

Data Expansion and Fine-Tuning for Melanoma Classification in Clinical Photographs

An exploratory study of generalization and skin-tone representation

Rohit Yokananth and Neel Gandhi

Research manuscript for author review | 25 September 2026 | Not peer reviewed

Abstract

Background. Clinical-image classifiers may perform differently across image sources and skin-tone groups. A useful improvement must be distinguished from a favorable data split or a change in the balance between missed melanomas and false alarms.

Objective. To investigate whether diagnosis balancing, encoder fine-tuning, and audited additions of PAD-UFES-20 and Fitzpatrick17k images improve melanoma-versus-other-lesion classification, while assessing the feasibility of darker-skin evaluation.

Methods. EfficientNet-B0 was adapted using 4,917 clinical MILK10k images, with exploratory DDI/MIDAS assessments and earlier PAD/Fitzpatrick expansion and fine-tuning experiments. A subsequent full-Fitzpatrick audit reserved 3,101 new representatives: 1,861 training, 624 validation and 616 assessment. Two matched classifiers used the same frozen encoder; only one received the added training images. Validation selected both models and thresholds before assessment. The primary follow-up endpoint was paired assessment AUROC difference.

Results. Earlier five-split expansion failed its practical consistency rule. The earlier encoder comparison reduced false positives from 393 to 280 at 29/36 detections, mainly through one fold. In the newly reserved 616-image Fitzpatrick assessment, adding 1,861 training representatives increased AUROC from 0.747 to 0.886 (paired difference +0.140; 95% bootstrap interval +0.098 to +0.185), increased detections from 57/73 to 63/73, and reduced false positives from 265/543 to 129/543. In the primary Fitzpatrick V-VI subgroup, detections changed from 2/3 to 3/3 and false positives from 23/36 to 13/36; no assessment melanoma was V-VI under both annotation sources.

Conclusion. The matched full-Fitzpatrick follow-up showed improved overall ranking and operating-point counts on a newly reserved within-source assessment. Earlier mixed outcomes remain relevant. Unverified patient independence, supplied labels and very small, inconsistently annotated darker-skin subgroups prevent clinical-utility or reliable darker-skin superiority claims.

Keywords: melanoma; transfer learning; EfficientNet; clinical photography; data quality; skin tone; grouped evaluation.

1. Introduction and study scope

Transfer learning allows a pretrained image model to be adapted through its final classifier or through additional updates to its feature-extracting layers [1]. EfficientNet provides an established convolutional architecture for such experiments [2]. A practical research question is whether adaptations improve held-out predictions under a controlled comparison, rather than merely producing a high accuracy on one dataset.

Dermatology datasets also require attention to representation and image provenance. The DDI study examined performance disparities using a curated collection spanning skin tones [3]. Fitzpatrick17k provides clinical photographs with skin-type annotations [4], while subsequent work documented duplicate, partitioning and annotation problems in dermatology datasets [5]. These observations motivate explicit audits before training and cautious subgroup interpretation.

This project investigated binary melanoma-label prediction from clinical photographs, not CT or MRI. Clinical views were intended model inputs; supplied Fitzpatrick metadata and image review do not independently establish every image's acquisition modality. The negative class included other malignant lesions as well as benign conditions, so a negative model output cannot be interpreted as absence of skin cancer.

The study developed sequentially: class weighting and MILK encoder adaptation; transfer to DDI and MIDAS; audited expansion and five PAD group splits; matched encoder adaptation; then the full-Fitzpatrick follow-up after complete dataset access. Protocols were written before each experiment, after earlier results were known. The latest assessment was newly reserved within Fitzpatrick, while earlier assessments had been reused. This is transparent adaptive exploration, not a preregistered clinical study.

The intended contribution is a reproducible comparison of modifications, including negative results, together with a melanoma-specific audit of darker-skin data availability. The study does not introduce a new neural-network architecture, compare against practicing clinicians, establish deployment readiness, or measure cost-effectiveness.

