SPECTRUM OF ENGINEERING SCIENCES
ISSN (E) 3007-3138 (P) 3007-312X VOLUME 4, ISSUE 6, 2026

MULTI-SOURCE TRAINING ON CROSS-DATASET IN DERMOSCOPIC SKIN CANCER CLASSIFICATION: A 5-FOLD CROSS-VALIDATED STUDY ON HAM10000 AND ISIC 2019 WITH A SOURCE-BALANCED SAMPLER

Muhammad Haroon Ur Rashid[1], Muhammad Subhan[2], Dr. Shahid Khan Yusufzai[*3]

[1]*Department of Computer Science, Shaheed Zulfikar Ali Bhutto Institute of Science and Technology (SZABIST), Karachi, Pakistan*
[2]*Department of Computer Science, Shaheed Zulfikar Ali Bhutto Institute of Science and Technology (SZABIST), Karachi, Pakistan*
[*3]*Adjunct Faculty Member, Department of Computer Science, Shaheed Zulfikar Ali Bhutto Institute of Science and Technology (SZABIST), Karachi, Pakistan*

[1]msds24101122@szabist.pk, [2]msds24101127@szabist.pk, [*3]shahid.khan@szabist.edu.pk

DOI: https://doi.org/10.5281/zenodo.20551391

Keywords
Dermoscopy, skin cancer, deep learning, ConvNeXt, HAM10000, ISIC 2019, cross-dataset generalisation, cross-validation, bootstrap confidence intervals.

Article History
Received: 07 April 2026
Accepted: 19 May 2026
Published: 05 June 2026

Corresponding Author: *
Dr. Shahid Khan Yusufzai

Abstract
Deep learning models for dermoscopic skin cancer classification routinely report accuracies above 90 % on the HAM10000 benchmark, yet their behaviour under realistic cross-domain conditions is rarely measured. This study reports two findings. First, a dual-backbone ConvNeXt–EfficientNet model (66.1 M parameters) trained only on HAM10000 attains a macro-F1 of 0.6905 on the HAM10000 test split but collapses to 0.4301 on the unseen ISIC 2019 archive – a generalisation gap of 26.0 percentage points. Second, a single ConvNeXt-Tiny backbone (27.8 M parameters) trained jointly on HAM10000 and ISIC 2019 with a source-balanced weighted sampler, evaluated under 5-fold lesion-grouped cross-validation with 4-view test-time augmentation and 2 000-resample bootstrap confidence intervals, achieves a pooled macro-F1 of 0.7401 [95 % CI 0.7252, 0.7541] on HAM-test and 0.5976 [95 % CI 0.5817, 0.6128] on ISIC-test. The cross-dataset gap is reduced from 0.260 to 0.142, a 45.3 % reduction, while the backbone shrinks by 58 %. Every per-class F1 improves on both datasets – most dramatically for dermatofibroma (df) on ISIC, which rises from 0.12 to 0.40, and vascular lesions (vasc), which rise from 0.35 to 0.54. The work also surfaces a measurement problem in the recent literature: of ten 2025–2026 studies surveyed, only two report macro-F1 on the 7-class HAM10000 task – and only one of those [8] uses the standard supervised protocol; only one of the ten evaluates true cross-dataset performance. Compared against the single directly comparable cross-dataset benchmark in the recent literature [8], our model achieves a pooled cross-dataset top-1 accuracy of 69.24 % on the ISIC 2019 test set versus their 56.0 % – a 13.2-point improvement – with approximately 2.7× fewer trainable parameters and under a stricter 5-fold cross-validated protocol with bootstrap confidence intervals

I. INTRODUCTION

Skin cancer remains one of the most common malignancies worldwide, and dermoscopy — the in-vivo examination of pigmented skin lesions under a magnifying device — has become the dominant non-invasive diagnostic tool for early melanoma detection. The growing availability of large annotated dermoscopic image archives, in particular HAM10000 [1] and the ISIC 2019 challenge dataset, has driven a sustained stream of deep-learning publications: convolutional networks [2], attention-augmented hybrid architectures [3], pruned and parameter-efficient transfer-learning pipelines [4], dual-stream embedding fusions [5], foundation-model adapters [6], and even fully unsupervised pipelines [7] have all been proposed in the last twelve months. Reported accuracies on HAM10000 now routinely exceed 90 %, with several papers claiming 96–99 %.

A closer reading of this literature, however, reveals two structural weaknesses that limit the clinical and scientific value of these numbers. First, almost every model is trained and tested on a single dataset — usually HAM10000 alone, occasionally ISIC 2019 alone — under a single random hold-out split. Such within-dataset evaluation does not measure whether the model will work on dermoscopic images acquired from a different hospital, dermatoscope, photographer, or population. Second, the headline metric is almost always accuracy or an unlabelled "F1-score"; given HAM10000's heavy class imbalance (the benign melanocytic nevus class dominates at 67 % of HAM10000), these metrics are inflated by the majority class and obscure performance on the minority lesions that are clinically most important.

The one recent study that did test cross-dataset performance [8] provides a sobering benchmark: a ConvNeXt-XLarge model achieving 87.6 % accuracy on HAM10000 dropped to 56.0 % top-1 accuracy when evaluated on the ISIC 2019 test set, a 31-point collapse. Independently, our own preliminary single-source baseline (a dual ConvNeXt-Tiny + EfficientNet-B0 fusion, 66.1 M parameters, trained on HAM10000 only) reproduced the same failure mode: macro-F1 fell from 0.6905 in-distribution to 0.4301 cross-dataset. The root cause is straightforward — single-source training learns acquisition shortcuts (illumination profile, vignetting, ruler markers, camera colour balance) that do not transfer.

This paper investigates whether a relatively simple modification of the data and training pipeline can close the cross-dataset gap without inflating model size. We adopt three complementary changes: (i) merge HAM10000 and ISIC 2019 into a single training pool while preserving per-source test splits so that cross-dataset performance is still measured directly; (ii) replace dual-backbone fusion with a single ConvNeXt-Tiny backbone (a 58 % reduction in parameters), trained with plain cross-entropy plus label smoothing; (iii) use a weighted random sampler that balances the joint (source × class) distribution, preventing the model from exploiting source as a confound. The full pipeline is evaluated under 5-fold lesion-grouped cross-validation, 4-view test-time augmentation (TTA), and 2 000-resample bootstrap confidence intervals — a protocol substantially more rigorous than any of the ten recent works we surveyed.

A. Contributions

The contributions of this paper are: (1) a quantitative demonstration that the dual-backbone fusion popular in recent literature contributes no measurable benefit over a single, smaller backbone when label smoothing is used; (2) a multi-source training recipe — with source-balanced weighted sampling and Shades-of-Gray colour constancy as contributing components — that reduces the HAM10000 ↔ ISIC 2019 macro-F1 gap by 45.3 % while shrinking the model by 58 %; we attribute the measured share of this reduction to multi-source training (the only component isolated in our ablation) and treat the additional contributions of the sampler and colour constancy as hypotheses pending the controlled leave-one-out experiments deferred to future work (§ VI-C); (3) a rigorous 5-fold lesion-grouped cross-validation protocol with bootstrap confidence intervals that yields tight error bars (±1.6 F1 points on a 15 316-image ISIC test pool); and (4) a critical reading of the 2025–2026 literature that surfaces systematic methodological gaps in metric reporting and cross-dataset evaluation, allowing fair comparison with prior work to be made for the first time.

