When Fairness Transfer Backfires: Dark-Skin Inversion in Dermatology Foundation Models and a Minimal In-Distribution Fix

Richard Adusei*, Alice Bagyiereyele Lakyiere, Kwabena Owusu-Agyemang, and Rose-Mary Owusuaa Mensah Gyening

Department of Computer Science, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana
radusei12@st.knust.edu.gh

Abstract. Dermatology foundation models detect skin cancer less accurately on dark skin (Fitzpatrick types V–VI) than on light skin (types I–II). Because biopsy-confirmed dark-skin data is scarce, a natural shortcut is to learn a fairness correction on a public, tone-labelled dataset and carry it over to the deployment population. We show that this shortcut can harm the group it is meant to help. A malignancy classifier trained on the Fitzpatrick17k benchmark and tested on DDI falls to or below chance on dark skin, with the area under the ROC curve (AUROC) at 0.50 or lower on five foundation models and significantly below chance on three, including a self-supervised model that reads no text. The failure is symmetric, reappearing when the two datasets switch roles. Its cause is a disagreement about disease rather than about skin tone. Most dark-skin cancers in DDI are mycosis fungoides, a lymphoma, whereas Fitzpatrick17k contains mostly carcinomas and melanomas. Changing the disease distribution alone does not reproduce the failure, nor does changing the imaging source alone; the failure appears only when both change together. Cross-dataset transfer is therefore not a viable fix on its own, and the reliable alternative is to stop transferring and calibrate locally instead. The fix is Per-tone few-shot calibration (PTFC), which refits a per-group readout from exactly twenty labelled examples per class drawn from the deployment population, raising dark-skin AUROC from 0.54 to 0.88 and turning the previously worst-performing group into the best-performing one. Tone-conditioned architecture does not reliably add to this, because the obstacle is label validity rather than model capacity.

Keywords: Fairness · Dermatology · Foundation models · Distribution shift · Skin tone.

1 Introduction

Deep learning models now detect skin cancer as accurately as dermatologists on standard benchmarks [4], and dermatology foundation models go further by

* Corresponding author.

learning general features from large image-text collections [21,29,25]. Needing little or no extra training, they are attractive for triage where specialists are scarce. Yet their performance is uneven. On dark skin (Fitzpatrick V–VI) they detect malignant lesions less reliably than on light skin (I–II) [3,8,1]. This is a safety problem, not a cosmetic one. A melanoma or carcinoma missed on a dark-skinned patient is a cancer left undetected, so closing this performance gap is a clinical priority.

In our experiments the gap is large. With no extra training, a frozen dermatology foundation model separates malignant from benign lesions at an area under the ROC curve (AUROC) of 0.69 on light skin but only 0.54 on dark skin, barely above chance. The traditional approach is to use more labelled examples from the affected group, but biopsy-confirmed dark-skin cases are scarce, so a natural shortcut is to take a public dataset that records skin tone, learn a fairness correction on it, and carry it over to the deployment population. We are not aware of prior work that documents how often this shortcut is taken in practice, only that the data scarcity motivating it is well documented [3,19].

Prior fairness corrections are trained and evaluated within a single dataset. To our knowledge, none has asked whether a malignancy correction trained on one tone-labelled dermatology benchmark still helps dark skin on a different one. We study cross-dataset fairness transfer in dermatology foundation models, asking two questions. Does a malignancy classifier trained on one tone-labelled benchmark transfer to another for dark skin? If not, what causes the failure, and what is the smallest reliable fix?

Cross-dataset transfer does not merely weaken on dark skin, it inverts. Trained on Fitzpatrick17k and tested on DDI, the malignancy score drops to or below chance on all five foundation models (significant on three) and returns symmetrically when the datasets switch roles. The cause is a disagreement about disease. In DDI most dark-skin cancers are mycosis fungoides, a lymphoma, while in Fitzpatrick17k most are carcinomas and melanomas. Neither this disease-mix difference nor a change of imaging source alone breaks the model; only the two together do. Because the cause is a label mismatch, not a feature deficit, transferring a correction across datasets cannot be patched; the reliable alternative is to stop transferring and calibrate on the deployment population directly. The fix is Per-tone few-shot calibration (PTFC), which re-fits a per-group readout from exactly twenty in-distribution labels per class, lifting dark-skin AUROC from 0.54 to 0.88, whereas a more elaborate tone-conditioned adapter does not reliably add to it. We contribute the inversion itself, shown to be symmetric, specific to dark skin, and not escaped by published mitigations; the controlled identification of a source-by-disease interaction as its cause; and PTFC, which recovers dark-skin AUROC where cross-dataset correction cannot.

