

Fair and Accurate Skin Disease Image Classification by Alignment with Clinical Labels

Aayushman¹, Hemanth Gaddey², Vidhi Mittal², Manisha Chawla² ✉
, and Gagan Raj Gupta²

¹ Indian Institute of Science Education and Research, Bhopal, India

² Indian Institute of Technology, Bhilai, India

manishachawla96@gmail.com

Abstract. Deep learning models have achieved great success in automating skin lesion diagnosis. However, the ethnic disparity in these models' predictions needs to be addressed before deployment of these models. We introduce a novel approach: PatchAlign, to enhance skin condition image classification accuracy and fairness through alignment with clinical text representations of skin conditions. PatchAlign uses Graph Optimal Transport (GOT) Loss as a regularizer to perform cross-domain alignment. The representations thus obtained are robust and generalize well across skin tones, even with limited training samples. To reduce the effect of noise/artifacts in clinical dermatology images, we propose a learnable Masked Graph Optimal Transport for cross-domain alignment that further improves the fairness metrics.

We compare our model to the SOTA model (FairDisCo) on two skin lesion datasets with different skin types: Fitzpatrick17k and Diverse Dermatology Images (DDI). Our proposed approach, PatchAlign, enhances the accuracy of skin condition image classification by 2.8% (in-domain) and 6.2% (out-domain) on Fitzpatrick17k and 4.2% (in-domain) on DDI compared to FairDisCo. In addition, it consistently improves the fairness of true positive rates across skin tones in all of our experiments.

Keywords: Machine Learning Fairness · Cross-Domain Alignment · Graph Optimal Transport

1 Introduction

Skin is among the six most common organs invaded by cancer and early detection and treatment can significantly improve the survival rate of patients [3]. Deep learning (DL) offers a promising approach for automated skin disease detection, eliminating the need for manual segmentation and feature extraction [23,16,20,18]. However, recent research [6,1] highlights the issue of fairness in these models : patients with darker skin tones experience significantly lower diagnosis accuracy compared to those with lighter skin. To address this fairness challenge, researchers have annotated clinical skin images by skin types (tones) and released benchmark datasets like Fitzpatrick17k [19] (see Fig. 1) and DDI [10].

To achieve fairness, DL models should learn representations of skin conditions, that are independent of the skin type. This is difficult because features such as lesion color and visual features that are critical for skin lesion classification can also discriminate an individual’s skin type. Furthermore, clinical skin images often have noise and artifacts (hair, black corners, ruler, etc.) [30].

In our proposed approach (Fig.2), PatchAlign, a clinical skin image is divided into patches and their representations (embeddings) are aligned with the text embeddings of clinical labels (114 in Fitzpatrick17k and 78 in DDI dataset) (Fig. 1a). We introduce a novel cross-domain alignment technique called **Masked Graph Optimal Transport (MGOT)**, building on the **Graph Optimal Transport (GOT)** framework [8]. GOT is designed for cross-domain alignment, excelling in differentiating between various skin conditions and understanding their inter-relationships. MGOT enhances this approach by employing a learnable mask to prioritize disease-relevant patches during alignment. This not only improves accuracy over GOT but also optimizes the GOT loss function. MGOT simultaneously learns to mask irrelevant image patches and update image representations to minimize the alignment cost. Given that most clinical skin images contain areas of healthy skin, incorporating the Eudermic skin label allows MGOT to distinguish between healthy and diseased regions, further enhancing alignment accuracy.

