Fitzpatrick Thresholding for Skin Image Segmentation

Duncan Stothers[1 [0000−0001−6873−851X]]✉, Sophia Xu[3], Carlie Reeves[4], and Lia Gracey[2 [0009−0007−6037−2518]]

[1] Independent Researcher, San Francisco, CA, USA *
duncanstothers@alumni.harvard.edu
[2] Vagelos College of Physicians and Surgeons, Columbia University, New York, NY, USA sx2400@cumc.columbia.edu
[3] University of Mississippi Medical Center, Jackson, MS, USA CReeves4@umc.edu
[4] Division of Dermatology, The University of Texas at Austin, Dell Medical School, Austin, Texas lia.gracey@austin.utexas.edu

Abstract. Accurate estimation of the body surface area (BSA) involved by a rash, such as psoriasis, is critical for assessing rash severity, selecting an initial treatment regimen, and following clinical treatment response. Attempts at segmentation of inflammatory skin disease such as psoriasis perform markedly worse on darker skin tones, potentially impeding equitable care. We assembled a psoriasis dataset sourced from six public atlases, annotated for Fitzpatrick skin type, and added detailed segmentation masks for every image. Reference models based on U-Net, ResU-Net, and SETR-small are trained without tone information. On the tuning split we sweep decision thresholds and select (i) global optima and (ii) per Fitzpatrick skin tone optima for Dice and binary IoU. Adapting Fitzpatrick specific thresholds lifted segmentation performance for the darkest subgroup (Fitz VI) by up to +31 % bIoU and +24 % Dice on UNet, with consistent, though smaller, gains in the same direction for ResU-Net (+25 % bIoU, +18 % Dice) and SETR-small (+17 % bIoU, +11 % Dice). Because Fitzpatrick skin tone classifiers trained on Fitzpatrick-17k now exceed 95 % accuracy, the cost of skin tone labeling required for this technique has fallen dramatically. Fitzpatrick thresholding is simple, model-agnostic, requires no architectural changes, no re-training, and is virtually cost free. We demonstrate the inclusion of Fitzpatrick thresholding as a potential future fairness baseline.

Keywords: Fitzpatrick · Psoriasis · Segmentation · BSA

1 Background

1.1 Significance

Skin rashes remain one of the most frequent reasons for new primary-care encounters, accounting for more than 13 million office visits annually in the United

* Present address: Vancouver, Canada

States and rising [3]. Diagnostic accuracy is unevenly distributed: both practicing dermatologists and trainees perform noticeably worse on images of darker skin tones [4, 8]. The accurate assessment of body surface area (BSA) affected by skin conditions, such as rashes, is crucial for clinical decision-making. Yet, physicians still rely on the outdated "1 palm = 1 percent BSA" method where BSA involved with a rash is estimated using the patient's palm size. This subjective measurement can lead to under- or over-treatment in the clinic. Additionally, a minimum threshold of BSA involvement is a criterion for payors in insurance coverage decisions, which makes accurate calculations imperative for a patient to be eligible for more advanced biologic treatments and for following treatment response. More specifically, BSA is an important calculation in the widely used Psoriasis Area and Severity Index that is most often deployed in clinical trial settings to assess baseline and treatment response for new therapeutics; these common measures are subjective and prone to human error [2]. No widely used tools exist to automate these important assessments in all skin types[16, 21, 26]. Any systematic error in segmenting lesions on dark skin therefore propagates directly into PASI scores, treatment eligibility, and ultimately patient outcomes.

1.2 Previous Work

Early ISIC Analyses Highlight Tone Bias The first wave of ISIC challenge papers demonstrated that convolutional networks trained almost exclusively on Fitzpatrick I–III images attained dermatologist-level accuracy on similarly light-skinned test sets, yet their performance degraded noticeably on darker tones [13, 11]. Follow-up studies on ISIC 2018, Fitzpatrick-17k, and DDI quantified AUROC and sensitivity gaps of 10–35 pp favoring light skin [13, 11, 5]. The consensus emerging from this literature is that distributional shift in pigmentation, not just lesion morphology, drives a substantial share of the error.