2 Related Work

2.1 Skin-tone disparities in dermatology AI

Automated skin lesion analysis now reaches specialist-level accuracy on curated dermoscopy benchmarks [4,23], but mostly on light skin, since public datasets are dominated by light-skinned patients and models are correspondingly unreliable on dark skin [14,1]. DDI, biopsy-confirmed and tone-balanced, shows accuracy dropping on Fitzpatrick V–VI [3], while Fitzpatrick17k shows that accuracy tracks the tones a model is trained on [8]; both measure the gap within one dataset, not across two.

2.2 Bias mitigation and whether it transfers across datasets

Methods to close group gaps reweight training toward the worst group [22], make features invariant to skin colour [20], or feed skin type to the classifier [14]. The most dependable option fine-tunes on labels from the target population and recovers much of the gap [3]. In dermatology, pruning-based corrections shrink the tone gap [24], but each of these is trained and tested on the same distribution. A separate line of work finds that fairness can break under distribution shift, a model fair on one dataset becoming unfair on another [26], and that medical-imaging models can exploit site-specific confounders that do not transfer [27], without explaining the cause or studying dermatology. We test that reuse here, and it fails for the group it was meant to help.

2.3 Foundation models, adaptation, and calibration

Every component we measure with is pre-existing, so the contribution is the diagnostic rather than the method. CLIP learns transferable features from image-text pairs [21], and medical versions follow the recipe, namely BiomedCLIP [29] and DermLIP [25]. SigLIP changes the loss [28], and DINOv2 learns from images alone, no text [18]. Since fine-tuning can distort pretrained features and hurt out-of-distribution accuracy [16], the common practice is to freeze the backbone and train a small module on top, such as a bottleneck adapter [12]. Calibration then turns a score into a decision at a threshold; temperature scaling does this cheaply [9], and calibrating each subgroup separately is the multicalibration principle [10]. We test five backbones spanning dermatology, biomedical, general, and text-free training, to show the inversion comes from the data, not from one model or from language.

3 Method

3.1 Datasets, backbones, and evaluation protocol

We study a single binary decision per image, whether a lesion is malignant or benign, and measure it separately for three Fitzpatrick groups [6], namely light

(I–II), intermediate (III–IV), and dark (V–VI). Pooling the six types into three keeps enough lesions per group for stable estimates.

Our held-out test set is DDI, with 656 biopsy-proven lesions balanced across tones (208/241/207 images; 49/74/48 malignant in light/intermediate/dark) [3]. For cross-dataset training we use the 3,734 Fitzpatrick17k images whose source URLs still resolved at download time (819/2,078/837 light/intermediate/dark) [8], with PAD-UFES-20 as a sanity check (11 dark-skin lesions) [19]. DDI is never used to select a model, threshold, or hyperparameter, and no hyperparameter search was run at all, every readout below using the same fixed settings given in the next paragraph.

The training datasets are multi-class, so before any analysis we map each disease class to malignant ($\mathcal{M}$) or benign ($\mathcal{B}$). This map follows each dataset's own malignancy grouping, is the same for every tone group, and is never tuned against DDI. For the text-prototype baseline, the malignancy score combines the two prototype sets as $s_{\text{mal}} = \log \sum_{c \in \mathcal{M}} e^{s_c} - \log \sum_{c \in \mathcal{B}} e^{s_c}$, where s_c is the image–text-prototype similarity for class c.

We run every probe on five frozen backbones chosen to span model families, namely DermLIP, trained on dermatology image-text pairs [25]; BiomedCLIP, trained on biomedical ones [29]; CLIP, a general model [21]; SigLIP, a general model with a different training loss [28]; and DINOv2, a self-supervised model that uses images only, no text [18]. Each backbone is frozen and its features cached once, so nothing in the backbone is trained.

