Fairness in Dermatology: MST-Grounded and Coverage-Aware Continuous Attribute Fairness

George Hu
Department of Computer Science
Stanford University
gehu@stanford.edu

Clara Everett
Department of Computer Science
Stanford University
ceverett@stanford.edu

Nicole Dundas
Department of Bioenginnering
UCSF and Berkeley
ndundas@berkeley.edu

Muhammad Yusuf Khattak
Department of Computer Science
Stanford University
khattmu@stanford.edu

Video Link: https://drive.google.com/file/d/1yH5b0n3XiVGRrMZnrYfkZnG8eW6K8n9b/view?usp=sharing

1 Introduction

Applying artificial intelligence to dermatology has the potential to increase the quality of care and access by either automating or providing a secondary opinion on analysis of skin lesions. Correctly determining the characteristics of skin lesions during early stages are essential for downstream treatments Hu et al. (2006), and various methodologies using machine learning techniques have been shown to be very effective Lopez et al. (2017) Zhang et al. (2019).

However, these systems have been found to perform poorly across demographic factors like race. Some have found that this can be due to inherent label bias, as dermatological practitioners producing the labels can have minimal experience viewing darker skin tone examples during medical school Hu et al. (2006). Therefore, machine learning algorithms trained perpetuate these disparities and underperform similarly to physicians when it comes to dark skin tones and uncommon lesions Daneshjou et al. (2022). These misdiagnoses in darker skin tone groups cause harm by leading to incorrect treatments and therefore poor and unfair outcomes regarding melanoma and other skin diseases. While we do not aim to directly tackle label bias directly in this paper, we aim to quantify these biases in order for future researchers to measure any disparate impacts and work towards fairer dermatology skin image analysis systems.

2 Related Work

2.1 Fairness in Dermatology

While considerations of skin tone do exist in dermatology and imaging software applications, they often rely upon socially constructed demographic data or are quite broad and coarse. The Fitzpatrick scale Andreassi et al. (1999), defining six skin tones based on studies of UV light physics and empirical responses, was historically used in various areas of dermatology, but modern applications tend to avoid the scale due to its Eurocentrism and thus inappropriate usage for many situations. Similar to the Fitzpatrick scale, the individual typology angle (ITA) scale 2 using luminance properties of perceptual lightness projection space, has been used to measure fairness, either in discrete buckets or as a continuous attribute Mary et al. (2019) Krishnapriya et al. (2022). These methods certainly improve fairness over previous baselines, but the ITA scale, like the Fitzpatrick scale, only concerns itself with a quite limited view of skin tone.

More recently, Dr. Ellis Monk, in collaboration with Google, has pioneered the Monk skin tone (MST) scale Monk (2019) which derives 10 MST orbs 1 from sociological studies, each containing a gradient possible of skin tone values. Defining fairness criteria using these orbs ***

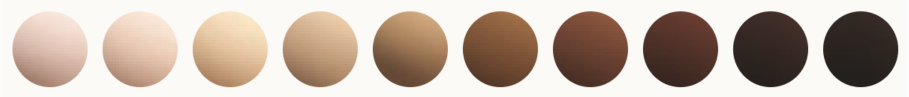


Figure 1: Orbs for skin palettes in the Monk skin tone scale

While the Monk scale is an improvement over previous algorithmic fairness considerations for skin tone, its usage is still limited to pre-defined buckets. Skin tones outside these orbs can arise from complex individual identities or medical conditions, and algorithmic fairness concerns should ensure greater universality of application Hanna et al. (2020). Therefore, we raise the central question: *If a group or individual's skin tone fails to closely match any MST orb, how should we apply fairness considerations?*

2.2 Representation Bias

Representation bias refers to the inadequacy of data across different demographic groups, and in the context of dermatology, tends to manifest as a lack of diverse data across darker skin tones. Common measures of fairness, such as equal parity, equalized odds, or calibration, all require little to no representation bias in the data Shahbazi et al. (2023); equalized odds between statistically insignificant subgroups does not really imply fairness at all. Moreover, this issue is pervasive in dermatology, as Adamson, Smith (2018) observe that even mainstream dermatology textbooks fail to meaningfully represent of different skin colors, and Kalb et al. (2023) find that common public skin cancer classification datasets crucial skin tone representation.

