Improving the Fairness of a Skin Disease Classifier Using Stable Diffusion

Olina Mukherjee
Taylor Allderdice High School
Pittsburgh, PA 15217
olina.mukherjee@gmail.com
https://github.com/olimu/fair-skin-disease-classifier-using-sd

Abstract

This study addresses the ethical implications of using neural networks for skin disease diagnosis. While such networks have demonstrated high accuracy, they tend to misdiagnose minority subpopulations because minorities are underserved and therefore underrepresented in current medical datasets. To prevent this, we augment a skin disease training dataset by adding stable diffusion-generated images depicting underrepresented dark-skinned patients with skin diseases. Additionally, prompt engineering and fine-tuning are utilized to optimize the realism and diversity of synthetic images. Notably, we achieve the first statistically significant increase in the fairness of a skin disease detector, to our knowledge. Specifically, fairness increased by 58%, as quantified by the average odds difference metric, helping to foster more equitable healthcare outcomes.

1 Introduction

Neural networks have shown promise in medical diagnosis applications. Lai et al. [1] trained a neural network on the International Skin Imaging Collective that diagnoses skin cancer with 95% accuracy. Yu et al. [2] trained a 89% accurate breast-cancer detector which took less than two seconds per diagnosis while clinicians were 30% accurate and took 314 seconds.

However, a recent Nature article found that artificially intelligent diagnostic assistants "exacerbated the gap in the diagnostic accuracy of generalists across skin tones" [3]. Due to systematic inequities in access to healthcare, people of color are underrepresented in many medical imaging datasets [4], creating race-based intraclass imbalance. Networks trained on these imbalanced datasets tend to overfit and misdiagnose underrepresented subpopulations [5], perpetuating inequities [3].

This study improves the fairness of a skin disease classifier by balancing its training dataset with stable-diffusion generated images of skin disease patients with underrepresented skin tones.

2 Related Works

2.1 Dataset

Fitzpatrick17K [6] is the largest publicly available medical dataset with annotations describing both skin tone and disease. Skin tone is evaluated using the Fitzpatrick scale, ranging from one (lightest) to six (darkest). Skin tones five and six make up less than 8% and 4% of the dataset respectively. Daneshjou et al. [7] found that networks trained on Fitzpatrick17K achieved an average AUROC score of 54% on skin tones 5 and 6, compared to 66% on tones 1 and 2. This shows how the dataset's imbalance might lead to disproportionate misdiagnosis of darker-skinned patients.

Preprint. Under review.

2.2 Methods to increase fairness

Various methods have been designed to increase neural network fairness. Some modify loss functions to optimize both fairness and accuracy. However, Xu et al. [8] note that without increasing the number of training images of minorities, networks struggle to improve accuracy in diagnosing minorities and instead sacrifice diagnostic accuracy for lighter skin tones in order to increase fairness.

Random oversampling (ROS) [9] duplicates randomly selected minority samples, balancing datasets. However, this reduces dataset diversity, leading to overfitting and decreased fairness [10, 11].

To combat overfitting, classical augmented oversampling (CAO) [12] balances datasets with diverse samples created by cropping, flipping, etc. existing samples. Velasco et al. [13] found that a skin disease classifier trained on an imbalanced dataset achieved 93.6% accuracy. ROS reduced accuracy to 91.8%, indicative of overfitting, while CAO increased it to 94.4% due to its diverse samples.

Deep-learning-based oversampling increases accuracy further by *generating* samples instead of altering existing ones, resulting in greater diversity. Generative Adversarial Networks (GANs) [14] create synthetic images from scratch with generators and filter out unrealistic images with discriminators. Douzas and Bacao [15] found that GANs outperformed CAO across 71 imbalanced datasets due to the diversity of GAN-generated images. Moreover, Burlina et al. [16] found that GAN-generated images improved the accuracy and fairness of a retina-diagnosing network. However, achieving convergence of discriminators and generators is challenging, making GANs unstable. If discriminators overfit, they reject diverse images, and if they underfit, they fail to filter out unrealistic ones [17]. Zhang et al. [18] encountered this while balancing Fitzpatrick17K by skin tone.

Unlike GANs, diffusion models [19] learn to generate images by gradually adding Gaussian noise to training images in the forward process and then learning to recover images by predicting and subtracting noise in the reverse process. Without a discriminator, mode collapse is avoided and diversity and quality of generated images is improved [20]. One pre-print [21] confirmed with clinical partners at Semmelweis University that images generated using a stable diffusion model fine-tuned on Fitzpatrick17K are realistic. Another pre-print [22] found that adding diffusion-generated images to Fitzpatrick17K, without balancing by skin tone, increased network accuracy by 11.9%.