On the cached features we fit four readouts, each a per-group L_2-regularized logistic regression on the frozen embedding (scikit-learn defaults, inverse regularization strength $C = 1$, 2000 iterations; every fitting set is class-balanced by construction, so no additional class weighting is applied). A zero-shot text-prototype score is the floor for the vision-language models. An in-domain oracle, a per-group linear probe trained and tested on DDI with 5-fold stratified out-of-fold cross-validation, is the best any readout of these frozen features could do in-distribution. The cross-dataset probe, the focus of the study, trains on Fitzpatrick17k and tests on DDI. A few-shot probe trains on a small labelled sample of DDI itself, drawn 300 times at random and averaged, so the reported score is not one lucky draw.

The benchmarks differ in two ways at once, imaging source and disease mix, so we separate them with a two-by-two control. One arm changes only disease (dark-skin carcinoma or melanoma vs mycosis fungoides, both inside DDI), the other only source (both datasets restricted to the shared carcinoma-and-melanoma group, squamous and basal cell carcinoma and melanoma, in-situ forms included). A failure that appears only when both are crossed, not either alone, marks an interaction rather than one factor.

AUROC is our primary metric because it is threshold-free, and a transferred threshold is itself one of the failures we report (Sect. 4.4). Metric-selection guidance recommends pairing such a ranking metric with a decision-relevant one [17], so we also report sensitivity at a fixed specificity. The dark-skin test set is small (48 malignant lesions), so the dominant source of variability across repeated

runs is which lesions happen to fall in the test set, not the pseudorandom seed used to initialise a probe fit or a resampling draw. Every per-group AUROC therefore carries a stratified bootstrap confidence interval ($\geq$1000 resamples within tone group); a below-chance AUROC is checked with a label-permutation test, and paired arms on the same lesions with a paired-difference bootstrap. Disease composition is reported as plain fractions of malignant lesions.

Three terms recur below, and we fix their meaning here. The *tone gap* is the difference in AUROC between two Fitzpatrick groups under the same training regime, the disparity that the fairness literature targets. The *transfer gap* is the difference, within a single group, between the in-domain oracle and the cross-dataset probe, that is, the cost of borrowing labels from another benchmark. *Inversion* is the specific case of a cross-dataset AUROC below 0.50, where the readout ranks malignant lesions below benign ones rather than merely ranking them weakly.

3.2 Per-tone few-shot calibration (PTFC)

Section 4 rules out the easy options: zero-shot scoring is near chance on dark skin, cross-dataset transfer inverts, and a per-group linear probe trained in-distribution already reaches the in-domain oracle, so extra architecture adds nothing beyond what the labels themselves provide. What remains is to re-learn the decision between malignant and benign in the deployment distribution itself, per tone group, from a few labels. We call this Per-tone few-shot calibration (PTFC).

PTFC has four steps, each built from a component cited in Sect. 2.3. (1) Cache features from a frozen backbone (DermLIP gives the highest ceiling, but all five work). (2) Collect exactly twenty biopsy-confirmed labels per class per tone group from the deployment population; Sect. 4.3 shows the learning curve saturates by this point, so the twenty we chose is close to the smallest budget that reaches the ceiling, not an arbitrary round number. (3) Fit one logistic readout per tone group on the frozen feature $z \in \mathbb{R}^d$, $p_g(\text{mal} \mid z) = \sigma(w_g^\top z + b_g)$, where σ is the logistic sigmoid and the per-group weight vector w_g and bias b_g are fit by L_2-regularized logistic regression (the same fitting procedure as the oracle and cross-dataset probes, Sect. 3.1). (4) Set each group's threshold on a held-out in-distribution split (disjoint from the twenty calibration labels and from the test lesions) to a target specificity, say 90%. PTFC thus adds no new component beyond a frozen feature extractor, a per-group linear classifier, and a per-group threshold. Its one requirement is that the labels come from the deployment population: the same budget of twenty labels drawn from another dataset instead collapses dark-skin AUROC back toward inversion (Sect. 4.3).