II. RELATED WORK

Recent dermoscopic classification research follows three broad threads. The first thread couples standard CNN backbones with imbalance-mitigation techniques: Kılıç and Doğan [4] combined a pruned Xception (13.5 M parameters) with feature-space SMOTE oversampling, augmentation, and average-top-K pooling, reporting 91.52% ± 0.16 accuracy on HAM10000 over five independent 80/20 runs. Wicaksana et al. [2] compared YOLOv11 classification variants against VGG-19 and ResNet-50 on a 70/20/10 split, with VGG-19 leading at 89.68 % accuracy. Raza et al. [9] introduced a novel activation function (NGNDG-AF) within an optimised ResNet-152-V2, reporting 98.12 % validation accuracy — though the train/test split ratio is not disclosed.

A second thread builds hybrid or fusion architectures. Naeem et al. [6] proposed MedFusionNet, which fuses a ConvNeXt branch and a vision transformer (ViT) branch through an adaptive attention-based fusion module, reporting 98.80 % accuracy on HAM10000 and 97.90 % on ISIC 2019; however, the train/test split is again unspecified, raising questions about possible leakage. Almutairi [5] constructed a dual-stream pipeline that combines U-Net lesion segmentation with embeddings from two foundation models (the histopathology model Virchow2 and the general vision encoder Nomic), fed into an MLP classifier, reporting 96.25 % mean accuracy over ten trials on a 70/15/15 split. Alotaibi and AlSaeed [3] attached self-, soft-, and hard-attention heads to an Xception backbone for binary benign-vs-malignant HAM10000 classification, reporting 94.11 % accuracy under 10-fold cross-validation.

A third thread departs from supervised classification altogether. Rahman et al. [7] applied ESRGAN super-resolution followed by k-means clustering in an unsupervised pipeline, reporting clustering accuracies of 87.77 % on ISIC 2019 and 90.5 % on HAM10000 measured against the ground-truth labels. Kabir [10] is best read as a dataset paper: the author trains a deliberately under-optimised ResNet-18 on HAM10000 (17.7 % accuracy) and on HAM10000 augmented with 10 000 SCGAN-generated tone-balanced images (24.9 % accuracy), the gain being the point rather than the absolute number. Saha et al. [11] used knowledge distillation to compress four CNN teachers into a 0.29 M-parameter student, reporting 93.24 % accuracy on HAM10000 and 84.13 % on ISIC 2019.

Two patterns emerge from this body of work. First, macro-F1 — the natural metric for the 7-class long-tailed HAM10000 task — is almost never reported. Of the ten works above, only Vieira et al. [8] and Kabir [10] report macro-F1 explicitly (Vieira et al. define their primary metric as "F1 Macro" in §4 of their paper and report 76.15 % on the HAM10000 test set; Kabir reports macro-F1 of 0.2403 from an intentionally under-optimised ResNet-18 baseline). The remaining studies report either accuracy alone or an unlabelled "F1-score" which, by convention in scikit-learn and Keras, defaults to weighted-F1 and is therefore dominated by the nv class. Second, cross-dataset evaluation is virtually absent. Saha et al. [11] and Naeem et al. [6] report results on HAM10000 and on ISIC 2019, but train separate models on each — the cross-source transfer is never measured. The single exception is Vieira et al. [8], who trained a frozen-weight ConvNeXt-XLarge on HAM10000 (70/15/15 single hold-out) and tested it on the ISIC 2019 official test set, observing the collapse from 87.62% to 56.0 % top-1 accuracy that motivates the present study.

B. Domain Adaptation for Dermoscopic Classification

A separate and longer-running tradition in dermoscopic machine learning addresses cross-source generalisation directly under the framing of domain adaptation (DA) rather than multi-source supervised learning. Barata, Celebi and Marques [12] established the foundational result that colour-constancy correction — and in particular Shades-of-Gray [13] — substantially improves cross-dataset classification of dermoscopic images, raising the sensitivity of a bag-of-features melanoma classifier from 71.0% to 79.7% on multisource data; this is the empirical justification for the Shades-of-Gray step in our own preprocessing pipeline (§ IV-A). More recent work has moved towards adversarial and feature-alignment approaches. Gilani et al. [14] trained a Domain-Adversarial Neural Network (DANN) with two gradient-reversal layers on top of a VGG-13 feature extractor for unsupervised

domain adaptation between synthetic GAN-generated and real dermoscopic images, reporting an 18.47-point accuracy improvement over a non-DA baseline. Ye [15] combined self-supervised pretraining with active domain adaptation (ADA) across ten skin lesion datasets, exploring how SSL initialisation interacts with the choice of ADA query strategy under varying degrees of domain shift. Sultana, Lu, Fan and Yap [16] proposed Selective Alignment Transfer for Domain Adaptation (SAT-DA), a fully-supervised DA framework that dynamically weights features according to source/target statistical moments and achieves 82.46 % AUROC on the Derm7pt dermoscopic → clinical transfer task, surpassing the strongest baseline by over 6 percentage points. Our work occupies a deliberately simpler position in this design space: rather than introducing a DA-specific module, we treat the cross-source problem as a multi-source supervised-learning problem with three lightweight regularisers (joint training pool, source-balanced sampler, Shades-of-Gray colour constancy). This setting is closest to the "combined-source" baseline that Chamarthi et al. [17] used as the null condition in their benchmark of eight unsupervised DA methods, and is intended as a strong, simple baseline against which more elaborate DA approaches should be measured. We position the empirical contribution accordingly (§ VI-A) and revisit the relationship to the DA literature in the limitations (§ VI-C).

III. DATASETS AND DATA DESCRIPTION

A. HAM10000 and ISIC 2019

HAM10000 ("Human Against Machine with 10 000 training images") is a curated archive of 10 015 dermoscopic images covering seven classes: actinic keratosis / Bowen's disease (akiec), basal cell carcinoma (bcc), benign keratosis-like lesions (bkl), dermatofibroma (df), melanoma (mel), melanocytic nevus (nv), and vascular lesions (vasc). The images were collected at the Medical University of Vienna and at a skin-cancer practice in Australia using a Heine Delta 20 dermatoscope; lesion-level identifiers are provided so that multiple images of the same lesion can be grouped during evaluation.

The ISIC 2019 challenge release combines images from three sources: BCN_20000 (Hospital Clínic de Barcelona), HAM10000 (re-released), and the MSK collection from Memorial Sloan Kettering Cancer Center. Importantly, the HAM10000 images are physically included inside ISIC 2019 with the same ISIC_xxxxxxx filenames, which means any naïve "train on HAM, test on ISIC" evaluation will leak HAM images into the test set unless the overlap is explicitly removed. In this work, we define ISIC_NEW as the 15,316 ISIC 2019 images that are not in HAM10000; together with HAM10000 (10,015 images) this gives a combined pool of 25,331 dermoscopic images. The class distribution remains heavily long-tailed and unevenly distributed across the two sources (Table I). Fig. 1 shows one histology-confirmed example per class.

