Enhancing Skin Disease Detection Accuracy and Fairness:

Mitigating Biases in Dermatological Diagnosis Models

CSC2529 Project Report

Xinran Zhang, Qianyi Li and Mingfei Li

Abstract—This article focuses on the impact of dataset imbalance on the results of the ResNet18 model and explores methods to address and mitigate imbalance. For the two types of imbalance in the Fitzpatrick dataset, we employed sampling, simple augmentation, and DCGAN augmentation to address the imbalance and enhance accuracy and efficiency. Additionally, we examined the data augmentation effects of DCGAN on both Fitzpatrick and HAM10000 datasets. Based on our experiment results, both simple and DCGAN augmentation have proven to be effective to mitigate bias in the model.

Index Terms—Computational Photography, Machine Learning, Convolutional Neural Networks, GAN, Clinical Images

1 Introduction

SUSPICIOUS skin conditions necessitate thorough medical examinations, as they may evolve into severe forms of skin cancer. Timely and accurate diagnosis is crucial, as early detection significantly improves the survival rates of skin cancer patients. To address the shortage of dermatology experts and expedite the diagnostic process, many healthcare institutions have begun integrating advanced machine learning (ML) techniques, particularly convolutional neural networks (CNNs), into their diagnostic workflows.

While these ML models have greatly aided medical professionals and yielded relatively accurate results for a broad patient demographic, a critical issue has surfaced. These models often reflect biases inherent in their training data, leading to potential inaccuracies and disparities in disease detection. A notable concern is the predominance of light skin tones in skin imaging datasets, which skews the effectiveness of these diagnostic models, especially for individuals with darker skin tones.

Our project specifically focuses on analyzing the Fitzpatrick17k dataset.A stark imbalance is evident within this dataset: the two lightest skin tones comprise nearly 40% of the data, whereas the darkest two skin tones represent a mere 13%, which can significantly impair the cancer detection accuracy for darker-skinned individuals. Imbalance also arises from label within the data. Fitzpatrick17k contains disproportionate number of non-cancerous cases, often exceeding 70% of the dataset. This imbalance can significantly influence the model's accuracy and fairness.

The primary aim is to confront these challenges directly. By thoroughly examining and addressing potential imbalances in the dataset, we intend to diminish the effects of these imbalances on the model's outcomes. Our goal is to rectify this imbalance within the dataset and, subsequently, enhance the accuracy of our models.

In our project, we implemented a variety of sampling and data augmentation techniques to address the imbalance in the dataset. Our approach included basic augmentation strategies like flipping and rotating images, as well as employing advanced machine learning techniques, specifically Generative Adversarial Networks (GANs), for image generation. The outcomes of these experiments demonstrated that both the simple and sophisticated techniques effectively reduced bias in the model. Additionally, they delivered commendable performance when compared to results obtained using the imbalanced dataset.

2 Related Work

Our project delves into racial disparities in dermatological conditions, drawing upon several key studies. A notable one, "The Ongoing Racial Disparities in Melanoma: An Analysis of the Surveillance, Epidemiology, and End Results Database (1975-2016)," provides essential insights into the growing melanoma-specific survival (MSS) disparities among minority groups compared to non-Hispanic whites (NHWs) since 2000 [1]. This study underscores the urgent need for better post-diagnosis management in minority populations, aligning closely with our project's focus.

Further contributing to our understanding is the systematic review, "Racial Disparities in Skin Tone Representation of Dermatomyositis Rashes." This review highlights the significant underrepresentation of darker skin tones in medical educational materials, with a majority of dermatomyositis rash images depicting very light skin. [2] This gap in resources underscores the diagnostic and treatment challenges for darker-skinned individuals and reinforces the necessity of our research in promoting inclusivity in dermatological education and practice.