From Complex Debiasing Schemes to Stratified Operating Points Most responses to the documented bias have focused on sophisticated data- or model-centric fixes—balanced resampling, adversarial representation learning, group-adaptive batch normalization, or fairness-guided pruning [17, 25, 24]. A conceptually simpler alternative, rooted in the equalized-odds post-processing of Hardt's 2016 approach [12], is to select a *separate decision threshold* for each Fitzpatrick group so that error rates align across tones.

The FPR–TPR Trade-off in Binary Classification Applying stratified thresholds to binary classification is not trivial: raising sensitivity for an under-served group often worsens its false-positive rate, and—by impossibility results—one cannot simultaneously satisfy perfect calibration and equalized odds once prevalence differs [14, 18]. Consequently, dermatology researchers have tended to pursue fairness during training [17, 25, 24], where the utility–equity trade-off is perceived as more controllable, rather than post-hoc calibration.

Segmentation: A Setting Where Tone-Specific Optima Exist Segmentation changes the landscape. Each image yields a dense probability map, and there is, in principle, a threshold that maximizes Dice or bIoU for every subgroup. If the score distributions for Fitzpatrick V–VI are shifted left—as empirical histograms repeatedly show [1] —a universal cut-off under-segments dark skin. Calibrating per-tone thresholds can therefore improve both subgroup Dice and *overall* performance, because each group operates closer to its own theoretical optimum. This observation motivates the present study, which evaluates Fitzpatrick-specific thresholding in the clinically consequential task of psoriasis BSA estimation.

1.3 Clinical Relevance of Precise BSA Estimation

Psoriasis management provides an ideal test-bed for tone-aware segmentation because small changes in the BSA assessment directly translate to different treatment pathways. The PASI scoring rubric weights percent-involved BSA in each anatomical region; a 5–10 percentage-point error may erroneously move a patient into a different disease severity category. In a recent review of machine learning BSA estimators, skin tone discussion was omitted from all segmentation approaches [15], with the sole exception that in one study it was shown that error modes exist where healthy darker skin regions are sometimes mis-classified as lesional [10]. Demonstrating that Fitzpatrick-specific thresholding can reduce this bias would offer a pragmatic, model-agnostic fairness intervention with potential uses both in clinical trials and photo-based tele-dermatology.

2 Methodology

2.1 Data Collection

We assembled a large publicly available psoriasis dataset by sourcing from six open dermatology repositories: Derm Atlas Brazil [20], DermIS [7], DermNet NZ [22], the Hellenic Dermatology Atlas [6], the Interactive Dermatology Atlas [23], and Fitzpatrick-17k [11]. Subtypes of psoriasis that were excluded included pustular variants and isolated nail disease, filtered out by keyword rules and manual dermatologist review. Duplicates were removed, and patient IDs were assigned to prevent leakage between train, tune, and test sets. The final dataset contained 754 psoriasis images from 631 patients.

2.2 Skin-Tone Annotation and Segmentation Labels

Each retained image was independently labeled with a Fitzpatrick type (I–VI) by a board-certified dermatologist. Pixel-level diseased-skin masks were produced using the VIA Image Annotator tool [9]. Three assistants (medical student, resident physician, and graduate research assistant) drew initial polygon masks; a board-certified MD–PhD dermatologist specializing in psoriasis revised every mask to ensure high quality segmentation masks, especially on difficult cases such as low contrast lesions on darker skin tones.

Fig. 1: Examples of high detail manual skin-disease labeling employed in the study.

Table 1: Per-dataset image counts by Fitzpatrick skin-type

Dataset	**I**	**II**	**III**	**IV**	**V**	**VI**	**Total**
Brazil	0	2	29	40	40	4	115
DermIS	22	67	2	3	3	0	97
DermNetNZ	27	171	68	32	7	5	310
Fitzpatrick 17k	35	57	8	11	7	1	119
Hellenic	1	29	16	6	0	0	52
Interactive	0	24	22	4	5	6	61
Column total	**85**	**350**	**145**	**96**	**62**	**16**	**754**

2.3 Data Split

Patient-level IDs were stratified by Fitzpatrick skin tone and, within each stratum, randomly permuted with a fixed seed (0). Stratified samples were then allocated to the training, tuning, and held-out test sets in a 30 / 30 / 40 proportion, ensuring balanced skin-tone representation and complete patient independence across partitions.