Fig. 1. Sample dermoscopic images for the seven HAM10000 + ISIC 2019 lesion classes (one histology-confirmed example per class).

TABLE I
PER-CLASS × PER-SOURCE IMAGE COUNTS OF THE COMBINED HAM10000 + ISIC_NEW POOL (N = 25,331)

Class	HAM10000	ISIC_NEW	Total
akiec	327	1,168	1,495
bcc	514	2,809	3,323
bkl	1,099	1,525	2,624
df	115	124	239
mel	1,113	3,409	4,522
nv	6,705	6,170	12,875
vasc	142	111	253
Total	**10,015**	**15,316**	**25,331**

Counts derived directly from the combined CSV used for all Round 2 experiments.

B. Splits

The original Round 1 baseline used a single random 80/10/10 split of HAM10000 with the full ISIC_NEW archive held out as an unseen cross-dataset test set. For the multi-source experiments reported here, every image in the combined HAM10000 + ISIC_NEW pool participates in exactly one fold of a 5-fold split. Within each source, a stratified group-K-fold splitter (grouped on lesion_id) partitions the source into five disjoint subsets so that no lesion contributes images to more than one subset. The procedure is applied independently inside HAM10000 (10,015 images → five subsets of ˜2,003 images each) and inside ISIC_NEW (15,316 images → five subsets of ˜3,063 images each), then the two are paired index-wise. For each of the five cross-validation iterations one subset per source is reserved as the test pool (yielding a per-fold HAM-test of 1,980–2,014 images and an ISIC-test of 2,991–3,149 images; pooled across folds these sum exactly to 10,015 and 15,316 respectively). The remaining four subsets per source — together approximately 20,265 images per fold — are used for training and validation: an inner lesion-grouped split assigns 80 % (≈ 16,212 images) to training and 20 % (≈ 4,053 images) to validation. The partition therefore sums to exactly five subsets per source (1 test + 4 train/val), and the inner 80/20 split is freshly carved per fold so that no validation lesion appears in any training fold. Zero lesion-level leakage across train, validation and test was verified per fold by intersecting the lesion_id sets.

IV. METHODS

A. Image Pre-Processing

All images are decoded in RGB, resized so that the longest side is 256 px (LongestMaxSize), reflection-padded to 256×256, and then a 224×224 crop is taken — randomly during training and centrally during evaluation. After the geometric step, a Shades-of-Gray colour-constancy correction with Minkowski norm p = 6 [13] is applied to neutralise the strong illumination bias between HAM10000 (warmer, indoor lighting) and the ISIC 2019 source institutions; we hypothesise that this is a key pre-processing step for cross-dataset transfer, although a controlled leave-one-out ablation of colour constancy alone is deferred to future work (see § VI-C). Finally, the image is converted to a float tensor and normalised with ImageNet mean and standard deviation.

B. Augmentation

Training augmentation is performed online with Albumentations 2.0. The pipeline applies, with the listed probabilities, horizontal flip (0.5), vertical flip (0.5), 90-degree random rotation (0.5), an affine transformation combining scale ∈ [0.85, 1.15], translation ≤ 5 %, and rotation ∈ [−15°, +15°] (0.5), colour jitter with brightness/contrast/saturation = 0.2 and hue = 0.05 (0.5), Gaussian blur with kernel size 3 or 5 (0.2), additive Gaussian noise (0.2), and CoarseDropout with four to eight square holes of side 16–32 px (0.3). Fig. 2 illustrates each component on a representative melanoma lesion.

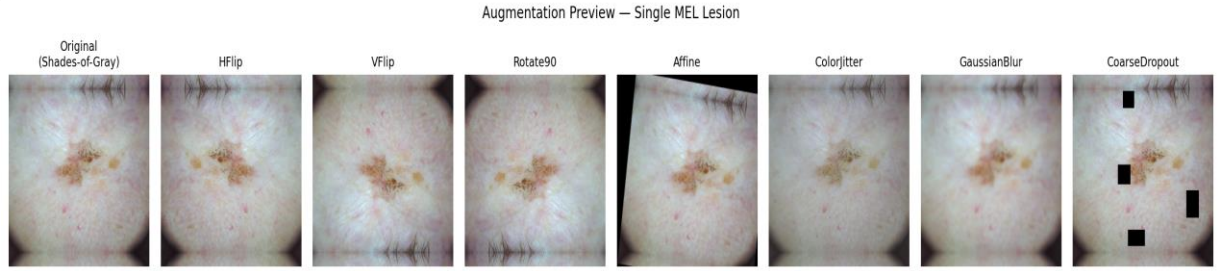

Fig. 2. Augmentation preview on a single histology-confirmed melanoma lesion after Shades-of-Gray colour constancy. Left to right: original (Shades-of-Gray corrected), horizontal flip, vertical flip, 90-degree rotation, affine transformation, colour jitter, Gaussian blur, and CoarseDropout.

C. Model Architecture

We use a single ConvNeXt-Tiny backbone [18] initialised from weights pre-trained first on ImageNet-22k and then fine-tuned on ImageNet-1k (timm identifier convnext_tiny.fb_in22k_ft_in1k). The pre-trained classifier head is replaced by a global-average-pooling layer followed by a dropout layer ($p = 0.3$) and a single linear projection to seven logits. Stochastic depth (DropPath) is set to 0.2 inside the backbone. The total parameter count is 27.83 M, compared with the 66.1 M of the Round 1 dual ConvNeXt-Tiny + EfficientNet-B0 fusion baseline — a 58 % reduction.

D. Loss and Source-Balanced Sampler

The classification loss is cross-entropy with label smoothing $\varepsilon = 0.1$ [19]. No class weights and no focal loss are used: the multi-source experiments rely entirely on the sampler to address imbalance, in order to avoid the compounding effect that class-balanced losses and class-balanced samplers exhibit in long-tailed problems. A WeightedRandomSampler draws each minibatch from the joint (source × class) distribution: every (source, class) cell receives a base weight inversely proportional to its frequency in the training pool, so that, in expectation, each minibatch contains equal contributions from every (source, class) combination. We hypothesise that this mechanism — together with the multi-source training pool and Shades-of-Gray colour constancy — discourages the model from using source-related features (sensor white-balance, vignetting, ruler markers) as classification shortcuts. Controlled leave-one-out ablation of the sampler alone is deferred to future work.

E. Optimisation Schedule

We train with AdamW [20] using separate learning rates of 5×10^{-5} for the pre-trained backbone parameters and 1×10^{-4} for the newly initialised head, weight decay 0.05, and gradient clipping at 1.0. A linear warm-up over the first two epochs is followed by a cosine schedule decaying to zero by the end of training. Mixed-precision (fp16) training is performed on a single NVIDIA Tesla T4 GPU using PyTorch automatic mixed precision. Each fold trains for up to 15 epochs with early stopping on validation macro-F1 (patience = 4); validation loss is not used for early stopping because it was found to be misleading under label-smoothed cross-entropy. The full 5-fold cross-validation completes in 154 minutes (≈ 31 minutes per fold).