This study applies stable diffusion [19] to balance Fitzpatrick17K by skin tone in order to increase fairness, capitalizing on stable diffusion's ability to generate realistic and diverse images even with few training images. To the best of our knowledge, this is the first study to apply stable diffusion to reduce skin tone intraclass imbalance in a medical dataset, as well as the first to achieve a statistically significant increase in fairness of a skin disease detector.

3 Methods

3.1 Dataset configuration

Fitzpatrick17K [6] is relatively small, averaging less than 11 images of skin tone six per disease. To bypass this and ensure there are enough images to train neural networks, skin tones one and two are grouped to form "light skin tone" and five and six are grouped to form "dark skin tone," as shown in Table 1, and we consider "dark skin tone" the minority class. In the future, results for the grouping of tones three and four will be found to increase fairness for patients of medium skin tones.

Only diseases with at least 40 images of both skin tones, 35 for testing and at least five for fine-tuning the diffusion model, are selected for use. This narrowed down the dataset to 11 diseases. Currently the proposed method has only been applied to three—folliculitis, lichen planus, and squamous cell carcinoma. The final dataset used is detailed in Table 1.

Table 1: Configured dataset

Disease	Train images		Test images	
	Light tone	Dark tone	Light tone	Dark tone
Folliculitis	110	7	35	35
Lichen planus	125	82	35	35
Squamous cell carcinoma	263	16	35	35

3.2 Dataset balancing

Fitzpatrick17K is balanced by skin tone using stable diffusion-generated images of skin diseases on underrepresented skin tones. The Stable Diffusion 2.0 model from Stability AI [23], pre-trained on LAION-5B [24], is fine-tuned on five Fitzpatrick17K images per skin disease/tone pairing using DreamBooth [25, 26]. This enables the generation of realistic disease images despite Fitzpatrick17K's small size. Additionally, disease name, disease description, and skin tone label are fused to create textual prompts for each fine-tuning image, e.g. "lichen planus (purplish, itchy, flat-topped bumps, lacy white patches sometimes with open sores) on skin tone 1 (light skin tone)."

After fine-tuning, prompt-engineering is used to increase image diversity. For 75% of the time, prompts only describe disease and skin tone, for 10% they also include "on an arm," for 10% they include "on a leg," and for 5% they include "on a face." Though fine-tuning images are not guaranteed to include images of these body parts, the foundational model has been pre-trained on such images.

Generated images are added to the imbalanced Fitzpatrick17K dataset until the number of images of each skin disease/tone pairing reaches the chosen threshold. Four thresholds are used: 0, 50, 100, and 150, resulting in four different datasets. In the future, higher threshold values will be tested.

3.3 Evaluation

For our evaluation, diagnostic neural networks are trained on the four datasets. As per the paper which introduced Fitzpatrick17K [6] a modified VGG-16 [27, 28] pre-trained on ImageNet is used. We code and train networks using Google Colab's free GPU resources, Keras, and TensorFlow [28]. The four networks are trained 17 times (n=17) each with different seeds and average results are reported.

After training, each network is tested on the test set created in section 3.1, and fairness is then evaluated using the following definitions. Demographic/statistical parity [29] states that classification rates should not vary for different values of a protected attribute, e.g. skin tone. This is quantified using statistical parity difference (SPD) where PR is the positive diagnosis rate of a given disease:

$$SPD = |PR_{minority} - PR_{majority}|$$

Equalized opportunity [29] states *correct* diagnosis rate should be constant across skin tones, as quantified by equal opportunity difference (EOD) which uses TPR (*true* positive rate) instead of PR:

$$EOD = |TPR_{minority} - TPR_{majority}|$$

Equalized odds [29] states that correct diagnosis *and misdiagnosis* rates should be constant across skin tones. Average odds difference (AOD) quantifies this where FPR is false positive rate:

$$AOD = \frac{(|TPR_{minority} - TPR_{majority}| + |FPR_{minority} - FPR_{majority}|)}{2}$$

These metrics measure disparity, so the smaller they are, the fairer the network. Lastly, AUROC [30, 28] scores are used to determine whether diagnostic performance is sustained.

4 Results

The generated synthetic images are diverse in appearance, due to the use of prompt engineering, as shown in Figure 1.