2.4 Model Architecture and Training Protocol

We benchmark three architectures chosen to represent successive stages in semantic segmentation design while remaining practical for a single-GPU medical study. U-Net [19] is the canonical encoder–decoder CNN against which most dermatology work is still compared. Our 256 × 256 implementation (four down-sampling stages, two 3 × 3 convs per block, batch-norm everywhere) contains 31.1 trainable parameters and therefore serves as a strong, yet widely recognizable, baseline. Residual U-Net [27] keeps the same overall topology and feature widths but replaces each plain block with a pre-activation residual pair plus a squeeze-and-excite (SE) channel attention gate. These lightweight additions raise the capacity only marginally to 33.1 M parameters, letting us test whether better optimization and local attention alone can reduce skin-tone bias. Finally, a 21M parameter reference implementation of SETR-small [28] swaps the convolutional encoder for a ViT-S/16 backbone (12 transformer layers, 6 heads, 384-D embeddings; positional tokens only) followed by a one-layer up sampling head. Because the encoder is fully self-attentional and translation-equivariant

only after training, SETR probes whether long-range context helps fairness on our limited psoriasis corpus. All three networks are trained with identical 256 $\times$ 256 inputs, vanilla SGD optimization with 0.9 momentum, learning rates of 0.01 for UNet and ResUNet (0.0004 for SETR for stability), an identical simple flip, rotate, and jitter data augmentation scheme, early-stopping watching the validation bIoU with a patience setting of 15, and an identical 3:1 weighted binary cross entropy + dice loss; thus any performance differences are most likely attributed to (i) architectural choice and (ii) the use of Fitz-specific versus global operating points, rather than to confounding hyper-parameter changes.

2.5 Operating-Point Search

To evaluate threshold sensitivity we swept decision cut-offs $\tau \in [0.001, 0.99]$ in steps of 0.001 on the *tuning* split and computed binary Intersection-over-Union (bIoU) and Dice:

$$\mathrm{Dice}(\tau) = \frac{2\,|\hat{M}(\tau) \cap M|}{|\hat{M}(\tau)| + |M|}, \quad \mathrm{bIoU}(\tau) = \frac{|\hat{M}(\tau) \cap M|}{|\hat{M}(\tau) \cup M|},$$

where M is the ground-truth mask and $\hat{M}(\tau) = \{\, p \geq \tau \,\}$. We recorded:

- Two *overall* optima, $\tau_{\mathrm{all}}^{\mathrm{Dice}}$ and $\tau_{\mathrm{all}}^{\mathrm{bIoU}}$, maximising performance across the entire tuning set.
- Twelve *tone-stratified* optima, $\tau_g^{\mathrm{Dice}}, \tau_g^{\mathrm{bIoU}}$ for $g \in \{\mathrm{I}, \ldots, \mathrm{VI}\}$, each maximising the metric within its Fitzpatrick subgroup.

All operating points were then frozen and evaluated once on the unseen test set to quantify gains from tone-specific calibration.

3 Results

Visual confirmation. Figure 2 plots validation bIoU and dice versus threshold for each tone. The curves illustrate a consistent left-shift for Fitzpatrick VI, explaining why the universal cut-off under-segments darker skin. The arrows mark the tone-specific optima chosen during calibration; note that lighter tones cluster around the global optimum, whereas tone VI requires substantially lower thresholds to maximize Dice.

Figure 3 illustrates example Fitz VI images that visually illustrate how the lower Fitz VI optimized operating points captures significantly more of the diseased skin than the globally optimized operating point which is dominated by lighter skin tones.

