Improving the Fairness of Artificially Intelligent Skin Disease Detectors Using Stable Diffusion

Word Count: 4016

April 29, 2024

***Abstract*—This study addresses the ethical implications of using artificially intelligent neural networks for medical diagnosis, with a focus on skin diseases. While such networks have demonstrated high accuracy, they have been found to misdiagnose minority populations at a disproportionately high rate due to imbalanced training datasets. To prevent this perpetuation of inequities, our study artificially balances a skin disease training dataset by adding stable diffusion-generated synthetic images depicting skin diseases on underrepresented skin tones. Additionally, prompt engineering and fine-tuning are utilized to optimize the photorealism and diversity of synthetic images. Notably, our study is the first to achieve a statistically significant increase in the fairness of a skin disease detector. Specifically, we increase fairness by over 50%, as quantified by the average odds difference metric, helping to foster more equitable healthcare outcomes.**

***Index Terms*—Stable diffusion, bias management, fairness**

I. Introduction

Artificially intelligent neural networks have shown great promise in medical diagnosis applications. Lai et al. [1] trained a neural network on the International Skin Imaging Collective and found it was 95% accurate in diagnosing skin cancer. Yu et al. [2] trained a network that is 89% accurate in diagnosing breast cancer in ultrasound images while clinicians are only 30% accurate. Additionally, neural networks are typically more efficient than human clinicians. Yu et al.'s network also took less than two seconds per diagnosis while clinicians took 314 seconds on average.

Though the accuracies and efficiencies of artificial intelligence have been researched, the ethics of these algorithms have not been explored to the same extent. One critical ethical concern is that artificially intelligent neural networks demonstrate bias against minority populations. In fact, an article recently published in Nature found that artificially intelligent diagnostic assistants "exacerbated the gap in the diagnostic accuracy of generalists across skin tones" [3]. This is largely because most neural networks are trained on imbalanced datasets where darker skin toned patients are underrepresented.

Datasets can be imbalanced between classes—interclass imbalance—or within classes—intraclass imbalance [4]. Interclass imbalance refers to datasets where the number of images in each class varies greatly, as shown in Figure 1A, causing the network to "learn" that the minority class is less common. Take a network that is trained to classify a skin lesion as folliculitis or carcinoma. If the network is trained on very few example images of carcinoma lesions, the network learns that for a given lesion, it is statistically more likely that the lesion is folliculitis. As a result, the network will likely not learn to recognize the visual features of carcinoma and instead classify most lesions as folliculitis based on probability. On the other hand, intraclass imbalance refers to datasets where subpopulations within each class are not evenly represented, as shown in Figure 1B, which can cause networks to disproportionately misdiagnose minorities as well. For example, if there are 10 times more examples of patients with lighter skin tones than darker skin tones, the network is not as aware of how darker skin tones appear and is therefore more likely to misdiagnose patients with darker skin tones [4].

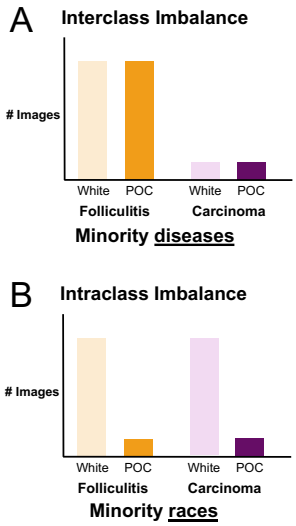

Fig 1. **A:** Interclass imbalance occurs when one class has fewer images than another, e.g. there are less images of carcinoma than folliculitis. **B:** Intraclass imabalance occurs when one subpopulation within classes has fewer images than another subpopulation, e.g. race.

Due to systematic inequities in access to healthcare, people of color are underrepresented in most training datasets [5], creating race-based intraclass imbalance. As a result, many medical diagnostic neural networks misdiagnose patients of color at a higher rate, perpetuating the existing inequities [6]. This study improves the fairness of artificially intelligent neural networks' diagnosis of patients of color by artificially balancing their training datasets so that patients of each skin tone are equally represented.

II. Related Works

A. Dataset

The Fitzpatrick17K dataset [7] is one of the only publicly available medical datasets with annotations describing skin tone as well as disease for each image. The dataset consists of about 17,000 images of 114 skin diseases. Additionally, the skin tone of the patient seen in each image is evaluated using the Fitzpatrick scale, with values ranging from one to six, one being the lightest and six being the darkest. The darkest skin tones, five and six, are vastly underrepresented, making up less than 8% and 4% of the dataset's images respectively.