Within representation bias, data coverage is the subfield concerned with exacting what data exists and does not exist among selected attributes Asudeh et al. (2021). Typical data coverage metrics for datasets involve combinations of density estimation, thresholding, and/or discretizing attribute spaces. In our work, we aim to incorporate data coverage directly into the fairness metric so that we capture representation bias as well as more traditional constraints.

2.3 Correlation-Based Fairness for Continuous Attributes

Let $U \in \mathcal{U}$ and $V \in \mathcal{V}$ be univariate random variables. The Hirschfeld-Gebelein-Renyi (HGR) Maximum correlation coefficient Rényi (1959) is defined as

$$\mathrm{HGR}(U, V) = \sup_{f,g} \rho(f(U), g(V))$$

for pearson correlation function ρ and measurable functions $f : \mathcal{U} \to \mathbb{R}$ and $g : \mathcal{V} \to \mathbb{R}$ such that

$$\mathbb{E}[f(U)] = \mathbb{E}[g(V)] = 0$$

$$\mathbb{E}[f(U)^2] = \mathbb{E}[g(V)^2] = 1$$

HGR $\in [0, 1]$ is desirable in that $\mathrm{HGR}(U, V) = 0$ if and only if U and V are independent and $\mathrm{HGR}(U, V) = 1$ if and only if U and V are strictly dependent.

As a measure of fairness, the HGR maximum correlation provides nice generalization of independence-based fairness to continuous attributes, and several works Mary et al. (2019) Grari et al. (2020) have derived tractable estimation methods for HGR between attribute S and predictions P to measure and/or impose fairness for machine learning algorithms. Correlation based fairness does not rely upon the labels themselves, so their usage on novel datasets does not require expensive labeling. Mary et al. (2019) and Grari et al. (2020) also note that HGR maximum correlation provides a better fairness notion than other correlation metrics (for example pearson or spearman correlation)—HGR accounts for common scenarios such as the association between U and V being positive for small U but negative for larger U through the maps f and g.

One critique of correlation-based fairness comes from the field of causal fairness, which notes that fairness with respect to latent indicators not directly observable in attribute S can fail to be satisfied

Carey, Wu (2022), even when independence between S and P exists. We address this problem by using a multi-dimensional skin tone attribute S with interpretable dimensions grounded in sociology.

3 Model Setup

We focus our approach on skin lesion classification. Consider a dataset $\mathcal{D}$ split into train data $\mathcal{D}_{tr} = (X_{tr}, Y_{tr})$ and evaluation data $\mathcal{D}_{tst} = (X, Y)$. Our method is generally designed to be used on the evaluation data, consisting of $n \in \mathbb{R}$ images, each from a unique patient. The images are $X = \{x_i\}_{i=1}^n$, $x_i \in [0,1]^{H \times W \times 3}$, and the labels are one-hot vectors $Y = \{y_i\}_{i=1}^n$, $y_i \in \{0,1\}^c$ for c classes of lesions. We aim to evaluate a classification predictor $p : X \to [0,1]^c$ that outputs class probabilities $\sum_{j=1}^c (p(x))_j = 1$ given an input image x.

3.1 Towards a MST-Grounded Skin Tone Attribute

As alluded to before, previous work on skin-tone fairness has either focused on distinct demographic categories or the linear individual typology angle (ITA) scale Krishnapriya et al. (2022). However, these methods ignore multi-dimensional facets of skin tone variation among different peoples, and instead we pursue a method grounded in sociology with the Monk skin tone (MST) scale Monk (2019).

