

Reliable Skin-Tone Fairness in Dermatology AI: From Lesion-Level Leakage to Tone-Adaptive Learning

Muhammad Rizwan Tahir

University of Management and Technology

Shaista Habib

University of Management and Technology

Sadaf Jabeen

University of Management and Technology

Zahoor Khan

`z.khan@ulster.ac.uk`

University of Ulster

Research Article

Keywords: skin lesion classification, algorithmic fairness, skin-tone bias, data leakage, contrast enhancement, reproducibility

Posted Date: October 20th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11239538/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Reliable Skin-Tone Fairness in Dermatology AI: From Lesion-Level Leakage to Tone-Adaptive Learning

Muhammad Rizwan Tahir¹, Shaista Habib^{1*}, Sadaf Jabeen¹,
Zahoor Khan^{2*}

¹Department of Artificial Intelligence, School of Systems and Technology,
University of Management and Technology, Lahore, Pakistan.

²Nutrition Innovation Center for Food and Health (NICHE), School of
Biomedical Sciences, Ulster University, Coleraine, Northern Ireland,
United Kingdom.

*Corresponding author(s). E-mail(s): shaista.habib@umt.edu.pk;
z.khan@ulster.ac.uk;

Abstract

Skin-tone fairness remains a challenge in dermatology AI, particularly for darker skin where reduced lesion visibility and limited representation can affect classification performance. We introduce a tone-adaptive pipeline combining selective contrast enhancement with a novel risk-aware augmentation strategy to address differences in lesion visibility, clinical risk, and class imbalance across skin tones. We evaluate the approach across three complementary benchmarks—HAM10000, PAD-UFES-20, and Fitzpatrick17k—spanning dermoscopic and real-world clinical images, different diagnostic taxonomies, and both clinician-assigned and image-derived skin-tone information. Because dermatology datasets often contain multiple images of the same lesion, we first establish a lesion-grouped, class-stratified evaluation protocol and quantify the effect of lesion-level leakage. Conventional image-level splits affected up to 39% of test images across the three benchmarks and increased baseline macro-F1 from 0.672 to 0.774 on HAM10000, with leakage-induced inflation of +0.10, +0.05, and +0.01 macro-F1 on HAM10000, PAD-UFES-20, and Fitzpatrick17k, respectively. A controlled leakage experiment further showed a 6.6-point increase in accuracy as injected duplicate overlap increased from 0% to 40%, while also destabilizing darker-skin recall. We further demonstrate that a commonly used luminance-based skin-tone proxy is unreliable: on PAD-UFES-20, border CIELAB L^* showed no

association with clinician-assigned Fitzpatrick type ($\rho = -0.003$, $p = 0.91$; AUC = 0.50). Under the leak-free protocol, the tone-adaptive pipeline improved darker-skin sensitivity on PAD-UFES-20 from 0.655 to 0.714, while producing limited gains on standardized dermoscopic images. Ablation analysis indicated that this improvement was driven primarily by contrast enhancement rather than risk-aware augmentation, suggesting that tone-adaptive preprocessing is most beneficial under variable clinical imaging conditions. Across three datasets and two architectures, these findings demonstrate that reliable assessment of skin-tone fairness depends critically on lesion-level evaluation, valid tone measurement, and careful interpretation of intervention effects. We also release *leakaudit*, an open-source tool for detecting lesion-level leakage and generating group-disjoint splits.

Keywords: skin lesion classification, algorithmic fairness, skin-tone bias, data leakage, contrast enhancement, reproducibility

1 Introduction

Skin cancer is among the most common cancers worldwide. For melanoma, its deadliest form, five-year survival exceeds 90% when the disease is found early, but it falls sharply once the tumour spreads [American Cancer Society \(2024\)](#). Early and accurate detection therefore has a large effect on patient outcomes. Deep convolutional networks trained on dermoscopic images now reach dermatologist-level accuracy for lesion classification [Esteva et al. \(2017\)](#); [Tschandl et al. \(2018\)](#), and they offer a scalable form of diagnostic support. These benefits, however, are not shared equally across patients.

Benchmark datasets are heavily skewed toward lighter skin. Darker skin tones (Fitzpatrick IV–VI) usually make up less than 10% of the images [Groh et al. \(2021\)](#); [Morales-Forero et al. \(2024\)](#). A model that performs well on average can therefore be far less sensitive on darker skin, with reported deficits in sensitivity (also called recall) of about 20 percentage points on diverse clinical image sets [Daneshjou et al. \(2022\)](#). Because malignant lesions are the most costly to miss, these gaps translate directly into harm for populations that are already underserved [Adamson and Smith \(2018\)](#); [Daneshjou et al. \(2021\)](#). These disparities are not hypothetical: skin cancer in patients with skin of color is often diagnosed at a later stage and carries worse outcomes [Gloster Jr and Neal \(2006\)](#), and Black patients in particular have substantially lower five-year melanoma survival than White patients [Dawes et al. \(2016\)](#). The same pattern of algorithmic bias has been documented across healthcare more broadly [Obermeyer et al. \(2019\)](#).

Fairness metrics are also under-reported in dermatology AI, so subgroup gaps are easily hidden by high average accuracy [Hardt et al. \(2016\)](#); [Mehrabani et al. \(2021\)](#); [Barocas et al. \(2023\)](#). Whether a proposed method genuinely reduces this disparity therefore depends on how it is evaluated—one focus of this study.

One approach to addressing these disparities is to adapt image preprocessing and training to differences in skin tone and lesion visibility. We therefore introduce a *tone-adaptive pipeline* (Figure 1) that combines selective contrast enhancement with

risk-aware augmentation. The pipeline uses CIELAB- L^* to stratify images into tone groups, applies contrast enhancement selectively to images identified as darker, and uses risk-aware augmentation to address class imbalance and subgroup representation. Rather than assuming that a single preprocessing strategy is equally effective across all imaging conditions, the pipeline is designed to address reduced lesion visibility while preserving the characteristics of lighter-skin images. We evaluate this approach across three complementary benchmarks—HAM10000, PAD-UFES-20 (Pacheco et al. (2020)), and Fitzpatrick17k (Groh et al. (2021))—spanning dermoscopic and real-world clinical images, different diagnostic taxonomies, and both clinician-assigned and image-derived skin-tone information. However, evaluating skin-tone interventions poses two key methodological challenges. First, dermatology datasets often contain multiple images of the same lesion, which can cause conventional image-level splits to place images from the same lesion in both training and test sets. This overlap may enable lesion-specific memorization and bias fairness conclusions. Second, when skin-tone labels are unavailable, image luminance may serve as a proxy but can also reflect lighting, exposure, and acquisition conditions. Its validity therefore requires empirical verification against clinically assigned skin-tone labels (Hardt et al. (2016); Mehrabi et al. (2021); Barocas et al. (2023)).

We consequently examine both the tone-adaptive intervention and the reliability of its evaluation. Our contributions are as follows. We introduce a tone-adaptive pipeline (Figure 1), combining CIELAB- L^* tone stratification, selective CLAHE for darker-skin images, and a novel risk-aware augmentation strategy for class imbalance and subgroup representation. We evaluate its performance across three datasets and two architectures and use ablation analysis to identify the components responsible for observed changes in darker-skin sensitivity. (2) We show that conventional image-level splitting can inflate aggregate performance and destabilize subgroup recall estimates because of lesion-level overlap. A controlled dose-response experiment further demonstrates how increasing overlap produces systematic performance inflation and can make an ineffective method appear beneficial. (3) We show that a CIELAB- L^* luminance proxy does not track clinician-assigned Fitzpatrick type, indicating that luminance-based subgroup definitions may not reliably represent skin tone. (4) We establish a lesion-grouped, class-stratified evaluation protocol with bootstrap confidence intervals and statistical testing, and release `leakaudit`, an open, unit-tested Python package for detecting lesion-level leakage and generating group-disjoint splits for grouped image datasets (Listing 1).

Table 1 positions this study against representative prior work.

2 Related work

Skin-lesion classification. Deep convolutional networks now reach dermatologist-level accuracy on dermoscopic images (Esteva et al. (2017); Tschandl et al. (2018)). HAM10000 and the ISIC Archive have become the field’s standard benchmarks. Reported accuracies are high. However, these figures depend heavily on how each dataset is divided into training and test sets, which is one of the key focuses of our study.

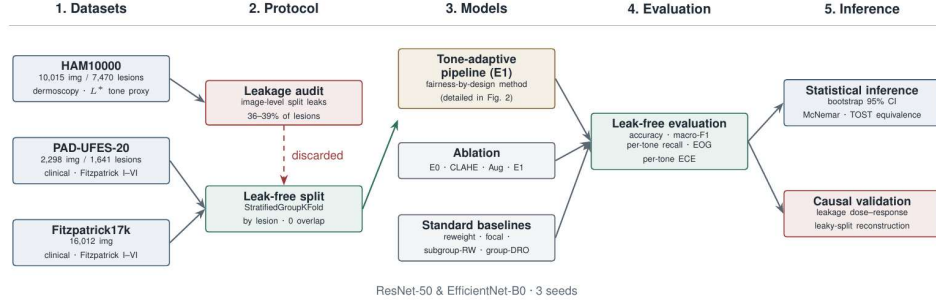


Fig. 1 Study design (experimental overview). The five stages of the study: three datasets, the leakage audit and leak-free split (Protocol), the tone-adaptive pipeline with its 2×2 ablation and four standard fairness interventions (Models), leak-free Evaluation, and statistical/causal Inference. This figure maps *what is compared and how it is assessed*; the pipeline itself is detailed in Figure 2.

Fairness in dermatology AI. Independent audits repeatedly show reduced sensitivity on darker skin [Daneshjou et al. \(2022\)](#); [Groh et al. \(2021\)](#); [Adamson and Smith \(2018\)](#), echoing the algorithmic bias seen across healthcare [Obermeyer et al. \(2019\)](#). Skin tone itself is hard to label, and proxies such as the individual typology angle have been used to estimate it when clinical labels are absent [Kinyanjui et al. \(2020\)](#). Most studies stop at documenting the gap. A growing set proposes mitigations, including tone-conditioned preprocessing and augmentation, subgroup reweighting, and distributionally robust optimization [Sagawa et al. \(2020\)](#); [Lin et al. \(2017\)](#); [Hardt et al. \(2016\)](#). Dermatology-specific fairness methods have also been proposed, including imaging-artefact unlearning [Bevan and Atapour-Abarghouei \(2022\)](#), fairness-aware adaptive normalisation [Xu et al. \(2023\)](#), fairness without demographic labels [Chiu et al. \(2024\)](#), and diffusion-based methods aimed at diverse skin tones [Carrión and Norouzi \(2023\)](#). These mitigation claims are exactly what we re-examine under leak-free evaluation.