• *Department of Computer Science, University of Toronto*

Additionally, "Common Dermatologic Disorders in Skin of Color: A Comparative Practice Survey" examines variations in skin disease diagnoses across races and ethnicities [3]. This study illuminates how genetic, environmental, socio-economic, and cultural factors contribute to these disparities. Its emphasis on the need for comprehensive data on skin conditions in diverse populations resonates with our project's aim to expand dermatological knowledge and ensure equitable healthcare access.

In shaping our methodology, we were inspired by Frid-Adar et al.'s use of Generative Adversarial Networks (GANs) to augment CNN performance in medical image classification. Their work [4], "GAN-based Synthetic Medical Image Augmentation for increased CNN Performance in Liver Lesion Classification", sets a precedent for employing GANs to create high-quality synthetic images. This technique was pivotal in applying GAN models to the Fitzpatrick and HAM10000 datasets to counteract data imbalances. The synthetic images generated by GANs enabled us to expand our training dataset, leading to a more balanced data distribution and enhancing the accuracy and fairness of skin disease detection. [5]

Radford, Metz, and Chintala's development of Deep Convolutional Generative Adversarial Networks (DCGANs) further informed our approach. Their 2016 study demonstrates DCGANs as an effective unsupervised learning framework, capable of learning detailed features from images [6]. The architectural constraints they proposed for DCGANs have improved training stability and informed our use of GANs for dataset augmentation in skin disease classification, addressing both imbalances and enhancing model accuracy.

3 Dataset

3.1 Fitzpatrick17k

In this project, we use the dataset Fitzpatrick17k [7] to evaluate the performance of various augmentation techniques. The Fitzpatrick17k, consisting of 16,577 clinical images, is collected from DermaAmin and Atlas Dermatologico, with additional Fitzpatrick scale labels [8]. The Fitzpatrick scale is a six-point measurement for sun reactivity of skin phenotype, and used in many research to assess the fairness of models or algorithms (it is also used for emoji skin colors).

The dataset includes labels for over 100 skin conditions, which are then categorized into nine subgroups, encompassing inflammatory, genodermatoses, and malignant epidermal conditions. These are further consolidated into three overarching categories: benign cancer, malignant cancer, and non-neoplastic (non-cancerous). To simplify and highlight the potential impact of biases within the dataset on model performance, we will combine the two types of cancers to have a binary label: cancerous and non-cancerous.

Through an analysis of the dataset, we identified some certain issues, including two types of extremely imbalance and uneven distribution in the content of the images.

1) **Imbalance in Fitzpatrick scales**
From figure 1.1, we observe a highly uneven distribution of the dataset across different scales. While over 66.74% of images fall into the Fitzpatrick I,II, and III scales, the V and VI scales only account for 13.08%. We speculate that due to the training data imbalance and lack of representation for individuals with darker skin tones, the models will inherit this bias and may exhibit poor performance when applied to this demographic [9].
2) **Imbalance in labels**
From figure 1.2, in the distribution of the three labels, we also identified a similar imbalance issue. The "non-neoplastic" category is overly represented in the dataset, constituting over 70% in all six scales. In contrast, the proportions of the other two labels are significantly smaller, particularly for the "benign" label. Such an imbalance can lead to suboptimal learning outcomes for the model, as a model might achieve decent results without truly learning from the dataset and regarding minority as outliers [10]. Nevertheless, in the context of skin cancer diagnosis, correctly identifying those people with cancers is the primary concern. With the learned model shows bias towards the predominant class, the non-cancerous patients, it may neglect the minority class, which defies our original purpose [11]. To make our lives easier and mitigate the imbalance to some extent, we combined labels 'benign' and 'malignant' into one label.
3) **Uneven distribution of images**
As shown in Figure 1.3, we provide sample images for each Fitzpatrick scale. From these images, we can observe that the patterns in this dataset do not follow a similar distribution – patients' affected areas are widely distributed, including hands, faces, small patches of skin, irrelevant backgrounds, and so on. This issue is persistent within each subgroup of the dataset. These noticeable content variations can introduce instability in the model. [12] Such a dataset poses a significant challenge for GAN, making it difficult to generate high-quality fake images [13].

