Enhancing Skin Disease Detection Accuracy and Fairness: Mitigating Biases in Dermatological Diagnosis Models

CSC2529 Project Proposal

Mingfei Li 1005217294 **Qianyi Li** 1010267382 **Xinran Zhang** 1010171181

November 16, 2023

1 Introduction & motivation

Suspicious skin conditions necessitate closer medical examinations as they might progress to severe skin cancers. Early diagnosis and treatment have shown to drastically increase the survival rate of skin cancer patients. To alleviate the scarcity of dermatologist professionals and speed up diagnosis of skin conditions, hospitals have been incorporate convolution neural networks (CNNs) into the diagnostic process [6] [10].

While the incorporation has undoubtedly assisted medical experts and delivered relatively accurate results for the general patient population, research has revealed that skin imaging data overwhelmingly comprises individuals with light skin tones, resulting in biases that impact the effectiveness of disease detection models.

The primary objective of this research project is to tackle this critical issue head-on. Our aim is to develop and compare various models for skin disease detection, with a particular emphasis on promoting fairness and reducing bias across different skin tones.

2 Related Works

In the work done by Torralba and Efros [9], the authors highlight that despite researchers' efforts to encapsulate a general view of the real world, the resulting datasets contain pronounced, inherited biases. This issue is especially severe in the field of medical imaging. Two significant contributions [2] [4] shed light on the stark underrepresentation of darker skin tones in medical education materials, with high percentage of dermatomyositis rash images depicting individuals with very light skin.

Moreover, our exploration has unveiled variations in skin disease diagnoses among different races and ethnicities [1]. Another pivotal study [7] has brought attention to worsening melanoma-specific survival (MSS) disparities among minority groups compared to non-Hispanic whites (NHWs), particularly since 2000. These studies points to the influence of genetic, environmental, socio-economic, and cultural factors on these disparities, underscoring the growing need for comprehensive data on disease processes in skin of color populations.

Furthermore, research conducted by Shankar et al.[8] discovered that a lack of diversity lead to noticeably unsatisfactory performance in image classification tasks for underrepresented subgroups. To combat the underrepresentation issue, several studies utilize Generative Adversarial Networks (GANs) due to its advanced capability in synthestic image generation and noise detection[5]. A related work demonstrates that the GAN-augmented image dataset can boost the classifier performance by increasing the training size. [3] We determined to incorporate these solutions into the model 3 and 4, elaborated in the section 4.

3 Overview and Goal

This project will focus on analyzing the Fitzpatrick skin image dataset, which encompasses six distinct skin tone categories. To underscore the stark imbalance within the dataset, it is noteworthy that the two lightest skin tones

account for nearly 40% of the data, while the darkest two skin tones represent only 13%. Such a pronounced disparity can have severe implications, especially for individuals with darker skin tones, as it may hinder the accuracy of disease detection.

Our overarching goal is to rectify this imbalance within the dataset and, subsequently, enhance the accuracy of our models. Importantly, we recognize that the data imbalance may not solely arise from differences in skin tones. Other potential imbalances within the dataset may also exist, and we intend to explore these nuances as part of our project.

4 Method

4.1 Methods to Solve Imbalance in Skin Tones

The methodology for this research project centers on the development and comparative analysis of three distinct models for skin disease diagnosis. Each model represents a different approach to handling and processing dermatological data, with the objective of evaluating their respective accuracies in diagnosing skin diseases across diverse skin tones. The following outlines the approach for each model:

1. **Baseline Model Development:** The first step involves creating a baseline model for skin disease detection. This model will be trained using the original Fitzpatrick dataset, which is a standard classification system for human skin color but has limitations in representing a diverse range of skin tones. The performance of this baseline model will provide a benchmark for comparison with the subsequent models.

2. **Model 2 with Simple/Naive Image Augmentation:** The second model will incorporate basic image augmentation techniques. These techniques, including cropping, flipping, and adding noise, are utilized to enhance the diversity of the dataset. The aim is to investigate whether these straightforward augmentation methods can improve the model' s accuracy in diagnosing skin diseases.

3. **Model 3 with Sophisticated Augmentation Generative Models for Synthetic Data Points:** The third model will explore the use of synthetic data. It will be trained on a combination of the original Fitzpatrick dataset and artificially generated images. These synthetic images are intended to fill the gaps in the representation of various skin tones, particularly those underrepresented in the original dataset. This model aims to test the hypothesis that synthetic data can significantly enhance the model' s diagnostic accuracy and fairness.