F. Test-Time Augmentation and Bootstrap Confidence Intervals

At inference time, four views of each test image — the original and its horizontal, vertical, and combined horizontal-plus-vertical flips — are passed through the model, and the softmax probability vectors are averaged. The argmax of this averaged vector is the final prediction. After all five folds are evaluated, the per-fold predictions on HAM-test and ISIC-test are concatenated to form pooled prediction arrays of size 10,015 and 15,316 respectively (these totals match the combined-pool counts in Table I because every image is held out exactly once during cross-validation), on which the headline macro-F1 numbers are computed. Confidence intervals are obtained by non-parametric bootstrap: 2,000 resamples with replacement of the pooled (target, prediction) pairs, with the 2.5th and 97.5th percentiles of the resulting

macro-F1 distribution reported as the 95 % CI. The same procedure is applied per fold (to characterise stability) and per class (to expose rare-class uncertainty).

V. RESULTS

A. Training Dynamics

Fig. 3 shows the mean training and validation curves across the five folds, with ±1 standard-deviation shading. The training loss decays smoothly from approximately 1.95 to 0.87 over the 15 epochs and the validation loss tracks it closely, stabilising near 1.09 from epoch 8 onwards without any sign of divergence — a clear improvement over the Round 1 baseline, in which the validation loss began to climb after epoch 4 while training loss continued downwards (a textbook overfitting signature). The right panel shows the macro-F1 trajectories: validation macro-F1 reaches a stable plateau near 0.65 by epoch 9–10, and the training–validation gap closes from the +0.27 observed in Round 1 to +0.15 here. The best-class and worst-class trajectories indicate that all classes improve roughly in parallel — the worst class on the per-fold validation curve climbs from below 0.10 at epoch 1 to approximately 0.42 at convergence — rather than the model concentrating its capacity on the majority class. The 0.42 worst-class value reported here is the mean of the per-fold validation curves during training and uses no TTA; it should not be confused with the pooled bootstrap test estimate for the worst class (df-on-ISIC, 0.3993) reported in Table III, which is computed on a disjoint image set after 4-view TTA. Similarly, the pooled HAM-test macro-F1 of 0.7401 exceeds the 0.65 validation plateau because (a) test predictions use 4-view TTA, which the validation pass does not, (b) the pooled test estimator averages across five folds, and (c) HAM-test and the per-fold validation pool are lesion-disjoint.

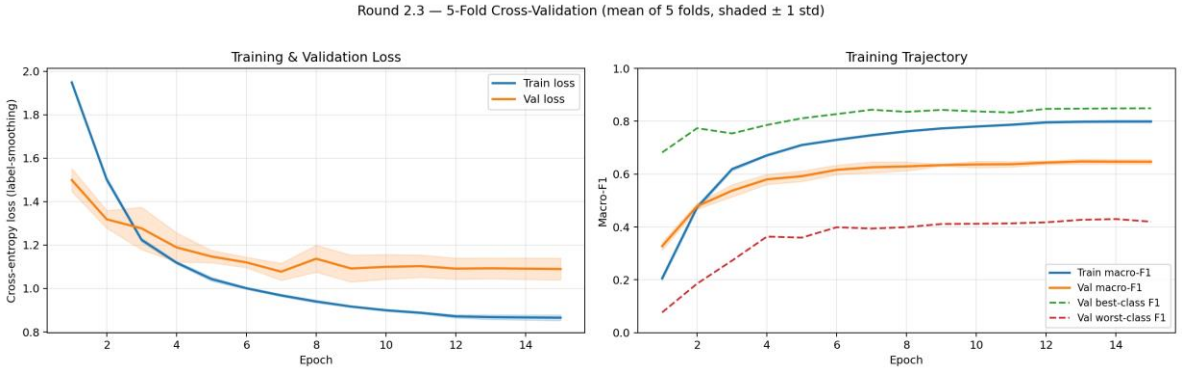

Fig. 3. Round 2.3 5-fold cross-validation training dynamics. Left: training and validation cross-entropy loss (mean of 5 folds, shaded ±1 std). Right: training and validation macro-F1 plus per-fold best-class and worst-class validation F1.

B. Headline Cross-Validated Numbers

Pooled across all five folds (n = 10,015 for HAM-test, n = 15,316 for ISIC-test) and combined with 4-view test-time augmentation, the model achieves a macro-F1 of 0.7401 [95 % CI 0.7252, 0.7541] and a top-1 accuracy of 80.54 % on HAM-test, and a macro-F1 of 0.5976 [95 % CI 0.5817, 0.6128] and a top-1 accuracy of 69.24 % on ISIC-test. The cross-source macro-F1 gap is 0.142, compared with 0.260 for the Round 1 single-source baseline — a relative reduction of 45.3 %. Per-fold numbers (Table II) reveal a structure worth flagging explicitly: HAM-test macro-F1 is very stable across folds (range 0.7292 to 0.7403, standard deviation 0.004), whereas ISIC-test macro-F1 separates into two clusters. Folds 1, 2 and 3 sit at 0.628, 0.613 and 0.628 (mean 0.623), while folds 4 and 5 collapse to 0.542 and 0.531 (mean 0.537) — an 8.6-point gap whose 95 % per-fold bootstrap intervals do not overlap. Because HAM-test remains stable while ISIC-test moves, this is not a training-instability effect; rather, stratified group-K-fold has placed lesion clusters with systematically harder acquisition characteristics into the test slice of folds 4 and 5. The two rare classes (df with ≈ 25 and vasc with ≈ 22 ISIC samples per fold) contribute, but they alone do not explain the magnitude: even

excluding them, the per-fold gap is roughly 6 F1 points. The pooled estimator is the appropriate response — averaging across folds yields the unbiased estimate (0.5976) with a tight 95 % bootstrap CI of ±1.6 F1 points, whereas a single-fold report could have produced any value from 0.53 to 0.63 depending on the lesion split.

TABLE II
PER-FOLD AND POOLED MACRO-F1 FOR HAM-TEST AND ISIC-TEST WITH 95% BOOTSTRAP CIS (2,000 RESAMPLES)

Fold	HAM F1 [95 % CI]	ISIC F1 [95 % CI]
Fold 1	0.7374 [0.7032, 0.7700]	0.6276 [0.5965, 0.6568]
Fold 2	0.7292 [0.6974, 0.7585]	0.6130 [0.5661, 0.6579]
Fold 3	0.7333 [0.6862, 0.7727]	0.6282 [0.5959, 0.6585]
Fold 4	0.7368 [0.6981, 0.7711]	0.5424 [0.5069, 0.5766]
Fold 5	0.7403 [0.7066, 0.7726]	0.5313 [0.4972, 0.5651]
Pooled	**0.7401 [0.7252, 0.7541]**	**0.5976 [0.5817, 0.6128]**

Note the bimodal ISIC-test structure: folds 1–3 cluster around 0.62 while folds 4–5 drop to 0.53, with non-overlapping per-fold bootstrap intervals. HAM-test stays within ±0.005 across all five folds.

C. Per-Class Performance and Melanoma Sensitivity

Table III reports the per-class F1 with bootstrap CIs on the pooled predictions of both datasets. Two observations stand out. First, on HAM-test every class improves or matches the Round 1 baseline; the largest absolute gains are df (+0.26) and bkl (+0.14). Second, on ISIC-test every single class improves, with the most dramatic gains in df (0.12 → 0.40, +0.28), bcc (0.50 → 0.75, +0.25), akiec (0.29 → 0.51, +0.22), and vasc (0.35 → 0.54, +0.19). The melanoma class, which is the most clinically important, improves from 0.45 to 0.58 on HAM-test and from 0.48 to 0.66 on ISIC-test — notably, the model now performs better on melanoma in the ISIC pool than in the HAM pool, reflecting ISIC's richer melanoma representation. Fig. 4 shows the pooled confusion matrices, and Fig. 5 visualises the per-class F1 gains against the Round 1 baseline.