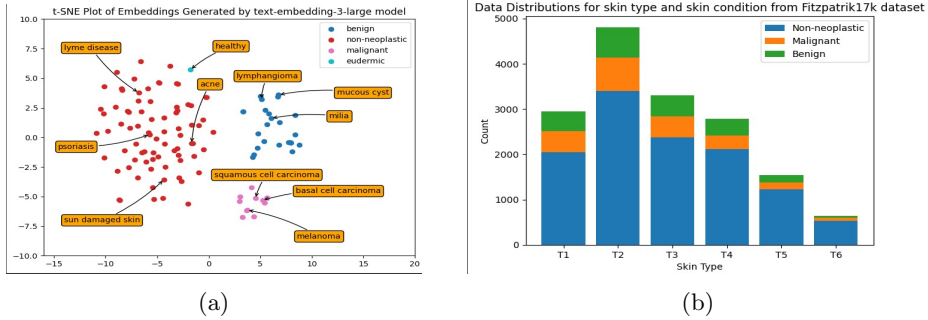


Fig. 1: (a) t-SNE plot of embeddings of 114 skin conditions in Fitzpatrick17k (b) Data distribution for skin type and skin condition from Fitzpatrick17k dataset.

Predictive Quality Disparity (PQD), Equality of Opportunity (EOM), and Demographic Disparity (DPM) are three key fairness metrics [14,15] for evaluation which promote fairness of accuracy, true positive rate, and positive rate prediction respectively across skin types.

The contributions of our paper are as follows:

- i. We propose a novel Masked Graph Optimal Transport (MGOT) method as an explicit domain alignment approach for noisy and limited clinical skin images. Our technique incorporates the Eudermic skin label to further enhance alignment accuracy.
- ii. Extensive experiments, including both in-domain and out-domain evaluations on Fitzpatrick17k and DDI benchmarks.

- iii. Our method, PatchAlign, enhances the accuracy of skin condition image classification by 2.8% (in-domain) and 6.2% (out-domain) on Fitzpatrick17k and 4.2% (in-domain) on DDI compared to SOTA model FairDisCo.
- iv. On DDI, it improves PQD by 11.4%, DPM by 2.8%, and EOM by 9.2% and on Fitzpatrick17k improves most metrics compared to FairDisCo.

2 Related Works

Recent research in skin cancer detection using deep learning [19] [15] has highlighted the crucial issue of fairness, with the latter proposing the FairDisCo framework to remove skin type bias from model representations. Additionally, [21] reviewed methods for skin tone classification, revealing limitations due to data and methodological inconsistencies. Moreover, [13] provided a broader perspective on fairness issues in deep learning, emphasizing interpretability and mitigation strategies. Efforts to address this challenge include automatic skin tone labeling, proposed by [5], and fairness-aware model architectures explored by [2]. Furthermore, [30] presented a domain generalization method (EPVT) to improve model performance across diverse settings. These related works provide a foundation for our research, which aims to contribute to the development of fair and effective skin cancer detection models.

Furthermore, unlike prior fairness-focused methods that often require extensive data normalisation [7] or additional human intervention [31], PatchAlign leverages a simpler and potentially more effective approach. Our work PatchAlign builds upon the GOT [8] where cross-domain alignment is formulated as a graph matching problem, by representing image patches and skin conditions into a dynamically-constructed graph. The learned model is sparse and interpretable. However, it gives equal importance to all skin patches during alignment. Since clinical skin images have many irrelevant patches, our proposed MGOT decreases their contribution to the cost function.

3 Model Architecture

In a multi-class skin lesion classification task with M classes, the model aims to predict a skin condition y based on an RGB skin image $\mathbf{X} \in \mathbb{R}^{H \times W \times 3}$. We consider the skin type as a sensitive attribute $s \in S$, comprising N groups with diverse types. The objective is to model $p(y|\mathbf{X})$ while minimizing the influence of s . We propose a model architecture (illustrated in Fig. 2) comprising an image input, a feature extractor ϕ , a classifier f_c , and a loss function L_c . The feature extractor ϕ processes the input image to generate a representation $\mathbf{z} = \phi(\mathbf{X})$, which is then utilized by the classifier f_c to make predictions regarding the skin condition $\mathbf{p}_c$. The model is trained using cross-entropy loss L_c .