3.2 HAM10000

Because of the third limitation in Fitzpatrick17k dataset, we introduced the use of another dataset, HAM10000, to assess the capability of using GANs for data augmentation.

HAM10000("Human Against Machine with 10000 training images") consists of 10015 dermatoscopic images which are released as a training set for academic machine learning purposes and are publicly available through the ISIC archive. [14]

As shown in Figure 2, we present some images from the HAM10000 dataset (with subgroup of Male with melanoma). In contrast to Figure 1.3, we can observe the advantage of data consistency in HAM1000 ——images in HAM10000 are all focused on displaying small patches of skin, in other words, images in HAM10000 adhering to a similar distribution within each subgroup, which is advantageous for GANs [13] to generate high-quality fake images based on the training set.

To demonstrate the benefits of Generative Adversarial Networks (GANs) in data augmentation, our study concentrated on four subgroups characterized by gender (Male or

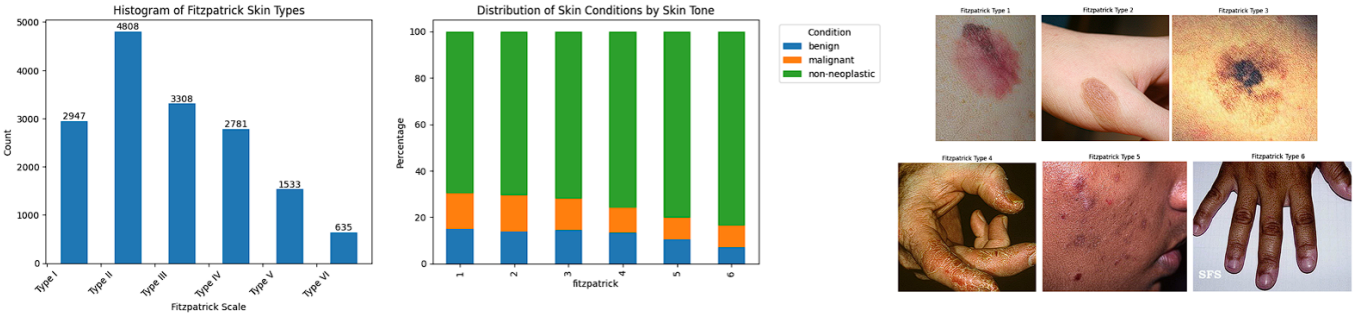

Fig1.1 imbalance in scales

Fig1.2 imbalance in labels

Fig1.3 sample images from the dataset

Fig. 1. Three main issues in Fitzpatrick17k dataset

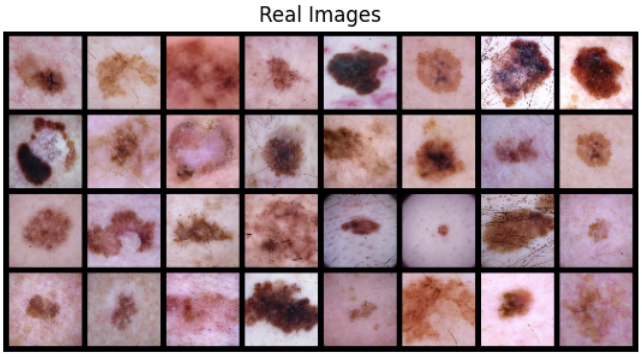

Fig. 2. Images from HAM10000 - male with Melanoma

Female) and skin conditions (Melanoma, a malignant skin cancer, and Benign Keratosis, a benign condition). Figure 3 reveals a slight gender imbalance in our dataset, whereas the distribution between two cancers is almost equal.