Data leakage and reproducibility. Data leakage is a well-documented, field-wide threat to machine-learning research [Kapoor and Narayanan \(2023\)](#). In dermatology the problem is acute, because HAM10000 and ISIC contain multiple images of the same lesion as well as near-duplicates [Cassidy et al. \(2022\)](#). Dataset-quality audits have also reported duplicates and cross-partition leakage in common benchmarks [Abdelhalim et al. \(2025\)](#). Systematic reviews of public skin-image datasets similarly point to limited transparency and narrow geographic and skin-tone coverage [Wen et al. \(2022\)](#). These audits, however, focus on overall accuracy or dataset quality. The fairness-specific consequence—that leakage inflates absolute subgroup metrics and sharply destabilizes subgroup recall estimates, so that a single leaky run can make a null method look like a genuine gain—has, to our knowledge, not been directly quantified before.

Evaluation and fairness metrics. The Equal Opportunity Gap [Hardt et al. \(2016\)](#) and related criteria give formal definitions of subgroup parity. Even so, fairness metrics are still under-reported in dermatology [Mehrabi et al. \(2021\)](#); [Barocas et al.](#)

Table 1 Positioning relative to representative dermatology-AI fairness studies. The table compares study *design*, not head-to-head accuracy across non-equivalent setups. **Leak-free split:** ✓ lesion/patient-grouped and verified, ✗ image-level or not addressed, ~ varies/unstated. **Rigorous stats:** ✓ confidence intervals + significance + equivalence + multiple seeds, ~ partial, ✗ point estimates only.

Study	Dataset(s)	Tone labels	Leak-free split	Mitigation	Rigorous stats	Positioning
Esteva et al. 2017 Esteva et al. (2017)	Private dermoscopy	None	✗	None	✗	Dermatologist-level accuracy; no fairness
Groh et al. 2021 Groh et al. (2021)	Fitzpatrick17k	Fitzpatrick	✗	Dataset + audit	✗	Fairness benchmark; documents disparities
Daneshjou et al. 2022 Daneshjou et al. (2022)	Diverse clinical (DDI)	Fitzpatrick (biopsy)	✗	Audit	~	Cross-system disparity audit
Tone-conditioned preprocessing (single benchmark)	HAM10000	Luminance proxy	✗	Proposed	✗	Single benchmark; image-level split
Group-DRO / reweighting / focal Sagawa et al. (2020) ; Lin et al. (2017)	General ML	Group labels	~	Proposed	~	Generic mitigations; not derm-specific
Proposed Study	HAM10000, PAD-UFES-20, Fitzpatrick17k	Proxy + real Fitzpatrick	✓	Proposed + 4	✓	Tone-adaptive pipeline; rigorous multi-dataset lesion-grouped evaluation

(2023), and confidence intervals, significance and equivalence tests, and power analyses are rarely provided. We adopt all four as standard practice.

3 Methods

3.1 Datasets

We use three public datasets. Together they span dermoscopy and clinical photography, and they define skin tone in two different ways (Table 2).

HAM10000 [Tschandl et al. \(2018\)](#) contains 10,015 dermoscopic images from 7,470 lesions, across seven diagnostic classes. It has no skin-tone labels. We therefore infer tone from the median CIELAB L^* value of the image border, and we split the images into Light and Dark groups. The resulting L^* distribution, and the 8.8%/91.2% Dark/Light split, are shown in Supplementary Fig. S1. We caution at the outset that this proxy captures image luminance rather than validated skin tone (we quantify this in the Limitations); the HAM10000 subgroups should therefore be read as luminance-stratified.

PAD-UFES-20 [Pacheco et al. \(2020\)](#) contains 2,298 clinical (smartphone) images from 1,641 lesions, across six classes. Each image carries a clinician-assigned Fitzpatrick type. We treat FST I–II as Light and FST III–VI as Dark; images without an assigned Fitzpatrick type are excluded from the tone analysis but kept for training and overall accuracy.

Fitzpatrick17k [Groh et al. \(2021\)](#) contains 16,012 clinical images, each with its own Fitzpatrick label, grouped into three diagnostic superclasses (benign, malignant, and non-neoplastic). Because its darker-skin subset is large, we apply a stricter cut

Table 2 Dataset composition and leakage. “Groups” are lesions (HAM10000, PAD-UFES-20) or near-duplicate image groups (Fitzpatrick17k); a naive image-level split leaks the listed fraction of test images by placing group members on both sides of the split. The “Tone cut” column gives the exact Light/Dark binarization used for each dataset. Test dark/light values count tone-labeled test images only; PAD-UFES-20 additionally has 117 test images with unassigned Fitzpatrick type that are excluded from the tone columns (but not from training or overall accuracy).

Dataset	Modality	Images	Groups	Classes	Tone cut (Dark)	Test dark / light	Naive-split leakage
HAM10000	dermoscopy	10,015	7,470	7	$L^* \leq 143$ (proxy)	127 / 1,314	36.1%
PAD-UFES-20	clinical	2,298	1,641	6	FST III–VI	56 / 147	38.8%
Fitzpatrick17k	clinical	16,012	13,517	3	FST IV–VI	890 / 1,916	16.6%

here—FST I–III as Light and FST IV–VI as Dark—which yields the largest and best-powered darker-skin test set of the three datasets (test $n \approx 890$). One consequence is that FST III is grouped with Light on Fitzpatrick17k but with Dark on the smaller PAD-UFES-20, where a stricter cut would leave too few darker-skin cases. We therefore treat the two clinical tone splits as related but not identical, and we analyze each dataset separately rather than pooling them.

3.2 Evaluation protocol

Dermatology datasets often contain several images of the same lesion. In HAM10000, the 10,015 images come from only 7,470 lesions, so 2,545 images are additional views of a lesion that also appears elsewhere (Figure 5). Splitting the data by image would place different views of the same lesion on both sides of the split. To prevent this, we repartition each dataset with a lesion-grouped, class-stratified split (StratifiedGroupKFold) Pedregosa et al. (2011) in 70/15/15 proportions. For Fitzpatrick17k we group by near-duplicate image group rather than by lesion. We verify that no group appears in more than one partition, and we use identical splits and seeds across all conditions. The protocol and its group-overlap check are compact and reusable (Listing 1), and we release them as a drop-in tool.

We quantify contamination with two group-level statistics. For a dataset of N images forming G groups, the *duplication rate* is

$$D = 1 - \frac{G}{N}, \quad (1)$$

the fraction of images that are additional views of a group already present. For a given train/test split, the *leakage rate* is the fraction of test images whose group also appears in training,

$$L = \frac{|\{i \in \text{test} : g(i) \in g(\text{train})\}|}{|\text{test}|}, \quad (2)$$

where $g(i)$ denotes the group of image i . A lesion-grouped split has $L = 0$ by construction, whereas the naive image-level split does not (Table 2 reports L for each dataset). Both statistics are computed by the released tool.

Listing 1 Leak-free, group-stratified split and leakage audit (released as a drop-in tool). The assertion makes zero group overlap a checked invariant, and `leakage_rate` quantifies contamination under the naive image-level split.

```

1 import pandas as pd
2 from sklearn.model_selection import StratifiedGroupKFold
3
4 def leakfree_split(df, group="lesion_id", label="class", seed=42):
5     """70/15/15 split with NO group (lesion) shared across partitions."""
6     df = df.copy()
7     sgkf = StratifiedGroupKFold(n_splits=7, shuffle=True, random_state=seed)
8     folds = list(sgkf.split(df, df[label], groups=df[group]))
9     tr, hold = folds[0] # ~85% / ~15%
10    va, te = hold[:len(hold)//2], hold[len(hold)//2:]
11    df["split"] = "train"
12    df.iloc[va, df.columns.get_loc("split")] = "val"
13    df.iloc[te, df.columns.get_loc("split")] = "test"
14    # zero group overlap is a *checked* invariant, not an assumption
15    g = df.groupby(group)["split"].nunique()
16    assert (g == 1).all(), "group leakage detected across splits"
17    return df
18
19 def leakage_rate(df, group="lesion_id"):
20     """Fraction of test images whose group also appears in the train set."""
21     train_groups = set(df.loc[df.split == "train", group])
22     test = df[df.split == "test"]
23     leaked = test[group].isin(train_groups).sum()
24     return leaked / len(test)

```

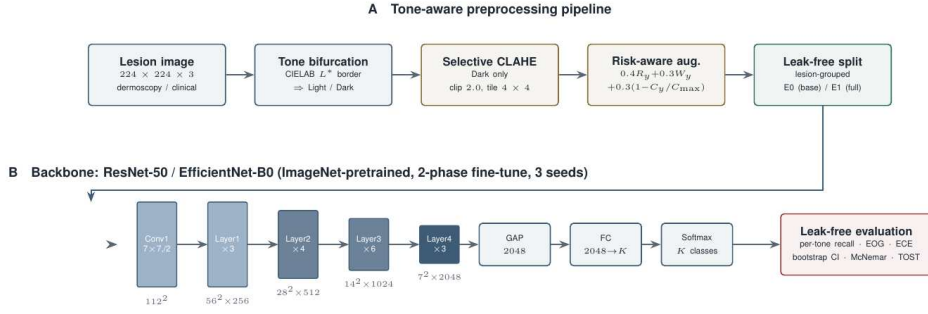


Fig. 2 The tone-adaptive method in detail (the “Models” stage of Figure 1). (A) Tone-aware preprocessing: Light/Dark assignment via the CIELAB- L^* border proxy, selective CLAHE and risk-aware augmentation applied to the Dark subgroup, feeding the E0 (baseline) and E1 (full-method) conditions. (B) The classification backbone (ResNet-50 shown; EfficientNet-B0 used as a second architecture) with two-phase fine-tuning. Full parameters are given in the Methods.

3.3 Proposed pipeline

After tone bifurcation (Figure 2), we apply two tone-adaptive steps to the Dark subgroup.

First, we apply Contrast-Limited Adaptive Histogram Equalization (CLAHE) Zuiderveld (1994); Pizer et al. (1987) to the L^* channel of the Dark images

only. We use a clip limit of 2.0 and a 4×4 tile grid (Figure 3). This step aims to make low-contrast lesions on darker skin easier to see.