PTFC assumes the tone group is known at test time, as DDI and Fitzpatrick17k do when recording a Fitzpatrick label from chart review or dermatologist consensus rather than automated measurement. In deployment that label could instead come from clinician assessment, self-report, or an automated estimator, and the three do not agree with each other as closely as a single ground truth would imply [13,15]. We do not test PTFC's sensitivity to a misassigned tone group, which would receive the wrong classifier and threshold; finer-grained

Table 1. Cross-dataset transfer of a malignancy readout, by backbone and Fitzpatrick group. AUROC on DDI. Zero-shot is the text-prototype floor (n/a for the text-free DINOv2); the in-domain oracle is a per-group probe trained and tested on DDI; cross-dataset trains on Fitzpatrick17k, CIs in brackets. p tests whether AUROC is significantly below chance (permutation test), reported only for V–VI, the only group whose point estimate falls below 0.50.

	Zero-shot			Oracle			Cross-dataset [95% CI]			
Backbone	I–II	III–IV	V–VI	I–II	III–IV	V–VI	I–II	III–IV	V–VI	p(V–VI)
DermLIP	0.69	0.74	0.54	0.77	0.83	0.90	0.62[.53-.70]	0.64[.55-.71]	0.39[.30-.48]	0.008
BiomedCLIP	0.63	0.62	0.55	0.75	0.80	0.76	0.66[.57-.74]	0.65[.57-.73]	0.45[.36-.55]	0.157
CLIP	0.57	0.52	0.44	0.75	0.81	0.86	0.65[.56-.73]	0.70[.63-.77]	0.39[.30-.49]	0.008
SigLIP	0.63	0.67	0.54	0.73	0.83	0.82	0.60[.50-.69]	0.67[.59-.74]	0.43[.33-.53]	0.072
DINOv2	n/a	n/a	n/a	0.68	0.79	0.76	0.62[.53-.70]	0.65[.57-.73]	0.37[.28-.47]	0.002

tone measures have been proposed partly to reduce this ambiguity [11]. Reliable tone assignment at test time is a precondition of PTFC, not a part of it we have validated.

4 Experiments

We report AUROC on the held-out DDI test set for all three Fitzpatrick groups and all five backbones, each with a stratified bootstrap confidence interval (Table 1), taking DermLIP as the headline backbone and DINOv2 as the text-free control.

4.1 Cross-dataset transfer inverts on dark skin

The tone gap on dark skin is present before any correction, and the cross-dataset fix makes it worse (Table 1): trained on Fitzpatrick17k and tested on DDI, dark-skin AUROC falls below the 0.50 chance level on all five backbones, significantly on three (DermLIP, CLIP, DINOv2). The text-free DINOv2 inverts hardest, so the failure cannot be linguistic; it lies in the image features, placing malignant dark-skin lesions below benign ones.

The inversion is specific to dark skin (Table 1 shows light and intermediate transferring at 0.60–0.70 on every backbone), and it is symmetric. Training on DDI and testing on Fitzpatrick17k inverts dark skin again (0.30–0.45), so neither dataset is uniquely at fault. Published mitigations do not rescue it. We train two of them on Fitzpatrick17k and test cross-dataset on DDI exactly as above: Group-DRO [22], which reweights the training objective toward whichever tone group currently performs worst (our implementation follows Sagawa et al.'s [22] exponentiated-gradient update, 200 epochs, model learning rate 0.02, group-weight learning rate 0.05), and tone-balanced reweighting, a single logistic regression fit with inverse-tone-frequency sample weights. Both still leave dark-skin transfer below 0.50 on all five backbones (group-DRO 0.40–0.47), since a worst-group training objective cannot fix a failure that comes from the labels describing a different disease, not from their weighting.

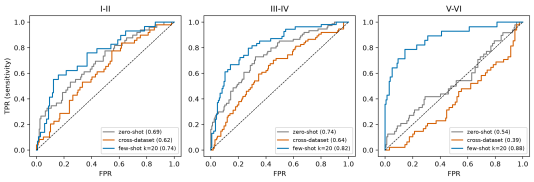

Fig. 1. Per-skin-tone ROC curves for DDI malignancy (DermLIP). For dark skin the cross-dataset curve (orange) falls below the diagonal, the geometric signature of an inverted classifier (Table 1); the few-shot curve (blue, Table 2) recovers all three groups.