Quantitative improvements of Fitzpatrick thresholding In Table **??** we can see applying a single global threshold already yields reasonable performance for all three networks, but re-tuning the operating point for each Fitzpatrick subgroup uncovers systematic gains that disproportionately benefit the darkest

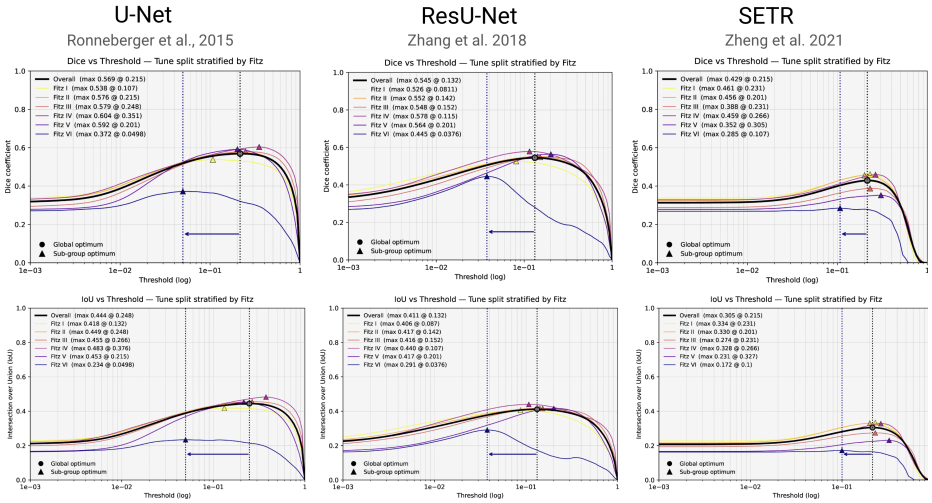

Fig. 2: Data plotted from the validation set. Arrows highlight that the optimal operating points for Fitz VI is consistently and significantly lower than for other Fitzpatrick tones. This observation is consistent between architectures: from all-conv UNet to all-attention SETR, as well as between metrics Dice (top row) and bIoU (bottom row). Fitz tones I-V have optimal operating points around the aggregate overall optimum which is shown in black.

Table 2: Segmentation performance by skin-tone subset. τ_g: global threshold; τ_{F}: Fitz-specific threshold.

		U-Net			**ResU-Net**			**SETR-small**		
Metric	**Subset**	τ_g	τ_{F}	Δ (%)	τ_g	τ_{F}	Δ (%)	τ_g	τ_{F}	Δ (%)
Dice	Overall	0.682	—	—	0.647	—	—	0.510	—	—
	Fitz I	0.569	0.575	+0.97	0.543	0.556	+2.41	0.472	0.470	−0.54
	Fitz II	0.584	0.584	0.00	0.587	0.585	−0.36	0.512	0.516	+0.66
	Fitz III	0.649	0.650	+0.05	0.619	0.622	+0.49	0.474	0.475	+0.26
	Fitz IV	0.691	0.657	−4.94	0.723	0.730	+1.03	0.597	0.575	−3.71
	Fitz V	0.557	0.563	+1.16	0.593	0.561	−5.26	0.449	0.442	−1.52
	Fitz VI	0.475	0.590	(+24.13)	0.556	0.656	(+18.01)	0.535	0.594	(+11.04)
bIoU	Overall	0.558	—	—	0.514	—	—	0.371	—	—
	Fitz I	0.424	0.434	+2.42	0.398	0.411	+3.24	0.338	0.336	−0.55
	Fitz II	0.457	0.457	0.00	0.454	0.452	−0.42	0.373	0.376	+0.92
	Fitz III	0.514	0.514	+0.01	0.478	0.480	+0.45	0.335	0.338	+0.72
	Fitz IV	0.564	0.530	−6.06	0.600	0.613	+2.31	0.456	0.438	−4.05
	Fitz V	0.414	0.424	+2.37	0.455	0.428	−6.00	0.311	0.306	−1.74
	Fitz VI	0.353	0.464	(+31.46)	0.423	0.527	(+24.63)	0.395	0.463	(+17.14)