3.1 Disentangled Representation Learning

In sensitive attribute branch we utilize disentangled representation learning by using two different losses. It minimizes a cross entropy loss L_s and confusion loss

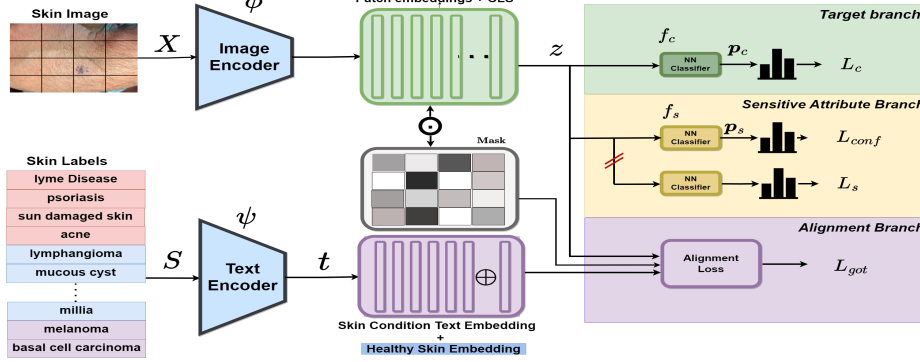


Fig. 2: PatchAlign: Our proposed alignment-based skin disease classifier

$L_{conf} = -\sum_{i=1}^N \frac{1}{N} \log(p_s^i)$ to effectively remove skin-type information from the learned representations [15] so that f_s (in Fig. 2) outputs equal probability p_s^i for all skin types i . Optimizing f_s with L_s prevents p_s to become a zero vector.

3.2 Alignment

For each batch in the data loader, we utilize OpenAI’s text-embedding-3-large model [26] (ψ in fig. 2) to encode all the clinical labels used for alignment. We also added a label for *Eudermic* skin. The alignment is performed between image patches and clinical labels leveraging **Graph Optimal Transport (GOT)**. Notably, we evaluated other text encoders, including BERT [11] and Medical BERT [28], for clinical label encoding. However, [26] achieved superior performance. As a pre-trained model, it requires a one-time computational cost for encoding and has minimal impact on overall model complexity (FLOPs).

We consider two discrete probability distributions $\mathbf{P}$ and $\mathbf{Q} \in \mathcal{P}(X)$ (for text embeddings and image embeddings) on space $X \in \mathbb{R}^d$. For dynamic construction of graph, we calculate $\{\cos(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j}$, then it is thresholded by $\max(\cos(\mathbf{x}_i, \mathbf{x}_j) - \tau, 0)$, where τ is a threshold hyper-parameter and if the value is positive an edge is added between node i and j . The **Graph Optimal Transport (GOT)** distance between $\mathbf{P}$ and $\mathbf{Q}$ is formulated as the optimization problem:

$$\mathcal{D}_{\text{GOT}}(\mu, \nu) = \min_{\mathbf{T} \in \Pi(\mathbf{u}, \mathbf{v})} \sum_{i,i',j,j'} \mathbf{T}_{ij} \left(\lambda c(\mathbf{x}_i, \mathbf{y}_j) + (1 - \lambda) \mathbf{T}_{i'j'} c(\mathbf{x}_i, \mathbf{y}_j, \mathbf{x}'_{i'}, \mathbf{y}'_{j'}) \right). \quad (1)$$

with $\mathbf{T}\mathbf{1}^m = \mathbf{u}$ and $\mathbf{T}^T\mathbf{1}^n = \mathbf{v}$, where $\mathbf{1}^m$ is the m -dimensional vector of ones. The transport cost $c_{ij} = c(\mathbf{e}_i, \mathbf{l}_j) \geq 0$ between points $\mathbf{e}_i$ and $\mathbf{l}_j$ is defined by a cost function $c(\cdot)$.

