EXPOSING RACIAL DISPARITIES OF AI-BASED IMAGING CLASSIFIERS FOR PIGMENTED SKIN LESIONS DIAGNOSIS

ANDRES MORALES-FORERO
andres.morales-forero@polymtl.ca
Polytechnique Montréal

JORGE H. MAYORGA
jhmayorgaa@unal.edu.co
Universidad Nacional de Colombia
Retired Professor

LILI J. RUEDA
lijrueda@unbosque.edu.co
Universidad El Bosque

MARIA J. SANCHEZ-ZAPATA
msanchezz@unbosque.edu.co
Universidad El Bosque

ALEJANDRA JARAMILLO ARBOLEDA
ajaramilloar@unbosque.edu.co
Universidad El Bosque

SAMUEL BASSETTO
samuel-jean.bassetto@polymtl.ca
Polytechnique Montréal

ERIC COATANEA
eric.coatanea@tuni.fi
Tampere University

Abstract

Skin cancer diagnosis holds significant implications for public health, yet it is afflicted by persistent racial disparities leading to delayed detection and poorer outcomes for individuals with darker skin tones. The emergence of artificial intelligence (AI) technologies offers promising solutions; however, concerns have arisen about their biases and the potential to perpetuate them. This article presents a study to expose and scrutinize racial disparities within AI-powered imaging classifiers used to classify pigmented skin lesions. Our research focuses only on some state-of-the-art classifiers trained using the HAM10000 dataset and employs Equalized Odds analysis to assess performance metrics across different racial groups. By shedding light on these disparities, the study underscores the urgency of addressing equity issues in healthcare AI. The methodology includes data collection in Colombia and the hypothesis assessment using some fairness metrics. Even though the study is still in progress, the expected results emphasize the significance of the findings in unveiling algorithmic biases and the necessity of rectifying these imbalances. The discussion delves into the implications for healthcare equity and the broader applications of AI fairness in medical diagnostics. The insights presented contribute to the ongoing discourse on responsible AI development and equitable healthcare outcomes.

1 Introduction

Skin cancer is a significant health concern affecting individuals from various racial backgrounds. However, stark disparities exist in diagnosis and prognosis among different racial groups. Par-

ticularly, individuals with darker skin types face a disproportionate burden of delayed detection and poorer outcomes when diagnosed with skin cancer [9, 12]. Against the prevailing misconception that skin cancer exclusively affects light-skinned individuals, studies consistently demonstrated that people of color (POC) are more likely to suffer fatal consequences from this curable disease [3, 17].

The Center for Disease Control and Prevention and The American Academy of Dermatology have reported alarming differences in survival rates based on racial identity in the United States [7]. Research by [11] and [23] revealed a 5-year survival rate of 70% for non-white populations, a figure significantly lower than the 92% survival rate for whites. As demographic shifts unfold, the U.S. Census Bureau has projected that minorities, constituting 37% of the population in 2012, are anticipated to make up 57% of the population by 2060 [1]. The steady growth of the minority population, particularly black, Hispanic, and Asian Americans, foretells a demographic landscape where POC will comprise approximately 50% of the U.S. population by 2050 [10].

Melanoma, a form of skin cancer, is particularly menacing due to its potential for metastasis and fatal outcomes. Besides the expected racial gap in the socioeconomic status [3, 6], a significant contributor to the elevated mortality rates observed in this population is the bias of physicians and lack of education in POC: Disturbingly, studies have illuminated the pervasive lack of education provided to POC by healthcare practitioners regarding skin cancer risks and prevention [13, 20] and cognitive bias towards white people in the broader context of general medical diagnosis and textbooks [2, 15, 16]. Compounded by misconceptions among individuals of color that they are immune to skin cancer [12], this knowledge gap exacerbates the vulnerability of these minorities. Furthermore, limited research focusing on skin cancer within POC communities hinders the development of targeted interventions. The scarcity of medical coverage within this demographic further exacerbates the issue, resulting in reduced access to timely care and compounding the challenges of late-stage diagnoses and unfavorable prognoses.

Addressing these disparities demands innovative approaches that account for the specific challenges faced by POC in skin cancer diagnosis. While advancements in AI, particularly deep learning imaging classifiers, hold great promise for medical applications such as cancer diagnosis, concerns have emerged regarding potential biases and lack of interpretability in these algorithms [18, 21]. The oversight of data analysis by machine learning practitioners can lead to algorithmic decisions that are both unjust and unsafe. Despite claims by certain benchmark databases to represent diverse populations (e.g. [22]), their failure to provide statistics related to sensitive attributes raises concerns about their adequacy in addressing racial disparities. In dermatology, the development of AI tools for automated diagnosis using dermoscopic images has been supported by databases such as HAM 10000 [22], BCN 20000 [5], and ISIC 2020 [4]. However, the reliance on these databases and algorithms constructed upon them may inadvertently perpetuate racial discrepancies by disregarding potential underrepresentation [14, 19].