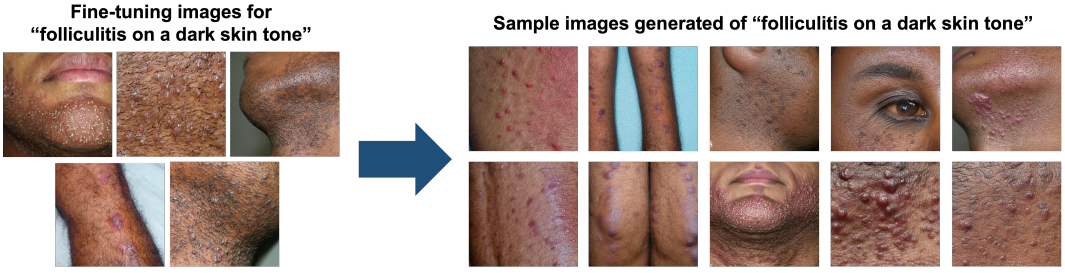

Figure 1: The five seed images used to fine-tune our diffusion model for generating images of folliculitis on a dark skin tone are on the left. Several generated images are on the right.

To evaluate synthetic images' realism, we conducted a two component principal component analysis [31], shown in Figure 2, on a dataset of all real and synthetic images. The large overlap of synthetic images with the real images indicates that they are highly realistic due to the efficacy of the diffusion model and DreamBooth. The fact that the synthetic images are more closely clustered than the real images indicates that they are less diverse as a whole.

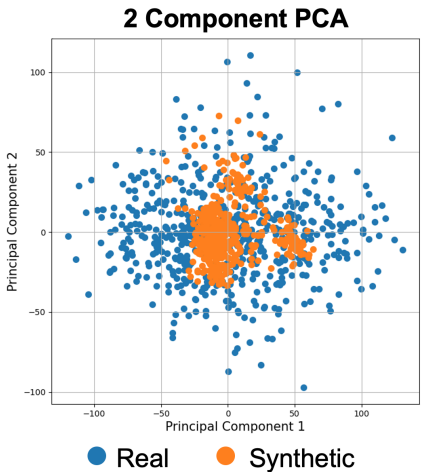

Figure 2: Each image is projected on the first two principal components of the dataset matrix.

As synthetic images are added SPD decreases by 44%, EOD decreases by 53%, and AOD decreases by 58%, shown in Figure 3, indicating increased fairness as discussed in section 3.3. Additionally, t-tests prove that these decreases are statistically significant with p values less than 0.01. As the number of synthetic images increased, AUROC scores increased by 4.5% from the control's 66.3%.

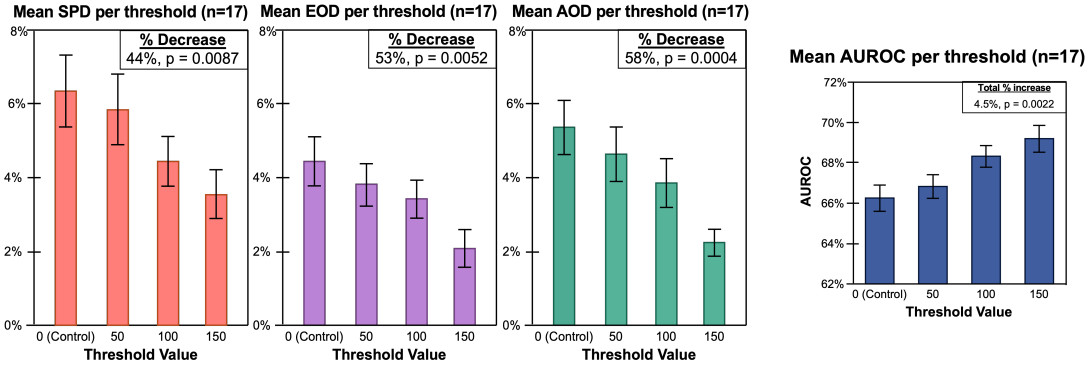

Figure 3: On the left, mean SPD, EOD, and AOD are plotted and on the right, mean AUROC scores are plotted with 95% confidence interval bars. Percent decreases from the control to the highest threshold value of 150 and corresponding p values are also provided.

5 Conclusions

In conclusion, to the best of our knowledge this is the first study to achieve a statistically significant increase in fairness of a skin disease-diagnosing neural network, making meaningful strides towards enhancing the equity of healthcare outcomes for patients of diverse backgrounds. This highlights 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 predictive policing algorithms—to increase fairness in these fields as well.

Acknowledgments and Disclosure of Funding

I would like to thank my mentor, Dr. William Cohen 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.