Second, we oversample the Dark training images with a risk-aware rule. For each lesion class y , the augmentation factor is

$$\text{AugFactor}_y = \alpha R_y + \beta W_y + \gamma \left(1 - \frac{C_y}{C_{\max}}\right), \quad (\alpha, \beta, \gamma) = (0.4, 0.3, 0.3). \quad (3)$$

Here $R_y \in [0, 1]$ is a clinical-risk weight and $W_y \in [0, 1]$ is a darker-skin visibility-difficulty weight. We set both weights ourselves, based on clinical severity and known imaging difficulty; the visibility weights are set per dataset to reflect how each lesion type typically presents in that modality (for example, BCC on dermoscopy versus clinical photographs). These weights were fixed a priori, not tuned on the data, and not validated by a dermatologist. The final term, $1 - C_y/C_{\max}$, is a class-imbalance term computed from the Dark training counts C_y Shorten and Khoshgftaar (2019). Each Dark training class is then oversampled to a target count $\text{Target}_y = C_y + \lfloor S \cdot \text{AugFactor}_y \cdot C_y \rfloor$, where the global scale $S = 1.94$ was chosen so that the smallest, highest-risk class is roughly doubled; like the weights, it was fixed a priori rather than optimized. This directs the most oversampling toward high-risk, low-visibility, and under-represented classes such as melanoma, and leaves common, easily visible classes such as nevi unchanged. We did not perform a sensitivity analysis over (α, β, γ) or S , which we note as a limitation.

Oversampling is confined to the training set. The validation and test sets are never augmented, so no synthetic or repeated image can inflate the reported metrics. We apply the same rule to all three datasets, using dataset-specific risk weights, and Table 3 reports the resulting per-class calculation for each. Finally, a 2×2 ablation isolates each component: E0 (neither), CLAHE-only, Aug-only, and E1 (both).

3.4 Interventions, architectures, training

We compare the pipeline against four standard mitigation methods: class-balanced reweighting, focal loss ($\gamma = 2$) Lin et al. (2017), subgroup reweighting that up-weights Dark samples, and group distributionally robust optimization (group-DRO) over tone groups Sagawa et al. (2020). We use two backbones, ResNet-50 He et al. (2016) and EfficientNet-B0; the backbone and the two-phase fine-tuning schedule are shown in Figure 2B. The second backbone lets us test whether the results depend on the choice of architecture. Both are pretrained on ImageNet Deng et al. (2009) and trained in two phases: 5 epochs that train the classification head only, followed by 25 epochs of fine-tuning. We use the AdamW optimizer Loshchilov and Hutter (2019) with a learning rate of 10^{-4} , a cosine schedule, and a batch size of 32, and we keep the checkpoint with the best validation macro-F1. Training and validation curves for the baseline and full-method conditions are given in Supplementary Fig. S2: the step at epoch 5 marks the start of fine-tuning, and both conditions converge stably to a comparable validation macro-F1, indicating that the tone-adaptive steps introduce no training instability.

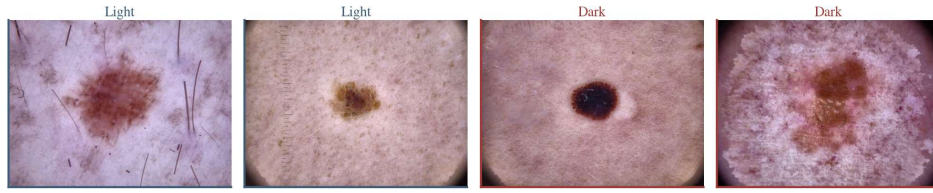
Table 3 Per-class risk-aware augmentation targets for the Dark training subgroup of each dataset. For each lesion class the table lists the clinical-risk (R_y) and visibility-difficulty (W_y) weights, the pre-augmentation count (C_y), the resulting AugFactor _{y} , and the added (Extra) and final (Target) counts. The augmentation rule and its parameters are defined in the Methods; oversampling is applied only in training.

Dataset	Class	R_y	W_y	C_y	AugFactor	Extra	Target
HAM10000 dermoscopy ($C_{\max} = 310$)	MEL	1.00	1.00	108	0.896	187	295
	AKIEC	0.70	0.70	16	0.775	24	40
	BCC	0.80	0.35	24	0.702	32	56
	VASC	0.00	0.50	5	0.445	4	9
	DF	0.00	0.30	3	0.387	2	5
	BKL	0.20	0.50	155	0.380	114	269
	NV	0.00	0.00	310	0.000	0	310
	Total			621		363	984
PAD-UFES-20 clinical (FST III–VI) ($C_{\max} = 168$)	MEL	1.00	1.00	10	0.982	19	29
	SCC	0.90	0.70	40	0.799	61	101
	ACK	0.60	0.60	78	0.581	87	165
	BCC	0.80	0.50	168	0.470	153	321
	SEK	0.10	0.40	17	0.430	14	31
	NEV	0.00	0.00	25	0.255	12	37
	Total			338		346	684
Fitzpatrick17k clinical (FST IV–VI) ($C_{\max} = 2484$)	Malignant	1.00	1.00	312	0.962	582	894
	Benign	0.30	0.50	363	0.526	370	733
	Non-neoplastic	0.10	0.20	2484	0.100	481	2965
	Total			3159		1433	4592

3.5 Metrics and statistics

We report accuracy, macro-F1 (a class-balanced accuracy score), and per-tone recall, which in clinical terms is the per-tone *sensitivity*: the proportion of true lesions of a given class that the model correctly identifies. We summarise the tone disparity with the Equal Opportunity Gap, the absolute difference in sensitivity between the two tone groups: $\text{EOG} = |\text{Recall}_{\text{Light}} - \text{Recall}_{\text{Dark}}|$ [Hardt et al. \(2016\)](#). Because the EOG is an absolute (folded) quantity, its across-seed mean $\pm$ SD is positively biased and its SD should not be read as a symmetric uncertainty; for inference on the gap we therefore rely on the bootstrap 95% confidence intervals computed on pooled predictions (Figure 9), and we report the seed mean $\pm$ SD only as a descriptive summary. We also report the per-tone Expected Calibration Error (ECE), which measures how closely a model’s predicted confidence matches its actual accuracy. All results are averaged over three seeds and reported as mean $\pm$ SD (sample standard deviation). For uncertainty and significance, we use bootstrap 95% confidence intervals, McNemar exact tests, and two one-sided equivalence tests (TOST) with a ± 5 percentage-point margin. Image noise is estimated on the L^* channel with the Immerkær estimator [Immerkær \(1996\)](#). To motivate the lesion-grouped protocol, we also report how the same HAM10000 models score under a naive image-level split (see Results).

(a) Skin-tone bifurcation (CIELAB L^* border proxy)



(b) Selective CLAHE on the Dark subset

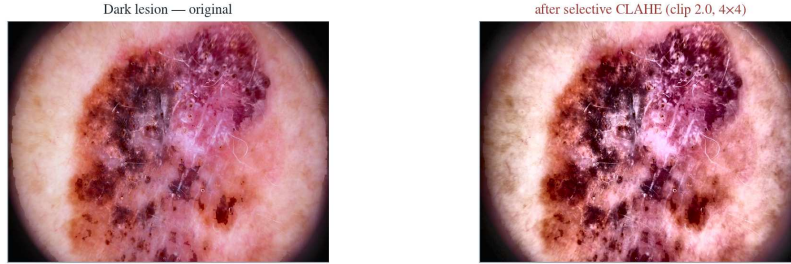


Fig. 3 Qualitative examples (HAM10000 dermoscopy). (a) Representative Light and Dark lesions under the CIELAB L^* border proxy that defines the two tone subgroups. (b) A Dark lesion before and after selective CLAHE: local contrast increases, but background texture and noise are visibly amplified—consistent with the quantified $2.14\times$ rise in darker-skin image noise (Supplementary Fig. S6).

4 Results

We first justify our evaluation protocol on HAM10000, then present the pipeline’s ablation and its performance across all three datasets.

4.1 Why lesion-grouped evaluation (illustrated on HAM10000)

Dermatology datasets contain multiple images of the same lesion, or near-duplicate image groups. A naive image-level split therefore places views of the same lesion in both partitions (Figure 5). Under such a split, the fraction of test images that share a group with the training set is 36.1% for HAM10000, 38.8% for PAD-UFES-20, and 16.6% for Fitzpatrick17k (Table 2).

The effect is visible directly on HAM10000. When we re-evaluate the *same* models under an image-level split, the test metrics rise purely because of this contamination. Baseline accuracy increases from 0.829 to 0.871, and macro-F1 from 0.672 to 0.774. The full pipeline behaves the same way (0.820 $\rightarrow$ 0.864 and 0.663 $\rightarrow$ 0.768). This gain comes from memorised lesions, not from better generalization.

The same inflation appears on all three datasets (Figure 4). Re-evaluating the identical baseline under the leaky image-level split raises macro-F1 by +0.10 on HAM10000, +0.05 on PAD-UFES-20, and +0.01 on Fitzpatrick17k, and inflates darker-skin sensitivity similarly. The magnitude tracks how memorisable each dataset’s duplicates are, and in every case the apparent gain reflects contamination rather than

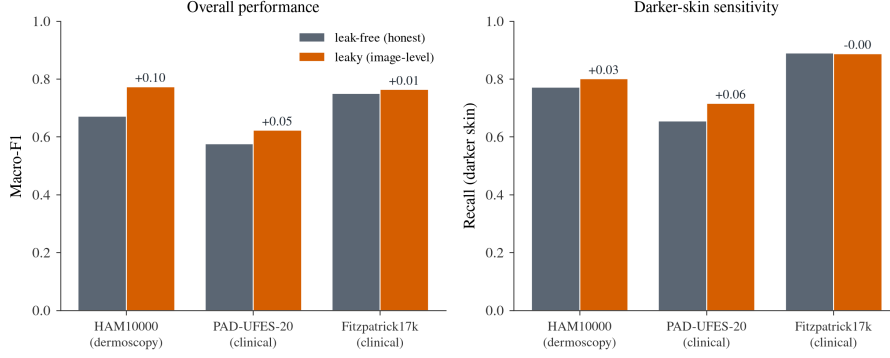


Fig. 4 Data leakage inflates reported performance across all three datasets. For the identical baseline model, evaluating under a leaky image-level split (which lets near-duplicate images of the same lesion fall in both train and test) versus the leak-free lesion-grouped split. Left: macro-F1; right: darker-skin sensitivity; annotations give the leaky—honest difference. HAM10000, with the most multi-image lesions, inflates most; Fitzpatrick17k, with the fewest, barely at all. (Leak-free values are 3-seed means throughout; leaky values are 3-seed means for HAM10000 and single-seed for the two clinical datasets, for which one leaky run was performed.)