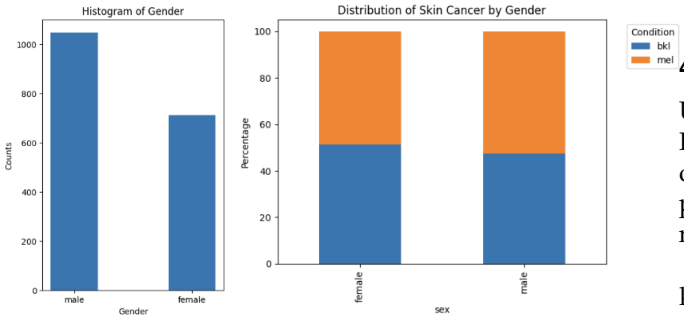

Fig. 3. HAM10000 Distributions

4 Proposed Method

4.1 ResNet

Residual Network [15], or ResNet, is a type of CNNs that aims to solve problems by training very deep networks.

The authors who introduced ResNet argued that adding more layers onto a shallow network should not degrade the model performance. They proposed the method to use residual blocks, which allows the inputs of a given block to be added to its outputs. This identity mapping makes sure the added layers do not degrade the performance. Moreover, instead of learning directly the underlying mapping of interests ($\mathcal{H}(x)$), the model learns the difference or residual of the mapping ($\mathcal{F}(x) := \mathcal{H}(x) - x$).

4.2 GAN

Generative Adversarial Networks [16], or GANs, is a class of unsupervised learning algorithm. It consists of two neural networks, one called generator and another discriminator. During the training, with inputs of random noise, the generator produces fake images that indistinguishable from the real images, whereas the discriminator aims to distinguish between fake and real images. Its structure is illustrated in Fig 6. The idea behind the algorithm is a zero-sum game. If the generator improves and outputs more realistic images (or images more similar to the real images), the discriminator will improve its ability to identify the fake and real. In the experiment, we use DCGAN, which specifically requires convolutional layers in both generator and discriminator.

4.3 Experiment Design

Upon recognizing two principal forms of imbalance in the Fitzpatrick dataset – specifically, the underrepresentation of certain skin tones and disparities in labeling – we hypothesize that rectifying these imbalances will significantly narrow the accuracy gap across different groups.

To address these issues, we have developed 2 different kinds of methods: sampling and data augmentation, aimed at mitigating these disparities. We input the processed dataset into the ResNet18 model [15] and then compare the results with the baseline model using accuracy as the evaluation metric.

The dataset was divided into training and testing sets using an 80:20 split ratio.

4.4 Baseline

First of all, the unprocessed Fitzpatrick training dataset is directly input into the ResNet18 model as **Model 1**, serving as a baseline for comparison.
With data size of 12809

4.5 Sampling

For two imbalances, we performed different sampling strategies on the dataset, constructing models tailored to address specific imbalance scenarios. These models were then compared against the baseline model for evaluation.

1) **Model 2: Adjusting Fitzpatrick Scales Imbalance**
   We identify the least represented skin type in the six-point Fitzpatrick scale, which in this case is skin type VI with 508 samples. We then equalize the sample size across all skin types by randomly reducing the number of samples in the other five categories to match that of skin type VI.
   With data size of 3096
2) **Model 3: Adjusting Label Imbalance**
   For simplicity, we reclassify the dataset into two groups: cancerous (encompassing both benign and malignant cancers) and non-cancerous. With approximately 75% of samples labeled as non-cancerous, we randomly reduce this group to equalize the count with the cancerous group.
   With data size of 6838
3) **Model 4: Combining Adjustments for Two Imbalances**
   We apply both Method 2 and Method 3. The dataset is same as the reduced data from Method 3, and then it is further divided according to the six-point Fitzpatrick scale, resulting in 6 groups. We then adjust the sample size in each group to match the group with the lowest sample count, which is Fitzpatrick VI.
   With data size of 1440

The purpose of these methodologies is to establish balance in the dataset. Nevertheless, although undersampling can create balanced data, the dataset inevitably decreases in size, which is not desired. An ideal dataset should be balanced in both Fitzpatrick and labels, and contain a reasonable number of data points.