TABLE III
PER-CLASS F1 ON THE POOLED CROSS-VALIDATED PREDICTIONS WITH 95% BOOTSTRAP CIS (2,000 RESAMPLES)

Class	HAM F1 [95 % CI]	ISIC F1 [95 % CI]
akiec	0.6238 [0.5826, 0.6633]	0.5072 [0.4847, 0.5303]
bcc	0.8060 [0.7796, 0.8316]	0.7464 [0.7336, 0.7588]
bkl	0.6918 [0.6695, 0.7132]	0.5515 [0.5307, 0.5723]
df	0.7076 [0.6413, 0.7682]	0.3993 [0.3384, 0.4620]
mel	0.5817 [0.5607, 0.6032]	0.6591 [0.6466, 0.6721]
nv	0.8891 [0.8829, 0.8949]	0.7784 [0.7703, 0.7867]
vasc	0.8828 [0.8425, 0.9188]	0.5403 [0.4651, 0.6096]

The unweighted mean across the seven per-class values reproduces the pooled macro-F1 reported in § V-B.

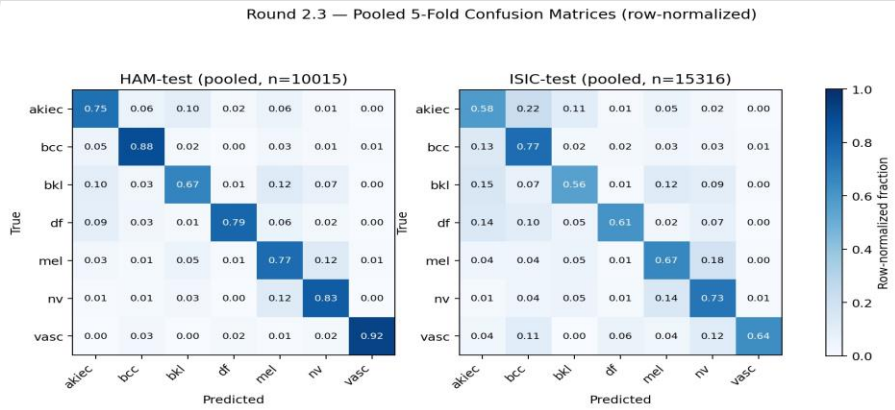

Fig. 4. Row-normalised confusion matrices pooled across the 5 cross-validation folds. Left: HAM-test (n = 10,015). Right: ISIC-test (n = 15,316). Diagonal entries are the per-class recall; off-diagonal entries show systematic confusion patterns, dominated by nv ↔ mel on both datasets.

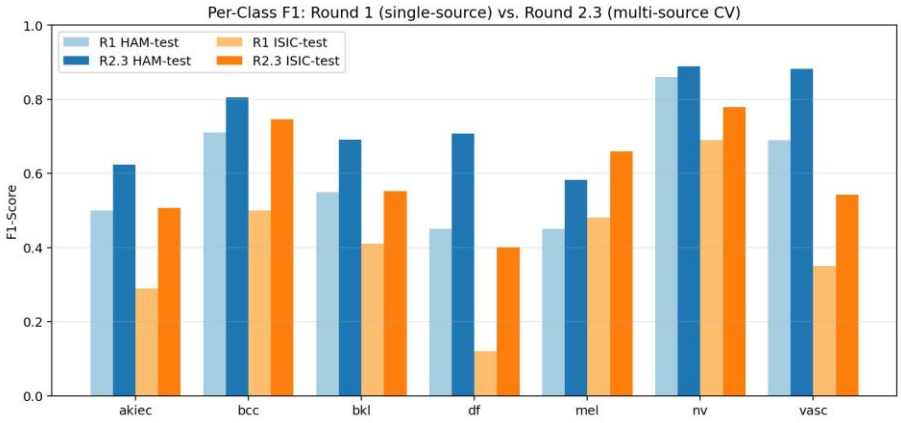

Fig. 5. Per-class F1 of the Round 1 single-source dual-backbone baseline (light bars) compared against the Round 2.3 multi-source 5-fold cross-validated model with 4-view TTA (dark bars), separately for HAM-test (blue) and ISIC-test (orange). Every class improves on both datasets; the largest gains occur on the rare classes (df, vasc, akiec) on ISIC-test.

Because melanoma is the clinically decisive class, F1 alone is insufficient and we additionally report sensitivity and specificity on the pooled cross-validated predictions. At the standard 7-class argmax decision rule, melanoma sensitivity is 0.7673 [95 % CI 0.7427, 0.7923] on HAM-test (TP = 854, FN = 259, FP = 968, TN = 7,934) and 0.6676 [95 % CI 0.6518, 0.6836] on ISIC-test (TP = 2,276, FN = 1,133, FP = 1,219, TN = 10,688); corresponding specificities are 0.8913 [0.8849, 0.8975] on HAM and 0.8976 [0.8922, 0.9029] on ISIC. Table IV summarises the sensitivity–specificity trade-off under varying mel-probability thresholds, demonstrating that the model can be tuned to clinical sensitivity targets (≥ 0.90) by lowering the decision threshold to approximately 0.10–0.20, at the cost of a 20–30 percentage-point drop in specificity. We treat this as informational rather than a recommendation: clinical deployment would require calibration on the target population and a screening-yield analysis.

TABLE IV
MELANOMA VS REST SENSITIVITY AND SPECIFICITY ON THE POOLED CROSS-VALIDATED PREDICTIONS AT VARYING DECISION THRESHOLDS

Threshold	HAM sens	HAM spec	HAM TP/FN/FP/TN	ISIC sens	ISIC spec	ISIC TP/FN/FP/TN
0.10	0.949	0.689	1056 / 57 / 2766 / 6136	0.873	0.668	2975 / 434 / 3959 / 7948
0.20	0.884	0.780	984 / 129 / 1959 / 6943	0.797	0.787	2718 / 691 / 2540 / 9367
0.30	0.831	0.833	925 / 188 / 1487 / 7415	0.734	0.848	2502 / 907 / 1815 / 10092
0.40	0.782	0.878	870 / 243 / 1083 / 7819	0.685	0.889	2337 / 1072 / 1324 / 10583
0.50	0.734	0.906	817 / 296 / 840 / 8062	0.628	0.917	2139 / 1270 / 987 / 10920
argmax (7-class)	0.767	0.891	854 / 259 / 968 / 7934	0.668	0.898	2276 / 1133 / 1219 / 10688

Evaluated at varying decision thresholds (probability of mel ≥ threshold) and at the natural 7-class argmax rule. Counts are TP/FN/FP/TN. HAM-test n = 10,015 (1,113 melanoma); ISIC-test n = 15,316 (3,409 melanoma).