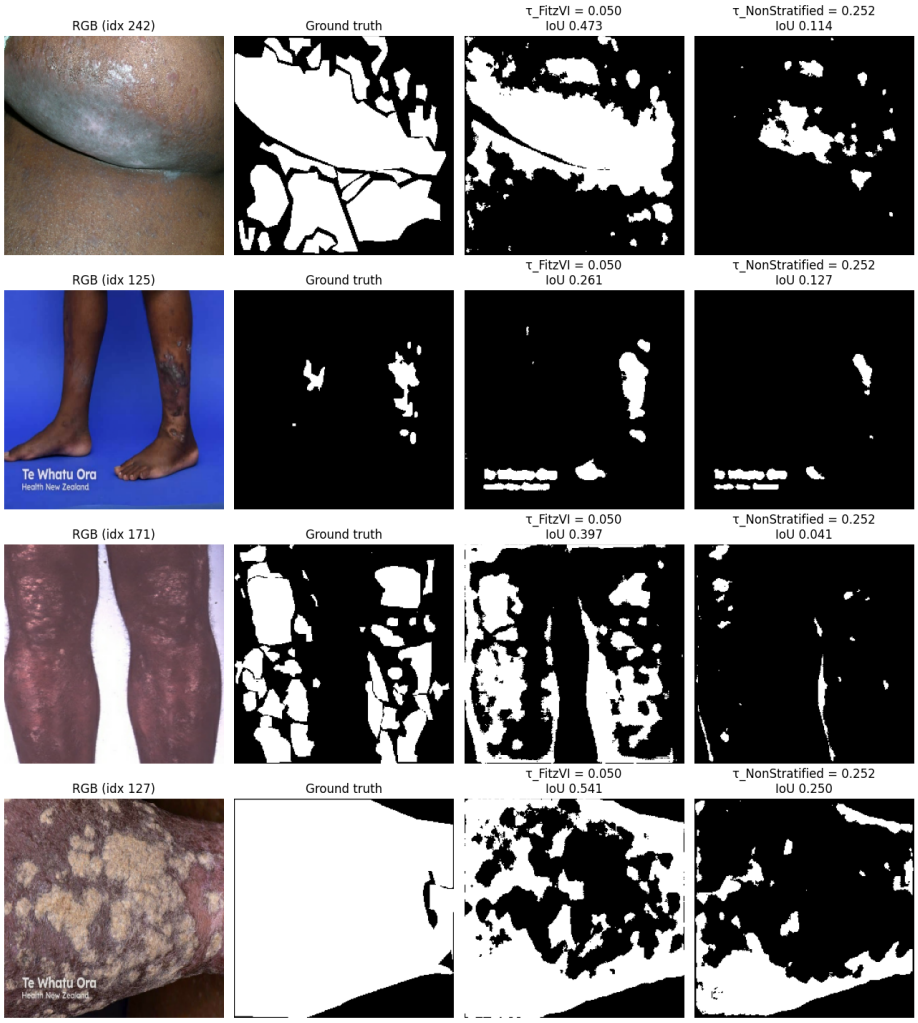

Fig. 3: The lower Fitz-VI optimized threshold (column 3) captures much more of the diseased skin than the global optimized operating point (column 4). Inference examples from U-Net.

skin tones. Because the Fitzpatrick specific operating points are only exercised within its own subgroup, the macro-average for both dice and bIoU across all tones moves minimally for every architecture.

4 Discussion

Labor intensive skin-tone annotation to verify equitable stratified performance is no longer a bottleneck. Historically, per-image Fitzpatrick labels required labour-intensive, subjective grading by board-certified dermatologists—prohibitive for million-image repositories. The advent of high-accuracy tone classifiers trained on Fitzpatrick-17k that reach $>95\%$ balanced accuracy in external validation across a broad cross section of dermatology diseases [11]. In practice, a lightweight classifier adds minimal inference time and can be applied retrospectively to every archive or even prospectively on-device.

A practical addition to the fairness toolbox. Per-group threshold calibration is an immediately deployable fairness lever—orthogonal to, and composable with, data balancing, representation alignment, FairAdaBN [25], or FairPrune [24]. We therefore recommend that future skin-segmentation studies:

1. Define in the metadata the skin tone of images in the pre-processing pipeline via an automated method such as a Fitzpatrick17k classifier.
2. Tune g on a validation split, or ideally on the set of predictions from a cross-fold validation.
3. Report both overall and tone-stratified metrics at those thresholds.