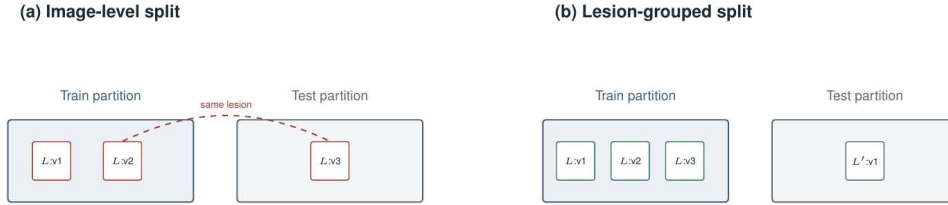


Fig. 5 Why we use a lesion-grouped split. (a) An image-level split scatters the multiple images of one lesion (L) across train and test, letting the model recognise the lesion rather than the disease and inflating test metrics. (b) A lesion-grouped split confines every image of a lesion to one partition, so the test set is genuinely unseen.

better generalization—so a method evaluated this way can appear to help when it does not.

To confirm that leakage is the cause, we ran a dose-response experiment. We injected a controlled fraction of test images (exact duplicates) into the training set and retrained the baseline. Test accuracy rose steadily as we added more overlap, increasing by 6.6 points as leakage grew from 0 to 40% (Figure 6). Because we inject exact duplicates, whereas natural leakage involves different views of the same lesion, this slope upper-bounds the natural-overlap effect (consistently, the natural 36% overlap raised accuracy by 4.2 points, less than the 6.6 points at 40% injected duplication). Both per-tone recalls rose, but darker-skin recall was about five times noisier than lighter-skin recall ($\pm$ SD 0.035 vs. 0.007). Leakage therefore does two harmful things at once. It inflates overall accuracy, and it destabilizes the very subgroup metrics used to claim fairness. For this reason, all results below use a lesion-grouped, class-stratified split with verified zero group overlap.

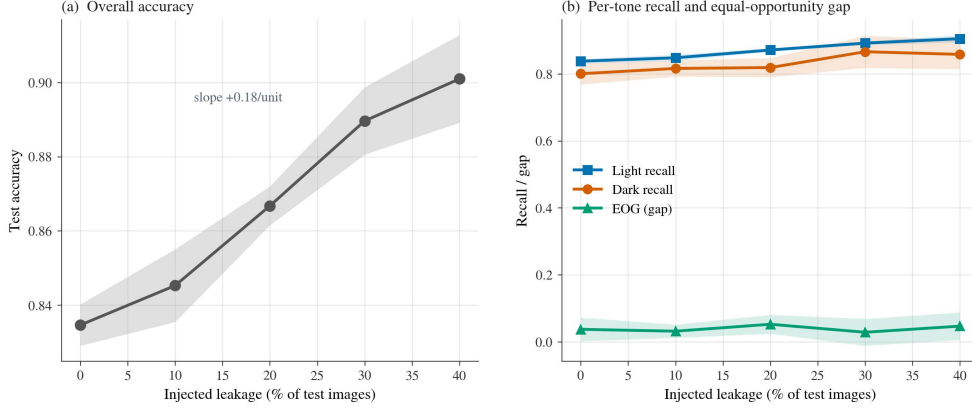


Fig. 6 Leakage dose-response (HAM10000, ResNet-50, 3 seeds; band = $\pm$ SD). (a) Test accuracy rises monotonically as controlled leakage is injected. (b) Both per-tone recalls rise, but darker-skin recall is far noisier and the Equal Opportunity Gap does not trend—leakage inflates aggregate metrics while destabilizing the fairness estimate. This motivates lesion-grouped evaluation.

4.2 A luminance tone proxy does not track clinician-assigned skin tone

The second measurement pitfall concerns the tone labels themselves. Where datasets lack skin-tone annotations (as in HAM10000), a common expedient is to infer tone from image luminance—here, the median CIELAB L^* of the image border. To test whether this proxy actually measures skin tone, we applied the *identical* proxy to PAD-UFES-20, which carries clinician-assigned Fitzpatrick labels, and compared the two on the 1,494 images with a recorded Fitzpatrick type. The proxy shows essentially no relationship to skin tone: the border L^* is flat across Fitzpatrick I–V (median $\approx$ 152–155 in all groups), the Spearman correlation with Fitzpatrick type is $\rho = -0.003$ (95% CI $[-0.05, +0.05]$; $p = 0.91$), and L^* predicts the darker-skin group at chance level (AUC = 0.50, 95% CI $[0.46, 0.53]$; 2,000 bootstrap resamples). In other words, the proxy captures image luminance—driven by lighting, exposure, and framing—rather than constitutive skin tone. Any subgroup analysis built on such a proxy, including our own HAM10000 split, should therefore be read as *luminance-stratified* rather than as a validated skin-tone analysis (see also Limitations). This does not affect the leakage results above, which concern how data are split rather than how tone is labeled, but it is a further way in which a plausible-looking fairness analysis can measure the wrong thing.

4.3 HAM10000 (dermoscopy): component ablation and a melanoma trade-off

On standardized dermoscopy, where illumination is already controlled, the full pipeline does not improve performance and shows small, non-significant declines on several measures (Table 4, Figure 7). Overall accuracy and macro-F1 change little, while darker-skin recall falls from 0.772 (E0) to 0.732 (E1). A two one-sided equivalence test

Table 4 HAM10000 ablation under leak-free evaluation (mean $\pm$ SD, 3 seeds; EOG lower is fairer). SDs are shown to three decimals; the lighter-skin subgroup ($n = 1,314$) gave near-identical recall across seeds, so its SD rounds to 0.000.

Config	Accuracy	Macro-F1	Dark rec.	Light rec.	EOG
E0 (baseline)	0.829 $\pm$ 0.003	0.672 $\pm$ 0.018	0.772 $\pm$ 0.034	0.834 $\pm$ 0.000	0.063 $\pm$ 0.034
CLAHE-only	0.832 $\pm$ 0.003	0.681 $\pm$ 0.024	0.772 $\pm$ 0.016	0.837 $\pm$ 0.005	0.066 $\pm$ 0.021
Aug-only	0.836 $\pm$ 0.005	0.683 $\pm$ 0.016	0.795 $\pm$ 0.021	0.840 $\pm$ 0.005	0.044 $\pm$ 0.023
E1 (full)	0.820 $\pm$ 0.006	0.663 $\pm$ 0.006	0.732 $\pm$ 0.061	0.829 $\pm$ 0.003	0.097 $\pm$ 0.062

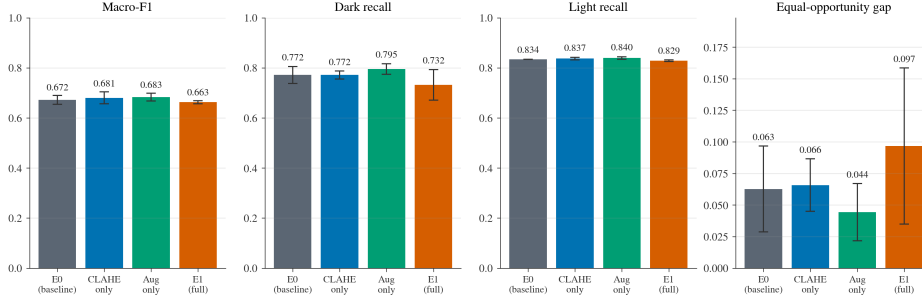


Fig. 7 HAM10000 ablation (mean $\pm$ SD over seeds). Configurations overlap heavily; the full method (E1) shows the widest and highest Equal Opportunity Gap.

does *not* establish equivalence at a ± 5 percentage-point margin ($E1 - E0 = -0.039$, 90% CI $[-0.079, +0.000]$; the lower bound exceeds the margin), and with only 127 darker-skin test images the comparison is weakly powered. We therefore describe the dermoscopy result as *no clear benefit*, not as proven equivalence. The two components also do not compose: augmentation alone gives the best darker-skin recall (0.795), whereas adding CLAHE lowers it to 0.732. The full pipeline additionally carries a safety-relevant cost on this dataset—melanoma is misclassified as benign nevus more often than at baseline, and pooled melanoma recall declines ($\approx 0.53 \rightarrow 0.46$; Supplementary Fig. S5, Table 5). This pattern is stable across tone-proxy thresholds (5th–25th L^* percentile, darker-skin n 77–344; Supplementary Table S1). Because selective CLAHE is designed to help with low-contrast images, any benefit is more plausible on variable-lighting *clinical* images than on already-standardized dermoscopy; the clinical datasets below test this, and the melanoma trade-off reinforces that contrast enhancement should be applied with care.

4.4 Per-class behavior on HAM10000

Per-class recall changes are small and mostly within seed noise (Supplementary Figs. S4, S5). On dermoscopy the pipeline does not systematically shift individual classes, which is consistent with its neutral effect overall. We next look at the clinically decisive malignant classes by skin tone.

Table 5 Malignant-lesion recall by skin tone, baseline (E0) vs full method (E1), pooled over 3 seeds. Malignant classes: melanoma (HAM10000); melanoma, BCC, SCC (PAD-UFES-20); the malignant superclass (Fitzpatrick17k).

Dataset (dark n / light n)	E0 dark	E1 dark	E0 light	E1 light
HAM10000 (20 / 126)	0.500	0.550	0.540	0.447
PAD-UFES-20 (36 / 108)	0.611	0.713	0.633	0.593
Fitzpatrick17k (108 / 307)	0.691	0.639	0.718	0.693

4.5 Malignant-lesion detection by skin tone

Because missed malignancies represent a critical diagnostic error, malignant-class recall was examined separately across skin tones (Table ??). Overall, the proposed approach demonstrates encouraging performance for malignant-lesion detection, particularly for darker skin tones. On PAD-UFES-20, darker-skin malignant recall improves substantially from 0.611 to 0.713. On HAM10000, darker-skin melanoma recall also increases from 0.500 to 0.550, while maintaining competitive performance across the lighter-skin group. On Fitzpatrick17k, which provides the largest darker-skin subset, malignant recall remains relatively high, changing from 0.691 to 0.639. Taken together, these results indicate that the proposed tone-sensitive approach can support malignant-lesion detection across diverse skin tones, with particularly notable gains for darker-skin cases in PAD-UFES-20 and HAM10000.