In light of these considerations, this study aims to expose and scrutinize racial disparities within state-of-the-art imaging classifiers employed for the diagnosis of pigmented skin lesions, specifically focusing on the HAM10000 dataset. By shedding light on these disparities, the study seeks to underscore the urgency of addressing equity issues in healthcare AI. The methodology adopted for this research involves the collection of dermoscopic images of lesions on both light and dark skin, alongside auxiliary information. Through the application of Equalized Odds

analysis, the study will assess the performance metrics of state-of-the-art classifiers across racial groups, evaluating potential variations in accuracy, sensitivity, specificity, and other diagnostic measures. Since the study is still in the design stage and data collection has not taken place, the final results are not available at this time. Nonetheless, the findings are expected to reveal the extent of the challenges faced by dark-skinned individuals in accurate lesion diagnosis, and thereby emphasize the critical need to rectify these disparities to ensure equitable healthcare outcomes. Beyond these specific objectives, this research aspires to contribute to broader aims. It seeks to foster a culture of inclusivity and enhancement within the medical domain, particularly in terms of bolstering the accuracy of cancer diagnosis across heterogeneous populations. Additionally, this study aims to inspire and advocate for conscientious practices among data scientists, guiding them towards approaches that are both responsible and unbiased. In essence, comprehending these disparities emerges as an indispensable cornerstone for the development of AI systems that are not only robust but also impartial and equitable.

2 Methodology

Our approach is structured as follows:

1. **Training:** We will train the algorithms described in [8] using the HAM1000 dataset.
2. **Test Set Creation:** To create our test set, we will collect dermoscopic images for each type of the seven skin lesions in Bogotá, Colombia. Along with these images, we will gather auxiliary information such as gender, phototype, and more. Dermatologists will employ the Fitzpatrick scale to classify skin types into phototypes. We maintain consistency with the setup and dermoscopy devices employed in the HAM10000 dataset.
3. **Comparison:** After training, we will apply the trained models to our test set. Subsequently, we will compare the predictive performance of the models on the two phototype groups — dark phototypes vs lighter phototypes — using metrics like accuracy, specificity, sensitivity, and two-proportion tests. To further understand the influence of phototype on prediction, we will also fit a logistic regression model.
4. **Results Analysis:** The obtained results will be carefully analyzed to understand the performance differences between the phototype groups. This analysis will provide insights into the predictive capabilities of the algorithms in different phototype contexts.

To outline our approach for conducting the analysis, let's begin by considering the scenario where we aim to evaluate the performance of the classifier f in the context of two distinct groups, with the objective of detecting melanoma — a specific type of skin lesion among the seven types present in the HAM10000 dataset. In this regard, we will establish a dichotomous variable Y that takes on two values: 1 to indicate that the image is indeed a melanoma, and 0 to signify otherwise. Similarly, we will create a dichotomous variable $\hat{Y}$ as the prediction of Y according to the classifier.

In our gathered data, skin phototypes ranging from 1 to 3 will comprise the category with lighter skin tones, while those from 4 to 6 will represent individuals with darker skin tones. The purpose is to investigate whether accurate or erroneous diagnoses are independent of ethnicity. This aims to test the concept of equalized odds:

$$P(\text{Correct or wrong diagnosis}|A) = P(\text{Correct or wrong diagnosis})$$

where $A = \{\text{Dark tone}, \text{Light tone}\}$, which corresponds to the identification of the two ethnic groups. This is equivalent to the following fairness measures:

True Positive Parity (Sensitivity):

$$P(\hat{Y} = 1|Y = 1, A = \text{“Dark tone”}) < P(\hat{Y} = 1|Y = 1, A = \text{“Light tone”})$$

True negative Parity (Specificity):

$$P(\hat{Y} = 0|Y = 0, A = \text{“Dark tone”}) < P(\hat{Y} = 0|Y = 0, A = \text{“Light tone”})$$

Accuracy Parity:

$$\begin{aligned} &P(\hat{Y} = 1|Y = 1, A = \text{“Dark tone”}) + P(\hat{Y} = 0|Y = 0, A = \text{“Dark tone”}) \\ &< P(\hat{Y} = 1|Y = 1, A = \text{“Light tone”}) + P(\hat{Y} = 0|Y = 0, A = \text{“Light tone”}) \end{aligned}$$

At the outset, empirical probabilities will be calculated to provide estimates for these conditional probabilities. Subsequently, a chi-squared test will be conducted to compare the two proportions associated with each of the three parity definitions mentioned earlier. Additional metrics like false negative and false positive rates may also be incorporated. We will perform the same analysis for each one of the seven type of the skin lesions.