Either the stratified performance with the global threshold is equitable across skin tones, or it may not be, in which case Fitzpatrick thresholding provides a lever to lower the under performance as demonstrated here in psoriasis segmentation. Doing so requires no architectural change, no re-training, and negligible runtime overhead, yet—as shown here—can significantly increase performance on the darkest skin tones. Given its simplicity and efficacy, Fitzpatrick-specific thresholding could become a standard baseline for ISIC fairness tracks and for any clinical deployment of dermatology segmentation models.

Acknowledgments. No sponsoring company was involved in the production of this work.

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

References

1. Benčević, M., Ljubić, A., Horvat, M., Zanchi, M.: Understanding skin colour bias in deep learning-based skin lesion segmentation. Computer Methods and Programs in Biomedicine **233**, 107593 (2024). https://doi.org/10.1016/j.cmpb.2024.107593

2. Bożek, A., Reich, A.: The reliability of three psoriasis assessment tools: Psoriasis area and severity index, body surface area and physician global assessment. Advances in Clinical and Experimental Medicine **26**(5), 851–856 (2017). https://doi.org/10.17219/acem/69804, http://dx.doi.org/10.17219/acem/69804
3. Centers for Disease Control and Prevention: NAMCS/NHAMCS — web tables, https://www.cdc.gov/nchs/ahcd/web$_t ables.htm$
4. Daneshjou, R., Vodrahalli, K., Novoa, R.A., Jenkins, M., Liang, W., Rotemberg, V., Ko, J., Swetter, S.M., Bailey, E.E., Gevaert, O.: Disparities in dermatology AI performance on a diverse, curated clinical image set. Science Advances **8**(32), eabq6147 (2022). https://doi.org/10.1126/sciadv.abq6147
5. Daneshjou, R., Vodrahalli, K., Novoa, R.A., Jenkins, M., Liang, W., Rotemberg, V., Ko, J., Swetter, S.M., Zou, J., Chiou, A.S.: Disparities in dermatology ai performance on a diverse, curated clinical image set. Science **376**(6594), 413–419 (2022). https://doi.org/10.1126/science.abj2097
6. of Dermatology, H.S., Venereology: Hellenic dermatological atlas. https://www.hellenicdermatlas.com/ (2025), online Greek atlas with 2 600 + annotated images. Accessed 8 Jul 2025
7. Department of Dermatology, U.o.E., of Heidelberg, U.: Dermis / dermatology information system. http://www.dermis.net/ (2025), dermatology Online Atlas (DOIA) with multilingual diagnostic database. Accessed 8 Jul 2025
8. Diao, J.A., Adamson, A.S.: Representation and misdiagnosis of dark skin in a large-scale visual diagnostic challenge. Journal of the American Academy of Dermatology **86**(4), 950–951 (2022). https://doi.org/10.1016/j.jaad.2021.03.088
9. Dutta, A., Zisserman, A.: The VIA annotation software for images, audio and video. In: Proceedings of the 27th ACM International Conference on Multimedia (MM '19). pp. 2276–2279. Association for Computing Machinery (2019). https://doi.org/10.1145/3343031.3350535, software available at https://www.robots.ox.ac.uk/ vgg/software/via/
10. George, Y., Aldeen, M., Garnavi, R.: Automatic psoriasis lesion segmentation in two-dimensional skin images using multiscale superpixel clustering. Journal of Medical Imaging **4**(4), 044004–044004 (2017). https://doi.org/10.1117/1.JMI.4.4.044004, https://doi.org/10.1117/1.JMI.4.4.044004, [PMC free article] [PubMed]
11. Groh, M., Harris, C., Soenksen, L., Ng, D., Abou Jassoum, H., Zhu, B., White, R., Zou, J.: Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPR W) 2021 – ISIC. pp. 364–373. IEEE (2021). https://doi.org/10.1109/CVPRW53098.2021.00041
12. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: Advances in Neural Information Processing Systems 29 (NeurIPS). pp. 3315–3323 (2016)
13. Kinyanjui, N.M., Odonga, T., Cintas, C., Bashyam, V., Gadodia, S., Müller, H., Baumgartner, C.F.: Fairness of classifiers across skin tones in dermatology. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, LNCS 12265. pp. 320–329. Springer (2020). https://doi.org/10.1007/978-3-030-59725-2_31
14. Kleinberg, J., Mullainathan, S., Raghavan, M.: Inherent trade-offs in the fair determination of risk scores. In: Proceedings of the 8th Innovations in Theoretical Computer Science Conference (ITCS). pp. 43:1–43:23 (2017). https://doi.org/10.4230/LIPIcs.ITCS.2017.43
15. Li, H., Chen, G., Zhang, L., Xu, C., Wen, J.: A review of psoriasis image analysis based on machine learning. Frontiers in Medicine **11**, 1414582 (2024). https://doi.org/10.3389/fmed.2024.1414582, https://doi.org/10.3389/fmed.2024.1414582