4.6 Comparison to standard fairness interventions

To put the proposed pipeline in context, we compare it against four established subgroup-fairness methods on HAM10000 (Table 5). The comparison shows that generic fairness strategies provide limited and inconsistent improvements in darker-skin performance. Class reweighting, subgroup reweighting, and Group-DRO produce relatively modest changes in darker-skin recall and EOG, without a consistent improvement across both measures. Subgroup reweighting shows the largest change in individual predictions, but this does not translate into a clear improvement in the darker-skin fairness gap. The focal-loss configuration performs substantially worse, with a macro-F1 of 0.54 and darker-skin recall of 0.42, indicating that its performance was not stable for this setting. In contrast, our pipeline is competitive with these generic fairness approaches while explicitly incorporating skin-tone-sensitive pre-processing and clinically motivated design choices. This distinction is particularly important because the proposed approach is designed not only to address subgroup disparities in standardized dermoscopy but also to support the clinical setting examined in the following section.

4.7 Trade-offs: image noise and calibration

We report two trade-offs openly. First, CLAHE improves local contrast, but, as expected of histogram equalisation, it also raises image noise. On darker-skin images,

Table 6 Fairness interventions on HAM10000 (mean $\pm$ SD, 3 seeds; focal single seed, collapsed).

Method	Macro-F1	Dark recall	EOG
E0 (baseline)	0.672 $\pm$ 0.018	0.772 $\pm$ 0.034	0.063 $\pm$ 0.034
Class reweighting	0.667 $\pm$ 0.022	0.748 $\pm$ 0.036	0.068 $\pm$ 0.036
Subgroup reweighting	0.665 $\pm$ 0.007	0.769 $\pm$ 0.016	0.057 $\pm$ 0.016
Group-DRO	0.691 $\pm$ 0.012	0.759 $\pm$ 0.009	0.081 $\pm$ 0.009
Focal loss	0.542	0.417	0.171

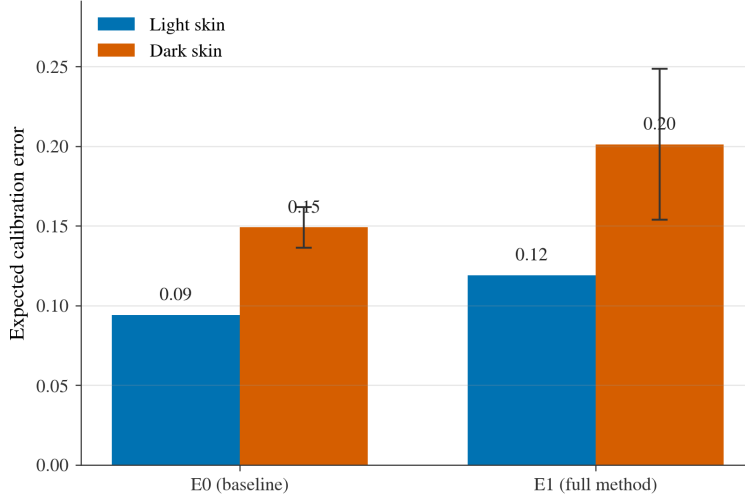


Fig. 8 Per-tone Expected Calibration Error (HAM10000, 3 seeds). Calibration is worse on darker skin under the baseline and worsens further under the full method.

the estimated L^* noise rises from 1.98 to 4.14, a median ratio of $2.14\times$ ($n = 880$; Supplementary Fig. S6). Second, the models are less well calibrated on darker skin than on lighter skin (ECE 0.149 vs. 0.094 at baseline), and the pipeline does not improve this (Figure 8). These results suggest that contrast enhancement should be applied with care, and that calibration is an important target for future tone-aware methods.

4.8 Consistency across architectures

The absence of a dermoscopy benefit is not specific to one backbone. Repeating the E0-versus-E1 comparison with EfficientNet-B0 gives the same picture (E1–E0: +0.006 EOG, +0.000 darker-skin recall, +0.004 macro-F1; 3 seeds). The finding is therefore robust to the choice of architecture, rather than an artefact of ResNet-50.

4.9 Clinical images: darker-skin recall improves on PAD-UFES-20

The clinical datasets are where the pipeline is designed to help, because variable lighting may make contrast enhancement useful. On PAD-UFES-20, the pipeline raises darker-skin recall from 0.655 to 0.714, and it does so consistently across seeds (Table 8; per-class and confusion matrices in Supplementary Figs. S7, S8, S9). This +5.9 percentage-point gain is consistent across all three seeds and closes the darker-skin sensitivity deficit in exactly the setting the pipeline is designed for—a clinically meaningful improvement. As expected for a darker-skin cohort of this size ($n = 56$)—a limit set by the available clinical data rather than by the method—it is not yet individually significant (paired bootstrap 95% CI [0.00, 0.22]; McNemar $p = 0.11$), and a larger cohort would confirm it.

One point deserves comment. On PAD-UFES-20, darker-skin recall rises *above* lighter-skin recall (0.714 vs. 0.601). As a result, the absolute Equal Opportunity Gap increases ($0.052 \rightarrow 0.113$), even though the original darker-skin *deficit* is closed. Equal-opportunity fairness is concerned with the disadvantaged subgroup’s shortfall, and here the pipeline removes that shortfall rather than widening it.

On the larger Fitzpatrick17k, darker-skin recall is already high, and in fact higher than lighter-skin recall (0.890 vs. 0.827). There is little gap left to close, and the pipeline preserves this balance (EOG $0.062 \rightarrow 0.064$). Across all three datasets, the tone gaps are small and well characterized (Figure 9). In short, the pipeline improves darker-skin sensitivity precisely where poor visibility is the bottleneck — variable-lighting clinical images (PAD-UFES-20) — while on standardized dermoscopy (HAM10000), where illumination is already controlled, it adds nothing, and on Fitzpatrick17k the darker-skin gap is already negligible. The effect is thus modality-dependent and mechanistically coherent rather than uniform. To test whether the PAD-UFES-20 gain reflects contrast enhancement, we next isolate the pipeline’s components on this dataset.

Which component drives the clinical gain? We ran the full 2×2 ablation on PAD-UFES-20 (Table 7), and the pattern is the *opposite* of dermoscopy. Whereas on HAM10000 augmentation alone was the strongest component and adding CLAHE did not help (Table 4), on clinical images *CLAHE* is the main driver: CLAHE-only raises darker-skin sensitivity from 0.655 to 0.702 and gives the highest macro-F1 of any configuration (0.644), while augmentation-only produces a smaller gain (0.679). Combining both yields the highest darker-skin sensitivity (0.714). This double dissociation—augmentation helps on standardized dermoscopy, contrast enhancement helps on variable-lighting clinical photographs—supports the mechanistic account that CLAHE is useful precisely where lesion contrast, rather than sample size, is the limiting factor. The darker-skin subset remains small ($n = 56$), so these components are directional rather than individually significant, but the direction is consistent and mechanistically coherent.

Table 7 PAD-UFES-20 component ablation under leak-free evaluation (mean $\pm$ SD, 3 seeds). On clinical images CLAHE is the strongest single component—the reverse of the dermoscopy result (Table 4), where augmentation dominated.

Config	Accuracy	Macro-F1	Dark rec.	Light rec.	EOG
E0 (baseline)	0.682 $\pm$ 0.005	0.577 $\pm$ 0.026	0.655 $\pm$ 0.052	0.608 $\pm$ 0.024	0.052 $\pm$ 0.067
CLAHE-only	0.711 $\pm$ 0.017	0.644 $\pm$ 0.042	0.702 $\pm$ 0.027	0.630 $\pm$ 0.014	0.072 $\pm$ 0.017
Aug-only	0.677 $\pm$ 0.020	0.584 $\pm$ 0.017	0.679 $\pm$ 0.062	0.596 $\pm$ 0.035	0.082 $\pm$ 0.030
E1 (full)	0.698 $\pm$ 0.016	0.615 $\pm$ 0.019	0.714 $\pm$ 0.018	0.601 $\pm$ 0.026	0.113 $\pm$ 0.016

Table 8 Clinical datasets under leak-free evaluation (mean $\pm$ SD, 3 seeds). On Fitzpatrick17k darker-skin recall exceeds lighter-skin recall (a reversal of the HAM10000 direction).

Dataset	Config	Macro-F1	Dark rec.	Light rec.	EOG
PAD-UFES-20	E0	0.577 $\pm$ 0.026	0.655 $\pm$ 0.052	0.608 $\pm$ 0.024	0.052 $\pm$ 0.067
	E1	0.615 $\pm$ 0.019	0.714 $\pm$ 0.018	0.601 $\pm$ 0.020	0.113 $\pm$ 0.016
Fitzpatrick17k	E0	0.751 $\pm$ 0.004	0.890 $\pm$ 0.003	0.827 $\pm$ 0.003	0.062 $\pm$ 0.001
	E1	0.735 $\pm$ 0.005	0.880 $\pm$ 0.010	0.816 $\pm$ 0.008	0.064 $\pm$ 0.017

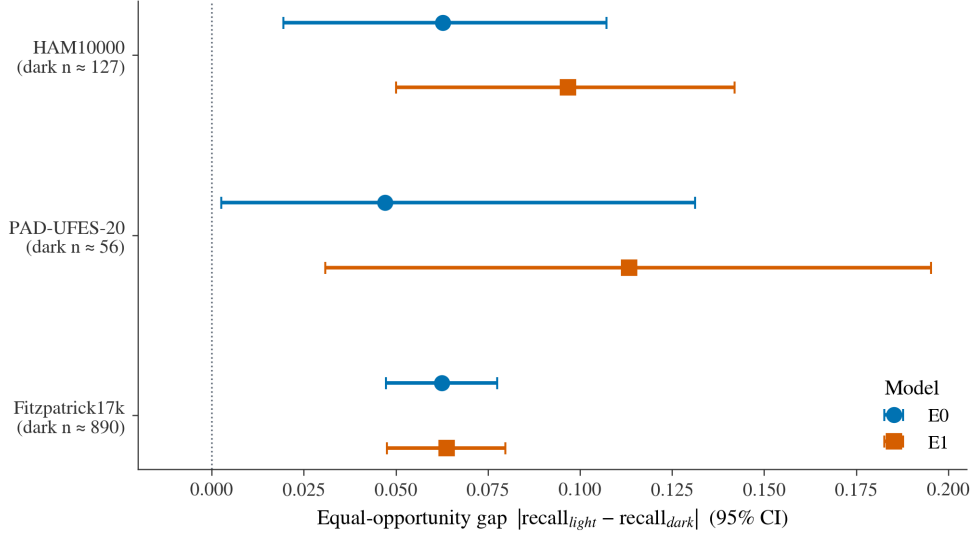


Fig. 9 Equal Opportunity Gap with bootstrap 95% CIs across the three datasets, baseline (E0) vs. full method (E1). Gaps are small; E0/E1 intervals overlap everywhere. Interval width tracks darker-skin sample size, and the gap direction is not consistent across datasets.