The authors of [7] have found that when training a diagnostic neural network on the Fitzpatrick17K dataset, it achieved an overall accuracy of 20.2%. This relatively low score suggests that the classes are very visually similar, making it more difficult to distinguish between different diseases, and the number of images is relatively small, leading to its inaccurate generalization.

Researchers from Stanford [6] also trained three neural network frameworks typically used for diagnosing skin diseases on the Fitzpatrick17K dataset and found that the networks achieved an average area under the receiver operating curve (AUROC) score of 54% on the darker skin tones of 5 and 6 compared to an average of 66% on the lighter tones of 1 and 2. This shows that the dataset's intraclass imbalance causes diagnostic networks to misdiagnose darker skin toned patients at a disproportionately high rate.

B. Fairness evaluation metrics

There are several definitions of fairness commonly used when evaluating diagnostic neural networks. Demographic or statistical parity [8] requires $P(\hat{Y} = 1|A = 0) = P(\hat{Y} = 1|A = 1)$ for a given network, $\hat{Y}$. In other words, the rate at which the network makes a given classification, or diagnosis, should not vary for different values of a protected attribute, e.g. different skin tones. This can be evaluated and quantified using statistical parity difference (SPD):

$$\left|PR_{minority} - PR_{majority}\right| \quad (1)$$

In this equation, PR is the positive diagnosis rate. However, this definition does not consider the accuracy of the given diagnosis and therefore does not directly address the fact that networks misdiagnose minorities at a disproportionately high rate.

Equalized opportunity [8] requires $P(\hat{Y} = 1 \,|\, A = 0, Y = 1) = P(\hat{Y} = 1 \,|\, A = 1, Y = 1)$, claiming that the probability of a fair network, $\hat{Y}$, making a *correct* diagnosis should not vary across skin tones. This is quantified by equal opportunity difference (EOD):

$$\left|TPR_{minority} - TPR_{majority}\right| \quad (2)$$

In this formula, TPR is the true or correct diagnosis rate. Using the true positive rate, instead of the overall positive rate, equalized opportunity ensures equal diagnostic accuracy for minorities and majorities alike. However, misdiagnoses are not considered.

Finally, the equalized odds metric [8] requires $P(\hat{Y} = 1 \,|\, A = 0,\ Y = y) = P(\hat{Y} = 1 \,|\, A = 1, Y = y),\ y \in \{0,1\}$. This means that a fair network should have the same correct diagnosis rate and misdiagnosis rates for all skin tones. In other words, it has the same true positive rate (TPR) and false positive rate (FPR) for all skin tones. To quantify this, average odds difference (AOD) is used:

$$\frac{\left|TPR_{minority} - TPR_{majority}\right| + \left|FPR_{minority} - FPR_{majority}\right|}{2} \quad (3)$$

The smaller the statistical parity difference, equal opportunity difference, and average odds difference, the fairer the network. three metrics are used in this study to evaluate fairness.

C. Artificial dataset balancing

There are four main approaches to artificially balancing training datasets: random under-sampling, random oversampling, classical augmented oversampling, and deep learning-based augmented oversampling. Random under-sampling [9] deletes random samples from majority groups until all groups have the same number of images, as shown in Figure 2A. Batista et al. [10] found success in balancing interclass imbalanced datasets using this method. However, they note that this method does not take advantage of all data available due to the deletion of samples.

This finding led to their study of random oversampling. Random oversampling [11], balances datasets by duplicating random samples from minority groups until all groups have the same number of images, as shown in Figure 2B.

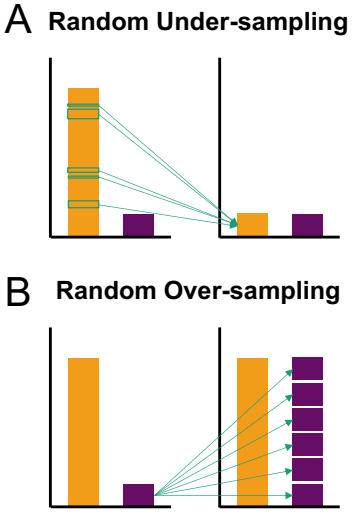

Fig 2. **A:** Random under-sampling deletes images from majority groups until all groups have the same number of images. **B:** Random oversampling duplicates images from minority groups until all groups have the same number of images.

Batista et al. [10] applied both random under-sampling and oversampling to 15 different interclass imbalanced datasets. They found that neural networks were almost always more accurate when trained on oversampled datasets than when trained on the original imbalanced datasets and the under-sampled datasets. This indicates that oversampled training datasets increase network accuracy in classifying minority groups. However, several authors [12], [13] note that random oversampling reduces the average diversity of datasets due to the exact duplication of minority samples. If used to an extreme this in turn often leads to overfit networks, or networks that are unable to generalize from their limited training data, and consequent decreases in accuracy.