Skin-tone or Fitzpatrick annotations were retained in their supplied form. They were not converted into race or ethnicity. MILK10k uses its own six-category tone scale, with zero darkest and five lightest [6]; those categories were not equated to Fitzpatrick I-VI.

2. Data and quality controls

Table 1. Final analysis cohorts and their roles. Counts refer to images or selected representatives, not necessarily independent patients.

Source	Eligible / melanoma	Role	Key restriction
MILK10k	4,917 / 448	Initial fitting and assessment	No supplied patient IDs
DDI	656 / 21	External assessment only	7 melanomas in each tone group
MIDAS	894 / 78	External assessment only	No melanoma in FST IV-VI
PAD-UFES-20	1,875 / 36	Expansion and grouped folds	No melanoma in FST V-VI
Fitzpatrick early subset	38 / 8	26 train, 2 validation, 10 test	Visual groups; uncertain labels
Fitzpatrick follow-up	3,101 / 376	1,861 train, 624 validation, 616 test	New groups; unverified patients

MILK10k supplies paired clinical and dermoscopic views [6]. This study retained clinical images with histopathology-confirmed, non-indeterminate labels, partitioned into 2,949 training, 984 validation and 984 assessment images (269, 90 and 89 melanoma labels). Known lesion and duplicate groups stayed within one partition. Patient independence could not be verified.

DDI melanoma mapping included melanoma in situ and specified melanoma subtypes; the cohort contained seven positive images in each supplied I-II, III-IV and V-VI group. MIDAS used the photographed lesion's pathology category, not personal melanoma history [7]. Its assessment retained exact filename matches to eligible six-inch clinical views and definite target labels. Unresolved record-ID semantics prevented patient-level inference.

PAD-UFES-20 originated as a smartphone clinical-image dataset [8]. All 2,298 downloaded images decoded and matched metadata. Twenty-six images were excluded because identical pixels had conflicting specific diagnoses, although their binary labels agreed. One highest-resolution representative was kept per supplied patient/lesion pair, followed by same-label duplicate consolidation. Standalone lesion IDs were not treated as globally unique. Supplied patient IDs and duplicate links defined conservative groups; age inconsistencies were not repaired.

Original-source Fitzpatrick downloads were matched to source hashes and reviewed against published cleaned-release membership [5]. Possible related views were grouped, conflicting skin-type annotations preserved, and one mouth-focused image excluded. The final 38 representatives contained eight melanoma-labeled visual groups, not eight verified patients. The 26 training representatives contained five melanoma labels; other-cancer comparisons comprised BCC/SCC examples.

Exact file/pixel checks and a conservative joint pHash/dHash screen were used across sources. Visual review found related-looking Fitzpatrick photographs missed by the hash screen. Absence of a detected match therefore does not establish absence of shared patients or related cases.

3. Model development and evaluation

Images were converted to RGB, resized to 256 pixels on the shorter side, center-cropped to 224 by 224, and ImageNet-normalized. EfficientNet-B0 produced 1,280 pooled features. Frozen-feature experiments fitted a training-only StandardScaler and L2 logistic classifier. No skin recoloring, synthetic darker-skin participants, or demographic inference was used.

The initial pilot compared unweighted, diagnosis-balanced, and diagnosis-plus-tone-balanced classifiers. Weights were inverse frequencies of training diagnoses or observed diagnosis/tone cells, normalized to mean one. Its regularization grid was $C = 0.01, 0.1, 1$. An initial matched fine-tuning experiment allowed up to two epochs on MILK training data; validation selected epoch zero for the head-only control and epoch one for the encoder-adapted model. Early operating thresholds maximized sensitivity subject to at least 80% validation specificity.

Expansion experiments used the selected MILK encoder with three matched training arms: current data, current plus PAD, and current plus PAD plus the fixed Fitzpatrick training subset. Diagnosis-balanced logistic fits selected $C = 0.001, 0.01, 0.1$ by the equal-weight mean of MILK and PAD validation AUROCs. Even the current-data arm used PAD validation for selection; this was adaptation rather than zero-shot transfer.