The optimal transport plan $\mathbf{T}$ is trained by minimizing the $\mathcal{D}_{\text{GOT}}$ using the iterative Sinkhorn algorithm [9,27] to solve (1) with an entropic regularizer [4] resulting in the iterative Sinkhorn algorithm to reduce the time complexity:

$$\min_{\mathbf{T} \in \Pi(\mathbf{u}, \mathbf{v})} \sum_{i=1}^n \sum_{j=1}^m \mathbf{T}_{ij} c(\mathbf{x}_i, \mathbf{y}_j) + \beta H(\mathbf{T}), \quad (2)$$

where $H(\mathbf{T}) = \sum_{i,j} \mathbf{T}_{ij} \log \mathbf{T}_{ij}$, and β is the hyper-parameter controlling the importance of the entropy term.

Masked GOT: It is the same as in eq: [1, 2] with ($\mathbf{T}$ replaced by $M \odot \mathbf{T}$) for both objective function and constraints. We generate the Mask weights from patch embeddings using a generator network with sigmoid activation function so that $M_{ij} \in [0, 1]$. The generated weights from the mask are then multiplied to patch embeddings. Thus, patches that align better with skin condition representations have higher weights than noisy patches.

Thus the final Loss function is:

$$L_{total} = L_c(\theta_\phi, \theta_{f_c}) + \alpha L_{conf}(\theta_\phi, \theta_{f_s}) + L_s(\theta_{f_s}) + \beta L_{GOT}(\theta_\phi, \theta_\psi, \mathbf{Mask}). \quad (3)$$

We use α and β to adjust contributions of confusion loss and alignment loss. Notice L_s is only used to optimize f_s .

3.3 Multi-Task Approach for learning better representations

In addition to PatchAlign, we investigate multi-task learning (MTL) [24] to learn high-quality representations for skin cancer detection. MTL involves training a model on multiple tasks simultaneously. In our case, the two tasks are: 1) predicting meta-labels and 2) predicting the skin condition.

We modify the GOT loss function from Eq. 3 for MTL. The GOT loss is replaced with a combination of two losses: a predictive cross-entropy loss for the skin conditions and a contrastive loss for the skin labels used for classification, similar to FairDisCo [15]. This approach implicitly encourages alignment between the tasks by leveraging shared representations. Here, MTL implicitly encourages alignment, whereas PatchAlign explicitly addresses alignment through the GOT loss, potentially leading to more robust model performance.

4 Implementation

4.1 Dataset Explanation

We study two skin lesion datasets: Fitzpatrick17k dataset [19] and Diverse Dermatology Images (DDI) dataset [10]. [19] contains 16,577 clinical images annotated with their Fitzpatrick17k skin-type labels which is a six-point scale initially developed for classifying sun reactivity of skin and adjusting clinical treatment according to skin phenotype [17]. There are 114 different skin conditions (clinical labels), and each one has at least 53 images. They further grouped these skin conditions into 3 categories: malignant, non-neoplastic, benign. See Fig. 1 for the data distribution. The DDI dataset contains 656 images with diverse skin types and pathologically confirmed skin condition labels, including 78 detailed disease labels and malignant identification. They grouped 6 Fitzpatrick scales into 3 groups: Fitzpatrick-12, Fitzpatrick-34, and Fitzpatrick-56, where each contains a pair of skin-type classes, i.e., $\{1,2\}$, $\{3,4\}$ and $\{5,6\}$, respectively.

4.2 Metrics

1. Predictive Quality Disparity (PQD) [14] is computed as the ratio between the lowest accuracy to the highest accuracy across each sensitive group j in S (where S is a set of skin type)

$$PQD = \frac{\min(acc_j, j \in S)}{\max(acc_j, j \in S)} \quad (4)$$

2. Demographic Disparity (DP) [14] computes the percentage diversities of positive outcomes for each sensitive group. We compute DPM across multiple skin conditions, $m \in \{1, 2, \dots, M\}$, as follows:

$$DPM = \frac{1}{M} \sum_{i=1}^M \frac{\min[p(\hat{y} = i | s = j), j \in S]}{\max[p(\hat{y} = i | s = j), j \in S]} \quad (5)$$