To increase the diversity of oversampled training datasets and combat overfitting, classical augmented oversampling was proposed. Classical augmented oversampling creates synthetic samples of minorities by altering existing samples. This approach traces back to the Synthetic Minority Oversampling Technique (SMOTE) [12], which interpolates similar minority samples using the k-Nearest Neighbors algorithm to generate synthetic samples. Since then, Krizhevsky et al.'s proposal of AlexNet [14] has popularized methods such as random cropping, flips, and color space transformations to augment original samples and generate synthetic samples, as shown in Figure 3.

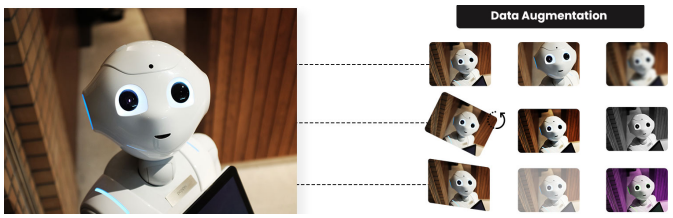

Fig 3. Augmentation techniques include cropping, turning, resizing, and color transformantions, all of which are shown above. Image from [13].

Velasco et al.'s [16] application of classical augmented oversampling to an interclass imbalanced skin disease dataset yielded impressive results, achieving a 94.4% accuracy compared to random oversampling's 91.8% and random under-sampling's 84.3% in classifying underrepresented objects.

Recently, generative artificial intelligence has vastly progressed, leading to the development of deep learning-based augmented oversampling. Introduced by Goodfellow et al., Generative Adversarial Networks (GANs) [17] use two competing networks, a generator that creates synthetic images and a discriminator that attempts to distinguish between the real training images and generated synthetic images. Once the discriminator consistently misclassifies the generated synthetic images as real, the generator is declared fully trained and able to generate realistic synthetic images, as shown in Figure 4.

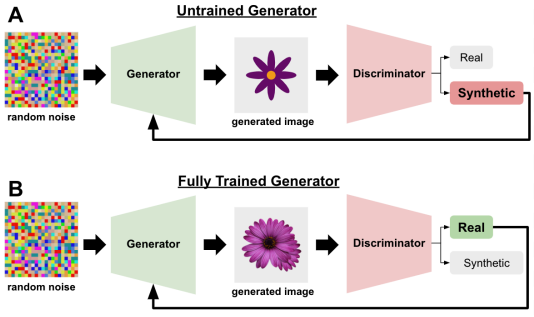

Fig 4. **A:** GAN generators are trained to generate realistic images. **B:** Discriminators are trained to determine whether images are "real" or "fake." Flower image from [18].

Douzas et al. [19] found that GANs outperformed classical augmented oversampling methods across 71 interclass imbalanced datasets. Moreover, [20] utilized a

GAN to balance an intraclass imbalanced dataset of retinal images by skin tone. Their findings revealed that this enhanced the accuracy and fairness of their network in diagnosing retinas of darker skin toned patients.

However, in a project report, student from the University of Toronto [21] attempted to use GANs to rebalance the intraclass imbalanced Fitzpatrick17K dataset by skin tone but found that the images the GAN generated were "subpar and unusable." The subtle visual differences between retinal images of light and dark-skinned patients, compared to the more pronounced disparities in images of skin diseases, likely contributed to the discrepancy in results of [20] and [21]. Achieving convergence of discriminators and generators is challenging, making this methodology unstable and undermining GAN accuracies, as seen in [21]. Discriminators often overfit, learning to reject every image, even truly real ones. This causes mode collapse as the generator can never be declared fully trained [22].

Unlike GANs, diffusion models [23] learn to generate synthetic images by gradually adding Gaussian noise to training images in the forward process and then attempting to recover the images in the reverse process. Once a diffusion model consistently recovers the images accurately, it is declared fully trained. Without a discriminator, mode collapse is avoided. Additionally, OpenAI's recent improvements to diffusion models [24] demonstrated their superiority to GANs in terms of the quality and diversity of their image generation capabilities. Lastly, diffusion models are trained with textual prompts that describe each training image. This enables prompt engineering and therefore, more personalized image generation.

