A Mathematically Rigorous Framework for Mitigating Annotation Bias Across Fitzpatrick Skin Types in Dermatology AI

Dr. Nitin Mishra[1], Riddhi Maheshbhai Patel[2], Mansi Jadav[3], Yash Jitin Bhesania[4], Panthkumar Nileshkumar Patel[5]

[1]Professor, Department of Computer Science and Engineering Parul Institute of Engineering & Technology (PIET), Parul University Vadodara, Gujarat, India

[2,3,4,5]Department of Computer Science and Engineering Parul Institute of Engineering & Technology (PIET), Parul University Vadodara, Gujarat, India

DOI: https://doi.org/10.51244/IJRSI.2026.1307000100

Received: 12 July 2026; Accepted: 17 July 2026; Published: 30 July 2026

ABSTRACT

Deep learning models in computational dermatology frequently exhibit reduced diagnostic performance on darker skin tones (Fitzpatrick skin types IV–VI) because of annotation bias, under-representation of darker skin images, and variability in annotator expertise. This paper presents a mathematically grounded framework that aims to mitigate these challenges through three complementary components: (1) a conditional generative adversarial network (cGAN) with structural texture constraints for generating clinically realistic synthetic anchor images, (2) a skin-invariant deep metric learning strategy for separating pathological features from skin-tone characteristics, and (3) a variational Bayesian adjudication model for estimating annotator reliability across demographic groups. Rather than claiming validated clinical performance, this work provides a conceptual and mathematical foundation for developing fairer dermatology AI systems. Future validation using diverse clinical datasets and expert dermatologist assessment will be necessary to establish the framework's effectiveness in real-world healthcare environments.

Keywords: Dermatology AI, Annotation Bias, Fitzpatrick Skin Types, Generative Adversarial Networks, Metric Learning, Bayesian Adjudication, Global Health Equity.

INTRODUCTION

The deployment of deep neural networks in automated medical diagnostics has progressed rapidly, yet its benefits remain severely unbalanced across global demographics. In computer vision systems trained for diagnostic dermatology, this discrepancy manifests as a critical drop in clinical efficacy when evaluating dark skin tones, specifically Fitzpatrick skin categories IV, V, and VI [5, 12]. This limitation does not stem from intrinsic architectural failures in neural modeling, but rather from systemic structural inequalities embedded within both data acquisition pipelines and human labeling loops [8].

A primary factor driving this challenge is the sheer disparity in population-wide data representations. Globally, populations with Fitzpatrick skin types IV–VI represent more than two-thirds of the world's population, yet open clinical repositories like the International Skin Imaging Collaboration (ISIC) dataset are heavily skewed toward European and light-skinned demographics (Fitzpatrick types I–III), where over 70% of historical samples reside [17, 2]. Furthermore, clinical annotators—often trained primarily on lighter skin samples—exhibit higher variance and elevated error rates when diagnosing rare dermatological conditions on darker skin [3]. This human annotation bias is baked directly into the training datasets, leading the deep learning models to systematically misclassify malignant lesions on darker skin tissues [7].

To address these challenges without relying solely on long-term data collection efforts, this paper proposes a three-component computational framework. The framework combines conditional adversarial image generation, skin-invariant metric learning, and Bayesian annotator reliability modelling to address data imbalance and

annotation bias simultaneously. The proposed methodology is intended as a mathematically rigorous foundation for future implementation and clinical validation toward fair and inclusive dermatology AI systems.

Global Demographic Disparities And Clinical Impact

Understanding the real-world scale of this bias requires examining the global distribution of skin types. The standard clinical metric for classification is the Fitzpatrick scale, which categorizes skin tones from Type I (pale white, always burns) to Type VI (deeply pigmented, never burns) [9]. Globally, the population breakdown does not match the data inside modern medical AI datasets, creating a severe mismatch between clinical tools and the global human population.

Table 1. Global Skin Type Distribution, Regional Density, and Baseline AI Error Rates

Skin Classification Category	Fitzpatrick Types	Approx. Global Population Share (%)	Primary Regional Densities	Baseline AI Relative Error Rate
Light Skin Tones	Types I–III	~31.0%	Europe, North America, Northern Asia	Baseline (1.0x)
Medium / Olive Tones	Types IV–V	~48.5%	South Asia (India), Middle East, Latin America	1.8x – 2.4x Elevation
Dark / Deep Tones	Type VI	~20.5%	Sub-Saharan Africa, Global African Diaspora	3.1x – 4.5x Elevation

As illustrated in Table 1, over 69% of the world population falls within the Fitzpatrick IV–VI spectrum. This includes the vast majority of populations in South Asia, Africa, and Latin America. However, historical dermatology datasets reflect a completely inverted reality, where over 80% of open-source repository samples originate from light-skinned populations [12]. When an algorithm trained on such skewed data is deployed globally, its clinical impact can be highly detrimental [4]. For instance, early-stage acral lentiginous melanoma or uncommon inflammatory dermatoses present differently depending on skin pigment [14]. Missing these markers due to data bias leads to delayed diagnoses, poor treatment paths, and worse patient outcomes globally [1]. This research directly targets these issues by rewriting how data is structured and validated before it reaches clinical deployment.