3. Equality of Opportunity (EO) [14] asserts that different sensitive groups should have similar true positive rates. We compute EOM across multiple skin conditions as follows:

$$EOM = \frac{1}{M} \sum_{i=1}^M \frac{\min[p(\hat{y} = i | y = i, s = j), j \in S]}{\max[p(\hat{y} = i | y = i, s = j), j \in S]} \quad (6)$$

where y is the ground-truth skin condition label and $\hat{y}$ is the model prediction. A model is fairer if it has higher values for the above three metrics. EOM, ensuring a consistent true positive rate across diverse skin types, stands out as a crucial fairness metric. In contrast, DPM focuses solely on achieving a comparable positive prediction rate, not necessarily true positive rate. Since we do not want to increase false positives, EOM is a more critical metric than DPM.

4.3 Model Training

For the Fitzpatrick17k dataset, we carry out a three-class classification and follow Groh et al. [19] in performing two experimental tasks: In-domain and Out-domain classification. The first is an in-domain classification, in which the train and test sets are randomly split in an 8:2 ratio (13261:3318). For each batch in the data loader, we utilize text-embedding-3-large model to generate embeddings of all clinical labels. In addition, we add the label *eudermic* skin because many of the patches have healthy skin. Source code is available at <https://github.com/aayushmanace/PatchAlign24>

For the DDI dataset, we perform an in-domain binary classification (malignant vs non-malignant) using the same train-test ratio of 8:2 (524:132). For all the models, we use a pre-trained ViT. Images are augmented through random cropping, rotation, and flipping to boost data diversity, then resized to $224 \times 224 \times 3$. We use Adam [25] optimizer to train the model with an initial learning rate 1×10^{-4} , which changes through a linear decay scheduler whose step size is 2, and decay factor γ is 0.8. We deploy models on a single NVIDIA RTX A6000 and train them with a batch size of 32. We set the training epochs for the Fitzpatrick17k dataset to 20 and the DDI dataset to 20.

5 Results

5.1 Overall Performance

For all the experiments, we report individual accuracies for all skin types, their average, and fairness metrics, as explained in Section 4.2 on DDI and Fitzpatrick17k Dataset. We employed several strong baseline models to ensure a comprehensive evaluation. We began with a simple pre-trained ResNet-18 architecture(BASE) as a foundation. The BASE model uses pre-trained ResNet-18 as the feature extractor and only cross-entropy loss on the final labels for optimization. We then incorporated the results from FairDisCo [15], a well-established approach, as a baseline for comparison. Finally, to demonstrate our strategy’s effectiveness, we included a multi-tasking model’s performance(MTL) as an additional baseline.

Table 1: In-Domain classification on DDI Dataset

Model	Accuracy (%) $\pm$ Std-dev.				Fairness Metrics		
	Avg	T12	T34	T56	<i>PQD</i>	<i>DPM</i>	<i>EOM</i>
BASE	82.4 ± 1.5	83.3 ± 1.0	74.6 ± 5.7	89.7 ± 2.2	77.0 ± 1.9	75.2 ± 13.3	58.7 ± 4.3
FairDisCo	83.8 ± 0.4	88.6 ± 0.1	71.7 ± 2.2	92.0 ± 2.8	78.0 ± 4.5	72.8 ± 12	63.7 ± 3.5
MTL	82.3 ± 0.4	79.4 ± 3.3	82.6 ± 3.1	85.6 ± 1.5	91.4 ± 2.7	57.7 ± 5.9	77.2 ± 4.6
PatchAlign	87.4 ± 1.2	89.6 ± 2.6	80.3 ± 5.7	92.3 ± 1.3	86.9 ± 6.1	74.9 ± 12	69.6 ± 1.7

DDI Dataset: From Table 1 we can see that PatchAlign consistently improves all metrics over FairDisCo. It improves average accuracy over FairDisCo by 4.2%, PQD by 11.4%, DPM by 2.8%, and EOM by 9.2%. Notably, since DDI contains a limited number of samples per skin condition, we conclude that PatchAlign can learn high-quality representations that are independent of the skin type.