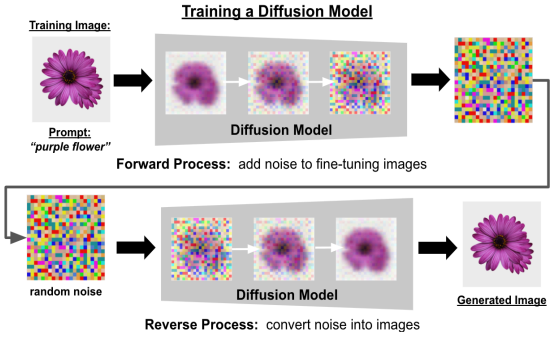

Fig 5. Diffusion models are trained with example images and corresponding textual prompts. As the models add noise and learn to recover images from noise, they learn to generate realistic images, similar to the example images, from noise.

The authors of one pre-print [25] fine-tuned a stable diffusion model on images from the Fitzpatrick17K dataset and confirmed the generated images' realism with their clinical partners at Semmelweis University, indicating that this methodology also excels in generating realistic images of skin diseases.

In another recent pre-print, Sagers et al. [26] used a diffusion model to address interclass imbalance in the Fitzpatrick17K dataset of skin diseases. They found that by training their diagnostic neural network on the dataset with the added synthetic data, the network's accuracy improved by 11.9%, indicating that diffusion models show promise in mitigating interclass imbalance.

Though these dataset balancing approaches successfully reduce interclass imbalance, they have not significantly reduced *intraclass* race imbalance in skin disease datasets. As a result, they fail to address the consequent misdiagnosis of racial minorities' skin diseases. This is the first study to

1. Apply state-of-the-art stable diffusion technology to reduce racial intraclass dataset imbalance.
2. Achieve a statistically significant increase in fairness of a skin disease detector.

III. Procedure

A. Dataset

The Fitzpatrick17K dataset is used in this study as its skin tone labels enable the direct quantification of intraclass imbalance and evaluation fairness for majority and minority skin tones. Sagers et al. [26] used the Fitzpatrick17K dataset and faced difficulties with its limited training data, as the average disease has less than 11 images of skin tone six. To bypass this, they paired skin tones one and two together, three and four together, and five and six together, creating a new skin tone labeling system with three categories instead of six. Similarly, we group skin tones one and two together to form "lightest skin tone" and tones five and six to form "darkest skin tone." Results for the grouping of tones three and four have not yet been found due to time constraints in this work.

Diagnostic neural networks require training as well as

testing data, which will be used in this study to evaluate fairness and overall diagnostic accuracy. Due to these data requirements, only diseases with at least 40 images of each of the skin tones, 35 for testing and at least five for fine-tuning the diffusion model, are selected for use. This narrows down the dataset to 11 diseases: darier disease, folliculitis, keloid, lichen planus, lupus erythematosus, neurofibromatosis, psoriasis, sarcoidosis, scabies, scleroderma, and squamous cell carcinoma. Currently the proposed method has only been applied to folliculitis, lichen planus, and squamous cell carcinoma. The final dataset used, after the modifications outlined above, is detailed in Table 1.

TABLE I. DISTRIBUTION OF MODIFIED DATASET

Disease	Lightest Skin Tone	Darkest Skin Tone
Folliculitis	145	42
Lichen planus	160	117
Squamous cell carcinoma	298	51

B. Diffusion model-based augmented sampling

After choosing and configuring the foundational dataset, it must be artificially balanced. To do so, we generate and add synthetic images of skin diseases on minority skin tones using a diffusion model. The Stable Diffusion 2.0 model from Stability AI [27] used is pretrained on LAION-5B [28], a dataset of 5 billion images with accompanying textual labels from the internet. As Fitzpatrick17K is a relatively small dataset, utilizing a pretrained model ensures that the model is trained on enough data to learn how to generate realistic images. However, LAION-5B has very few images of skin diseases, so to generate realistic images of skin diseases we fine-tune, or retrain, the diffusion model on images from the Fitzpatrick17K dataset using a state-of-the-art fine-tuning technique, DreamBooth [29]. Five training images of each disease on each skin tone are selected as examples for the fine-tuning. Additionally, the disease label, description of the disease, and skin tone label are fused to create textual prompts for each example image, like, "lichen planus (purplish, itchy, flat-topped bumps, lacy white patches sometimes with open sores) on skin tone 1 (lightest skin tone)." Then, DreamBooth is used to retrain the pretrained model on the example images and prompts. With the aim of learning to generate images that depict the inputted textual prompts from scratch, it gradually deconstructs the images by adding Gaussian noise to them and then learns to reconstruct the images by reversing that noising process. As it repeats this reverse process for the five example images of each skin tone-disease combination, the model learns how to construct realistic images from random noise, as shown in Figure 6.

