tone images. The results are effective even in this backwards cycle where we feed the majority lighter skin tone images to appear as darker skin tone images, despite the large discrepancy in available darker skin tone images to lighter skin tone images.

3.3 Mobile System

A mobile application (Figure 6) designed to be used on iOS devices was also developed to use the model pipeline previously outlined to provide assessments on how similar a mole resembled that of skin cancer. With this capacity to provide an informed and individual recommendation, the app made a modest contribution to solving the problem of equal access to early skin cancer detection. Features were developed to allow a user to track their mole over time, as cancers such as melanoma can develop as quick as 6-weeks [12]. Therefore an archive that allowed them to track the change of their moles was created. Furthermore, an index of common skin cancer types that showed images, facts and statistics, and where on their body to monitor for each skin cancer type was provided to help educate a user.

Implementing the classification pipeline required the use of CoreML libraries to export a tensorflow model, the minority to majority generator from the CycleGAN, as a CoreML model file extension to be used by iOS devices for augmenting user captured photos. Furthermore, exporting the Inception V3 CNN classification model was done in a similar manner using CoreML libraries. With the CoreML files the models were easily implemented by using the

Figure 6: Screenshots from mobile app

iOS device's camera to output an image which would be fed into the generator and that generator's output would be converted into an image to be fed into the classification model.

4 RESULTS

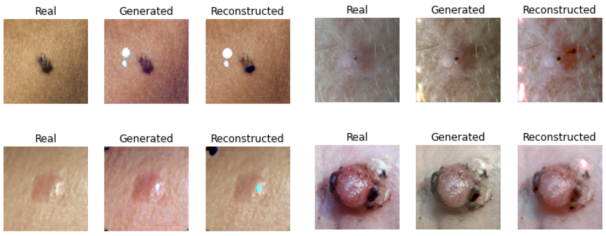

Figure 7: Regenerated and Constructed skin condition images. Minority to majority generated images and the majority to minority, respectively from left to right.

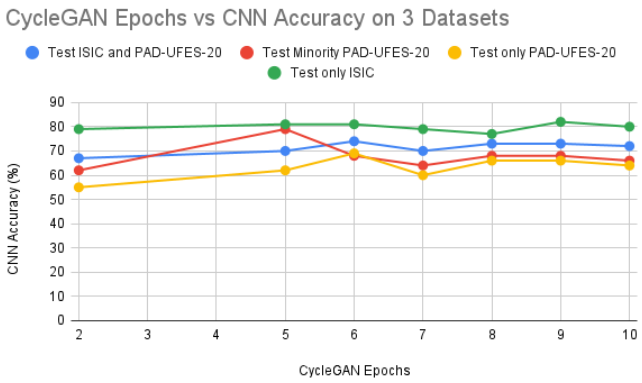

Figure 8: CycleGAN Epochs v/s. CNN model Accuracy

The results of the CycleGAN are described and shown in Figure 7, the minority to majority generated images and the majority to minority respectively.The generator to make the minority data look like the majority is shown with great results, as the skin lesion and its contents barely change, yet the darker skin tone of the minority image is changed to appear lighter, as we hoped for. The generator to make majority lighter tone skin lesions look like minority darker skin tone images can also be shown with great results. Again the skin lesion itself rarely changes, yet the skin tone becomes visibly darker. The figures show from left to right the real data sample, what the data sample looks like when augmented by the respective generator, and the reconstructed sample, which is what happens when the augmented sample is passed into the generator of the original domain, which should augment the augmented image to appear back into its previous domain.

As shown in Figure 8 and tables 1 and 2, when training and testing the classification model on the majority data and augmented minority data we were able to achieve a jump from 50% accuracy for the minority datasets without CycleGAN augmentation to 79% accuracy with a 0.71 F1 score for the benign classification and a 0.83 F1 score for the malignant classification, when tested on the minority dataset augmented using the CycleGAN. This shows our success in domain adaptation, however, interestingly, the amount

CNN with 5-Epoch CycleGAN		
Metrics	Lighter Skin Tone	Darker Skin Tone
Accuracy	81%	79%
Precision	Benign: 0.76 Malignant: 0.90	Benign: 0.67, Malignant 0.86
Recall	Benign: 0.92 Malignant: 0.71	Benign: 0.75 Malignant: 0.81
F1 Score	Benign: 0.83 Malignant: 0.79	Benign: 0.71 Malignant: 0.83

Table 1: Performance metrics of domain adaptation approach

CNN without Domain Adaptation		
Metrics	Lighter Skin Tone	Darker Skin Tone
Accuracy	85%	50%
Precision	Benign: 0.82 Malignant: 0.88	Benign: 0.61, Malignant 0.27
Recall	Benign: 0.89 Malignant: 0.80	Benign: 0.65 Malignant: 0.24
F1 Score	Benign: 0.85 Malignant: 0.84	Benign: 0.63 Malignant: 0.25

Table 2: Performance metrics of baseline approach

of epochs that the CycleGAN was trained varied the results when tested using the classification model. More specifically, after training the CycleGAN model for 10 epochs, individual instances of the CycleGAN trained at different epochs were used to create multiple variations of augmented minority data, which was tested to train and test the classification model. What we found was that the 5th epoch CycleGAN resulted in the smallest gap between classification accuracy for minority darker skin tone images and majority lighter skin tone images. The classification model trained and tested on minority data that was augmented with the 5th epoch CycleGAN resulted in our best results of 79% accuracy for minority darker skin tones and 81% accuracy for majority lighter skin tones, yielding the smallest gap between accuracies that we achieved.