The five-split follow-up stratified 1,371 supplied-patient/duplicate PAD groups using seed 20260920. For assessment fold k , fold $(k+1)$ modulo five supplied validation data and the remaining three folds supplied fitting data. MILK training/validation partitions and Fitzpatrick training membership stayed fixed. Each PAD representative received exactly one assessment prediction per arm. This was five-fold assessment with a rotating inner holdout, not complete inner cross-validation [9].

Expansion and representation experiments selected the highest single threshold reaching at least 80% sensitivity in each validation source. This experimental rule differs from the early pilot and does not guarantee 80% assessment sensitivity. Raw confusion counts are compared only within matched experiments using the same rule.

For the earlier representation adaptation, both branches began from the same fold-specific expanded-data classifier and MILK encoder. One complete epoch used batch eight, diagnosis-weighted binary cross-entropy, AdamW, head learning rate 10^{-4} , encoder learning rate 10^{-5} , weight decay 10^{-4} and gradient clipping at five.

Batch-normalization statistics and stochastic layers stayed fixed; no augmentation was applied. Only the full-encoder branch updated image features. Validation selected epoch zero or one separately for each branch. All ten selections were saved before assessment scoring.

4. Earlier experiments and data-expansion results

Table 2. Early transfer-learning experiments. MILK and DDI values are AUROC; the later row reused the same assessment cohorts.

Experiment	MILK AUROC	DDI AUROC	Interpretation
Frozen, unweighted	0.876	0.449	Initial reference
Frozen, diagnosis balanced	0.872	0.465	No internal improvement
Frozen, diagnosis + tone	0.853	0.450	Weighting did not help internally
Initial full-encoder tuning	0.886	0.491	Small internal gain; weak transfer

The initial encoder adaptation detected 75/89 MILK melanomas versus 72/89 for its matched epoch-zero head-only control, with false positives falling from 197 to 189. The exploratory known-group AUROC-difference interval was -0.0024 to +0.0239, including zero. On MIDAS, AUROC changed from 0.6112 to 0.6192, detections from 54/78 to 56/78, and false positives from 425 to 462. These results demonstrated limited and source-dependent transfer.

In the first reserved PAD comparison (376 images, eight melanoma labels), original-data, +PAD and +PAD+Fitzpatrick classifiers each detected 5/8 melanomas, with 69, 55 and 40 false positives. AUROCs were 0.8903, 0.9185 and 0.9117. Both sequential 95% paired known-group bootstrap intervals included zero. Those encouraging operating-point counts motivated the five-split stability check; they were not treated as final independent proof.

Table 3. Five-split data expansion. Out-of-fold counts include each of 1,875 representatives once; AUROC is the mean of the five within-fold values.

Training data	Mean AUROC	TP / 36	FP / 1,839
Current	0.8980	29	399
Current + PAD	0.9026	28	374
Current + PAD + Fitzpatrick	0.9000	30	388

The combined addition detected one more melanoma and produced 11 fewer false positives overall than current data, but was favorable in only three of five folds. Each primary sequential addition reduced false positives without reducing detections in just one fold. The prespecified practical continuation gate required that pattern in at least four folds, improved aggregate counts, and no lower mean AUROC. Both sequential additions failed. The representation experiment was separately authorized as a new hypothesis; this failure was not reclassified as success.

4.1. External and supplementary expansion assessments

Table S1. All secondary assessments from the initial matched expansion experiment. These outcomes were not used to fit classifiers or select thresholds.