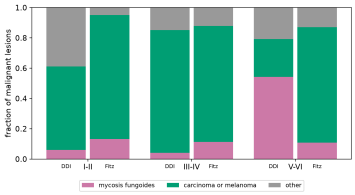

Fig. 2. Malignant-lesion composition by skin-tone group. Only for dark skin do the benchmarks disagree, DDI being 54% mycosis fungoides and Fitzpatrick17k 76% carcinoma or melanoma.

4.2 The cause is a disease-distribution mismatch

The in-domain oracle shows the failure is about transfer, not features (Table 1): the signal is in the frozen features, but it does not arrive from the other dataset, whose dark-skin lesions occupy a different region of feature space (Fig. 3). The reason is that the two benchmarks sample different cancers on dark skin. In DDI, 54% of dark-skin malignancies are mycosis fungoides, a lymphoma, and 25% are carcinomas or melanomas, while in Fitzpatrick17k the proportions reverse to 11% and 76% (Fig. 2). The balance, 21% and 13%, is other malignancies such as Kaposi sarcoma and acral lentiginous melanoma. The light and intermediate groups have the same disease mix in both datasets, so they transfer where dark skin does not.

A two-by-two control over source and disease shows neither difference alone is enough. Reading the dark-skin cells from the in-domain baseline of 0.90, swapping only the disease (inside DDI) gives 0.81, swapping only the source (disease matched) gives 0.62 on twelve malignant lesions, and crossing both gives 0.39, the inversion. Those single-factor cells are small, so the claim we certify is the powered one, that matching the disease removes the inversion on all five backbones (paired-difference bootstrap), the disease-matched control recovering to 0.68 pooled across tone groups (99 malignant lesions). The bottleneck is the interaction, not either factor alone. The below-chance score is not a sign flip a single negation would repair: the same readout scores DDI mycosis at 0.25 to 0.30 but carcinoma or melanoma at 0.50 to 0.66, so no single orientation is right for both.

4.3 Per-tone calibration requires in-distribution labels

With exactly twenty labelled examples per class per tone group, PTFC raises dark-skin AUROC from 0.54 to 0.88 and makes the worst group the best (Table 2). The benefit comes early, with two labels per class already lifting dark skin to 0.74 and the curve nearing its ceiling by twenty (Fig. 4). Recovery repeats on all five backbones and, on every backbone, the largest in-distribution gain is on the

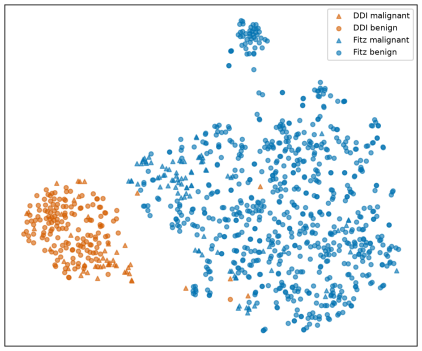

Fig. 3. Dark-skin (V–VI) lesions in DermLIP feature space (t-SNE). DDI and Fitzpatrick17k occupy separate regions, the feature-space picture behind Table 1's cross-dataset column.

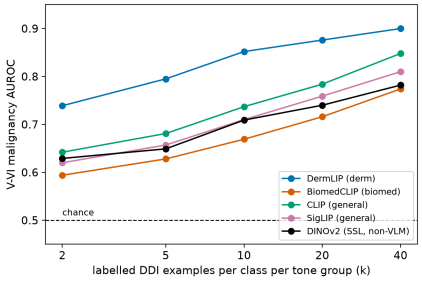

Fig. 4. Dark-skin (V–VI) AUROC against in-distribution label budget, all five backbones. Even two to five labels help well above chance; near its ceiling by twenty, PTFC's budget (Sect. 3.2).

Table 2. Matched-budget comparison of label sources, by backbone and Fitzpatrick group. AUROC on DDI at a budget of exactly twenty labels per class per tone group, varying only the label source (in-distribution DDI vs cross-dataset Fitzpatrick17k), with 95% CIs. The in-distribution column at V–VI is PTFC's headline result; the cross-dataset column shows the same recipe fails when the labels are borrowed.