16. Mogawer, R.M., Mostafa, W.Z., Elmasry, M.F.: Comparative analysis of the body surface area calculation method used in vitiligo extent score vs the hand unit method used in vitiligo area severity index. J Cosmet Dermatol **19**(10), 2679–2683 (2020). https://doi.org/10.1111/jocd.13311
17. Pakzad, M., Kawesh, O., Lohweg, V., Nazari, M.: CIRCLe: Color invariant representation learning for unbiased classification of skin lesions. In: Computer Vision – ECCV 2022 Workshops (Skin Image Analysis) (2022), arXiv:2208.13528
18. Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., Weinberger, K.Q.: On fairness and calibration. In: Advances in Neural Information Processing Systems 30 (NeurIPS). pp. 5680–5689 (2017)
19. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. pp. 234–241. Springer International Publishing (2015). https://doi.org/10.1007/978-3-319-24574-4_28
20. Silva, S.F.d.: Dermatology atlas brazil (atlas dermatológico). https://www.atlasdermatologico.com.br/ (2024), online atlas of 12 000+ clinical photographs. Accessed 8 Jul 2025
21. Silverberg, J.I., Lei, D., Yousaf, M., et al.: Measurement properties of the product of investigator's global assessment and body surface area in children and adults with atopic dermatitis. J Eur Acad Dermatol Venereol **35**(1), 180–187 (2021). https://doi.org/10.1111/jdv.16846
22. Trust, D.N.Z.: Dermnet nz—the world's leading free dermatology resource. https://dermnetnz.org/ (2025), over 25 000 high-resolution images and accompanying clinical topics. Accessed 8 Jul 2025
23. Usatine, R.P., Madden, B.D.: Interactive dermatology atlas. https://www.dermatlas.net/ (2025), 1 000+ photos cross-referenced by lesion attributes. Accessed 8 Jul 2025
24. Wu, Q., Li, J., Yu, B.W., Zhang, Y., Zhang, H., Xu, C.: Fairprune: Achieving fairness through pruning for dermatological disease diagnosis. arXiv preprint **arXiv:2203.02110** (2023)
25. Xu, C., Song, L., Zhang, X., Li, Q., Chen, C., Yan, S.: Fairadabn: Mitigating unfairness with adaptive batch normalization in dermatological disease classification. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023 (2023), arXiv:2303.08325
26. Yoo, K.H., Jeong, G.J., Park, J.H., Park, S.H., Li, K.S.: Estimation error of the body surface area in psoriasis: a comparative study of physician and computer-assisted image analysis (imagej). Clin Exp Dermatol **47**(7), 1298–1306 (2022). https://doi.org/10.1111/ced.15148
27. Zhang, Z., Liu, Q., Wang, Y.: Road extraction by deep residual U-net. IEEE Geoscience and Remote Sensing Letters **15**(5), 749–753 (2018). https://doi.org/10.1109/LGRS.2018.2802944
28. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H.S., Zhang, L.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6877–6886. IEEE (2021). https://doi.org/10.1109/CVPR46437.2021.00681