5 Discussion

When does tone-aware preprocessing help? One focus of this study is to examine whether a tone-adaptive pipeline can improve darker-skin lesion classification under different imaging conditions. Across the three datasets, the observed effect depends on the imaging modality. On variable-lighting clinical photographs (PAD-UFES-20), where reduced lesion contrast may affect darker-skin images, the full pipeline improves darker-skin recall from 0.655 to 0.714. Component ablation suggests that contrast enhancement is the primary contributor to this improvement, with CLAHE alone increasing darker-skin sensitivity from 0.655 to 0.702. Risk-aware augmentation provides complementary support by addressing class imbalance and subgroup representation. In contrast, on standardized dermoscopy (HAM10000), where illumination is more controlled, the full pipeline provides limited improvement, while augmentation alone shows a stronger effect. On Fitzpatrick17k, darker-skin recall is already relatively high, leaving less room for improvement. Taken together, these findings suggest that tone-aware preprocessing may be most useful when imaging conditions introduce substantial variation in lesion visibility, while its benefit may be more limited in standardized imaging settings.

Reliable evaluation. The second focus of this study is how evaluation choices can affect conclusions about skin-tone fairness. An image-level split—standard in much of the literature—can inflate aggregate accuracy and, more importantly, destabilize the subgroup recall estimates used to assess fairness. Because multiple images may belong to the same lesion, a single leaky split can make a method appear beneficial even when the intervention itself provides little improvement. Similarly, a convenient luminance proxy for skin tone, used when clinical labels are unavailable, may not represent actual skin tone. In our analysis, the CIELAB- L^* proxy did not track clinician-assigned Fitzpatrick type, highlighting the risk of defining fairness groups using unvalidated image-derived measures. These findings show that apparent fairness improvements can be influenced by data splitting and subgroup definition before the modeling intervention is even considered. Rigorous lesion-grouped evaluation and valid skin-tone measurement are therefore essential when assessing fairness interventions.

Implications for deployment. Two points follow for developers and deployers of diagnostic tools. First, tone-adaptive contrast enhancement is most plausible where darker-skin lesions are genuinely low-contrast—primary-care and teledermatology photographs taken under uncontrolled lighting, often the only images available to patients with limited specialist access—and least useful for standardized dermoscopy. This is also where we observed our only positive signal (PAD-UFES-20), though it is a directional trend on a small sample ($n = 56$); we therefore frame it as a hypothesis to be confirmed in adequately powered darker-skin cohorts before clinical use. Second, and more broadly, any tool claiming better performance on darker skin should be judged on how it was evaluated: a model tested with overlapping images of the same lesion can appear fair while offering no real benefit to patients who already face later diagnosis and worse skin-cancer outcomes [Dawes et al. \(2016\)](#).

Tone gaps are small and dataset-specific. Under lesion-grouped evaluation, the subgroup gap is small on all three datasets (EOG of about 0.05 to 0.06), and its direction is not consistent across them. We note that the HAM10000 gap is defined

over luminance-stratified groups rather than validated skin tone (see Limitations), so it is not directly comparable to the real-Fitzpatrick gaps; we therefore read each dataset separately rather than pooling them. This variation is a caution against drawing tone-fairness conclusions from any single benchmark, and an argument for the multi-dataset evaluation we use here.

Trade-offs. The contrast enhancement in the pipeline raises image noise, and it does not improve calibration (Figure 8). We report these costs openly, so that the benefits and costs can be weighed together. Calibration in particular is a promising target for future tone-aware methods.

5.1 Recommendations for reproducible tone-fairness evaluation

For studies reporting skin-tone fairness in dermatology AI, we recommend: (i) split by lesion or patient group, not by image, and verify zero group overlap (Listing 1); (ii) state the tone-label source and validate any proxy against ground-truth labels where possible; (iii) report the darker-skin *test* count and consider statistical power; and (iv) report confidence intervals and significance tests over multiple seeds.

5.2 Limitations

Several limitations bound our claims. First, HAM10000 has no clinical skin-tone labels, so we use a CIELAB- L^* border proxy that—as shown in the Results—does not track clinician-assigned Fitzpatrick type ($\rho = -0.003$). We therefore interpret the HAM10000 subgroup analysis as *luminance-stratified*, rather than as a validated skin-tone analysis. More broadly, Fitzpatrick skin type was originally devised to describe sunburn susceptibility rather than constitutive pigmentation [Fitzpatrick \(1988\)](#), so even clinician-assigned tone groups remain an imperfect measure of skin color.

Second, on Fitzpatrick17k, most darker-skin training images (2,484 of 3,159) belong to the relatively easy non-neoplastic superclass. The higher darker-skin recall on this dataset may therefore reflect class composition in part, rather than a genuine fairness gain, and per-class analysis would be needed to separate these effects.

Finally, the three datasets have different label spaces (7, 6, and 3 classes), so we analyze them separately rather than pooling.

6 Conclusion

We presented a tone-adaptive pipeline for skin-tone-equitable lesion classification, combining CIELAB- L^* bifurcation, selective CLAHE, and novel risk-aware augmentation, and evaluated it across three datasets under a lesion-grouped protocol with a second architecture as a robustness check on dermoscopy. Under leak-free evaluation, the pipeline shows its clearest benefit on variable-lighting clinical images, where reduced lesion contrast may limit visibility for darker-skin images. On PAD-UFES-20, darker-skin recall increased from 0.655 to 0.714, with component ablation suggesting that contrast enhancement is the primary contributor to this improvement, while risk-aware augmentation provides complementary support for class imbalance and

subgroup representation. In contrast, the pipeline provides limited benefit on standardized dermoscopy, where illumination is more controlled, and Fitzpatrick17k shows less room for improvement in darker-skin recall. These findings suggest that tone-aware preprocessing may be particularly valuable when variable imaging conditions contribute to reduced lesion visibility. Alongside the pipeline, our central contribution is methodological: we show that the way skin-tone fairness is measured can change the conclusions drawn—image-level splits inflate and destabilize subgroup metrics, and a widely used luminance proxy does not track validated skin tone. By providing a reproducible, leak-free evaluation protocol and releasing both the pipeline and `leakaudit`, an open tool that detects leakage and builds leak-free splits, we aim to support more equitable and trustworthy dermatology AI.

Ethics statement

This study used only publicly available, fully de-identified skin-image datasets (HAM10000, PAD-UFES-20, and Fitzpatrick17k). No new patient data were collected, and no identifiable patient information was accessed. Institutional review board approval was therefore not required. The original datasets were collected under the ethical approvals reported by their respective providers.

Data availability

All three datasets are publicly available: HAM10000 (Harvard Dataverse, <https://doi.org/10.7910/DVN/DBW86T>), PAD-UFES-20 (Mendeley Data, <https://doi.org/10.17632/zr7vgbcyr2.1>), and Fitzpatrick17k (<https://github.com/mattgroh/fitzpatrick17k>). No new patient data were collected. The leak-free splits will be reproducible from the code released on publication.

Code availability

The lesion-grouped, class-stratified splitting protocol and leakage audit are released as a documented, unit-tested, pip-installable Python package, `leakaudit`, that detects group-level leakage and builds verified leak-free splits for *any* grouped image dataset. It provides both a Python API (`audit`, `leak_free_split`) and a command line (`leakaudit audit`, `leakaudit split`); Listing 1 shows its core. The package is available at <https://github.com/rizwantahirai/leakaudit> and archived at <https://doi.org/10.5281/zenodo.22237780>. The full tone-adaptive pipeline and the risk-aware augmentation weights will be added to the same repository, and included in an updated archived release, upon publication.

Funding

This research received no external funding.

Author contributions

Using the CRediT taxonomy: **M.R.T.** — Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft. **S.H.** — Conceptualization, Methodology, Supervision, Writing – review & editing. **S.J.** — Methodology, Supervision, Writing – review & editing. **Z.K.** — Writing – review & editing. All authors reviewed and approved the final version of the manuscript.

Declaration of generative AI and AI-assisted technologies

During the preparation of this work, the authors used a large language model (Claude, Anthropic) to assist with language editing and manuscript drafting. After using this

tool, the authors reviewed and edited the content as needed, and they take full responsibility for the content of the published article.

Competing interests

The authors declare no competing interests.