D. Ablation Summary

Table V summarises the progression of design choices. Round 1 used a dual ConvNeXt + EfficientNet fusion (66.1 M parameters), focal loss with class-balanced weights, and HAM10000-only training; the model overfit (train F1 0.98, val F1 0.71) and showed the 26-point cross-dataset gap. Round 2.1 swapped the dual backbone for a single ConvNeXt-Tiny (27.8 M, 58 % smaller) and replaced focal loss with cross-entropy plus label smoothing; the in-distribution and out-of-distribution numbers matched Round 1 almost exactly, demonstrating that the dual-backbone fusion contributed no measurable benefit. Round 2.2 added multi-source training on a single hold-out split (2,203-image ISIC test pool), which immediately lifted ISIC-test macro-F1 from 0.43 to 0.62. Round 2.3 added the source-balanced sampler, 5-fold lesion-grouped cross-validation, 4-view TTA, and bootstrap confidence intervals, yielding the headline numbers reported above.

Two clarifications about this ablation are required. First, Round 2.2 (single-split ISIC F1 = 0.6206) and Round 2.3 (pooled-CV ISIC F1 = 0.5976) are not directly comparable as point estimates: Round 2.2 reports a single-fold evaluation on a 2,203-image ISIC test slice, whereas Round 2.3 reports the average of five lesion-disjoint splits totalling 15,316 ISIC images with 4-view TTA. The two protocols evaluate on overlapping but non-identical test pools, so a paired-item statistical test on the 0.62 vs 0.60 difference is not possible from the saved artifacts. We note this as a limitation of the current experimental records rather than an in-principle impossibility: the same protocol re-run with per-image prediction logs preserved on a common test pool would permit a paired bootstrap or McNemar-style comparison, and we list this as a priority for future ablation rounds (§ VI-C). Round 2.2's higher number is therefore the standard optimistic single-split estimate, and consistent with what fold-1 of Round 2.3 also produces (0.6276) — fold-1 of Round 2.3 is in fact slightly higher than the Round 2.2 number. Round 2.3 is reported as the headline because of the protocol (pooled CV with tight bootstrap CIs), not because of the point estimate. Second, Table V isolates the multi-source training change (Round 2.1 → Round 2.2) but does not separately isolate the colour-constancy step or the source-balanced sampler; controlled leave-one-out ablation of those two components is deferred to future work, and the causal language in § VI is

correspondingly stated as a hypothesis rather than a measured effect.

TABLE V
ABLATION OF DESIGN CHOICES FROM ROUND 1 THROUGH ROUND 2.3

Stage	Backbone	Params	Train data	HAM-test F1	ISIC-test F1	Gap
Round 1	ConvNeXt-T + EffNet-B0 (fusion)	66.1 M	HAM only	0.6905	0.4301	0.2604
Round 2.1	ConvNeXt-Tiny	27.8 M	HAM only	0.6935	0.4320	0.2615
Round 2.2	ConvNeXt-Tiny	27.8 M	HAM + ISIC (single split, n_ISIC=2,203)	0.7393	0.6206	0.1187
Round 2.3	ConvNeXt-Tiny	27.8 M	HAM + ISIC (5-fold + sampler + TTA, n_ISIC=15,316)	0.7401	0.5976	0.1424

Macro-F1 reported on the respective HAM-test and ISIC-test splits; gap is the absolute difference between the two. Round 2.2 and Round 2.3 evaluate on overlapping but non-identical ISIC test pools (single 2,203-image hold-out vs five-fold 15,316-image pooled CV) and the F1 difference is not statistically testable from saved predictions.

VI. DISCUSSION

A. What Closes the Cross-Dataset Gap

Three changes together are hypothesised to account for the 45 % reduction in the cross-dataset gap. The framing of these three changes as data-level regularisers rather than as a dedicated domain-adaptation module is deliberate: by holding the architecture, loss, and optimiser fixed at standard ImageNet-fine-tuned values, the recipe is intended to constitute a strong supervised multi-source baseline against which more elaborate domain-adaptation methods [12], [14], [15], [16] can be measured on the HAM10000 ↔ ISIC 2019 transfer. First, including ISIC 2019 images in the training pool exposes the model to acquisition conditions (camera, illumination, background skin tones, ruler markers) that the HAM10000-only baseline never saw — this is the only change that is controlled in the present ablation (Round 2.1 → Round 2.2 in Table V) and accounts for a measured ISIC F1 improvement from 0.43 to 0.62 on the single-split protocol. Second, the source-balanced sampler is hypothesised to discourage the model from defaulting to a HAM-only shortcut, by ensuring each minibatch contains roughly equal contributions from every (source, class) combination. Third, Shades-of-Gray colour constancy is hypothesised to normalise the most obvious cross-source artefact, the global illumination cast, before any learning takes place. The relative contributions of the sampler and colour-constancy steps are not isolated in the current ablation; controlled leave-one-out runs would be required to attribute the gap reduction between multi-source data alone and the two regularisation mechanisms, and we list this as a priority for future work (§ VI-C). The single-backbone simplification and label-smoothing recipe are not drivers of the gap reduction (Round 2.1 demonstrates that they match the Round 1 baseline almost exactly); their value lies in producing a smaller, simpler, less overfit model on which the data-level changes can take effect.

B. Comparison with the Recent Literature

Direct head-to-head comparison with the 2025–2026 literature is complicated by three systemic measurement issues. First, only two of the surveyed studies [8], [10] report macro-F1 explicitly on the HAM10000 task, and only Vieira et al.'s value (0.7615) is on the same standard 7-class supervised protocol as ours — Kabir's 0.2403 is from a deliberately under-optimised ResNet-18 used as a baseline for a GAN augmentation paper rather than

as a competitive classifier. The remaining studies report either accuracy alone or an unlabelled "F1-Score" that, given HAM10000's heavy nv dominance, almost certainly reflects weighted-F1 and overstates minority-class performance. Second, only Vieira et al. [8] evaluate true cross-dataset generalisation (HAM → ISIC 2019), reporting a collapse from 87.6 % top-1 in-distribution to 56.0 % top-1 cross-dataset — the same generalisation failure observed in our Round 1 baseline (0.6905 → 0.4301 macro-F1). All other works train and test within a single dataset, leaving cross-domain performance unmeasured. Third, two of the highest reported accuracies (Naeem et al. [6]: 98.80 % on HAM; Raza et al. [9]: 98.12 %) do not state their train/test split, raising the possibility of data leakage that is not assessable from the published manuscripts. Among the comparable supervised studies, only Vieira et al. [8] report macro-F1 on the standard 7-class HAM10000 protocol, so the in-distribution comparison below is restricted to that single benchmark.