PROPOSED METHODOLOGY

The proposed methodology replaces traditional uncalibrated data workflows with an integrated, three-tiered framework. This mathematical system balances the dataset, isolates disease features from skin color, and corrects for human annotator error.

Method 1: Conditional Adversarial Generation with Structural Texture Constraints

To bridge the extreme data scarcity for rare dermatological conditions on darker skin, we implement a Conditional Generative Adversarial Network (cGAN) [10]. The generator network is optimized using a combined loss function that incorporates both a classical adversarial loss and a structural texture coherence constraint. This ensures that the generated skin lesions maintain exact clinical realism without introducing artifact distortions.

Let x represent the input source image, y represent the target Fitzpatrick skin type label, and z represent a latent random noise vector. The optimization objective function for the conditional generator is formulated as:

$$\min_G \max_D L_{cGAN}(G, D) = E_{x,y}[\log D(x, y)] + E_{x,z}[\log(1 - D(G(x, z), y))] + \lambda L_{text}(G) \quad (1)$$

Where L_{text} represents a structural localized loss that measures high-frequency textural boundaries, preventing the generator from blending out important lesion borders [13]. These generated synthetic anchor images act as controlled calibration standards to benchmark the accuracy of our human annotators.

Method 2: Skin-Invariant Deep Metric Learning

Standard classification networks often bundle skin tone and pathological patterns together, leading to biased representations. To decouple these elements, we employ a deep metric learning paradigm using a modified triplet loss structure [16]. This forces the underlying ResNet-50 embedding backbone to minimize distances between identical pathologies across different skin colors, while maximizing distances between completely different pathologies.

Given an anchor sample a (e.g., Melanoma on Type II skin), a positive sample p (Melanoma on Type VI skin), and a negative sample n (Psoriasis on Type VI skin), the metric learning loss is defined as follows:

$$L_{metric} = \sum \max(0, ||f(a) - f(p)||^2 - ||f(a) - f(n)||^2 + \alpha) \quad (2)$$

Here, $f(\cdot)$ denotes the high-dimensional feature embedding map generated by the network, and α represents the strict margin enforced between distinct pathological boundaries. By optimizing this distance metric, the model learns to identify diseases independently of the surrounding melanin density [8].

Method 3: Variational Bayesian Adjudication for Annotator Confidence Weighting

Human annotators show varying levels of accuracy when classifying lesions across unfamiliar skin tones. Instead of treating all annotator opinions equally, we implement a Bayesian adjudication framework to infer the hidden reliability of each annotator cohort [6]. This reliability is modeled as a skin-type-dependent random variable.

Let A_k represent the reliability matrix of the k-th annotator, and s represent the underlying Fitzpatrick skin category. The true clinical status probability distribution $P(\theta \mid Y)$, given the observed multi-annotator label matrix Y, is solved using variational inference to compute a dynamic confidence weight:

$$P(\theta \mid Y, s) \propto P(Y \mid \theta, A_k, s) \cdot P(\theta \mid s) \quad (3)$$

By executing this variational inference step, the system dynamically decreases the voting weight of an annotator on Type VI skin images if their measured error rate on our synthetic anchor standards is high. This approach effectively shields the AI model from learning incorrect human biases.

Global Impact And Deployment Considerations

The implementation of this framework has broad implications across multiple sectors of global health and medical technology:

- Demographic Inclusion: By improving accuracy across Fitzpatrick Types IV–VI, the framework ensures that modern medical AI serves the global majority safely and effectively.
- Scalable Dataset Curation: Using synthetic anchor images drastically lowers the costs and logistical barriers of searching for rare disease data in underrepresented communities.
- Standardized Clinical Training: The generated anchor images double as high-quality training tools for medical students and clinicians globally, raising diagnostic standards across all skin tones.

However, deployment within a real-world software pipeline requires careful engineering management. Training conditional GANs with structural texture constraints demands significant computational power and computing infrastructure. Furthermore, the synthetic images must undergo strict validation by expert dermatologists to ensure absolute clinical accuracy before being integrated into active training loops.