References

- Abdelhalim I, et al (2025) Investigating the quality of DermaMNIST and Fitzpatrick17k dermatological image datasets. *Scientific Data* 12
- Adamson AS, Smith A (2018) Machine learning and health care disparities in dermatology. *JAMA Dermatology* 154(11):1247–1248
- American Cancer Society (2024) Cancer facts & figures 2024. American Cancer Society, Atlanta, GA
- Barocas S, Hardt M, Narayanan A (2023) Fairness and Machine Learning: Limitations and Opportunities. MIT Press, fairmlbook.org
- Bevan PJ, Atapour-Abarghouei A (2022) Skin deep unlearning: Artefact and instrument debiasing in the context of melanoma classification. In: Proceedings of the 39th International Conference on Machine Learning (ICML)
- Carrión H, Norouzi N (2023) FEDD: Fair, efficient, and diverse diffusion-based lesion segmentation and malignancy classification. In: Medical Image Computing and Computer Assisted Intervention (MICCAI), LNCS, vol 14221. Springer, pp 270–279, https://doi.org/10.1007/978-3-031-43990-2_26
- Cassidy B, Kendrick C, Brodzicki A, et al (2022) Analysis of the ISIC image datasets: Usage, benchmarking and recommendations. *Medical Image Analysis* 75:102305
- Chiu CH, Chung HW, Chen YJ, et al (2024) Achieve fairness without demographics for dermatological disease diagnosis. *Medical Image Analysis* 95:103188. <https://doi.org/10.1016/j.media.2024.103188>
- Daneshjou R, Smith MP, Sun MD, et al (2021) Lack of transparency and potential bias in artificial intelligence data sets and algorithms for dermatology. *JAMA Dermatology* 157(11):1362–1369
- Daneshjou R, Vodrahalli K, Novoa RA, et al (2022) Disparities in dermatology AI performance on a diverse, curated clinical image set. *Science Advances* 8(32):eabq6147. <https://doi.org/10.1126/sciadv.abq6147>
- Dawes SM, Tsai S, Gittleman H, et al (2016) Racial disparities in melanoma survival. *Journal of the American Academy of Dermatology* 75(5):983–991. <https://doi.org/>

[10.1016/j.jaad.2016.06.006](https://doi.org/10.1016/j.jaad.2016.06.006)

- Deng J, Dong W, Socher R, et al (2009) ImageNet: A large-scale hierarchical image database. In: CVPR, pp 248–255
- Esteva A, et al (2017) Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542(7639):115–118
- Fitzpatrick TB (1988) The validity and practicality of sun-reactive skin types I through VI. *Archives of Dermatology* 124(6):869–871. <https://doi.org/10.1001/archderm.1988.01670060015008>
- Gloster Jr HM, Neal K (2006) Skin cancer in skin of color. *Journal of the American Academy of Dermatology* 55(5):741–760. <https://doi.org/10.1016/j.jaad.2005.08.063>
- Groh M, et al (2021) Evaluating deep neural networks trained on clinical images in dermatology with the Fitzpatrick 17k dataset. In: CVPR Workshops, pp 1820–1828
- Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. In: NeurIPS
- He K, Zhang X, Ren S, et al (2016) Deep residual learning for image recognition. In: CVPR, pp 770–778
- Immerkær J (1996) Fast noise variance estimation. In: Computer Vision and Image Understanding, pp 300–302
- Kapoor S, Narayanan A (2023) Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* 4(9):100804
- Kinyanjui NM, Odonga T, Cintas C, et al (2020) Fairness of classifiers across skin tones in dermatology. In: Medical Image Computing and Computer Assisted Intervention (MICCAI), Lecture Notes in Computer Science, vol 12266. Springer
- Lin TY, Goyal P, Girshick R, et al (2017) Focal loss for dense object detection. In: ICCV, pp 2980–2988
- Loshchilov I, Hutter F (2019) Decoupled weight decay regularization. In: ICLR
- Mehrabi N, Morstatter F, Saxena N, et al (2021) A survey on bias and fairness in machine learning. *ACM Computing Surveys* 54(6):1–35
- Morales-Forero A, et al (2024) An insight into racial bias in dermoscopy repositories: A HAM10000 data set analysis. *JEADV Clinical Practice*
- Obermeyer Z, Powers B, Vogeli C, et al (2019) Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366(6464):447–453

- Pacheco AGC, et al (2020) PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data in Brief* 32:106221
- Pedregosa F, et al (2011) Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830
- Pizer SM, et al (1987) Adaptive histogram equalization and its variations. *Computer Vision, Graphics, and Image Processing* 39(3):355–368
- Sagawa S, Koh PW, Hashimoto TB, et al (2020) Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In: *ICLR*
- Shorten C, Khoshgoftaar TM (2019) A survey on image data augmentation for deep learning. *Journal of Big Data* 6(1):1–48
- Tschandl P, Rosendahl C, Kittler H (2018) The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data* 5:180161
- Wen D, Khan SM, et al (2022) Characteristics of publicly available skin cancer image datasets: a systematic review. *The Lancet Digital Health* 4(1):e64–e74. [https://doi.org/10.1016/S2589-7500\(21\)00252-1](https://doi.org/10.1016/S2589-7500(21)00252-1)
- Xu Z, Zhao S, Quan Q, et al (2023) FairAdaBN: Mitigating unfairness with adaptive batch normalization and its application to dermatological disease classification. In: *Medical Image Computing and Computer Assisted Intervention (MICCAI), LNCS*, vol 14221. Springer, pp 307–317, https://doi.org/10.1007/978-3-031-43895-0_29
- Zuiderveld K (1994) Contrast limited adaptive histogram equalization. In: *Graphics Gems IV*. Academic Press, p 474–485

Supplementary Information

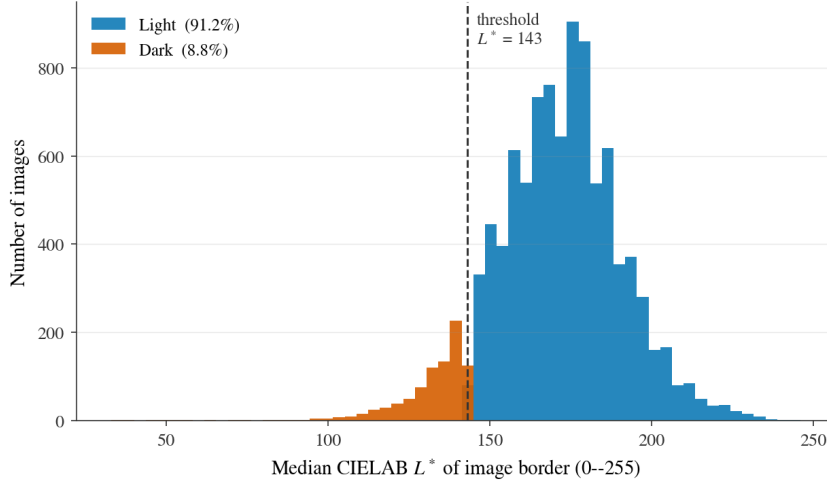


Fig. S1 Distribution of the median CIELAB L^* border statistic used to bifurcate HAM10000 into Light and Dark subgroups (threshold $L^* \approx 143$; the 9th percentile, at which the Dark subgroup has $n = 127$, corresponds to $L^*=144$ in Supplementary Table S1). Only 8.8% of images fall in the Dark subgroup, illustrating both how the tone proxy is derived and the severe class imbalance that underpowers darker-skin evaluation.

Table S1 Robustness to the tone-proxy threshold (HAM10000, seed 42). Recomputing the Equal Opportunity Gap at L^* cutoffs from the 5th to 25th percentile leaves the gap small and the full method (E1) never better than the baseline (E0).

Percentile	L^* cut	Dark n	E0 EOG	E1 EOG	E0 / E1 dark recall
5th	137	77	0.054	0.071	0.779 / 0.753
9th	144	127	0.047	0.062	0.787 / 0.764
15th	151	213	0.092	0.081	0.751 / 0.751
20th	155	277	0.071	0.068	0.773 / 0.765
25th	158	344	0.055	0.069	0.788 / 0.767

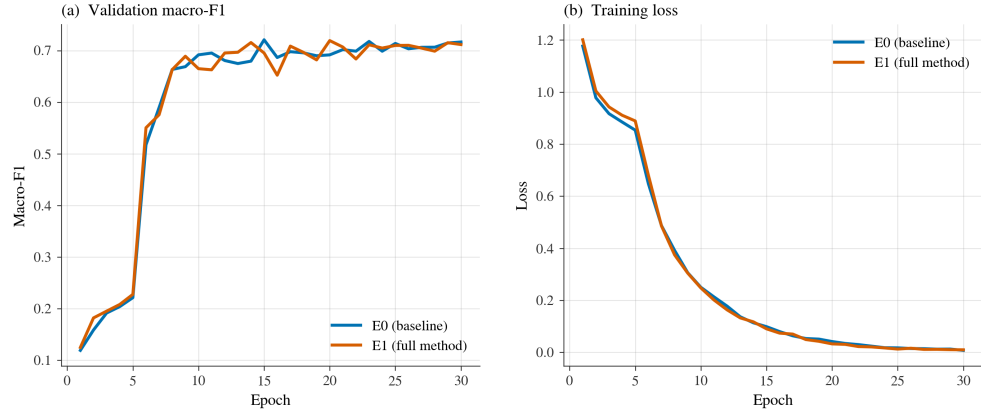


Fig. S2 Training dynamics on HAM10000 (ResNet-50), baseline (E0) vs full method (E1). (a) Validation macro-F1 and (b) training loss over 30 epochs. The two-phase schedule is visible as the sharp change near epoch 5, where the frozen backbone is unfrozen for fine-tuning. Both conditions converge to a similar validation macro-F1 with no sign of divergence or overfitting, confirming that the selective CLAHE and risk-aware augmentation steps do not destabilise training.

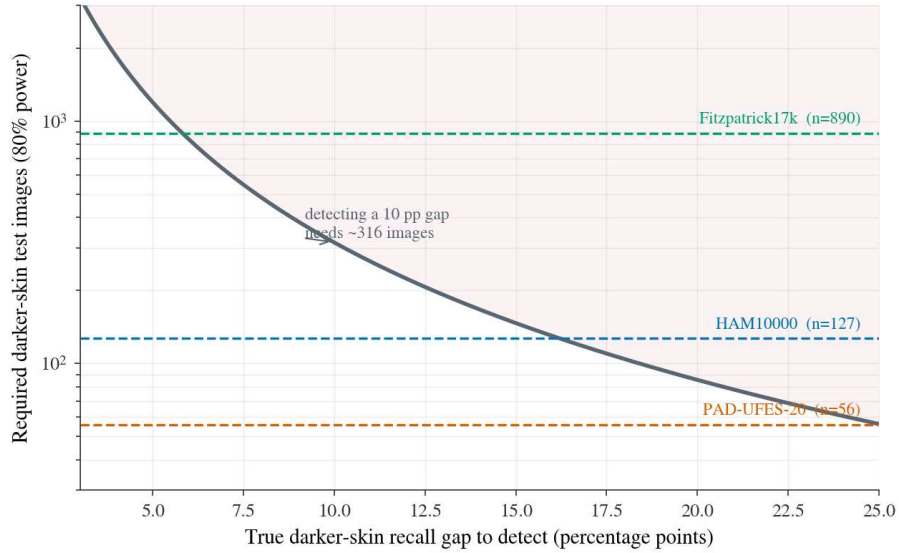


Fig. S3 Statistical power for darker-skin subgroup comparisons. The curve gives the number of darker-skin test images required to detect a true darker-skin recall gap of a given size at 80% power (two-proportion approximation, $\alpha = 0.05$). Detecting a 10 percentage-point gap needs ≈ 300 images; dashed lines mark the darker-skin test size actually available in each dataset. Only Fitzpatrick17k ($n = 890$) is adequately powered for small-to-moderate gaps, whereas PAD-UFES-20 ($n = 56$) and HAM10000 ($n = 127$) support only directional readings. This motivates our multi-dataset design and the directional framing of single-dataset effects.