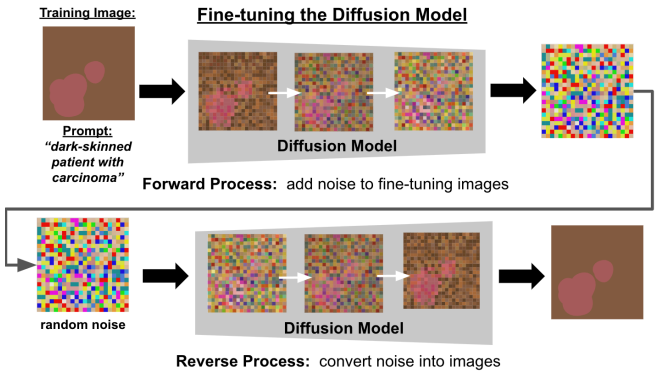

Fig 6. The pre-trained diffusion model is fine-tuned on images of skin diseases on various skin tones from Fitzpatrick17K.

After fine-tuning, images can be generated. For each image, the diffusion model starts from random noise and a textual prompt that describes what the image should depict. From this, the model applies the learned reverse process, constructing realistic images "from scratch." Additionally, to increase the diversity of the images, the generation prompts are methodically engineered. For 75% of the images, the image generation prompts only described the disease and skin tone, for 10% that original prompt was used with the addition of "on an arm," for 10% that original prompt was used with the addition of "on a leg," and for 5% the original prompt was used with the addition of "on a face." This controls the image generation process to ensure realism while varying the appearance of the images to increase dataset diversity.

Lastly, all images generated are added to the original dataset until it is artificially balanced. For example, suppose there are F_L images of light skin toned patients with folliculitis and C_D of dark skin toned patients with squamous cell carcinoma. To ensure each skin tone-disease combination has the same number of images, N, $N - F_L$ images of light skin toned patients with folliculitis and $N - C_D$ images of dark skin toned patients with squamous cell carcinoma are added to the dataset. The value of N is also optimized for fairness and overall diagnostic performance: we tested values of 50,

100, and 150, as in Figure 7. If the number of images of a skin tone-disease combination exceeds N, no images need to be added. This effectively creates four datasets, N=0, which acts as a control, N=50, 100, and 150.

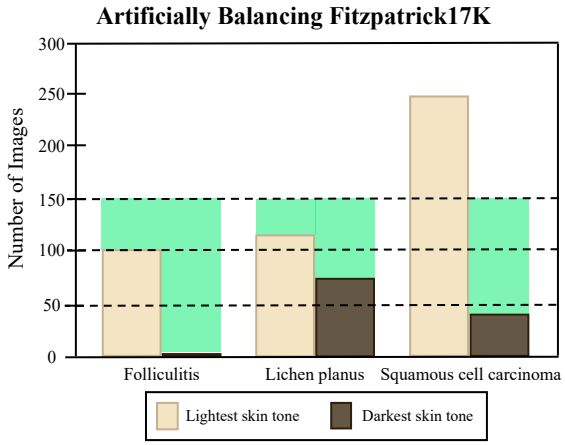

Fig 7. For each threshold value, the green area below it represents the number of synthetic images that are added.

C. *Diagnostic neural networks*

Finally, to evaluate the fairness of the methodology, a diagnostic neural network is trained on the four effective datasets. In accordance with [7], the paper in which the Fitzpatrick17K dataset was first introduced, a VGG-16 neural network architecture [30] pretrained on ImageNet is used. As the paper suggested, the last fully connected layer is replaced with the following five layers: a fully connected layer of 256 units, ReLU activation, a dropout layer with a 40% chance of dropping, a dense layer with the number of disease labels as the number of units, and SoftMax activation. This framework is represented visually in Figure 8. The networks are trained using Keras and TensorFlow [31] on Google Colaboratory.

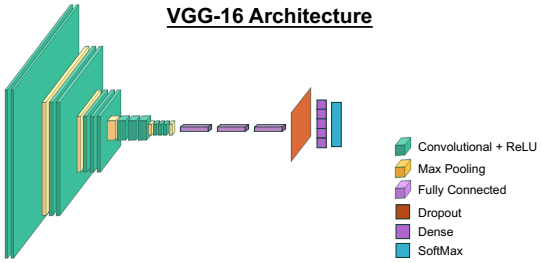

Fig 8. The VGG-16 architecture used consists of alternating convolutional and max pooling layers followed by three fully connected, a dropout, a dense, and SoftMax activation layers.