Fitzpatrick17k Dataset: Table 2 presents a comprehensive comparison of various baseline and other existing fairness methods RESM [22], REWT[22]and ATRB[29] on Fitzpatrick17k dataset. RESM (Resampling Algorithm) organizes samples with similar skin types and conditions into unified groups and then oversamples minorities and undersamples majorities to construct a balanced dataset. REWT (Reweighting Algorithm) aims to make skin types and conditions independent of each other. ATRB (Attribute-aware Algorithm) adds skin type information to the model so that its prediction will not be dominated by major groups. From Table 2, we see a significant increase of 4.2% in overall accuracy and fairness metrics, especially DPM by 8.4% and EOM by 14.7% using PatchAlign. Moreover, we achieved the highest average test accuracy using MTL. We also note that our methods perform significantly better even on Darker skin tones that have a much smaller number of samples in the training set. Notably, PatchAlign achieves the highest accuracy (91.4%) on skin type T6, increasing the accuracy score from all other previously proposed models by more than 7%.

Table 2: In-Domain classification on Fitzpatrick17k Dataset

Model	Accuracy (%) $\uparrow$							Fairness Metrics (%)		
	Avg	T1	T2	T3	T4	T5	T6	PQD	DPM	EOM
BASE(ResNet-18)	85.0	82.6	81.7	85.0	88.8	90.7	82.8	88.9	52.5	64.0
RESM	85.1	82.5	81.8	86.2	89.0	91.2	82.3	89.0	50.7	62.0
REWT	85.6	83.4	83.0	86.2	89.0	90.6	83.5	90.8	50.0	62.9
ATRB	84.9	82.4	82.0	85.4	88.9	90.4	84.0	90.0	46.7	60.4
FairDisCo	85.1	82.2	82.1	86.2	89.4	90.0	83.2	90.2	51.2	63.8
MTL	<u>88.3</u>	86.0	86.3	88.9	91.5	91.3	<u>90.0</u>	93.7	50.5	71.2
PatchAlign	88.6	82.3	85.6	<u>89.2</u>	<u>91.4</u>	91.3	91.4	90.0	<u>55.5</u>	74.8
BASE(ViT)	86.0	83.5	82.9	86.9	90.5	90.5	84.6	89.4	47.0	59.2
BASE(ViT)+GOT	86.6	81.8	84.2	88.4	91.4	91.3	83.3	89.4	47.0	59.2
FairDisCo [#]	87.0	83.2	84.7	89.1	90.8	92.6	88.4	89.8	48.3	68.1
PatchAlign [⊙]	87.4	<u>84.1</u>	84.7	90.3	90.7	<u>90.0</u>	84.2	<u>92.7</u>	62.7	<u>72.6</u>

5.2 Ablation Study

We conduct ablation studies to analyze various components of our different models, as detailed in the last four rows of Table 2. For a fair comparison, all models utilize the pre-trained ViT [12] as the image encoder. We begin with the Baseline Model, which employs only the cross-entropy loss (BASE ViT). Subsequently, we integrate GoT with the BASE ViT, and finally, we experiment with FairDisCo using the ViT encoder. Furthermore, we have also compared with PatchAlign[⊙], which does not use *eudermic* skin type during cross-domain alignment. Upon introducing alignment between labels and patches, PatchAlign increases the average accuracy score along with the fairness metrics. We also observe that the addition of the *eudermic* label results in the highest EOM Metric. Some patches of the image that have healthy skin align with the *eudermic* label and not with the labels of other diseases, as was happening earlier. Thus, false positives decrease, in turn increasing true positives and the EO Metric. We have also done an ablation study for entropy value β and found 0.8 to yield maximum accuracy.

