Enhancing Diagnostic Equity: A Deep Learning Framework for Fair Skin Lesion Classification

Rajeev Ranjan Dwivedi[1] RAJEEV22@IISERB.AC.IN
[1] *Indian Institute of Science Education and Research Bhopal, India*
Monika Sharma[2] MONIKA.SHARMA1@TCS.COM
[2] *TCS Research, India*
Amit Sangroya[2] AMIT.SANGROYA@TCS.COM

Editors: Under Review for MIDL 2025

Abstract

Algorithmic bias remains a critical challenge in dermatological diagnosis, especially as deep learning models often underperform for underrepresented populations. In this work, we present a novel framework that integrates bias mitigation directly into the training process for skin lesion classification. Motivated by the chronic underrepresentation of darker skin tones (Fitzpatrick types V–VI) in standard dermatology resources, our approach employs a composite loss function that jointly optimizes disease classification and skin-tone prediction. By incorporating cosine dissimilarity regularization, the method encourages the learning of disentangled, robust feature representations, while a Gradient Reversal Layer ensures that these features remain invariant to skin tone. Evaluated on both the Fitzpatrick-17k and ISIC datasets, our framework demonstrates significant improvements in fairness and accuracy, paving the way for more equitable diagnostic tools in medical imaging.

Keywords: Fairness, Skin Lesion Diagnosis, Bias Mitigation

1. Introduction

Dermatological diagnosis is a critical field where algorithmic bias can affect clinical outcomes. Deep neural networks, especially convolutional neural networks (CNNs), have excelled in classifying skin lesions—often matching experienced dermatologists (Esteva et al., 2017). However, when applied to diverse populations, these systems can exhibit systematic biases related to skin tone, gender (Daneshjou et al., 2021), surgical markings and rulers (Bevan and Atapour-Abarghouei, 2021), and other latent factors.

The underrepresentation of darker skin tones in training datasets leads to disparities in diagnostic accuracy. In standard dermatology datasets, images of darker skin (Fitzpatrick types V–VI) account a very small fraction, perpetuating bias in algorithmically trained systems (Daneshjou et al., 2021). The Fitzpatrick scale (Fitzpatrick, 1988), categorizes skin into six types based on ultraviolet response; despite its widespread use, it has been criticized for its Eurocentric bias. Nevertheless, the Fitzpatrick-17k dataset (Groh et al., 2021) remains a key benchmark by combining disease and skin-tone labels to assess fairness.

Figure 1: Fitzpatrick Scale

While several bias reduction strategies exist—such as post-processing (Hardt et al., 2016) and in-training mitigation (Zemel et al., 2013)—recent work leveraging VAE-derived latent representations (Pundhir et al., 2024) still tends to address bias through data balancing or post-hoc fixes, overlooking critical learning dynamics. Motivated by domain adaptation, we propose developing skin color-invariant disease classifiers by embedding bias mitigation directly into training. Our approach employs a composite loss that simultaneously optimizes classification and bias prediction, with a cosine dissimilarity term discouraging feature alignment by skin color, and a Gradient Reversal Layer (GRL) (Ganin et al., 2016) that reinforces the extraction of robust, disease-specific features across demographics.

2. Methodology and Dataset

We train our proposed model on dataset $\mathcal{D} = \{(x_i, y_i, b_i)\}_{i=1}^{N}$, where x_i are RGB skin images, y_i denote disease labels, and $b_i \in \{0, \ldots, 5\}$ indicate Fitzpatrick skin-tone categories (see Figure 1). The dataset is split into training and validation sets via stratification of disease label to ensure balanced representation, which minimizes further bias during training. The model begins with a shared encoder E_ϕ, that extracts feature vectors $h_i = E_\phi(x_i)$. These features are fed into two parallel branches with 3 MLP and a classification layer:

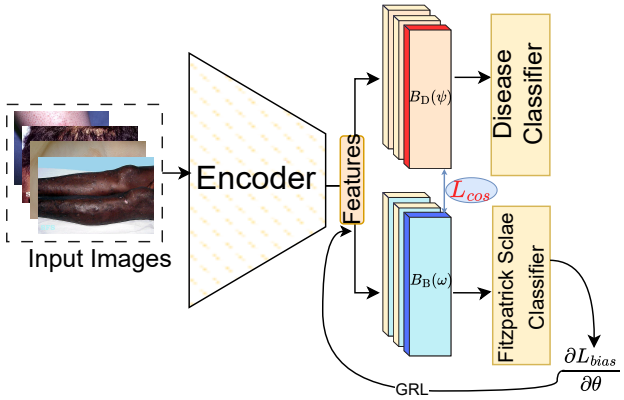

Figure 2: Proposed Model

- The **debiased branch** $B_{\mathrm{D}}(\psi)$ produces features $z_i^{(D)}$ for the disease classification head $H_{\mathrm{cls}}(\gamma)$, focusing on disease-related patterns while ignoring skin-tone cues.
- The **biased branch** $B_{\mathrm{B}}(\omega)$ outputs features $z_i^{(B)}$ for the skin-tone prediction head $H_{\mathrm{bias}}(\delta)$, thus informing the adversarial process.