Limitations and Future Validation

This framework is presented as a mathematical and architectural blueprint rather than a fully validated clinical system, and several limitations should guide its next stage of development. First, the specific training and

evaluation datasets, and their demographic composition, are not detailed here; without this information, it is difficult to establish how well the framework generalizes across different populations, imaging devices, and clinical settings. Second, while the conditional GAN component is intended to offset data scarcity for darker skin tones, no evidence is yet presented that the synthetic anchor images faithfully reproduce the clinical variation seen in real patients; this must be confirmed through direct, blinded comparison of synthetic and real images by experienced dermatologists before the anchors are used for annotator calibration or model training. Third, because Method 3 relies on these same anchors to score annotator reliability, any unrepresentativeness in the synthetic images would propagate into miscalibrated confidence weights, making independent clinical validation of Method 1 a prerequisite for trusting Method 3.

Future work should therefore prioritize: (i) assembling and reporting on larger, more geographically and ethnically diverse clinical datasets rather than relying primarily on synthetic augmentation; (ii) dermatologist-led validation studies comparing synthetic anchor images against real patient images for clinical fidelity; and (iii) evaluating the full framework across a broader range of dermatological conditions, disease severity levels, and healthcare deployment environments, rather than the idealized setting assumed in this paper. Addressing these points is necessary before the framework can be considered ready for real-world clinical deployment.

CONCLUSION

This paper presents a mathematically rigorous framework intended to reduce annotation bias and improve representation across Fitzpatrick skin types in dermatology AI. By integrating conditional adversarial image generation, skin-invariant deep metric learning, and variational Bayesian adjudication, the proposed methodology addresses several important challenges associated with data imbalance and annotation variability. At present, the framework should be regarded as a conceptual architecture requiring comprehensive experimental validation. Future research should evaluate the framework using large, geographically diverse clinical datasets, dermatologist-led assessment of synthetic images, and extensive benchmarking across multiple dermatological conditions. If validated, the proposed framework has the potential to contribute toward more equitable, reliable, and clinically robust AI-assisted dermatology systems.

REFERENCES

1. Adamson, A. S., & Smith, A. (2018). Machine learning and health care disparities in dermatology. Journal of the American Medical Association (JAMA) Dermatology, 154(11), 1247–1248.
2. Alipour, N. (2025). Skin Type Diversity in Image Datasets. Doctoral dissertation, Technological University Dublin.
3. Cu, C. W., et al. (2025). Validity of two subjective skin tone scales and its implications on healthcare model fairness. Journal of Medical Systems, 49(2), 112–121.
4. Daelos, R. L. (2025). Artificial intelligence predicts Fitzpatrick skin type, pigmentation, redness, and wrinkle severity from color photographs of the face. BMC Dermatology, 25(1), 45–56.
5. Daneshjou, R., et al. (2022). Disparities in dermatology AI performance on a diverse, curated clinical image set. Science Advances, 8(32), eabq6147.
6. Dawid, A. P., & Skene, A. M. (1979). Maximum likelihood estimation of observer error-rates using the EM algorithm. Journal of the Royal Statistical Society: Series C (Applied Statistics), 28(1), 20–28.
7. Dowie, T., et al. (2025). Exploring the diagnostic capability of artificial intelligence in dermatology for darker skin tones: A narrative review. Cureus, 17(3), e78412.
8. Du, S., Hers, B., Bayasi, N., Hamarneh, G., & Garbi, R. (2023). FairDisCo: Fairer AI in dermatology via disentanglement contrastive learning. Lecture Notes in Computer Science, 13848, 185–202.
9. Fitzpatrick, T. B. (1988). The validity and practicality of sun-reactive skin types I through VI. Archives of Dermatology, 124(6), 869–871.
10. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems, 27, 2672–2680.

11. Groh, M. (2022). Towards transparency in dermatology image datasets with skin tone annotations by experts, crowds, and an algorithm. Proceedings of the ACM on Human-Computer Interaction, 6(CSCW2), 1–23.
12. Groh, M., Harris, C., Soenksen, L., Lau, F., Han, R., Kim, A., Koochek, A., & Badri, O. (2021). Evaluating deep neural networks trained on clinical images in dermatology with the Fitzpatrick 17k dataset. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 1820–1828.
13. Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. IEEE Conference on Computer Vision and Pattern Recognition, 1125–1134.
14. Kinyanjui, N. M., et al. (2020). Fairness of classifiers across skin tones in dermatology datasets. International Conference on Medical Image Computing and Computer-Assisted Intervention, 451–460.
15. López-Pérez, M., Hauberg, S., & Feragen, A. (2025). Are generative models fair? A study of racial bias in dermatological image generation. arXiv preprint arXiv:2501.11752.
16. Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. IEEE Conference on Computer Vision and Pattern Recognition, 815–823.
17. Shabu, A. M., et al. (2026). A generative AI approach for reducing skin tone bias in skin cancer classification. arXiv preprint arXiv:2602.14356.