4.6 Data Augmentation

With the empirical evidence suggesting that large, balanced data yields considerably better results, our next objective is to expand the dataset. This expansion aims to enhance the accuracy of our models by enlarging the dataset and maintaining a balanced representation in terms of skin color and labeling.

To further enhance the dataset, we employed two data augmentation methods on 3 subgroups: **scale V with label 1** (F5b1), **scale VI with label 1** (F6b1) and **scale VI with label 0** (F6b0), to expand the database, aiming for improved results. We chose these 3 subgroups since they are significantly smaller than others. Here is the procession flowchart in figure 4.

4.6.1 Simple Augmentation

Model 5: *With data size of 6396*
Basic techniques such as flipping and rotation are applied to the training dataset. We used simple augmentation to generate new images for the three subgroups with the smallest quantities mentioned above —— (F5b1), (F6b1), and (F6b0).

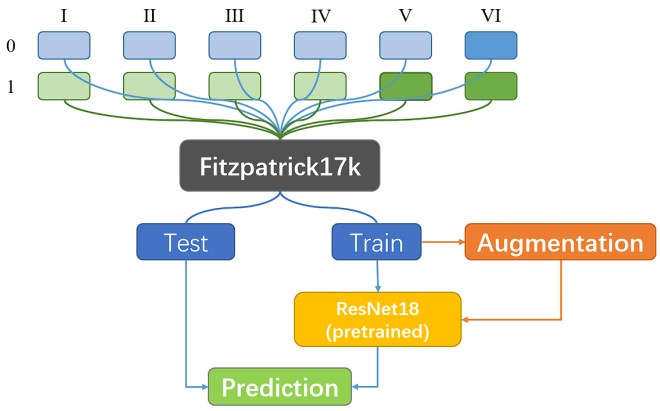

Fig. 4. Data Augmentation procession flowchart

The number of images before and after generation is shown in Table 1.

	F5b1	F6b1	F6b0	F5b0	other subgroups
# images before	246	83	433	533	NA
# images generated	492	581	433	0	NA
sampling	533				

TABLE 1
number of new images generated by simple augmentation methods

Subgroup (f5b0) comprises 533 data points, the smallest number among the subgroups that do not require augmentation. The number 533 is set to be a threshold.

Since the other 8 subgroups contain more images, we randomly sample them to match the threshold to achieve a balance in the dataset.

After generating new images, using the quantity of F5b0 as the baseline, we sampled 533 images from each of the 12 subgroups. The final dataset consists of 6396 images, which is closest to training set used in the **Model 3**. We then input this dataset into the ResNet18 network to obtain **Model 5**.

Here are some samples of new images generated by simple augmentation methods in Figure 5.

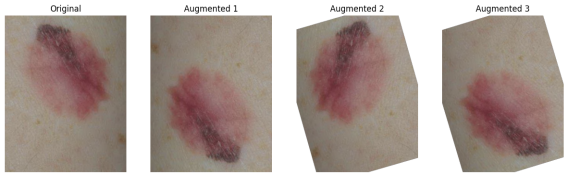

Fig. 5. data augmentation by simple methods

4.6.2 Synthetic Photo Generation

In contrast to simple augmentation, which involves variations of the same image, we aimed to expand our dataset with distinct images. We opted to use DCGAN to generate new images, and the learning process of this model is illustrated in the following figure 6.

The strategy for applying DCGAN on the Fitzpatrick17k dataset, including the number of generated images, aligns with that of simple augmentation.