	In-distribution labels [95% CI]			Cross-dataset labels [95% CI]		
Backbone	I–II	III–IV	V–VI	I–II	III–IV	V–VI
DermLIP	0.74[.66-.81]	0.80[.75-.84]	0.88[.82-.93]	0.60[.48-.69]	0.64[.53-.75]	0.39[.28-.50]
BiomedCLIP	0.69[.58-.78]	0.76[.70-.82]	0.72[.62-.81]	0.60[.50-.71]	0.60[.46-.72]	0.45[.35-.56]
CLIP	0.70[.60-.78]	0.78[.72-.82]	0.79[.68-.88]	0.64[.50-.73]	0.67[.53-.78]	0.40[.30-.50]
SigLIP	0.70[.60-.77]	0.79[.74-.84]	0.76[.66-.84]	0.57[.44-.68]	0.62[.48-.75]	0.42[.32-.53]
DINOv2	0.61[.48-.71]	0.71[.61-.79]	0.75[.67-.81]	0.60[.48-.70]	0.63[.48-.74]	0.42[.28-.58]

dark-skin group specifically (Table 2), matching where the inversion was worst (Table 1).

The objection that any twenty labels would do is refuted by a matched-budget test. Holding the label count, readout, and test set fixed and changing only the label source, dark-skin AUROC collapses from 0.88 with in-distribution labels to 0.39 with cross-dataset labels on all five backbones (Table 2). The gap is concentrated on dark skin (0.27 to 0.50) and small on the light groups (0.01 to 0.17), matching what the disease-mismatch finding in Sect. 4.2 predicts, so what carries PTFC is the in-distribution data, not the label count.

4.4 Clinical operating points and over-correction

For triage the relevant quantity is sensitivity at a fixed specificity, which Table 3 reports at 90% specificity on DermLIP for all three groups and both the in-distribution and the transferred threshold. An in-distribution PTFC threshold lifts sensitivity in every group, but by very different amounts, decisively on dark skin (18% to 73%, separated intervals) and only marginally on light skin (33%

Table 3. Sensitivity at a fixed operating point, by Fitzpatrick group. Sensitivity at 90% specificity on DDI (DermLIP), with 95% CIs. "PTFC (in-dist. threshold)" sets the operating point on a held-out split of the deployment population, as in Sect. 3.2; "PTFC (cross-source threshold)" instead transfers the threshold from Fitzpatrick17k, holding the score fixed.

Group	Zero-shot [95% CI]	PTFC, in-dist. thr.	PTFC, cross-src. thr.
I–II	33% [13, 57]	38% [14, 66]	19% [0, 69]
III–IV	35% [19, 55]	52% [31, 75]	38% [3, 75]
V–VI	18% [0, 38]	73% [50, 93]	14% [0, 37]

to 38%, intervals almost entirely overlapping). A threshold transferred from the other dataset instead drops dark-skin sensitivity below zero-shot, so both the score and the threshold must be local. PTFC can also overshoot, with dark skin overtaking light at the same operating point, though the overlapping intervals in Table 3 make the reversal only suggestive. Part of that reversal belongs to the test population, not to PTFC, since the in-domain oracle is already higher on dark skin than on light skin on all five backbones (Table 1), where DDI's dark-skin malignancies are dominated by one distinctive lymphoma. Equalizing across tones therefore means splitting the label budget per group.

Architecture does not robustly help either, since a tone-conditioned adapter adds at most 0.03 to 0.04 over the plain per-group probe, which already matches the in-domain oracle. The lever is the in-distribution labels, not the module.

5 Discussion

Two explanations for the dark-skin gap are easy to confuse. One is too little signal in the frozen features, a model problem, and the other is borrowed labels describing a different disease, a data problem. The in-domain oracle settles which one is correct, reaching 0.90 on dark skin from labels in the deployment population, so the gap is about which labels the model is given, not what it can represent, and a more elaborate adapter cannot beat a plain per-group probe. What helps is PTFC, a few in-distribution labels and a per-group operating point.

The mismatch also carries a lesson for fairness benchmarks. A tone-balanced set is not enough when two benchmarks match on tone yet differ on disease, so reporting the disease mix within each tone group would expose such mismatches before deployment.

What we document is a form of annotation shift, where two datasets attach the same label to different underlying concepts [2,5]. Named that way, the failure is unsurprising, since if malignant means something different in each benchmark then no model should be expected to transfer between them. Our point is that nothing in the standard fairness audit reveals this annotation shift. Both benchmarks record Fitzpatrick labels, both are tone-balanced by design, both use the same binary malignant-benign schema, and the light and intermediate groups do transfer at 0.60 to 0.70. The incompatibility is confined to the dark-skin stratum, invisible in every summary a practitioner would ordinarily inspect before

reusing a correction, and it becomes visible only when disease composition is broken out within tone group (Fig. 2).