To determine if network fairness improves with the addition of synthetic images, we use statistical parity difference (SPD), equal opportunity difference (EOD), and average odds difference (AOD). The smaller the metrics, the fairer the network, as described in the Related Works section.

Additionally, Area Under the Receiver Operating Curve (AUROC) [32] scores are collected for each network across all skin tone-disease combinations to evaluate the networks' overall performance and ensure that increasing the networks' fairness did not compromise overall diagnostic performance. Higher AUROC scores represent better performance and random classifiers achieve scores of 50%, as shown in Figure 9. Lastly, the four networks, N=0, 50, 100, and 150, were trained 17 times each and their results were averaged to control for variability.

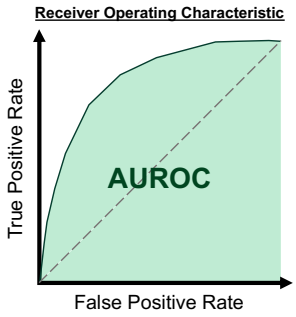

Fig 9. AUROC scores depict a true vs. false positive rate tradeoff.

IV. Results

A. *Image generation with stable diffusion*

Images generated are diverse in appearance, due to the use of prompt engineering, and realistic, due to the efficacy of the diffusion model and DreamBooth. Figure 10 shows several fine-tuning and generated images.

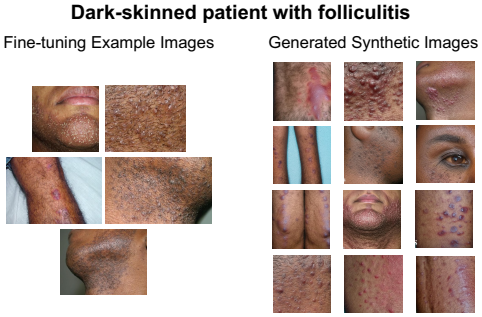

Fig 10. The five fine-tuning images depicting the skin tone-disease combination of darkest skin tone and folliculitis are on the left. Several of the images generated after fine-tuning are on the right. Among the generated images, disease location varies while the realism of the skin tone and disease appearance is maintained.

B. *Evaluating fairness*

As the Figure 11 shows, SPD, EOD, and AOD decrease as the number of synthetic images added increases. SPD decreases by 44%, EOD decreases by 53%, and AOD decreases by 58%. Additionally, t tests were used to prove that the decreases are statistically significant, with p-values of 0.0087, 0.0052, and 0.0004.

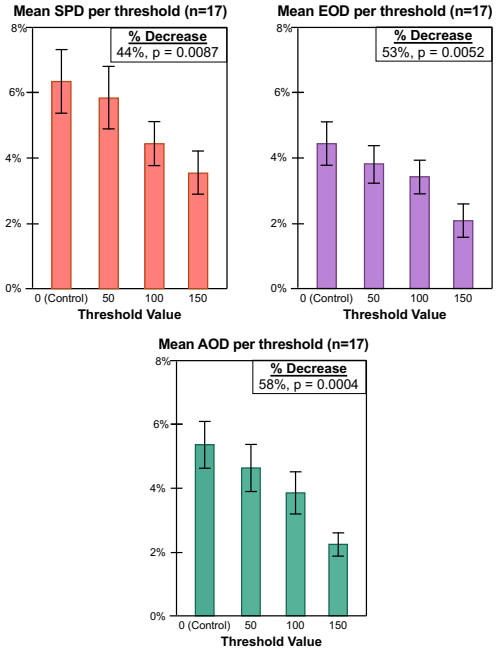

Fig 11. For each threshold, 17 networks were trained. The means of their SPD, EOD, and AOD have been plotted with 95% confidence interval bars. The percent decrease as well as the p-values obtained from the t tests are shown as well.

C. *Overall performance*

As the number of synthetic images increased, the networks were also found to have higher AUROC scores, as shown in Figure 12. This percent increase was by 4.5% from the control's 66.3%.

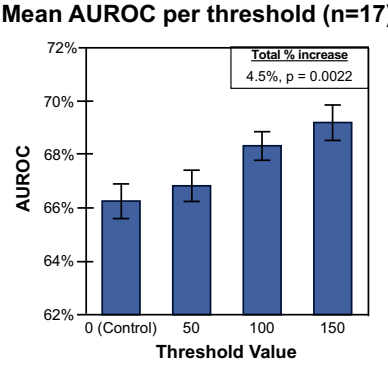

Fig 12. For each network type, 17 networks were trained, and the means of their AUROC scores have been plotted with 95% confidence interval bars.