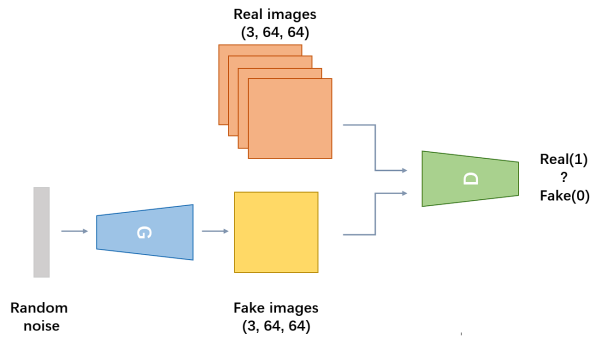

Fig. 6. How GANs generate new fake images

	Fitzpatrick I	II	III	IV	V	VI
Baseline						
Model1	0.7066	0.7141	0.7224	0.7629	0.8053	0.8067
Sampling						
Model2	0.6820	0.6885	0.7066	0.7594	0.8020	0.8151
Model3	0.7459	0.7357	0.7476	**0.7825**	0.8119	**0.8319**
Model4	0.7082	0.6988	0.7224	0.7415	0.7789	0.7983
Data Augmentation						
Model5	**0.7672**	**0.7408**	**0.7697**	0.7736	**0.8284**	0.8152

TABLE 2
Accuracy Across Different Fitzpatrick Scales for Various Models

However, due to the third limitation of this dataset – Inconsistent Image Distribution, the performance of DCGAN on the Fitzpatrick is unsatisfactory. Thus, we also conducted tests on the performance of DCGAN as a data augmentation technique using the HAM10000 dataset. In HAM10000 dataset, we chose genders and disease labels as a scenario where imbalance might occur.

1) **Model 1: Baseline Model**
From the earlier description of the HAM10000 dataset, we can infer that, in the gender and labels, the dataset itself is relatively balanced. Therefore, we used the original balanced data as input for ResNet18, and the obtained results serve as the baseline for this dataset.
With data size of 1761
2) **Model 2: Imbalance in gender**
To assess the improvement in model performance through data augmentation using DCGAN later, we artificially created an imbalance scenario. We randomly sampled and removed 70% of male examples, thus generating an imbalanced dataset with uneven male and female representation, doubling the difference between the two gender groups. This imbalanced dataset was then used as input for ResNet18.
With data size of 1027
3) **Model 3: Rebalancing through DCGAN**
We trained DCGAN to generate fake images with labels (for example, using images of 'male with melanoma' during training to generate images labeled as both 'male' and 'melanoma'). These generated images were then incorporated back into the dataset. Through this process, we obtained a balanced dataset once again.
With data size of 1727

5 Experimental Results

We used accuracy as our metric for evaluation. For the Fitzpatrick dataset, we provide accuracy scores for different models based on the Fitzpatrick scales. Regarding the HAM10000 dataset, we present accuracy scores for different models separately for each gender.

5.1 Fitzpatrick

The accuracy scores are shown in table2.

1) **Model 1: Baseline Model**
This model's performance is consistent with the proportions of non-cancerous cases in the dataset. For example, for Fitzpatrick VI, 83% images are labelled as non-neoplastic, and other groups exhibit similar behavior. This hints that its accuracy might stem from predicting the majority class, rather than learning the underlying patterns. The significant variation in accuracy between skin types, exceeding 10%, and the large dataset size contribute to inefficient training.
2) **Model 2: Adjusting Fitzpatrick Scales Imbalance**
Despite being balanced in the Fitzpatrick scale, the model 2 neither enhances accuracy nor reduces discrepancy. Presumably, this could be attributed the facts that (1) the dataset is reduced to a quarter of its original size and (2) the imbalance in labels is still present and has more significant impacts onto the bias in the model. Resolving the Fitz imbalance alone seems insufficient.
3) **Model 3: Adjusting Label Imbalance**
While Model 3 still shows the disparity across the six Fitz scale, its performance has significantly improved compared to previous models. Notably, the current dataset is balanced in labels, so the high accuracy indicated that the model has learned the underlying pattern of images to some extent. This supports our previous presumption that label imbalance contributes more the bias in the model. Moreover, the Model 3 is more efficient as it is trained on a dataset that's about half of the original dataset and achieved enhanced performance.
4) **Model 4: Adjustments for Two Imbalances**
Model 4, balanced in both Fitzpatrick scale and labels, does not show further performance improvement. However, like Model 3, its balanced labels and high accuracy indicate that the model is learning image patterns instead of predicting the majority class. It's important to consider that Model 4 was trained on the smallest dataset, reduced to just 1440 data points (about 1/10th of the original size) due to two phases of undersampling. This limited data size might account for the slight drop in accuracy. But this in turn shows the improved efficiency obtained from balanced data.
5) **Model 5: Simple Augmentation**
We observed that the 4/6 of the accuracies in the