A second objection is that the low AUROC values we report are too low to matter clinically, and that a stronger pipeline with augmentation and fine-tuning would not fail this way. The in-domain oracle answers the first half, since fitting the same frozen features on DDI labels reaches 0.90 AUROC on dark skin (Table 1). The features already carry the signal, so the low cross-dataset numbers measure where the labels come from, not what the backbone can represent. We hold the readout fixed and cheap deliberately, because augmentation and backbone fine-tuning would raise every arm at once and confound the one comparison the study needs, namely label source. We therefore claim nothing about clinical deployability at these absolute operating points, only about the direction and the cause of the change between arms.

6 Limitations and Future Work

Four limitations bound the paper's three claims, the inversion, its cause, and PTFC's recovery. First, evidence for all three rests on a single tone-balanced test benchmark. DDI is the only biopsy-proven dermatology dataset with enough dark-skin lesions for a per-group claim (PAD-UFES-20 has eleven), so generality is carried by the five backbones and the disease-matched control, not by a second test population. Second, several key quantities are estimated on small samples, the dark-skin malignant set being 48 lesions and the disease-matched control 12 to 16; that control is therefore certified by a paired-difference test and the pooled estimate rather than the dark-skin cell alone, and we do not correct for testing five backbones, so the inversion claim is the conservative one, significant on three of the five. Third, the PTFC fix is not label-free, since it needs in-distribution biopsy-confirmed labels from the deployment population.

Fourth, we cannot rule out patient-level leakage within DDI. DDI's public documentation reports 656 images from 570 unique patients,[1] but the released metadata carries only an image identifier, skin tone, malignancy label, and disease, no patient identifier, and the dataset's reference loader does not expose one either. We therefore cannot confirm that our within-DDI splits (the oracle's cross-validation folds and PTFC's few-shot draws, Tables 2 and 3) are patient-disjoint; at most 86 of 656 images (13%) could be an additional image from an already-represented patient, an upper bound, not an estimate of the effect. This risk is confined to those two arms. The headline cross-dataset inversion (Table 1) trains only on Fitzpatrick17k and evaluates only on DDI, so no split of DDI is involved there.

These limits set the agenda for future work. The most useful next measurement is a second tone-labelled test population, such as the PASSION cohort [7]. Beyond it, whether the disease mismatch we identify governs other protected groups, imaging domains, or trainable backbones remains open, as does replacing the Fitzpatrick scale with newer skin-tone measures [11].

[1] `https://ddi-dataset.github.io`, accessed 2026-08-06.

7 Conclusion

We tested a tempting shortcut, training a dark-skin fairness correction on one tone-labelled benchmark and deploying it on another, and found the opposite of help. The transfer inverts for the group it was meant to protect, scoring it below chance, and does so symmetrically, only on dark skin, and on all five foundation models including one that reads no text. The cause is a disagreement between the benchmarks about dark-skin disease, an interaction of imaging source and disease mix that a controlled experiment isolates. Because that correction cannot be patched, PTFC abandons cross-dataset transfer and calibrates on a small set of in-distribution labels with a per-group operating point, raising dark-skin AUROC from 0.54 to 0.88. This recasts the dark-skin gap as a problem of label validity and data budget, not model design.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article. Code will be available at `https://github.com/nanarich/PTFC`.