The resulting metrics using our domain adaptation method are shown more clearly in comparison to the baseline resulting metrics in tables 1 and 2.

Unexpectedly, in Figure 8 it appears as though the CycleGAN method loses its effectiveness as it's trained for more epochs. However, that could simply be due to a local minima in the back-propagation of the generators, or is maybe due to the limits in our dataset sizes. Either way, we need more generator data points to truly understand, which serves as an opportunity for future research.

5 DISCUSSION

What we showed was a great success when using a small dataset of minority darker skin tone images for domain adaptation. Using this process with larger datasets of minority darker skin tone images will further yield better results and more robust tests. Using the CycleGAN's generator to create synthetic minority data may help to create a large enough synthetic dataset of fake minority data that could be used in addition to the real minority data and real majority data to train and test another classification model.

6 CONCLUSION

Through this work we demonstrated a machine learning based mobile skin cancer detector that would be able to almost equalize the fidelity for dark skin tones versus white skin tones. We demonstrated through a domain adaptation process that it is possible to create an inclusive ML model that is trainined by augmenting the missing skin tones from other racial ethnicity, thereby leading a process to mitigate the racial biases ML based vision based skin cancer detection applications. By using a CycleGAN to augment the minority data and correct the learned differences between the domains, we were able to train and test a classification model that achieved 79% accuracy, with a 0.71 F1 score for benign classifications, and a 0.83 F1 score for malignant classifications when tested on minority darker skin tone data, thereby improving inclusiviness of diverse skin tones for machine learning based skin cancer detection.

REFERENCES

[1] [n. d.]. AI is transforming medicine – but it can only work with proper sharing of data. https://theconversation.com/ai-is-transforming-medicine-but-it-can-only-work-with-proper-sharing-of-data-129515. Accessed: 2023-08-09.

[2] 2023. Skin Cancer Facts and Statistics. http://www.skincancer.org/skin-cancer-information/skin-cancer-facts.

[3] R. Agarwal, M. Bjarnadottir, L. Rhue, M. Dugas, K. Crowley, J. Clark, and G. Gao. 2023. Addressing algorithmic bias and the perpetuation of health inequities: An AI bias aware framework. *Health Policy and Technology* 12, 1 (2023), 100702. https://www.sciencedirect.com/science/article/pii/S2211883722001095

[4] Bell. 2023. 4 steps to detecting skin cancer early. https://www.centerforsurgicaldermatology.com/4-steps-to-detecting-skin-cancer-early/#:~:text=In%20fact%2C%20one%20in%20two,diagnosed%20and%20treated%20early%20enough

[5] Jack Breen, Kieran Zucker, Nicolas M. Orsi, and Nishant Ravikumar. 2022. Assessing domain adaptation techniques for mitosis detection in multi-scanner breast cancer histopathology images. *Biomedical Image Registration, Domain Generalisation and Out-of-Distribution Analysis* (2022), 14–22.

[6] Elise Brunsgaard, Jakob Jensen, and Douglas Grossman. 2022. Melanoma in skin of color: Part II. racial disparities, role of UV, and interventions for earlier detection. *Journal of the American Academy of Dermatology* (2022).

[7] Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks* 106 (2018), 249–259.

[8] Roxana Daneshjou, Kailas Vodrahalli, Roberto A. Novoa, Melissa Jenkins, Weixin Liang, Veronica Rotemberg, Justin Ko, Susan M. Swetter, Elizabeth E. Bailey, Olivier Gevaert, and et al. 2022. Disparities in dermatology AI performance on a diverse, curated clinical image set. *Science Advances* 8, 32 (2022).

[9] Ahmed Ech-Cherif, Mohammed Misbhauddin, and Mohammed Ech-Cherif. 2019. Deep neural network based mobile dermoscopy application for triaging skin cancer detection. *2019 2nd International Conference on Computer Applications & Information Security (ICCAIS)* (2019).

[10] Lisa N. Guo, Michelle S. Lee, Bina Kassamali, Carol Mita, and Vinod E. Nambudiri. 2022. Bias in, bias out: Underreporting and underrepresentation of diverse skin types in machine learning research for skin cancer detection—A scoping review. *Journal of the American Academy of Dermatology* 87, 1 (2022), 157–159.

[11] Amita Kapoor. 2019. *Hands-on Artificial Intelligence for IOT.* Packt Publishing.

[12] Department of Health & Human Services. 2001. Melanoma. https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/melanoma

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

[14] Eman Rezk, Mohamed Eltorki, and Wael El-Dakhakhni. 2022. Improving skin color diversity in cancer detection: Deep Learning Approach. *JMIR Dermatology* 5, 3 (2022).

[15] David Wen, Saad M Khan, Antonio Ji Xu, Hussein Ibrahim, Luke Smith, Jose Caballero, Luis Zepeda, Carlos de Blas Perez, Alastair K Denniston, Xiaoxuan Liu, and et al. 2022. Characteristics of publicly available skin cancer image datasets: A systematic review. *The Lancet Digital Health* 4, 1 (2022).

[16] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. *2017 IEEE International Conference on Computer Vision (ICCV)* (2017).