Assessment	Training arm	AUROC	Detected	False alarms
MILK	Current	0.897	76/89	174/895
MILK	+ PAD	0.895	84/89	218/895
MILK	+ PAD + Fitzpatrick	0.894	78/89	188/895
DDI	Current	0.508	10/21	296/635
DDI	+ PAD	0.539	12/21	314/635
DDI	+ PAD + Fitzpatrick	0.557	9/21	244/635
MIDAS	Current	0.654	60/78	481/816
MIDAS	+ PAD	0.677	57/78	442/816
MIDAS	+ PAD + Fitzpatrick	0.668	54/78	389/816
FITZ	Current	0.667	2/3	2/7
FITZ	+ PAD	0.667	2/3	3/7
FITZ	+ PAD + Fitzpatrick	0.667	2/3	2/7

MILK, DDI and MIDAS had already been assessed in earlier experiments, so repeated results are exploratory. The small Fitzpatrick assessment held ten representatives, including three melanoma-labeled groups. None of these sources provided a new, adequately powered darker-skin melanoma confirmation cohort.

The expanded classifiers changed the sensitivity-specificity balance differently across sources. On MIDAS, the combined-data model detected 54/78 melanoma images versus 60/78 for current data, while reducing false positives from 481 to 389. On DDI overall, detections decreased from 10/21 to 9/21, with false positives decreasing from 296 to 244. These trade-offs accompany, and limit the interpretation of, the favorable single-split PAD result.

The tiny Fitzpatrick assessment detected 2/3 positives under each arm. Its melanoma-versus-BCC/SCC composition and uncertain related-case identities prevent broad diagnostic conclusions. Complete tone-stratified counts are retained in the accompanying numerical experiment records; undefined subgroup sensitivity remains undefined when no positive examples are present.

5. Earlier matched representation experiment

The earlier matched head-only and full-encoder models detected 29/36 and 29/36 melanomas, with 393 and 280 false positives among 1,839 other-lesion representatives. Their mean fold AUROCs were 0.9011 and 0.9194, respectively.

Table 4. Earlier matched comparison on the same expanded fitting data and PAD assessment folds.

Version	Mean AUROC	TP / FN	FP / TN
Earlier expanded classifier (reference)	0.900	30 / 6	388 / 1451
Head-only continuation	0.901	29 / 7	393 / 1446
Full-encoder continuation	0.919	29 / 7	280 / 1559

Relative to head-only continuation, full-encoder adaptation changed total true positives by +0, false positives by -113, and mean within-fold AUROC by +0.0183. Fewer false positives without fewer detections occurred in 3/5 folds. No cross-fold statistical significance claim is made.

Fold 2 accounted for 109 of the net 113 fewer false positives; fold 3 added six and fold 1 was unchanged. Relative to the earlier expanded classifier, full-encoder continuation reduced false positives from 388 to 280 but detected 29 rather than 30 melanomas.

Table 5. Every assessment fold: head-only / full-encoder. Epoch zero retains the starting model.

Fold	AUROC	TP	FP	Selected epoch
1	0.885 / 0.884	6 / 6	36 / 36	1 / 0
2	0.945 / 0.964	7 / 7	208 / 99	1 / 1
3	0.793 / 0.865	4 / 4	26 / 32	1 / 1
4	0.928 / 0.930	6 / 6	64 / 63	1 / 0
5	0.955 / 0.954	6 / 6	59 / 50	1 / 0

The earlier classifier is a descriptive reference; the primary comparison uses the matched head-only control. Epoch-zero eligibility means trained weights were not necessarily retained. This one-epoch experiment cannot rule out gains from other training budgets.