References

1. Adamson, A.S., Smith, A.: Machine learning and health care disparities in dermatology. JAMA Dermatology, vol. 154, no. 11, pp. 1247-1248 (2018).
2. Castro, D.C., Walker, I., Glocker, B.: Causality matters in medical imaging. Nature Communications **11**(1), 3673 (2020).
3. Daneshjou, R., Vodrahalli, K., Novoa, R.A., et al.: Disparities in dermatology AI performance on a diverse, curated clinical image set. Science Advances **8**(31), eabq6147 (Aug 2022).
4. Esteva, A., Kuprel, B., Novoa, R.A., et al.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**(7639), 115–118 (2017).
5. Filbrandt, G., Kamnitsas, K., Bernstein, D., et al.: Learning from partially overlapping labels: Image segmentation under annotation shift. arXiv preprint arXiv:2107.05938 (2021).
6. Fitzpatrick, T.B.: The validity and practicality of sun-reactive skin types I through VI. Archives of Dermatology, 124(6), 869-871 (1988).
7. Gottfrois, P., Gröger, F., Andriambololoniaina, F.H., et al.: PASSION for dermatology: Bridging the diversity gap with pigmented skin images from Sub-Saharan Africa. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024 (LNCS vol. 15012) (2024).
8. Groh, M., Harris, C., Soenksen, L., et al.: Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitzpatrick 17k Dataset (2021).
9. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On Calibration of Modern Neural Networks (2017).
10. Hébert-Johnson, Ú., Kim, M., Reingold, O., Rothblum, G.: Multicalibration: Calibration for the (Computationally-Identifiable) Masses. In: Proceedings of the 35th International Conference on Machine Learning (ICML), PMLR 80:1939-1948 (2018), `https://proceedings.mlr.press/v80/hebert-johnson18a.html`
11. Heldreth, C.M., Monk, E.P., Clark, A.T., et al.: Which skin tone measures are the most inclusive? an investigation of skin tone measures for artificial intelligence. ACM Journal on Responsible Computing, 1(1), Article 7 (2024).

12. Houlsby, N., Giurgiu, A., Jastrzebski, S., et al.: Parameter-Efficient Transfer Learning for NLP (Jun 2019).
13. Howard, J.J., Sirotin, Y.B., Tipton, J.L., Vemury, A.R.: Reliability and validity of image-based and self-reported skin phenotype metrics. arXiv preprint arXiv:2106.11240 (2021).
14. Kassem, M.A., Hosny, K.M., Damaševičius, R., Eltoukhy, M.M.: Machine Learning and Deep Learning Methods for Skin Lesion Classification and Diagnosis: A Systematic Review. Diagnostics **11**(8), 1390 (Jul 2021).
15. Kinyanjui, N.M., Odonga, T., Cintas, C., et al.: Estimating skin tone and effects on classification performance in dermatology datasets. arXiv preprint arXiv:1910.13268 (2019).
16. Kumar, A., Raghunathan, A., Jones, R., et al.: Fine-tuning can distort pretrained features and underperform out-of-distribution. In: International Conference on Learning Representations (ICLR) (2022), https://arxiv.org/abs/2202.10054
17. Maier-Hein, L., Reinke, A., Godau, P., et al.: Metrics reloaded: Recommendations for image analysis validation. Nature Methods **21**(2), 195–212 (2024).
18. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., et al.: DINOv2: Learning Robust Visual Features without Supervision. Transactions on Machine Learning Research (2023).
19. Pacheco, A.G.C., Lima, G.R., Salomão, A.S., et al.: PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones (2020).
20. Pakzad, A., Abhishek, K., Hamarneh, G.: CIRCLe: Color Invariant Representation Learning for Unbiased Classification of Skin Lesions (2022).
21. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., et al.: Learning Transferable Visual Models From Natural Language Supervision. In: Proceedings of the 38th International Conference on Machine Learning. pp. 8748–8763. PMLR (2021).
22. Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P.: Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. In: International Conference on Learning Representations (ICLR). ICLR (2020).
23. Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data, vol. 5, art. 180161 (2018).
24. Wu, Y., Zeng, D., Xu, X., et al.: FairPrune: Achieving fairness through pruning for dermatological disease diagnosis. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2022 (LNCS vol. 13431) (2022).
25. Yan, S., Hu, M., Jiang, Y., et al.: Derm1M: A Million-scale Vision-Language Dataset Aligned with Clinical Ontology Knowledge for Dermatology (Mar 2025).
26. Yang, Y., Zhang, H., Gichoya, J.W., et al.: The Limits of Fair Medical Imaging AI In The Wild (2023).
27. Zech, J.R., Badgeley, M.A., Liu, M., et al.: Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLOS Medicine, 15(11): e1002683 (2018).
28. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid Loss for Language Image Pre-Training (2023).
29. Zhang, S., Xu, Y., Usuyama, N., et al.: BiomedCLIP: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs (Mar 2023).