model are the highest among the five models and the remaining 2 are comparable with the highest. This underscores the effectiveness and capability of simple augmentation. Notably, the discrepancy across the different Fitz scale decreases to 5% from 10% in the baseline, supporting our hypothesis that balance in dataset would reduce the performance gap among the subgroups. Furthermore, it reinforces our previous notion that the while the balance in data significantly influences model performance, the data size (6396) is also a pivotal factor to consider.

We attempted to apply DCGAN to the Fitzpatrick dataset; however, as previously discussed, the limitations of this dataset hindered DCGAN's ability to generate high-quality fake images. Figure 7 are samples of fake images generated by DCGAN using the Fitzpatrick dataset. Despite adjusting the input real images dimensions to 128*128, the performance remained subpar and unusable.

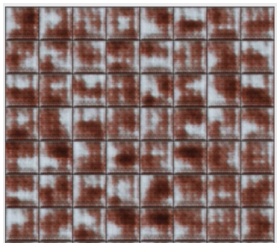

Fig. 7. Fake images generated by DCGAN based on Fitzpatrick dataset

5.2 HAM10000

1) **Model 1**
Model 1 was trained on the original dataset, which is balanced in labels and shows only a mild gender imbalance. According to the results in Table 2, this model does not exhibit gender bias in its performance. In terms of label accuracies, while the accuracy for Keratosis is not entirely satisfactory, the model demonstrates high performance in identifying Melanoma cases, a critical factor given melanoma's high fatality rate.
2) **Model 2**
The dataset for Model 2 was deliberately altered to create a gender imbalance. As expected, the model trained on this imbalanced data shows bias in its results, with a notable decrease in accuracy for males, the minority group. Interestingly, label accuracies are similar to Model 1, likely because the labels remained balanced.
3) **Model 3**
By incorporating images generated by DCGAN, the dataset achieves balance once again. Gender accuracies not only recover to levels comparable to Model 1, but also becomes more balanced. This outcome highlights DCGAN's advanced ability to expand datasets, thereby enhancing model performance and reducing bias, as seen in our study. In addition, the label accuracies are still closer to their counterparts in Model 1 and 2. Again, it is likely that the labels remained balanced throughout. It further reinforces our notion discussed in Fitzpatrick results: imbalance in labels appears to be more influential.

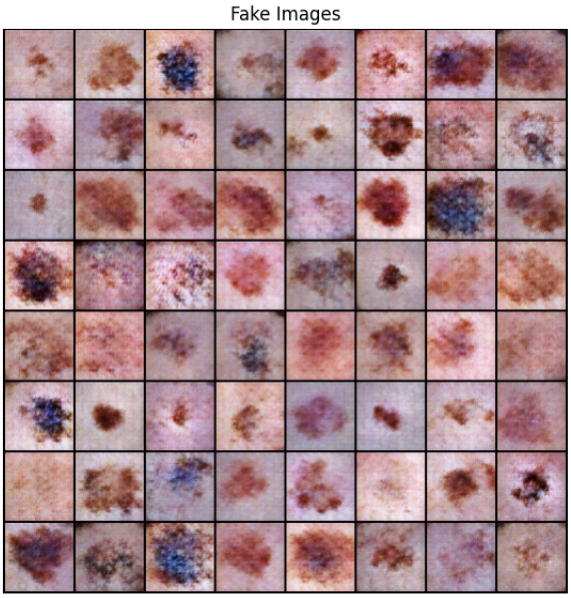

Fig. 8. Fake images generated by DCGAN based on HAM10000 dataset

	model1	model2	model3 (with GAN)
Female	0.7029	0.6514	0.6971
Male	0.7105	0.5451	0.6616
Melanoma	0.8364	0.8551	0.8505
Benign Keratosis	0.5815	0.5639	0.5286