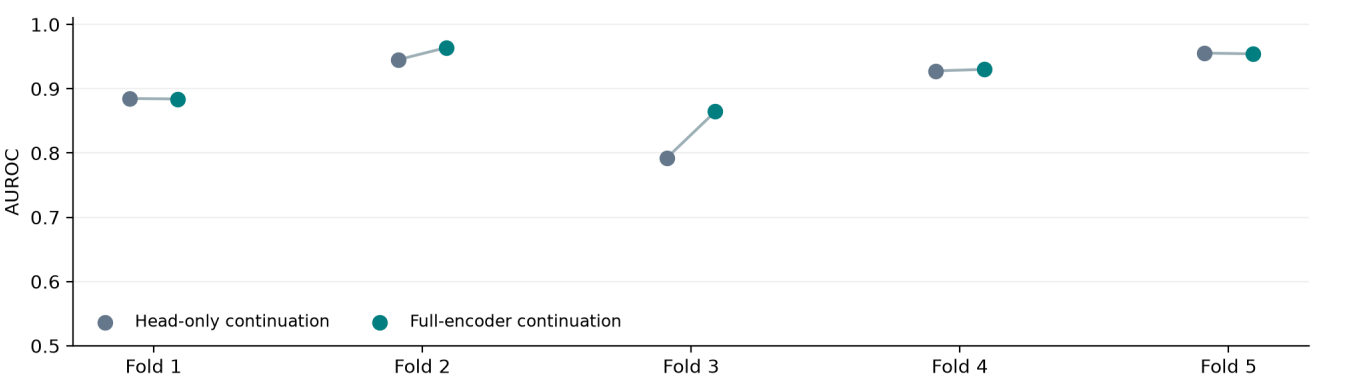


Figure 1. Paired fold-level AUROC. Each pair uses the same assessment representatives; folds share portions of their fitting data and are not independent replications.

6. Full-Fitzpatrick follow-up: audit and design

Full authorized Fitzpatrick17k access enabled a separate follow-up on 25 September 2026. The archive contained 16,577 images. Of these, 3,190 met the recorded eligibility rule: retained in the published cleaned-membership artifact [5] and assigned a benign or malignant lesion category. Non-neoplastic/inflammatory/genetic categories were excluded. Melanoma-category labels included melanoma in situ/lentigo maligna; other lesions included other cancers. Cleaned membership did not independently verify diagnoses.

Exact decoded-pixel matches, joint pHash/dHash distances of at most six, and documented visual related-view links formed conservative connected groups across the full archive. All 573 original melanoma-category thumbnails and 42 flagged images were reviewed with AI assistance, with selected enlarged comparisons. Non-melanoma images were not exhaustively visually reviewed. Composites, intraoral/lip-focused images, ocular-focused views and tissue specimens were excluded when identified. This was not clinical adjudication.

Groups touching previously acquired/reviewed Fitzpatrick candidates were excluded from the new cohort, as were unsuitable groups and cross-source similarity candidates. One cross-source candidate was visually judged distinct but conservatively omitted. Among eligible images, 69 were excluded and 20 additional repeated views collapsed, leaving 3,101 representatives. One largest eligible image represented each group. Neither the grouping nor the source metadata established independent patients.

Table 6. New follow-up partitions. Supplied Scale V-VI is the primary skin-type definition.

Partition	All / melanoma	Scale V-VI all / melanoma	Both V-VI all / melanoma
Training	1,861 / 226	122 / 11	65 / 1
Validation	624 / 77	42 / 4	24 / 2
Assessment	616 / 73	39 / 3	19 / 0

Seed 20260925 fixed approximately 60/20/20 group partitions within binary-label, source-atlas and Scale skin-type strata. Tiny strata used deterministic largest-remainder allocation. There was no result-based reroll. Conflicting or missing group annotations were retained as such; Scale, Centaur and agreement of both sources were reported separately.

Both arms used the original MILK-adapted EfficientNet-B0 encoder, frozen, with ImageNet preprocessing, a training-only StandardScaler and binary-class-balanced L2 logistic regression. Reference fitting used 4,091 existing MILK/PAD/early-Fitzpatrick images (295 melanoma labels); expanded fitting added 1,861 new images, for 5,952 (521 melanoma labels). Neither arm used the earlier fold-specific continuation models.

6.1. Follow-up selection and overall results

Both arms selected C from 0.001, 0.01 and 0.1 by equal-weight mean MILK, PAD and new Fitzpatrick validation AUROC, with smaller-C ties. Both selected C = 0.001. A single threshold per arm was the highest meeting at least 80% sensitivity in each validation source. Selected thresholds were 0.193511 (reference) and 0.368853 (expanded). Both models and thresholds were saved before extracting features or scoring new assessment images. The reference also used new Fitzpatrick validation, making this matched adaptation rather than zero-shot transfer.