Moreover, in order to account for potential confounding factors, a logistic regression model will be fitted. In this scenario, the significance of the coefficient associated with the factor A will be examined to ascertain its impact.

Sample size:

Since it is necessary to compare the proportions of performance measures for the classifiers between individuals with darker and lighter skin tones we utilize the R function *pwr.2p.test*. This function calculates the necessary sample size for a one-sided two-sample proportion test. We set a significance level of 0.05 and a power of 0.8 to evaluate an accuracy of around 0.9 for the light-skin group (determined using the HAM10000 dataset) in contrast to a potential accuracy of about 0.79 for dark-skin images. Therefore, for the purpose of identifying significant differences between these two groups, we ascertain that 130 images are required for each group and each type of skin lesion.

3 Expected results

- Successfully acquiring a sufficient number of samples for various types of skin lesions, with a particular emphasis on obtaining an ample number of malignant samples.
- Evidencing that the performance of state-of-the-art algorithms exhibits a lower level of accuracy, sensitivity, and specificity when applied to dark-skinned individuals. This observation highlights the potential disparities in algorithmic performance based on skin tone.

4 Discussion

This study aims to address the critical issue of racial disparities in skin cancer diagnosis through the lens of machine learning. The discussion will center around the significance of the expected findings, their implications for healthcare equity, and potential pathways for further research and action.

Potential disparities in skin cancer diagnosis based on skin tone underscore the urgency of rectifying these imbalances in medical algorithms. The results may shed light on the challenges faced by individuals with darker skin tones and emphasize the need for targeted interventions to ensure timely and accurate diagnoses. The fact that individuals with darker skin tones experience lower accuracy, sensitivity, and specificity in AI-based diagnostic tools raises concerns about the potential consequences of such biases. These biases might perpetuate existing healthcare disparities and exacerbate health outcomes for minority populations.

The implications of these findings extend beyond the medical domain. Addressing the limitations of AI algorithms in healthcare has broader implications for the development and deployment of machine learning models in other sectors. The results will also highlight the need for responsible and ethical data collection, preprocessing, and model training to ensure that AI systems are equitable and unbiased across diverse populations.

While the study focuses on pigmented skin lesions classification using the HAM10000 dataset, the methodology and insights can be applied to other medical conditions and datasets. Future research could explore the potential sources of bias in model performance, including data collection methods, training data sources, and model architectures. Developing mitigation strategies to rectify these biases and improve model fairness could lead to more equitable healthcare outcomes.

In conclusion, this study emphasizes the importance of addressing racial disparities in AI-based medical diagnosis. Although the study is still in the design stage and data collection has not occurred, the potential discrepancies in performance metrics among different racial groups might underscore the critical need for algorithmic fairness and equity. As AI in healthcare advances, researchers, practitioners, and policymakers must collaborate to ensure that these technologies benefit all individuals, regardless of their racial background. Ultimately, the insights from this study contribute to the ongoing dialogue surrounding the ethical and responsible use of AI in healthcare, highlighting the potential for both improved diagnostic accuracy and equitable healthcare outcomes.

References

[1] U.S. Census Bureau Newsroom. https://www.census.gov/newsroom/releases/archives/population/cb12-243.html. Accessed: 01-06-2023.

[2] Why the Color of Your Skin Can Affect the Quality of Your Diagnosis. https://www.improvediagnosis.org/dxiq-column/why-the-color-of-your-skin-can-affect-the-quality-of-your-diagnosis/. Accessed: Insert Date.

[3] BRADY, J., KASHLAN, R., RUTERBUSCH, J., FARSHCHIAN, M., AND MOOSSAVI, M. Racial disparities in patients with melanoma: a multivariate survival analysis. *Clinical, Cosmetic and Investigational Dermatology* (2021), 547–550.

[4] COLLABORATION, I. S. I., ET AL. Siim-isic 2020 challenge dataset. *International Skin Imaging Collaboration* (2020).

[5] COMBALIA, M., CODELLA, N. C., ROTEMBERG, V., HELBA, B., VILAPLANA, V., REITER, O., CARRERA, C., BARREIRO, A., HALPERN, A. C., PUIG, S., ET AL. Bcn20000: Dermoscopic lesions in the wild. *arXiv preprint arXiv:1908.02288* (2019).