References

[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 Report*, 13(1), 2023. doi: 10.1038/s41598-023-45039-w.

[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*, 134(4), 2021. doi: 10.1097/CM9.0000000000001329.

[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*, 30(2):573–583, 2024. doi: 10.1038/s41591-023-02728-3.

[4] R. Levi and D. Gorenstein. AI in medicine needs to be carefully deployed to counter bias – and not entrench it. NPR, Jun 2023. URL `https://www.npr.org/sections/health-shots/2023/06/06/1180314219/artificial-intelligence-racial-bias-health-care`. Accessed: Oct 15, 2023.

[5] 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*, 8(27), 2021. doi: 10.1186/s40537-021-00414-0.

[6] 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 *Conference on Computer Vision and Pattern Recognition Workshops*, pages 1820–1828, Los Alamitos, CA, USA, Jun 2021. doi: 10.1109/CVPRW53098.2021.00201. This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License. To view a copy of this license, visit `https://creativecommons.org/licenses/by-nc-sa/3.0/`.

[7] 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, R. Sahasrabudhe, J. A. C. Allerup, U. Okata-Karigane, J. Zou, and A. S. Chiou. Disparities in dermatology AI performance on a diverse, curated clinical image set. *Science Advances*, 8(32), 2022. doi: 10.1126/sciadv.abq6147.

[8] Z. Xu, S. Zhao, Q. Quan, Q. Yao, and S. K. Zhou. FairAdaBN: Mitigating unfairness with adaptive batch normalization and its application to dermatological disease classification. In *Medical Image Computing and Computer Assisted Intervention*, pages 307–317, 2023. doi: 10.1007/978-3-031-43895-0_29.

[9] 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*, 6(1), 2004. doi: 10.1145/1007730.1007735.

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16(1):321–357, 2002. doi: 10.5555/1622407.1622416.

[11] 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*, pages 179–186, 1997. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.43.4487.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. doi: 10.1145/3065386.

[13] 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*, page 2632–2637, 2019. doi: 10.30534/ijatcse/2019/116852019.

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. doi: 10.1145/3422622.

[15] G. Douzas and F. Bacao. Effective data generation for imbalanced learning using conditional generative adversarial networks. *Expert Systems with Applications*, 91:464–471, 2018. doi: 10.1016/j.eswa.2017.09.030.

[16] P. Burlina, N. Joshi, W. Paul, K. D. Pacheco, and N. M. Bressler. Addressing artificial intelligence bias in retinal diagnostics. *Translational Vision Science & Technology*, 10(13), 2021. doi: 10.1167/tvst.10.2.13.

[17] Z. Zhang, M. Li, and J. Yu. On the convergence and mode collapse of GAN. In *SIGGRAPH Asia 2018 Technical Briefs*, 2018. doi: 10.1145/3283254.3283282.

[18] X. Zhang, Q. Li, and M. Li. Enhancing skin disease detection accuracy and fairness: Mitigating biases in dermatological diagnosis models. Technical report, Department of Computer Science, University of Toronto, 2023.

[19] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

[20] P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf.

[21] 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. Technical report, arXiv, 2023.

[22] 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. Technical report, arXiv, 2023.

[23] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. This work is licensed under a CreativeML Open RAIL++-M License. To view a copy of this license, visit https://huggingface.co/stabilityai/stable-diffusion-2/blob/main/LICENSE-MODEL.

[24] 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 *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

[25] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. This work is licensed under a MIT License. To view a copy of this license, visit `https://github.com/XavierXiao/Dreambooth-Stable-Diffusion/blob/main/LICENSE`.

[26] Huggingface, 2022. URL `https://github.com/huggingface/notebooks/blob/main/diffusers/sd_dreambooth_training.ipynb`. Accessed: Oct 15, 2023. This work is licensed under an Apache 2.0 License. To view a copy of this license, visit `https://github.com/huggingface/diffusers/blob/main/LICENSE`.

[27] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations*, 2015.

[28] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, Y. Jia M. Isard, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. Technical report, Google Research, 2015. This work is licensed under an Apache 2.0 License. To view a copy of this license, visit `https://github.com/tensorflow/tensorflow/blob/master/LICENSE`.

[29] N. Mehrabi, F. Morsrarrer, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), 2022. doi: 10.1145/3457607.

[30] A. P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997. doi: 10.1016/S0031-3203(96)00142-2.

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. This work is licensed under a BSD 3-Clause License. To view a copy of this license, visit `https://github.com/scikit-learn/scikit-learn/blob/main/COPYING`.