In the newly reserved 616-image Fitzpatrick assessment, adding 1,861 training representatives increased AUROC from 0.747 to 0.886 (paired difference +0.140; 95% bootstrap interval +0.098 to +0.185), increased detections from 57/73 to 63/73, and reduced false positives from 265/543 to 129/543.

Table 7. Matched results on the same 616 new assessment representatives.

Arm	AUROC / AP	Sensitivity / specificity	TP / FN	FP / TN
Reference	0.747 / 0.339	0.781 / 0.512	57 / 16	265 / 278
Expanded	0.886 / 0.667	0.863 / 0.762	63 / 10	129 / 414

The expanded arm detected six additional melanoma-category examples and produced 136 fewer false positives (51.3% fewer than the reference). The prespecified paired AUROC interval excluded zero. It used 2,000/2,000 valid conservative-group bootstrap samples, seed 20260926, with one representative per group. The interval does not account for unidentified shared patients, source-label errors or the earlier adaptive research process. AUROC is a ranking measure, not accuracy.

Added training data: AUROC change +0.140 (95% interval +0.098, +0.185)

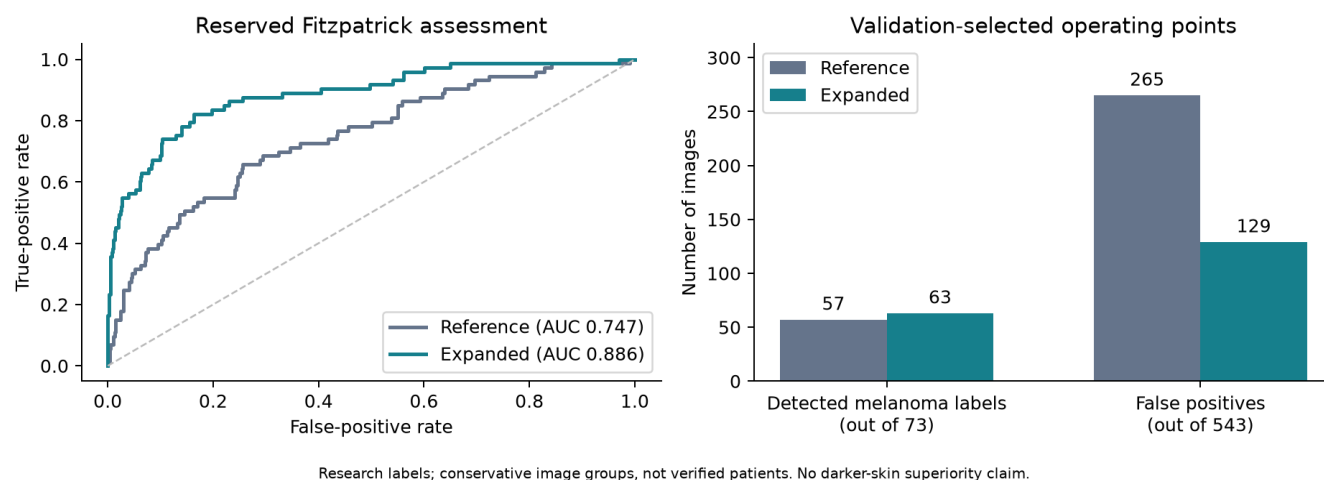


Figure 2. Newly reserved Fitzpatrick assessment. Both arms use the same frozen encoder and validation-selection rules. Only the expanded arm receives the additional training partition.

This follow-up meets the recorded operating-point criterion of fewer false positives without fewer detections and supports an overall benefit within this controlled source collection. It does not isolate whether particular skin-type groups caused the gain or establish transfer to a new clinical setting. This assessment is now used; subsequent tuning informed by it would require another independent confirmation set.

6.2. Follow-up skin-type results and uncertainty

The primary Scale V-VI assessment included only three melanoma-category and 36 other-lesion representatives. Reference and expanded models detected 2/3 and 3/3, with 23/36 and 13/36 false positives. A single changed positive prediction changes sensitivity by 33.3 percentage points. The corresponding 95% Wilson sensitivity intervals were 0.208-0.939 and 0.439-1.000. These are descriptive observations, not reliable evidence of darker-skin superiority.