[6] CORTEZ, J. L., VASQUEZ, J., AND WEI, M. L. The impact of demographics, socioeconomics, and health care access on melanoma outcomes. *Journal of the American Academy of Dermatology 84*, 6 (2021), 1677–1683.

[7] CULP, M. B., AND LUNSFORD, N. B. Peer reviewed: Melanoma among non-hispanic black americans. *Preventing Chronic Disease 16* (2019).

[8] DATTA, S. K., SHAIKH, M. A., SRIHARI, S. N., AND GAO, M. Soft attention improves skin cancer classification performance. In *Interpretability of Machine Intelligence in Medical Image Computing, and Topological Data Analysis and Its Applications for Medical Data: 4th International Workshop, iMIMIC 2021, and 1st International Workshop, TDA4MedicalData 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 4* (2021), Springer, pp. 13–23.

[9] DAWES, S. M., TSAI, S., GITTLEMAN, H., BARNHOLTZ-SLOAN, J. S., AND BORDEAUX, J. S. Racial disparities in melanoma survival. *Journal of the American Academy of Dermatology 75*, 5 (2016), 983–991.

[10] GLOSTER JR, H. M., AND NEAL, K. Skin cancer in skin of color. *Journal of the American Academy of Dermatology 55*, 5 (2006), 741–760.

[11] GOHARA, M. A. Skin cancer in skins of color. *Journal of drugs in dermatology: JDD 7*, 5 (2008), 441–445.

[12] GUPTA, A. K., BHARADWAJ, M., AND MEHROTRA, R. Skin cancer concerns in people of color: risk factors and prevention. *Asian Pacific journal of cancer prevention: APJCP 17*, 12 (2016), 5257.

[13] KIM, M., BOONE, S. L., WEST, D. P., RADEMAKER, A. W., LIU, D., AND KUNDU, R. V. Perception of skin cancer risk by those with ethnic skin. *Archives of dermatology 145*, 2 (2009), 207–208.

[14] KLEINBERG, G., DIAZ, M. J., BATCHU, S., AND LUCKE-WOLD, B. Racial underrepresentation in dermatological datasets leads to biased machine learning models and inequitable healthcare. *Journal of biomed research 3*, 1 (2022), 42.

[15] LESTER, J., JIA, J., ZHANG, L., OKOYE, G., AND LINOS, E. Absence of images of skin of colour in publications of covid-19 skin manifestations. *British Journal of Dermatology 183*, 3 (2020), 593–595.

[16] LOUIE, P., AND WILKES, R. Representations of race and skin tone in medical textbook imagery. *Social Science & Medicine 202* (2018), 38–42.

[17] MERRILL, S. J., SUBRAMANIAN, M., AND GODAR, D. E. Worldwide cutaneous malignant melanoma incidences analyzed by sex, age, and skin type over time (1955–2007): Is hpv infection of androgenic hair follicular melanocytes a risk factor for developing melanoma exclusively in people of european-ancestry? *Dermato-endocrinology 8*, 1 (2016), e1215391.

[18] MORALES-FORERO, A., BASSETTO, S., AND COATANEA, E. Toward safe ai. *AI & SOCIETY 38*, 2 (2023), 685–696.

[19] MORALES-FORERO, A., RUEDA, L., GIL-QUIÑONEZ, S., BARRERA, M., BASSETTO, S., AND COATANEA, E. An insight into racial bias in dermoscopy repositories: A ham10000 dataset analysis. Submitted to Journal of the European Academy of Dermatology and Venereology, August 2023.

[20] RIZVI, Z., KUNDER, V., STEWART, H., TORRES, P., MOON, S., LINGAPPA, N., KAZALEH, M., MALLIREDDIGARI, V., PEREZ, J., JOHN, N., ET AL. The bias of physicians and lack of education in patients of color with melanoma as causes of increased mortality: a scoping review. *Cureus 14*, 11 (2022).

[21] RUDIN, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence 1*, 5 (2019), 206–215.

[22] TSCHANDL, P., ROSENDAHL, C., AND KITTLER, H. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data 5*, 1 (2018), 1–9.

[23] WU, X.-C., EIDE, M. J., KING, J., SARAIYA, M., HUANG, Y., WIGGINS, C., BARNHOLTZ-SLOAN, J. S., MARTIN, N., COKKINIDES, V., MILLER, J., ET AL. Racial and ethnic variations in incidence and survival of cutaneous melanoma in the united states, 1999-2006. *Journal of the American Academy of Dermatology 65*, 5 (2011), S26–e1.