Table 3: Out-Domain Classification on the Fitzpatrick17k Dataset

E	Model	Accuracy							Fairness Metric		
		Avg	T1	T2	T3	T4	T5	T6	PQD	DPM	EOM
A	FairDisCo	79.5	-	-	80.2	79.4	78.1	79.0	96.9	73.5	71.2
	PatchAlign	84.0	-	-	84.2	84.0	83.1	84.6	98.3	61.7	68.2
B	FairDisCo	78.3	73.4	77.9	-	-	86.7	83.7	84.6	53.7	64.6
	PatchAlign	82.4	78.6	82.1	-	-	89.5	86.1	87.8	70.7	75.1
C	FairDisCo	71.5	65.3	68.9	73.1	80.4	-	-	81.2	59.3	77.5
	PatchAlign	77.6	74.4	75.6	78.2	83.5	-	-	89.1	74.7	78.1

5.3 Out Domain Classification on Fitzpatrick17k Dataset

We perform an out-domain classification where we train the model on all the images of two skin types and test on other skin types. From Table 3, we can see PatchAlign performs significantly better, reporting a significant increase in all three cases. Our proposed approach, PatchAlign, enhances the accuracy of skin condition image classification by 6.2% in out-domain on Fitzpatrick17k compared to FairDisCo.

6 Conclusion

In this paper, we have proposed a novel Masked Graph Optimal Transport (MGOT) as an explicit domain alignment approach for noisy and limited clinical skin images. Our method, PatchAlign, enhances the accuracy and fairness of skin condition image classification on both Fitzpatrick17k and DDI datasets compared to SOTA model FairDisCo. Significantly, PatchAlign performed well even with less data. Its performance on under-represented T56 images indicates its effectiveness in capturing and learning from diverse data distributions, which is crucial for real-world applications where such variations are common.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Adegun, A., Viriri, S.: Deep learning techniques for skin lesion analysis and melanoma cancer detection: a survey of state-of-the-art. *Artificial Intelligence Review* **54**(2), 811–841 (2021)
2. Almuzaini, A.A., Dendukuri, S.K., Singh, V.K.: Toward fairness across skin tones in dermatological image processing. In: 2023 IEEE 6th International Conference on Multimedia Information Processing and Retrieval (MIPR). pp. 1–7. IEEE (2023)
3. Beddingfield, 3rd, F.C.: The melanoma epidemic: res ipsa loquitur. *Oncologist* **8**(5), 459–465 (2003)
4. Benamou, J.D., Carlier, G., Cuturi, M., Nenna, L., Peyré, G.: Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing* **37**(2), A1111–A1138 (2015)
5. Bevan, P.J., Atapour-Abarghouei, A.: Detecting melanoma fairly: Skin tone detection and debiasing for skin lesion classification. *arXiv preprint arXiv:2202.02832* (2022)
6. Bhardwaj, A., Rege, P.P.: Skin lesion classification using deep learning. In: *Advances in Signal and Data Processing*, pp. 575–589. Springer (2021)
7. Bissoto, A., Valle, E., Avila, S.: Debiasing skin lesion datasets and models? not so fast. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 3192–3201 (2020). <https://doi.org/10.1109/CVPRW50498.2020.00378>
8. Chen, L., Gan, Z., Cheng, Y., Li, L., Carin, L., Liu, J.: Graph optimal transport for cross-domain alignment. *ArXiv abs/2006.14744* (2020), <https://api.semanticscholar.org/CorpusID:220128101>

9. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems* **26** (2013)
10. Daneshjou, R., Vodrahalli, K., Liang, W., Novoa, R.A., Jenkins, M., Rotemberg, V., Ko, J., Swetter, S.M., Bailey, E.E., Gevaert, O., et al.: Disparities in dermatology ai: Assessments using diverse clinical images. *arXiv preprint arXiv:2111.08006* (2021)
11. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: *North American Chapter of the Association for Computational Linguistics* (2019), <https://api.semanticscholar.org/CorpusID:52967399>
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale (2021)
13. Du, M., Yang, F., Zou, N., Hu, X.: Fairness in deep learning: A computational perspective. *IEEE Intelligent Systems* **36**(4), 25–34 (2020)
14. Du, M., Yang, F., Zou, N., Hu, X.: Fairness in deep learning: A computational perspective. *IEEE Intelligent Systems* **36**(4), 25–34 (2021). <https://doi.org/10.1109/MIS.2020.3000681>
15. Du, S., Hers, B., Bayasi, N., Hamarneh, G., Garbi, R.: Fairdisco: Fairer ai in dermatology via disentanglement contrastive learning. In: *European Conference on Computer Vision*. pp. 185–202. Springer (2022)
16. Esteve, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. *nature* **542**(7639), 115–118 (2017)
17. Fitzpatrick, T.B.: The validity and practicality of sun-reactive skin types i through vi. *Archives of dermatology* **124**(6), 869–871 (1988)
18. Gessert, N., Nielsen, M., Shaikh, M., Werner, R., Schlaefer, A.: Skin lesion classification using ensembles of multi-resolution efficientnets with meta data. *MethodsX* **7**, 100864 (2020)
19. Groh, M., Harris, C., Soenksen, L., Lau, F., Han, R., Kim, A., Koochek, A., Badri, O.: 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 (CVPR)*. pp. 1820–1828 (2021)
20. Harangi, B.: Skin lesion classification with ensembles of deep convolutional neural networks. *Journal of biomedical informatics* **86**, 25–32 (2018)
21. Kalb, T., Kushibar, K., Cintas, C., Lekadir, K., Diaz, O., Osuala, R.: Revisiting skin tone fairness in dermatological lesion classification. In: *Workshop on Clinical Image-Based Procedures*. pp. 246–255. Springer (2023)
22. Kamiran, F., Calders, T.: Data pre-processing techniques for classification without discrimination. *Knowledge and Information Systems* **33** (10) (2011). <https://doi.org/10.1007/s10115-011-0463-8>
23. Kawahara, J., BenTaieb, A., Hamarneh, G.: Deep features to classify skin lesions. In: *2016 IEEE 13th international symposium on biomedical imaging (ISBI)*. pp. 1397–1400. IEEE (2016)
24. Kawahara, J., Daneshvar, S., Argenziano, G., Hamarneh, G.: Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE journal of biomedical and health informatics* **23**(2), 538–546 (2018)
25. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. *ICLR* (2015)
26. Neelakantan, A., Xu, T., Puri, R., Radford, A., Han, J.M., Tworek, J., Yuan, Q., Tezak, N., Kim, J.W., Hallacy, C., Heidecke, J., Shyam, P., Power, B., Nekoul, T.E.,

- Sastry, G., Krueger, G., Schnurr, D., Such, F.P., Hsu, K., Thompson, M., Khan, T., Sherbakov, T., Jang, J., Welinder, P., Weng, L.: Text and code embeddings by contrastive pre-training (2022)
27. Peyré, G., Cuturi, M., et al.: Computational optimal transport. Center for Research in Economics and Statistics Working Papers (2017-86) (2017)
 28. Rasmy, L., Xiang, Y., Xie, Z., Tao, C., Zhi, D.: Med-bert: pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction. CoRR **abs/2005.12833** (2020), <https://arxiv.org/abs/2005.12833>
 29. Xu, T., White, J., Kalkan, S., Gunes, H.: Investigating bias and fairness in facial expression recognition (2020)
 30. Yan, S., Liu, C., Yu, Z., Ju, L., Mahapatra, D., Mar, V., Janda, M., Soyer, P., Ge, Z.: Epvt: Environment-aware prompt vision transformer for domain generalization in skin lesion recognition. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 249–259. Springer (2023)
 31. Yan, S., Yu, Z., Zhang, X., Mahapatra, D., Chandra, S.S., Janda, M., Soyer, P.H., Ge, Z.: Towards trustable skin cancer diagnosis via rewriting model’s decision. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 11568–11577 (2023), <https://api.semanticscholar.org/CorpusID:257279770>