Table 8. Skin-type strata: reference / expanded counts. The last row overlaps the supplied-annotation strata; rows must not be summed across definitions.

Definition	All / melanoma	TP ref / exp	FP ref / exp	AUROC ref / exp
Scale I-IV	557 / 68	54 / 59	237 / 114	0.754 / 0.888
Scale V-VI	39 / 3	2 / 3	23 / 13	0.620 / 0.926
Scale unknown	20 / 2	1 / 1	5 / 2	0.778 / 0.861
Centaur I-IV	535 / 69	55 / 61	225 / 114	0.763 / 0.901
Centaur V-VI	30 / 0	0 / 0	20 / 4	Undefined
Centaur unknown	51 / 4	2 / 2	20 / 11	0.527 / 0.553
Both V-VI	19 / 0	0 / 0	13 / 4	Undefined

No assessment melanoma was V-VI under the Centaur annotation or agreement of both sources; sensitivity and AUROC were undefined in those groups. Both-V-VI contained 19 other-lesion examples. The annotation disagreement and low positive counts persist despite full archive access. All defined groups, including empty conflicting-annotation groups, and Wilson intervals are retained in the supplementary results record.

The added training partition included eleven primary Scale V-VI melanoma labels, but only one was V-VI under both annotation sources. Fitzpatrick categories were not converted into race or ethnicity, and neither annotation source was treated as independently established skin tone. These data support a representation audit and an exploratory subgroup analysis, not a validated claim for people of color.

The complete follow-up record retains all six hyperparameter candidates, both model selections, thresholds, individual predictions and bootstrap settings. Independent reconciliation verified all 20 overall/subgroup tables, fitting membership, selection order and the paired bootstrap. A file-reader issue was corrected before feature extraction or fitting; the recorded split and analysis protocol were unchanged.

7. Discussion and limitations

In the newly reserved 616-image Fitzpatrick assessment, adding 1,861 training representatives increased AUROC from 0.747 to 0.886 (paired difference +0.140; 95% bootstrap interval +0.098 to +0.185), increased detections from 57/73 to 63/73, and reduced false positives from 265/543 to 129/543. The paired result is encouraging for within-source adaptation. It adds a distinct finding to the earlier source-dependent results, rather than overturning their failures or proving clinical effectiveness.

This study found that the interpretation of an improvement depends on what was changed, how the operating threshold was selected, and which source supplied the assessment images. The initial PAD comparison showed a large reduction in false alarms at an unchanged detection count. The broader five-split analysis reduced that apparent benefit to a small aggregate difference and exposed substantial variation between splits. Reporting all folds therefore materially changed the conclusion. An always-non-melanoma predictor would achieve 98.08% accuracy on these PAD representatives while missing all 36 melanomas, illustrating why accuracy alone is inadequate here.

Full-encoder adaptation produced fewer false positives without reducing the aggregate detection count relative to the matched head-only control. The earlier representation experiment isolated encoder adaptation under a matched one-epoch budget, using the same images, initialization, loss, ordering and selection rules. Its results should be read alongside the individual fold counts rather than reduced to a single headline accuracy. AUROC describes ranking across thresholds; the chosen operating-point counts determine how many labels were missed or flagged in this analysis.

Darker-skin evaluation remained the central data limitation. DDI supplied only seven melanoma images in V-VI; one changed prediction alters that subgroup's sensitivity by about 14 percentage points. MIDAS supplied no melanoma positives in IV-VI, and PAD supplied none in V-VI. The early Fitzpatrick review reduced 17 melanoma photographs to eight possible related-case groups across all partitions, with only five training representatives. Neither downloaded image volume nor broad dermatology diversity established an adequate melanoma-specific subgroup sample.