V. Discussion

The proposed method improved network fairness and performance, increasing the fairness of diagnostic neural networks in diagnosing minorities. Additionally, performance is not compromised while fairness increases—it in fact improved by 4.5%. This indicates that the methodology generates realistic medical images and is effective in artificially balancing datasets, even when starting with just five example fine-tuning images and generating images with very subtle variations in disease and skin tone appearance.

One of the drawbacks of this diffusion model-based method is its large memory requirement during fine-tuning. Due to the limited free resources of Google Colaboratory, only two skin tones and three diseases have been used in this study. In the future, the methodology should be applied to a larger range of diseases and skin tones to ensure generalizability of this solution.

Though this is the first study to address racial intraclass imbalance using diffusion model-based augmentative oversampling, several other papers support our conclusions. Our results of increased fairness and accuracy align with [18] which found that deep learning-based oversampling by skin tone improved the accuracy and fairness of networks trained on an imbalanced dataset of retinal images. Similarly, [22] found that diffusion-model based oversampling increased the accuracy of a network trained on Fitzpatrick17K. Lastly, [25] had human clinicians verify that diffusion models can generate realistic images of skin diseases from the Fitzpatrick17K dataset, which reflects the increase in network accuracy and fairness that we found.

VI. Conclusions

In conclusion, we have found that stable diffusion, fine-tuning, and prompt engineering are key tools for generating the realistic, diverse images necessary for dataset balancing. This indicates that this technology is superior to generative adversarial networks and should be further studied in order to potentially increase the diversity and realism of images generated, and in turn increase fairness, even further.

Additionally, this study is the first to find a

statistically significant increase in fairness of a skin disease-diagnosing neural network, making significant strides towards enhancing the equity of healthcare outcomes for patients of diverse backgrounds. Our findings highlight the potential of artificial intelligence to address biases inherent in current imbalanced medical datasets, thereby promoting a more equitable healthcare system. As intraclass imbalance exists outside the context of medicine, this method can and should be evaluated on other applications with underrepresented minorities—such as facial recognition, hiring algorithms or predicting policing algorithms—to increase fairness in these fields as well.

ACKNOWLEDGEMENTS

I would like to thank my mentor, Dr. William Cohen Principal Scientist at Google AI and SCS Consulting Professor at Carnegie Mellon University. His advice throughout the process of completing this project and writing this paper was incredibly helpful. I also thank my AP Research teacher, Mr. Saul Straussman, for connecting me with Dr. Cohen and reading through countless draft papers and presentations.

GLOSSARY OF ACRONYMS

AOD – Average Odds Difference
AUROC – Area Under the Receiver Operating Curve
EOD – Equal Odds Difference
GAN – Generative Adversarial Network
SMOTE – Synthetic Minority Oversampling Technique
SPD – Statistical Parity Difference

VII. BIBLIOGRAPHY

[1] W. Lai, M. Kuang, X. Wang, P. Ghafariasl, M. H. Sabzalian and S. Lee, "Skin cancer diagnosis (SCD) using Artificial Neural Network (ANN) and Improved Gray Wolf Optimization (IGWO)," *Scientific Reports,* vol. 13, no. 19377, 2023.

[2] T. Yu, W. He, C. Gan, M. Zhao, Q. Zhu, W. Zhang, H. Wang, Y. Luo, F. Nie, L. Yuan, Y. Wang, Y. Guo, J. Yuan, L. Ruan, Y. Wang, R. Zhang, H. Zhang, B. Ning, H. Song, S. Zheng, Y. Li and Y. Guang, "Deep learning applied to two-dimensional color Doppler flow imaging ultrasound images significantly improves diagnostic performance in the classification of breast masses: a multicenter study," *Chinese Medical Journal,* vol. 134, no. 4, 2021.

[3] M. Groh, O. Badri, R. Daneshjou, A. Koochek, C. Harris, L. R. Soenksen, P. M. Doraiswamy and R. Picard, "Deep learning-aided decision support for diagnosis of skin disease across skin tones," *Nature medicine,* vol. 30, pp. 573-583, 2024.

[4] V. Sampath, I. Maurtua and J. J. Aguilar , "A survey on generative adversarial networks for imbalance problems in computer vision tasks," *Journal of Big Data,* vol. 8, no. 27, 2021.

[5] R. Levi and R. Gorenstein, "AI in medicine needs to be carefully deployed to counter bias – and not entrench it," NPR, 6 June 2023. [Online]. Available: npr.org/sections/health-shots/2023/06/06/1180314219/artificial-intelligence-racial-bias-health-care. [Accessed 15 10 2023].