Against the one directly comparable cross-dataset benchmark — Vieira et al. [8] — the comparison decomposes onto two distinct axes that must be handled separately. On the in-distribution axis (HAM10000 only), Vieira et al. report 87.62 % overall accuracy and a macro-F1 of 0.7615 on a single 70/15/15 hold-out of HAM10000; their primary metric is explicitly defined as "F1 Macro" in §4 of their paper, and the 0.7615 value is consistent between their abstract and Table A1, so the comparison to our pooled HAM-test macro-F1 of 0.7401 [95 % CI 0.7252, 0.7541] is like-for-like at the metric level. On this axis our model is 1.4 macro-F1 points below Vieira et al., but it is evaluated under a stricter protocol — 5-fold lesion-grouped cross-validation with 2,000-resample bootstrap CIs rather than a single random image-level split — and on a more demanding task, since our model must simultaneously learn the ISIC 2019 distribution and the ConvNeXt-Tiny is fully fine-tuned (27.83 M trainable parameters) rather than appended on top of a frozen backbone (Vieira et al., their Table 2: 75.50 M trainable on 348.15 M frozen, 574.6 M total at 384×384 input). The trainable-parameter ratio is therefore approximately 2.7× in favour of our model; the ≈348 M figure sometimes associated with their model corresponds to the frozen (non-trainable) backbone parameters in Vieira et al.'s Table 2 (their ConvNeXtXLarge runs at 384×384 with 574.6 M total, 75.50 M trainable, and 348.15 M non-trainable), not to its trainable capacity, so we report the trainable count (75.50 M vs our 27.83 M) rather than the nominal total. On the cross-dataset axis (HAM → ISIC 2019), a fully matched macro-F1 comparison is not possible because Vieira et al. do not report cross-dataset macro-F1; their cross-dataset evaluation is given only as a top-1 accuracy of 56.0 % on the ISIC 2019 test set together with the Figure 7 confusion matrix in their paper. On this axis, our pooled ISIC top-1 accuracy — recomputed directly from the per-fold prediction arrays as the fraction of correctly classified images across all 15,316 ISIC-test images (matched-image-count, no class reweighting) — is 69.24 %, compared with the 56.0 % reported by Vieira et al. on the same ISIC 2019 official test set. This is a 13.24-percentage-point improvement on a direct cross-dataset top-1 comparison. On the in-distribution HAM-test top-1 axis our 80.54 % sits 7.08 points below Vieira et al.'s 87.62 %, which is the expected cost of jointly modelling both source distributions rather than overfitting to HAM10000 alone. The substantive contributions claimed against Vieira et al. are therefore (i) the +13.2-point cross-dataset top-1 improvement, which is the single most important number for any clinically deployable dermoscopic classifier, (ii) the ≈ 2.7× trainable-parameter reduction, (iii) cross-validated evaluation with bootstrap confidence intervals in place of a single random hold-out, and (iv) the explicit per-class and per-source breakdowns that their single-headline-number protocol does not produce. The 1.4-point macro-F1 deficit in-distribution and the 7.1-point HAM top-1 deficit are the trade made to obtain these properties. Table VI places the comparable numbers in their full context.

TABLE VI
COMPARISON WITH RECENT (2025–2026) SKIN-CANCER CLASSIFICATION LITERATURE ON HAM10000 AND ISIC 2019

#	Study (Year, Venue)	Method	Train data	Eval protocol	Headline	Macro-F1	Cross-dataset
1	Rahman et al. (2025)	ESRGAN + k-means (unsupervised)	HAM, ISIC 2019 (independent)	Unsupervised clustering	Acc 90.5 / 87.8	not reported	No
2	Raza et al. (2025)	CNN + ResNet152V2 + NGNDG-AF	HAM10000	Single hold-out (ratio unstated)	Acc 98.12	not reported	No
3	Vieira et al. (2025)	ConvNeXtXLarge (frozen)	HAM10000	Single split + cross-dataset	Acc 87.6 (HAM) / 56.0 (ISIC)	0.7615 (HAM)	Yes
4	Almutairi (2025)	U-Net + Virchow2 + Nomic + MLP	HAM10000	70/15/15, 10 trials	Acc 96.25	not reported (F1w 93.8)†	No
5	Saha et al. (2025)	KD student (0.29 M params)	HAM, ISIC 2019 (independent)	80/20 single split	Acc 93.2 / 84.1	not reported	No
6	Alotaibi & AlSaeed (2025)	Xception + attention	HAM10000	10-fold CV	Acc 94.11 (binary)	N/A (binary)	No
7	Naeem et al. (2025)	MedFusionNet (ConvNeXt + ViT)	HAM, ISIC 2019	Split unstated	Acc 98.8 / 97.9	not reported	No
8	Kılıç & Doğan (2026)	Pruned Xception + SMOTE + Avg-TopK	HAM10000	80/20, 5 runs	Acc 91.52 ± 0.16	not reported (F1w 91.5)†	No
9	Kabir (2026)	ResNet-18 + SCGAN (dataset paper)	HAM / HAM + SCGAN	Dataset paper	Acc 24.86	0.2403	No
10	Wicaksana et al. (2025)	YOLOv11x / VGG-19	HAM10000	70/20/10	Acc 89.68 (VGG) / 84.74 (YOLO)	not reported	No
—	Ours, Round 1	Dual ConvNeXt + EffNet (66 M)	HAM10000	Single hold-out	Acc 76.3 / 54.3	0.6905 / 0.4301	Yes
—	Ours, Round 2.3	ConvNeXt-Tiny full fine-tune (27.83 M trainable)	HAM + ISIC 2019 (joint)	5-fold + TTA + bootstrap CI	Acc 80.5 / 69.2	0.7401 / 0.5976	Yes

"Acc" = accuracy in percent; "F1w" = weighted-F1; "Cross-dataset" indicates whether true cross-dataset evaluation was performed (train on one dataset, test on a different dataset). The two values in our rows are HAM-test / ISIC-test. †F1w values for Almutairi (2025) and Kılıç & Doğan (2026) are inferred from per-class reports in the respective papers, not reported as a single headline metric.

C. Limitations

Several limitations should temper interpretation of these results. First, the two smallest classes — dermatofibroma (n = 239 combined; 124 in ISIC_NEW) and vascular lesions (n = 253 combined; 111 in ISIC_NEW) — contribute only 22–25 samples per fold to the ISIC-test split, giving rise to per-class bootstrap confidence intervals as wide as ±0.12. The macro-F1 numbers therefore retain a meaningful contribution from sampling noise in these classes; pooling across folds (as done in the headline numbers) tightens but does not eliminate that uncertainty. Second, the bimodal per-fold ISIC structure noted in § V-B — folds 1–3 averaging 0.62 while folds 4–5 collapse to 0.54 — indicates that the cross-source generalisation is sensitive to which specific lesion clusters end up in the test fold, and the pooled CV estimate is the appropriate way to report under these conditions but does not eliminate that sensitivity. Third, the present ablation (Table V) controls only the single-source vs multi-source training change; it does not separately isolate the contribution of Shades-of-Gray colour constancy or of the source-balanced sampler. The causal language in § VI-A is therefore stated as a hypothesis, and controlled leave-one-out runs of each component are the most important next experiment. Fourth, we used a single backbone family (ConvNeXt-Tiny) and did not run a separate hyper-parameter search on the multi-source pipeline; both the learning-rate and dropout configurations were inherited from the Round 1 baseline. A modest additional gain might therefore be available with backbone-specific tuning. Fifth, the evaluation pool is restricted to HAM10000 and ISIC 2019; generalisation to PH^2 [21], Dermofit [22], or in-clinic photographic datasets has not been measured and remains an important open question — the cross-domain story established here would gain considerably more weight from a true external test of that kind. Sixth, the model is purely image-based; metadata such as age, sex, and lesion localisation are available in both source datasets and have been shown to add diagnostic value [23] but are not used. Finally, our results were obtained on a single NVIDIA Tesla T4 GPU using fp16 mixed precision; whether the same numbers are reproducible under fp32 has not been tested.