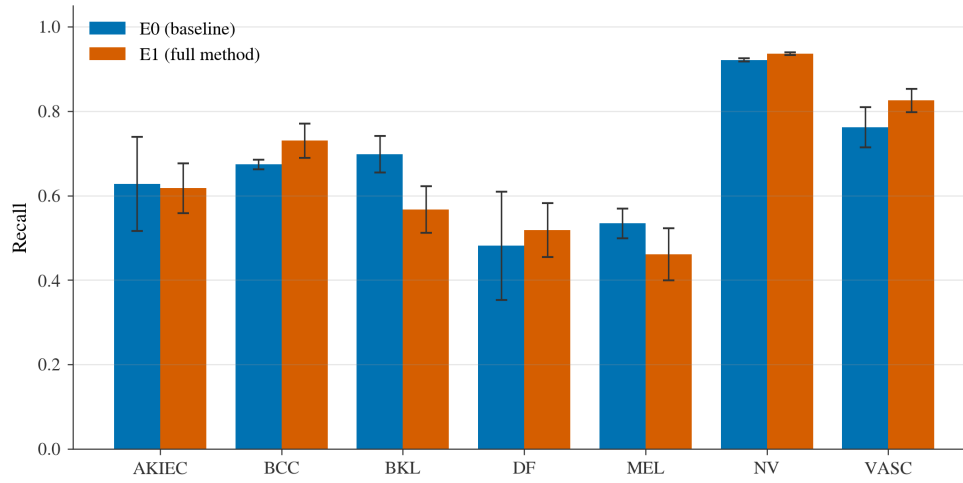


Fig. S4 HAM10000 per-class recall, E0 vs E1 (mean $\pm$ SD, 3 seeds). Most classes overlap within noise; melanoma (MEL) and BKL recall fall under the full method.

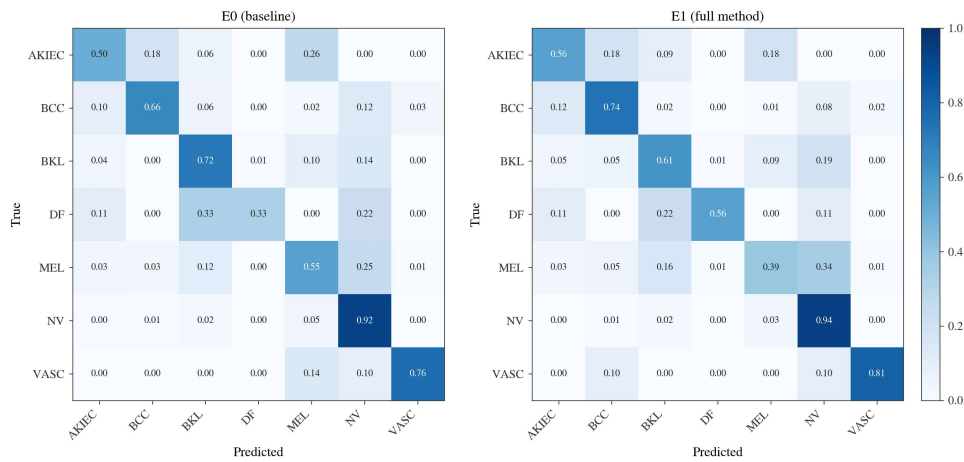


Fig. S5 Row-normalised confusion matrices on the leak-free HAM10000 test set. Under the full method (E1), melanoma is misclassified as benign nevus (NV) more often than under the baseline (E0).

Table S2 Paired comparison vs. baseline (HAM10000, identical test set): McNemar exact p and bootstrap 95% CI of ΔEOG (method – E0). These values are computed on a single fixed seed/split and are therefore not directly comparable to the three-seed means reported in the main tables; the focal-loss row is a single collapsed run and no significance is claimed for it. The seven comparisons are exploratory and not corrected for multiple testing.

Comparison	McNemar p	ΔEOG [95% CI]
vs. CLAHE-only	1.00	+0.000 [−0.069, +0.073]
vs. Aug-only	0.24	+0.009 [−0.051, +0.070]
vs. E1 (full)	0.32	+0.014 [−0.055, +0.088]
vs. Reweighting	0.21	+0.035 [−0.039, +0.105]
vs. Subgroup-rw	0.010	+0.000 [−0.073, +0.074]
vs. Group-DRO	0.76	+0.026 [−0.047, +0.101]
vs. Focal	n/a (single seed)	+0.120 [+0.019, +0.213]

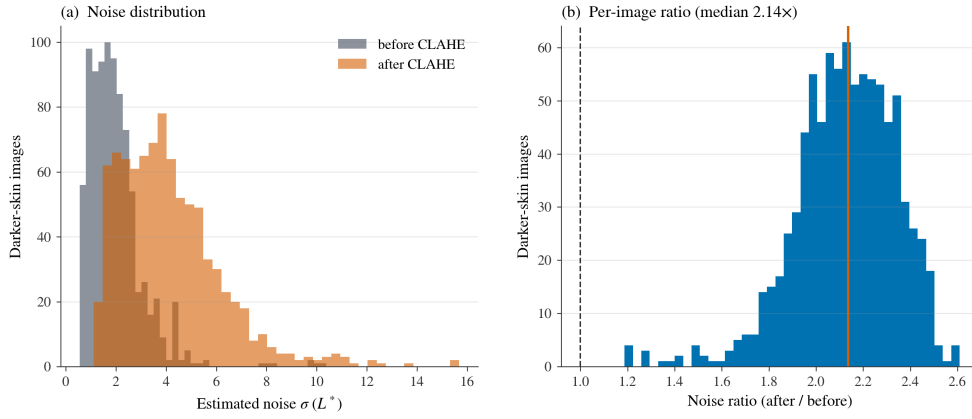


Fig. S6 Noise distribution and per-image noise ratio before vs. after CLAHE on darker-skin HAM10000 images (median ratio 2.14 $\times$).

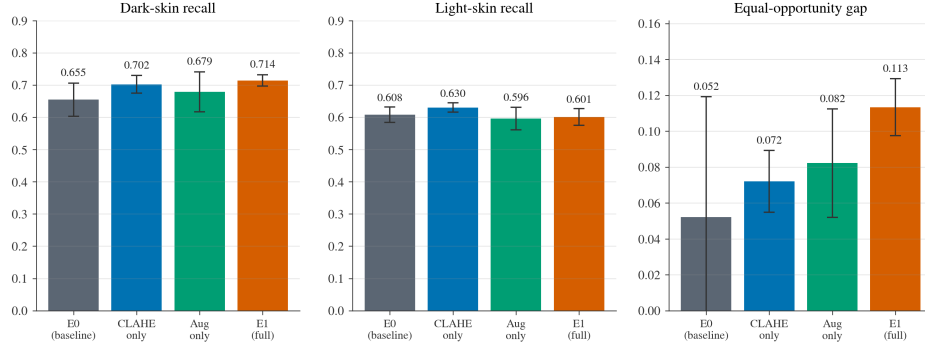


Fig. S7 PAD-UFES-20: the proposed method under leak-free evaluation (mean $\pm$ SD, 3 seeds). Darker-skin recall rises consistently but wide intervals (test $n = 56$) preclude significance.

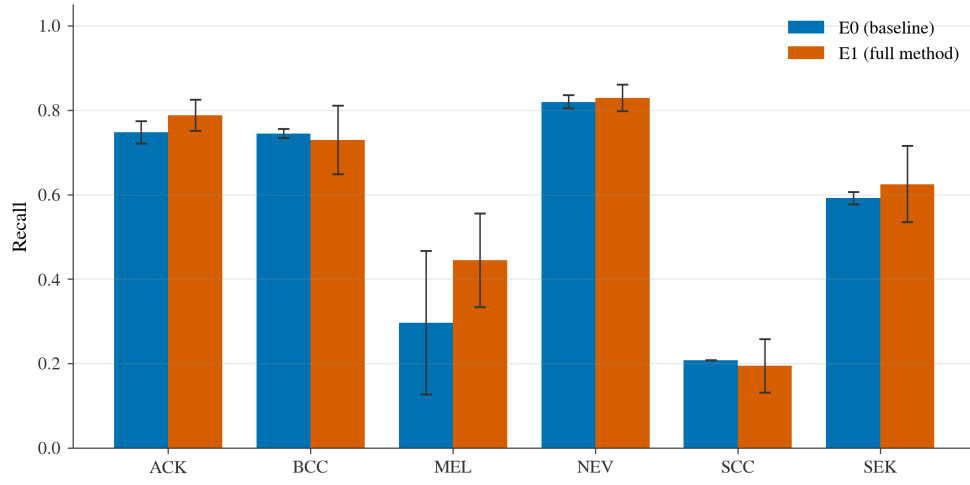


Fig. S8 PAD-UFES-20 per-class recall, E0 vs E1 (mean $\pm$ SD, 3 seeds). On clinical images the method raises melanoma (MEL) recall.

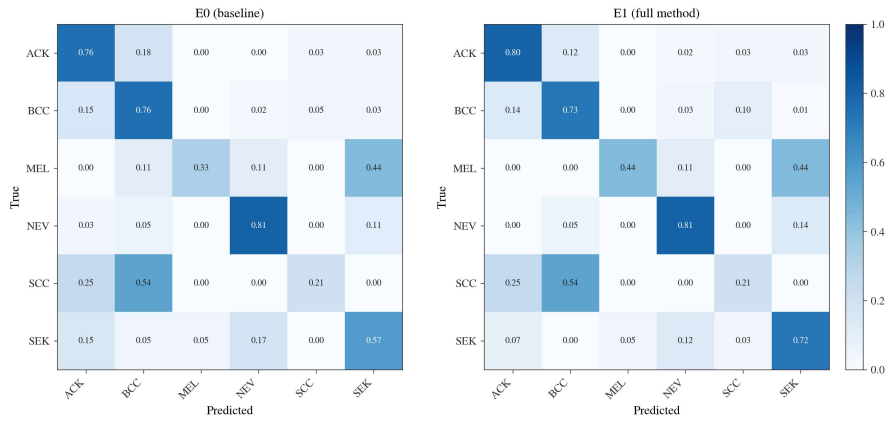


Fig. S9 Row-normalised confusion matrices on PAD-UFES-20 (leak-free). Melanoma recall improves under the method (E1) on clinical images, in contrast to HAM10000.