[6] R. Daneshjou, K. Vodrahalli, R. A. Novoa, M. Jenkins, W. Liang, V. Rotemberg, J. Ko, S. M. Swetter, E. E. Bailey, O. Gevaert, P. Mukherjee, M. Phung, K. Yekrang, B. Fong and Sahasrab, "Disparities in dermatology AI performance on a diverse, curated clinical image set," *Science Advances,* vol. 8, no. 32, 2022.

[7] M. Groh, C. Harris, L. Soenksen, F. Lau, R. Han, A. Kim, O. Koochek and Badri, "Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitzpatrick 17k Dataset," in *Computer Vision and Pattern Recognition*, 2021.

[8] N. Mehrabi, F. Morsrarrer, N. Saxena, K. Lerman and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning," *ACM Computing Surveys,* vol. 54, no. 6, 2022.

[9] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent Data Analysis,* vol. 6, no. 5, 2002.

[10] G. E. Batista, R. C. Prati and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newsletter,* vol. 6, no. 1, 2004.

[11] C. X. Ling and C. Li, "Data Mining for Direct Marketing: Problems and Solutions," in *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, New York, 1998.

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer , "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research,* vol. 16, no. 1, 2002.

[13] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: One sided selection," in *In Proceedings of the Fourteenth International Conference on Machine Learning*, Nashville, 1997.

[14] A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems,* p. 1097–1105, 2012.

[15] D. Madhugiri, "What is Data Augmentation in Image Processing?," Amygb.ai, 17 August 2022. [Online]. Available: https://www.amygb.ai/blog/what-is-data-augmentation-in-image-processing. [Accessed 12

February 2024].

[16] J. Velasco, C. Pascion, J. W. Alberio, J. Apuang, J. S. Cruz, M. A. Gomez, B. Molina, L. Tuala, A. Thio-ac and R. Jorda, "A Smartphone-Based Skin Disease Classification Using MobileNet CNN," *International Journal of Advanced Trends in Computer Science and Engineering,* vol. 8, no. 5, 2019.

[17] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative Adversarial Nets," in *NIPS*, 2014.

[18] *Purple flower isolated on transparent background PNG.* [Art]. Similar PNG.

[19] G. Douzas and F. Bacao, "Effective data generation for imbalanced learning using conditional generative adversarial networks," *Expert Systems with Applications,* vol. 91, pp. 464-471, 2018.

[20] X. Zhang, Q. Li and M. Li, "Enhancing Skin Disease Detection Accuracy and Fairness: Mitigating Biases in Dermatological Diagnosis Models," Department of Computer Science, University of Toronto, Toronto, 2023.

[21] P. Burlina, N. Joshi, W. Paul, K. D. Pacheco and N. M. Bressler, "Addressing Artificial Intelligence Bias in Retinal Diagnostics," *Translational Vision Science & Technology,* vol. 10, no. 13, 2021.

[22] Z. Zhang, M. Li and J. Yu, "On the convergence and mode collapse of GAN," in *SIGGRAPH Asia Technical Briefs*, 2018.

[23] J. Ho, A. Jain and P. Abbeel, "Denoising Diffusion Probabilistic Models," in *NIPS*, 2020.

[24] P. Dhariwal and A. Nichol, "Diffusion Models Beat GANs on Image Synthesis," in *NIPS*, 2021.

[25] M. Akrout, B. Gyepes, P. Holló, A. Póor, B. Kincsö, S. Solis, K. Cirone, J. Kawahara, D. Slade, L. Abid, M. Kovács and I. Fazekas, "Diffusion-based Data Augmentation for Skin Disease Classification: Impact Across Original Medical Datasets to Fully Synthetic Images," arXiv, 2023. [Online]. Available: https://arxiv.org/pdf/2301.04802v1.pdf.

[26] L. Sagers, J. Diao, L. Melas-Kyriazi, M. Groh, P. Rajpurkar, A. S. Adamson, V. Rotemberg, R. Daneshjou and A. K. Manrai, "Augmenting Medical Image Classifiers with Synthetic Data from Latent Diffusion Models," *arXiv,* 2023.

[27] R. Rombach, A. Blattmann, D. Lorenz, P. Esser and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," in *Computer Vision and Pattern Recognition*, 2022.

[28] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," in *Machine Learning Research*, 2021.

[29] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein and K. Aberman, "DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation," in *Computer Vision and Pattern Recognition*, 2023.

[30] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *International Conference on Learning Representations*, 2015.

[31] M. Abadi and et al., "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems," Google Research, 2015.

[32] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition,* vol. 30, no. 7, pp. 1145-1159, 1997.