Recall that the MST contains 10 orbs to represent skin tone communities. We do the following procedure to determine MST-grounded skin tone attributes:

1. Let $\mathcal{O} = \{O_1, O_2, \ldots, O_{10}\}$ where each O_i are the flattened pixels for the 10 orbs. We represent the first principal component for each set of pixels in each orb as $\{v_1, \ldots, v_{10}\}$, with $v_i \in \mathbb{R}^3$. Let $V \in \mathbb{R}^{10 \times 3}$ be defined as the matrix containing the principal vectors for each orb as row vectors.
2. For each input image x_i, we use Otsu thresholding Otsu (1979) to determine the pixels corresponding to skin rather than lesion, as done in Kalb et al. (2023) and Loaiza (2020). To do this, we first convert the RGB image to HSV, then apply Otsu thresholding, and return the original RGB pixels corresponding to the background as $x_i^{(s)} \in [0,1]^{o_i \times 3}$ for o_i number of pixels returned by the Otsu mask.
3. While previous literature Kalb et al. (2023) would often take the channel-wise mean of $x_i^{(s)}$ to represent the image, we aim to capture the variation within each image, so instead we uniformly sample m pixels from each masked image $x_i^{(s)}$ to get our set of RGB skin tone attributes
$$S^{(RGB)} = \{s_{ij}^{(RGB)}\}_{1 \le i \le n, 1 \le j \le m}$$
4. For each $1 \le i \le n$ and $1 \le j \le m$, we let
$$s'_{ij} = \text{argmin}_{s \in (\mathbb{R}^+)^{10}} ||V^T s - s_{ij}^{(RGB)}||_2 + \lambda ||s||_1$$
$$s_{ij} = \frac{s'_{ij}}{||s'_{ij}||_2}$$
for some regularization parameter λ. Note that we use 1-norm regularization to induce sparsity in s_{ij} to help with interpretation. We consider $S = \{s_{ij}\}_{1 \le i \le n, 1 \le j \le m}$ as the set of all MST attribute vectors.

Our approach therefore aims to measure a classifier f trained on (X_{tr}, Y_{tr}) based upon attributes S and predictions $P = \{p_{ij} \leftarrow p(x_i)\}_{1 \le i \le n, 1 \le j \le m}$ (note that p_{ij} is the same when varying j because predictions are image level).

4 Fairness Definition

To ensure our fairness metric is coverage-aware, we use the simple formulation $f = f_c f_f$ to calculate the fairness $f \in [0,1]$ based upon the coverage factor $f_c \in [0,1]$ and the fairness factor $f_f \in [0,1]$.

4.1 MST-Grounded Coverage

We design the coverage factor f_c to be directly interpretable and grounded in the MST, with the idea that clinical practitioners can directly figure out what additional data to acquire.

Let $r > 0$ be the *coverage radius* and $t \in \mathbb{N}$ be the *coverage threshold.* Let e_a, $1 \leq a \leq 10$ be the unit vector in the a-th dimension of S. We define the set

$$M_a(r) = \{s_{ij} \in S : ||s_{ij} - e_a||_2 \leq r\}$$

to represent the set of sampled pixel MST attribute values within r distance of the MST community defined by e_a. Now let

$$f_c = 1 - 0.1 \sum_{a=1}^{10} \mathbb{1}\{|M_a(r)| \geq mt\}$$

Thus, $f_c = 0.1d$ means that d MST communities are represented and have sufficient coverage—the number of expected individuals is at least t. If for some a, $|M_a(r)| < mt$, the dataset can be made more representative by collecting $\left\lceil \frac{mt - |M_a(r)|}{m} \right\rceil$ or more relevant skin lesion images from individuals belonging to MST community a.

4.2 HGR Correlation Fairness

At a high level, we want our correlation fairness notion to ensure that no MST attribute is correlated with any output class. Thus, for maximum correlation matrix $R \in [0, 1]^{10 \times c}$, we define the fairness factor as

$$f_f = 1 - \max(R)$$

To determine R such that r_{ab} is the HGR maximum correlation coefficient between MST dimension a and class dimension b, various methods that require density estimation of S or using an upper bound for the HGR have been proposed, but we employ a method that solves for f and g directly in $\text{HGR}(U, V) = \sup_{f,g} \rho(f(U), g(V))$ (for measurable f and g with zero mean and unit variance).