4. **Model 4 with Combined Simple and Sophisticated Augmentation:** The fourth model will apply both simple and sophisticated augmentation methods. This model strategically combines the merits of Models 2 and 3 to yield superior results. By leveraging a mix of fundamental and generative augmentation techniques, Model 4 aims to further enhance diagnostic accuracy and mitigate biases in skin disease detection across a wide spectrum of skin tones.

4.2 Explore Potential Imbalances

Our project also encompasses an investigation into potential imbalances within the input image data. To achieve this, we will employ vector embedding techniques to transform the image data into numerical vectors. Subsequently, we will apply clustering algorithms, such as K-Nearest Neighbors (KNN), to these vectors in order to identify clusters within the data.

If we observe the emergence of substantial and distinct clusters during this process, we will proceed to compare these clusters with the classification based on skin tones. This comparative analysis will help us determine if the clusters align with the traditional skin tone classifications.

In the event that the clustering results differ from the classification by skin tones, we may delve deeper into this potential imbalance in the future work. Our approach will involve conducting a detailed analysis, employing a logic similar to the methods previously described. This comprehensive examination will enable us to uncover and address any underlying data disparities that may impact our skin disease detection models.

5 Challenges

Data Bias: While our research focuses on the Fitzpatrick skin image dataset, it's important to acknowledge the presence of other potential imbalances and biases within this dataset. Factors such as age, gender, and sample location may also exhibit disparities that are beyond the scope of this project to address comprehensively.

Synthetic Data Limitations: In Model 3, the use of synthetic data to address dataset imbalances introduces certain limitations. The generated synthetic images may not always accurately represent the diversity of skin tones and may struggle to fully capture the complexity of various skin conditions as encountered in the real world.

Clinical Validation: Although our research may result in improved model accuracy, the importance of clinical validation cannot be overstated. While we aim to create models that achieve higher accuracy, it is essential to emphasize that the real-life applicability and effectiveness of these models must be rigorously tested in clinical scenarios. Drawing definitive conclusions about the models' performance without such validation would be premature.

Clustering Ambiguity: The process of identifying clusters within the data through vector embeddings can be subjective and may not consistently yield unequivocal results. Ambiguities in cluster interpretation may arise, requiring careful consideration and robust methodologies for cluster analysis.

6 Timeline & milestones

Nov 15: Project proposal
Nov 22: Train the first two models
Nov 29: Apply the model 3 and model 4 and compare results of these four models
Dec 6: Explore other potential imbalances and prepare the final presentation
Dec 7: Presentation

References

[1] Andrew F Alexis, Amanda B Sergay, and Susan C Taylor. Common dermatologic disorders in skin of color: a comparative practice survey. *Cutis*, 80(5):387–394, 2007.

[2] Sofia Babool, Salman F Bhai, Collin Sanderson, Amber Salter, and Lisa Christopher-Stine. Racial disparities in skin tone representation of dermatomyositis rashes: a systematic review. *Rheumatology*, 61(6):2255–2261, 2022.

[3] Hubert Cecotti and Ganesh Jha. Training dataset extension through multiclass generative adversarial networks and k-nearest neighbor classifier. In K. C. Santosh and Ravindra S. Hegadi, editors, *Recent Trends in Image Processing and Pattern Recognition*, pages 596–610, Singapore, 2019. Springer Singapore.

[4] Patricia Louie and Rima Wilkes. Representations of race and skin tone in medical textbook imagery. *Social Science & Medicine*, 202:38–42, 2018.

[5] Dipali Pattanayak and Kuntal Patel. Generative adversarial networks: Solution for handling imbalanced datasets in computer vision. In *2022 International Conference for Advancement in Technology (ICONAT)*, pages 1–6, 2022.

[6] Tri-Cong Pham, Chi-Mai Luong, Muriel Visani, and Van-Dung Hoang. Deep cnn and data augmentation for skin lesion classification. In *Intelligent Information and Database Systems: 10th Asian Conference, ACIIDS 2018, Dong Hoi City, Vietnam, March 19-21, 2018, Proceedings, Part II 10*, pages 573–582. Springer, 2018.

[7] Yingzhi Qian, Paul Johannet, Amelia Sawyers, Jaehong Yu, Iman Osman, and Judy 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*, 84(6):1585–1593, 2021.

[8] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world, 2017.

[9] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

[10] ZHE Wu, Shuang Zhao, Yonghong Peng, Xiaoyu He, Xinyu Zhao, Kai Huang, Xian Wu, Wei Fan, Fangfang Li, Mingliang Chen, et al. Studies on different cnn algorithms for face skin disease classification based on clinical images. *IEEE Access*, 7:66505–66511, 2019.