TABLE 3
Results of Augmentation on HAM10000 using DCGAN

6 Conclusion

The conclusion of our project underscores the pivotal role of data diversity in developing fair and accurate machine learning models for skin disease detection. Through a comparative analysis of five distinct models (baseline, sampling and simple augmentation), we have determined that Model 5, with its data-augmented approach, stands out not only by improving overall accuracy but also by significantly reducing the discrepancy in performance across different skin types.

By attempting DCGAN on two different datasets, we found that while GANs can indeed effectively augment the dataset by generating synthetic images similar to the original ones, this method imposes higher requirements on the dataset. It necessitates that the images in the dataset follow a similar distribution.

This project illustrates the necessity of addressing both quantity and variety of data to mitigate bias inherent in dermatological diagnosis models. Our findings advocate for the continued pursuit of inclusive and comprehensive datasets that reflect the diversity of the real-world population, ensuring that advancements in AI-driven diagnostics benefit all individuals equitably.

For code, poster, and other files, please check this link

ACKNOWLEDGMENTS

The authors express their gratitude to Dr. David Lindell for providing guidance and support during CSC2529: Computational Imaging at the University of Toronto. Additionally, the authors appreciate the valuable ideas and suggestions offered by Teaching Assistants Anagh Malik and Parsa Mirdehghan from the inception of this project.

REFERENCES

[1] Y. Qian, P. Johannet, A. Sawyers, J. Yu, I. Osman, and J. Zhong, "The ongoing racial disparities in melanoma: An analysis of the surveillance, epidemiology, and end results database (1975-2016)," *Journal of the American Academy of Dermatology*, vol. 84, no. 6, pp. 1585–1593, 2021.

[2] S. Babool, S. F. Bhai, C. Sanderson, A. Salter, and L. Christopher-Stine, "Racial disparities in skin tone representation of dermatomyositis rashes: a systematic review," *Rheumatology*, vol. 61, no. 6, pp. 2255–2261, 2022.

[3] A. F. Alexis, A. B. Sergay, and S. C. Taylor, "Common dermatologic disorders in skin of color: a comparative practice survey," *Cutis*, vol. 80, no. 5, pp. 387–394, 2007.

[4] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321–331, 2018.

[5] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[6] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[7] M. Groh, C. Harris, L. Soenksen, F. Lau, R. Han, A. Kim, A. Koochek, and O. Badri, "Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1820–1828.

[8] M. Groh, C. Harris, R. Daneshjou, O. Badri, and A. Koochek, "Towards transparency in dermatology image datasets with skin tone annotations by experts, crowds, and an algorithm," *arXiv preprint arXiv:2207.02942*, 2022.

[9] G. Kleinberg, M. J. Diaz, S. Batchu, and B. Lucke-Wold, "Racial underrepresentation in dermatological datasets leads to biased machine learning models and inequitable healthcare," *Journal of biomed research*, vol. 3, no. 1, p. 42, 2022.

[10] L. Gao, L. Zhang, C. Liu, and S. Wu, "Handling imbalanced medical image data: A deep-learning-based one-class classification approach," *Artificial Intelligence in Medicine*, vol. 108, p. 101935, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S093336572030261X

[11] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, no. 1, pp. 1–54, 2019.

[12] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Medical Image Analysis*, vol. 58, p. 101552, Dec. 2019. [Online]. Available: http://dx.doi.org/10.1016/j.media.2019.101552

[13] Y. Hong, U. Hwang, J. Yoo, and S. Yoon, "How generative adversarial networks and their variants work: An overview," *ACM Computing Surveys (CSUR)*, vol. 52, no. 1, pp. 1–43, 2019.

[14] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific data*, vol. 5, no. 1, pp. 1–9, 2018.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[16] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014.