Determining the optimal transformation between input variable S (in this case the attribute) and response variable P to maximize correlation can be done iteratively through the alternating conditional expectations (ACE) algorithm Breiman, Friedman (1985). Let $S^{(a)} = \{s_{ij}^{(a)} : \forall i, j\}$ be the sampled MST data points in the a-th dimension in the attribute space and $P^{(b)} = \{p_{ij}^{(a)} : \forall i, j\}$ be the corresponding predicted probability for class b. Our method determines transformation functions (f_{ab}, g_{ab}) using ACE so that

$$\text{HGR}(S^{(a)}, P^{(b)}) = \rho(f_{ab}(S^{(a)}), g_{ab}(P^{(b)})$$

For all a and all b,

- Run the ACE algorithm to get f_{ab} and g_{ab} based on $S^{(a)}$ and $P^{(b)}$. Details can be found in Appendix A.1.
- With the derived transforms f_{ab} and g_{ab}, set R_{ab} as

$$R_{ab} \leftarrow \rho(f_{ab}(S^{(a)}), g_{ab}(P^{(b)}) = \mathbb{E}[f_{ab}(S^{(a)}) g_{ab}(P^{(b)})]$$

With this, we have $f_f = 1 - \max(R)$ to capture the degree of independence between the skin tone attribute S and the predictions P.

5 Discussion and Next Steps

5.1 Discussion

The proposed methodology presents a significant advancement in addressing fairness concerns in dermatology. By grounding the skin tone attributes in a sociological scale like MST, this approach moves beyond simplistic categorizations and captures the complexity of skin tone variation among different individuals. Moreover, HGR maximum correlation-based fairness relying only on predictions avoids issues of label bias affecting fairness metrics. The incorporation of coverage-awareness ensures that the dataset represents diverse demographic groups adequately, addressing the issue of representation bias common in healthcare datasets.

However, there are certain limitations and areas for improvement in the proposed methodology. While our coverage factor can help account for and determine gaps in skin ton data coverage for an evaluation dataset, our method is not comprehensive. It can be the case that dropping some set of biased predictions in dimension a of the Monk attribute space leads to a small decrease in f_c but a large increase in f_f, which is undesirable behavior. Different arithmetic methods of combining f_c and f_f could help prevent such situations such as letting $f = f_c^{\alpha} f_f^{2-\alpha}$, for some hyperparameter $\alpha > 0$, but this explodes the hyper-parameter space even more. Our method in practice also relies upon convergence of the ACE algorithm for all a and b, which may be slow or computationally intractable. Breiman, Friedman (1985) show that convergence is guaranteed for all finite datasets, but for large complex datasets, this may require many iterations.

Additionally, while the Monk skin tone scale provides a sociology-based framework for skin tone, it might not capture all nuances of skin tone variation. Heldreth et al. (2023) have found that the Fenty scale, used in the cosmetics industry, slightly outperforms the MST scale in individual survey responses for how well a skin tone represents someone. Incorporating the finer granularity of such could potentially enhance the fairness metric's ability to capture skin tone diversity. However, this might come with the trade-off of increased computational complexity or be biased towards consumer cosmetic preferences, which warrants further investigation.

5.2 Next Steps

Empirical analyses are essential to validate the effectiveness of the proposed methodology. Conducting experiments on real-world dermatological datasets, such as the 2020 SIIM-ISIC Melanoma classification challenge Rotemberg et al. (2021), with various commonplace deep learning approaches would provide insights into how the approach performs in practice. Furthermore, exploring synthetic data scenarios could help identify potential failure modes of fairness metrics and refine the methodology accordingly.

Incorporating a direct form of interpretability and explainability techniques into the fairness metric could enhance the transparency of the model's decisions, fostering trust among stakeholders particularly when combined with real life empirical datasets. Furthermore, investigating downstream implications such as treatments resulting from diagnoses or fairness-induced access to care through clinical case studies could also improve trust and outcomes for marginalized communities.