A Gradient Reversal Layer (GRL) is applied before H_{bias}. The GRL acts as an identity in the forward pass but reverses and scales gradients during backpropagation, enabling E_ϕ to learn skin-tone invariant representations. In addition to this, a cosine regularization term $\mathcal{L}_{\cos}$ is applied between the penultimate layer of the two branches ($B_{\mathrm{D}}(\psi)$ and $B_{\mathrm{B}}(\omega)$), and it is intended to enforce feature disentanglement between the debiased and biased branches: The network is optimized using a composite loss:

$$\mathcal{L} = \mathcal{L}_{\text{main}} + \lambda_a \mathcal{L}_{\text{aux}} + \lambda_{\cos} \mathcal{L}_{\cos},$$

where $\mathcal{L}_{\text{main}}$ is the weighted cross-entropy loss ($\ell_{CE}(\hat{y}_i, y_i)$)for disease classification, $\mathcal{L}_{\text{aux}}$ is the cross-entropy loss for skin-tone prediction, and $\mathcal{L}_{\cos} = (1 - \overline{\cos\theta})$, where $\overline{\cos\theta} = 1/|B| \sum_{i \in B} \cos(\theta_i)$ regulates the average cosine similarity between disease features $z_i^{(d)}$ and

bias features $z_i^{(b)}$ to ensure their distinctness at the penultimate layer. λ_a and λ_{cos} are constants multiplied to control the level of regularization happening. This framework aims for discriminative and fair modelling, inspired by multi-task learning approaches (Kendall et al., 2018).

Dataset: We evaluate our method (in Figure 2) on the ISIC (Rotemberg et al., 2021) and FitzPatrick-17k (Groh et al., 2021) datasets. For ISIC, the Fitzpatrick skin type is computed using the Individual Typology Angle (ITA) method (Chardon et al., 1991). While ISIC includes eight disease classes, we restrict our analysis to three—benign, malignant, and neo-plastic—in FitzPatrick-17k. Additionally, since a dedicated validation set is not provided for FitzPatrick-17k, we create one by performing a random 20% split.

Table 1: Performance comparison on the Fitzpatrick-17k and ISIC datasets using different base encoders. T1, ⋯ T6 are Fitzpatrick types. 1

Dataset	Exp.	Base Encoder	Acc.	Acc. on Fitzpatrick Scale [1]					
				T1	T2	T3	T4	T5	T6
FitzPatrick	ERM	EfficientNet-B3	73.28	83.92	83.83	75.76	72.14	70.58	69.02
	LNTL	EfficientNet-B3	81.77	86.51	87.93	86.13	82.74	78.94	**80.42**
	Ours	EfficientNet-B3	**83.87**	**88.11**	**90.43**	**87.43**	**85.54**	**80.84**	79.52
	ERM	ResNet18	73.52	78.32	82.84	75.22	71.54	71.52	71.94
	LNTL	ResNet18	82.16	85.11	92.30	83.91	82.74	77.29	78.49
	Ours	ResNet18	**83.66**	**87.41**	**93.40**	**86.71**	**84.64**	**79.79**	**79.69**
ISIC	ERM	EfficientNet-B3	46.60	47.62	45.68	49.64	55.18	49.31	40.62
	LNTL	EfficientNet-B3	65.38	65.99	63.19	66.69	74.89	76.19	59.85
	Ours	EfficientNet-B3	**67.58**	**67.49**	**65.99**	**67.79**	**77.39**	**77.99**	**61.15**
	ERM	ResNet18	46.70	46.49	45.22	50.00	56.63	51.61	41.29
	LNTL	ResNet18	62.79	60.73	61.64	65.11	74.37	67.54	57.90
	Ours	ResNet18	**64.69**	**63.43**	**63.04**	**67.21**	**75.97**	**70.44**	**59.10**

3. Results and Discussion

In our evaluation, we compared the proposed method with baseline models (ERM and LNTL (Kim et al., 2019)) on the Fitzpatrick-17k (Groh et al., 2021) and ISIC (Rotemberg et al., 2021) datasets using EfficientNet-B3 (Tan and Le, 2019) and ResNet18 (He et al., 2015) as base encoders. Our method consistently outperformed the baselines by improving both overall disease classification accuracy and fairness across diverse Fitzpatrick skin types. As shown in Table 1 and illustrated by Figures 2, integrating cosine dissimilarity regularization with a Gradient Reversal Layer leads to balanced performance—achieving up to 83.87% accuracy on Fitzpatrick-17k and 67.58% on ISIC using EfficientNet-B3 and ResNet18. This approach not only refines the extraction of disease-relevant features but also minimizes the impact of skin-tone discrepancies, ensuring a more robust and fair classification system. Although the method depends on accurate bias annotations and requires careful tuning of the hyperparameters λ_a and $\lambda_{\cos}$, these promising results highlight its potential for broader applications in medical imaging tasks where bias labels are available or can be reliably estimated. The method overall proposes a promising direction in enhancing bias invariant diagnostic abilities.

References

Peter J Bevan and Amir Atapour-Abarghouei. Skin deep unlearning: Artefact and instrument debiasing in the context of melanoma classification. *arXiv preprint arXiv:2109.09818*, 2021.

Alain Chardon, Isabelle Cretois, and Colette Hourseau. Skin colour typology and suntanning pathways. *International journal of cosmetic science*, 13(4):191–208, 1991.

Roxana Daneshjou, Kailas Vodrahalli, Roberto A Novoa, Melissa Jenkins, Weixin Liang, Veronica Rotemberg, Justin Ko, Susan M Swetter, Elizabeth E Bailey, Olivier Gevaert, et al. Disparities in dermatology ai: Assessments using diverse clinical images, 2021.

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

Thomas B Fitzpatrick. The validity and practicality of sun-reactive skin types i through vi. *Archives of dermatology*, 124(6):869–871, 1988.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1820–1828, 2021.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. corr abs/1512.03385 (2015), 2015.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018.

Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9012–9020, 2019.

Anshul Pundhir, Sanchit Verma, and Balasubramanian Raman. Towards ethical dermatology: Mitigating bias in skin condition classification. In *IJCNN*, pages 1–8, 2024. URL https://doi.org/10.1109/IJCNN60899.2024.10650487.

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

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR, 2013.