The reused DDI V-VI comparison in the single-split expansion experiment detected 5/7 melanomas for every arm, while false positives were 102/200, 108/200 and 81/200. That descriptive pattern was not validated by a new darker-skin cohort. The earlier PAD cross-validation cannot confirm it because the relevant positive cases are absent. No model was shown to be superior specifically for people of color, and tone annotations were not interpreted as racial identity.

Additional limitations include mixed clinical versus histopathologic ground truth for some negative labels, source-specific diagnosis definitions including melanoma in situ, uncertain patient identities outside PAD, supplied PAD identifier inconsistencies, remaining related-case overlap, differing image acquisition conditions, narrow early Fitzpatrick diagnosis coverage, incomplete full-release modality review, inconsistent tone annotations, and center-crop information loss. The study did not measure calibration, clinician performance, workflow benefit, prospective outcomes or cost-effectiveness.

All later hypotheses were informed by earlier results on these datasets. Patient-group separation within a run reduces known leakage but does not erase this adaptive reuse. Reported bootstrap intervals from earlier single-split experiments do not account for all research-selection decisions. Five overlapping fitting sets were not treated as independent studies, and no pooled AUROC was calculated from differently fitted fold models.

8. Conclusion, reproducibility and references

Audited data expansion and limited fine-tuning produced benefits and trade-offs that depended on the assessment design. In the newly reserved 616-image Fitzpatrick assessment, adding 1,861 training representatives increased AUROC from 0.747 to 0.886 (paired difference +0.140; 95% bootstrap interval +0.098 to +0.185), increased detections from 57/73 to 63/73, and reduced false positives from 265/543 to 129/543. This controlled follow-up supports the potential of training-data expansion to improve image-label classification. Earlier negative comparisons, unknown patient independence and severe subgroup scarcity limit its scope: clinical usefulness and reliable darker-skin superiority remain unestablished. Confirmation requires a new clinical cohort with verified diagnoses, reliable patient grouping and sufficient darker-skin melanoma representation.

Reproducibility. Source hashes, cohort membership, duplicate decisions, grouped folds, candidate scores, selected model weights, thresholds and individual predictions are retained locally. Reports were generated from saved results and independently reconciled against labels and confusion counts. Key protocols are Pilot_Protocol, Finetune_v1_Protocol, Expansion_v1_Protocol, Expansion_CV_v1_Protocol Expanded_Finetune_v1_Protocol and Fitzpatrick_Followup_v1_Protocol. The full-Fitzpatrick assessment was newly reserved; older assessments remained exploratory. Python 3.12, PyTorch 2.2.2, torchvision 0.17.2 and scikit-learn 1.9.1 were used. A local CPU performed training; no new dataset or clinical cohort was collected.

Data access and manuscript status. Images are available through their original providers under the applicable access and license terms; restricted images are not included in this manuscript. This is a research manuscript for author review, not a peer-reviewed publication or a diagnostic product. AI assistance was used for programming, analysis workflow and manuscript drafting. No additional clinical validation or institutional ethics approval is claimed by this secondary-data project.

[1] PyTorch. Transfer Learning for Computer Vision Tutorial. [Source](#).

[2] Tan M, Le Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. ICML, 2019. [Source](#).

[3] Daneshjou R, et al. Disparities in dermatology AI performance on a diverse, curated clinical image set. Science Advances. 2022;8:eabq6147. [Source](#).

[4] Groh M, et al. Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitzpatrick 17k Dataset. CVPR Workshops, 2021. [Source](#).

[5] Abhishek K, Jain A, Hamarneh G. Investigating the Quality of DermaMNIST and Fitzpatrick17k Dermatological Image Datasets. Scientific Data. 2025;12:196. [Source](#).

[6] MILK study team. MILK10k. ISIC Archive. Dataset DOI: 10.34970/648456. [Source](#).

[7] Stanford AIMI. MRA-MIDAS: Multimodal Image Dataset for AI-based Skin Cancer. [Source](#).

[8] Pacheco AGC, et al. PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. Data in Brief. 2020;32:106221. [Source](#).

[9] Scikit-learn. Cross-validation: evaluating estimator performance. [Source](#).