Overall, the proposed methodology lays a strong foundation for advancing fairness in applying machine learning techniques to dermatology. Continued research and collaboration with healthcare practitioners and policymakers are crucial to ensure that the proposed approach translates into meaningful improvements in patient care and outcomes across all strata.

References

Adamson Adewole S., Smith Avery. Machine Learning and Health Care Disparities in Dermatology // JAMA Dermatology. 11 2018. 154, 11. 1247–1248.

Andreassi L., Flori M. L., Rubegni P. Sun and Skin // Rheumaderm: Current Issues in Rheumatology and Dermatology. Boston, MA: Springer US, 1999. 469–475.

Asudeh Abolfazl, Shahbazi Nima, Jin Zhongjun, Jagadish HV. Identifying insufficient data coverage for ordinal continuous-valued attributes // Proceedings of the 2021 international conference on management of data. 2021. 129–141.

Breiman Leo, Friedman Jerome H. Estimating optimal transformations for multiple regression and correlation // Journal of the American statistical Association. 1985. 80, 391. 580–598.

Buster Kesha J., Ledet Johnathan J. Photoprotection and Skin of Color // Principles and Practice of Photoprotection. Cham: Springer International Publishing, 2016. 105–124.

Carey Alycia N, Wu Xintao. The causal fairness field guide: Perspectives from social and formal sciences // Frontiers in big Data. 2022. 5. 892837.

Daneshjou Roxana, Vodrahalli Kailas, Novoa Roberto A., Jenkins Melissa, Liang Weixin, Rotemberg Veronica, Ko Justin, Swetter Susan M., Bailey Elizabeth E., Gevaert Olivier, Mukherjee Pritam, Phung Michelle, Yekrang Kiana, Fong Bradley, Sahasrabudhe Rachna, Allerup Johan A. C., Okata-Karigane Utako, Zou James, Chiou Albert S. Disparities in dermatology AI performance on a diverse, curated clinical image set // Science Advances. VIII 2022. 8, 32.

Grari Vincent, Lamprier Sylvain, Detyniecki Marcin. Fairness-Aware Neural Rényi Minimization for Continuous Features // Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20. 7 2020. 2262–2268. Main track.

Hanna Alex, Denton Emily, Smart Andrew, Smith-Loud Jamila. Towards a critical race methodology in algorithmic fairness // Proceedings of the 2020 conference on fairness, accountability, and transparency. 2020. 501–512.

Heldreth Courtney M, Monk Ellis P, Clark Alan T, Schumann Candice, Eyee Xango, Ricco Susanna. Which Skin Tone Measures are the Most Inclusive? An Investigation of Skin Tone Measures for Artificial Intelligence. // ACM Journal on Responsible Computing. 2023.

Hu Shasa, Soza-Vento Rita M., Parker Dorothy F., Kirsner Robert S. Comparison of Stage at Diagnosis of Melanoma Among Hispanic, Black, and White Patients in Miami-Dade County, Florida // Archives of Dermatology. 06 2006. 142, 6. 704–708.

Kalb Thorsten, Kushibar Kaisar, Cintas Celia, Lekadir Karim, Diaz Oliver, Osuala Richard. Revisiting Skin Tone Fairness in Dermatological Lesion Classification // Workshop on Clinical Image-Based Procedures. 2023. 246–255.

Krishnapriya KS, Pangelinan Gabriella, King Michael C, Bowyer Kevin W. Analysis of manual and automated skin tone assignments // Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2022. 429–438.

Loaiza K. The skin tone problem in artificial intelligence // 1st Congress of Women in Bioinformatics and Data Science Latin America. 2020.

Lopez Adria Romero, Nieto Xavier Giro-i, Burdick Jack, Marques Oge. Skin lesion classification from dermoscopic images using deep learning techniques // 2017 13th IASTED international conference on biomedical engineering (BioMed). 2017. 49–54.

Mary Jérémie, Calauzenes Clément, El Karoui Noureddine. Fairness-aware learning for continuous attributes and treatments // International Conference on Machine Learning. 2019. 4382–4391.

Monk Ellis. Monk Skin Tone Scale. 2019.

Otsu N. A threshold selection method from gray-level histograms, IEEE T. Syst. Man Cyb., 9, 62–66. 1979.

Rényi Alfréd. On measures of dependence // Acta mathematica hungarica. 1959. 10, 3-4. 441–451.

Rotemberg Veronica, Kurtansky Nicholas, Betz-Stablein Brigid, Caffery Liam, Chousakos Emmanouil, Codella Noel, Combalia Marc, Dusza Stephen, Guitera Pascale, Gutman David, others . A patient-centric dataset of images and metadata for identifying melanomas using clinical context // Scientific data. 2021. 8, 1. 34.

Shahbazi Nima, Lin Yin, Asudeh Abolfazl, Jagadish HV. Representation bias in data: a survey on identification and resolution techniques // ACM Computing Surveys. 2023. 55, 13s. 1–39.

Zhang Jianpeng, Xie Yutong, Xia Yong, Shen Chunhua. Attention residual learning for skin lesion classification // IEEE transactions on medical imaging. 2019. 38, 9. 2092–2103.

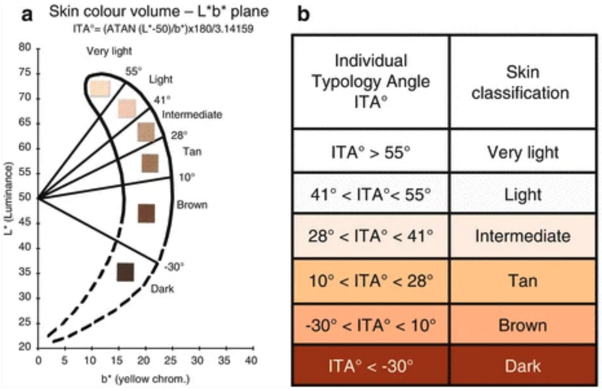
A Appendix

A.1 ACE Algorithm

For dimensions a and b (corresponding to attribute $S^{(a)}$ and class probability $P^{(b)}$) and convergence threshold $\epsilon > 0$, the ACE algorithm determines transformations f and g to maximize the correlation as follows:

1. Let $f_{ab}^{(0)}(S^{(a)}) \leftarrow \frac{S^{(a)}-\mathbb{E}[S^{(a)}]}{||S^{(a)}-\mathbb{E}[S^{(a)}]||_2}$
2. For $k = 1, 2, \ldots$:
   (a) Update $g_{ab}^{(k)}$ as:
   $$g_{ab}^{(k)}(Y^{(b)}) \leftarrow \frac{\mathbb{E}[f_{ab}^{(k-1)}(S^{(a)})|P^{(b)}]}{||\mathbb{E}[f_{ab}^{(k-1)}(S^{(a)})|P^{(b)}]||_2}$$
   (b) Update $f_{ab}^{(k)}$ as:
   $$f_{ab}^{(k)}(S^{(a)}) \leftarrow \frac{\mathbb{E}[g_{ab}^{(k)}(P^{(b)})|S^{(a)}]}{||\mathbb{E}[g_{ab}^{(k)}(P^{(b)})|S^{(a)}]||_2}$$
   (c) Stop if $\left|\mathbb{E}[f_{ab}^{(k)}(S^{(a)})g_{ab}^{(k)}(P^{(b)})] - \mathbb{E}[f_{ab}^{(k-1)}(S^{(a)})g_{ab}^{(k-1)}(P^{(b)})]\right| < \epsilon$
3. Return $f_{ab}^{(k)}, g_{ab}^{(k)}$

A.2 ITA Scale



Individual Typology Angle ITA°	Skin classification
ITA° > 55°	Very light
41° < ITA°< 55°	Light
28° < ITA° < 41°	Intermediate
10° < ITA° < 28°	Tan
-30° < ITA° < 10°	Brown
ITA° < -30°	Dark

Figure 2: ITA scale commonly used in computation for skin tone fairness. Figure from Buster, Ledet (2016)