D. Implications

Two implications follow directly from the results. First, the strong evidence that fusion of two CNN backbones contributes nothing measurable when a properly regularised loss is used suggests that recent "hybrid" architectures [6], [5] may be carrying complexity that is not justified by performance — at least on HAM10000 alone. Second, the literature's near-universal focus on within-dataset accuracy is producing models that cannot be deployed across hospitals or dermatoscope brands without expensive re-training. The 26-point in-distribution-to-out-of-distribution macro-F1 drop observed in our Round 1 baseline, and the 31-point top-1 accuracy drop independently observed by Vieira et al. [8], are not edge cases: they are the expected behaviour of any model trained on a single source. Cross-dataset evaluation should become a standard requirement for dermoscopic classification publications.

VII. CONCLUSION

This work investigated whether a simpler model trained on multi-source data — with a source-balanced sampler and colour constancy as supporting components — could close the cross-dataset generalisation gap that has been reported for the HAM10000 → ISIC 2019 setting. A single ConvNeXt-Tiny backbone (27.83 M parameters), trained jointly on HAM10000 and ISIC 2019 with a source-balanced weighted sampler, Shades-of-Gray colour constancy, plain cross-entropy with label smoothing, 4-view test-time augmentation, and 5-fold lesion-grouped cross-validation, achieves a pooled macro-F1 of 0.7401 [95 % CI 0.7252, 0.7541] on HAM-test and 0.5976 [95 % CI 0.5817, 0.6128] on ISIC-test. The cross-dataset gap is reduced from 0.260 to 0.142 — a 45 % reduction —

while the backbone shrinks by 58 % relative to a dual-backbone fusion baseline. Every per-class F1 improves on both datasets. Against the one directly comparable cross-dataset benchmark in the recent literature [8], the proposed model achieves a pooled cross-dataset top-1 accuracy of 69.24 % on the ISIC 2019 test set versus their 56.0 % — a 13.24-percentage-point improvement — with approximately 2.7× fewer trainable parameters (27.83 M trainable, fully fine-tuned, versus 75.5 M trainable on top of a frozen ≈ 348 M backbone in their Table 2) and under a stricter cross-validated protocol with bootstrap confidence intervals. The trade is a 7.08-point HAM-in-distribution top-1 deficit (80.54 % vs 87.62 %) and a 1.4-point macro-F1 deficit on the same in-distribution axis, both of which are the expected cost of jointly modelling two source distributions rather than overfitting to HAM10000 alone (§ VI-B). The work also surfaces a broader measurement issue: of ten 2025–2026 publications surveyed, only two report macro-F1 on the 7-class HAM10000 task (and only one of those uses the standard supervised protocol), and only one of the ten evaluates true cross-dataset performance. We argue that macro-F1 and explicit cross-dataset evaluation should become standard requirements for dermoscopic classification research.

E. Future Work

Three follow-up directions seem most promising. The first is to extend the cross-dataset evaluation to genuinely external archives — PH^2, Dermofit, and in-clinic photographic collections — to test whether the multi-source + source-balanced recipe also closes the gap to those domains, or whether a domain-adaptation step is additionally required. The second is to fuse the image classifier with the available patient metadata (age, sex, lesion localisation) through a simple late-fusion head, replicating the Liu et al. [23] finding in a cross-dataset setting. The third is to explore self-supervised pretraining on the combined unlabelled dermoscopic image pool before the multi-source supervised fine-tuning, building on the observation that the bcc, df, and vasc classes — which see the largest cross-dataset improvements in our experiments — also have the most distinctive visual features that self-supervised pretraining is well placed to capture.

REFERENCES

[1] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, p. 180161, 2018. https://doi.org/10.1038/sdata.2018.161

[2] B. P. Wicaksana, A. Pratama, and S. A. Wibowo, "Skin lesion classification using YOLOv11 on the HAM10000 dataset compared to VGG-19 and ResNet-50," *Inform*, vol. 10, no. 1, 2025. https://doi.org/10.25139/inform.v10i1.9310

[3] A. Alotaibi and D. AlSaeed, "Optimizing skin cancer screening with convolutional neural networks and attention mechanisms," *Diagnostics*, vol. 15, no. 1, p. 99, 2025. https://doi.org/10.3390/diagnostics15010099

[4] H. Kılıç and B. Doğan, "A pruned and parameter-efficient Xception framework with SMOTE and average top-K pooling for skin lesion classification," *PLoS ONE*, vol. 21, no. 3, p. e0341227, 2026. https://doi.org/10.1371/journal.pone.0341227

[5] K. M. Almutairi, "A dual-stream deep learning framework for skin cancer classification using foundation model embeddings and segmentation," *Scientific Reports*, vol. 15, p. 32301, 2025. https://doi.org/10.1038/s41598-025-01319-1

[6] A. Naeem, T. Anees, R. Shams, R. Latif, and A. Khan, "MedFusionNet: An adaptive attention-based feature fusion network of ConvNeXt and ViT for medical image classification," *Scientific Reports*, vol. 15, p. 43811, 2025. https://doi.org/10.1038/s41598-025-31816-2

[7] M. M. Rahman, M. T. I. Khan, M. F. Kabir, S. Akhter, and F. Chowdhury, "Advancing skin cancer detection: Integrating a novel unsupervised classification with ESRGAN-based image enhancement," *CAAI Transactions on Intelligence Technology*, 2025. https://doi.org/10.1049/cit2.12410

[8] J. Vieira, F. Mendonça, and F. Morgado-Dias, "Deep learning approaches for skin lesion detection," *Electronics*, vol. 14, no. 14, p. 2785, 2025. https://doi.org/10.3390/electronics14142785

[9] R. Raza, M. J. Iqbal, H. T. Rauf, and M. J. Awan, "An optimized convolutional neural network with novel activation function NGNDG-AF for skin lesion classification," *PLoS ONE*, vol. 20, no. 3, p. e0317181, 2025. https://doi.org/10.1371/journal.pone.0317181

[10] S. Kabir, "SCGAN: Generated skin-tone-balanced extension of the HAM10000 dataset," *PLoS ONE*, vol. 21, no. 4, p. e0331081, 2026. https://doi.org/10.1371/journal.pone.0331081

[11] S. Saha, M. T. Hossain, F. Sultana, and M. S. Hossain, "Phase-wise knowledge distillation for efficient skin lesion classification using a compact student network," *Healthcare Technology Letters*, vol. 12, p. e12120, 2025. https://doi.org/10.1049/htl2.12120

[12] C. Barata, M. E. Celebi, and J. S. Marques, "Improving dermoscopy image classification using color constancy," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 3, pp. 1146–1152, 2015. https://doi.org/10.1109/JBHI.2014.2336473

[13] G. D. Finlayson and E. Trezzi, "Shades of gray and colour constancy," in *Color and Imaging Conference*, vol. 2004, no. 1, pp. 37–41, 2004.

[14] S. Q. Gilani, M. Umair, M. Naqvi, O. Marques, and H.-C. Kim, "Adversarial training based domain adaptation of skin cancer images," *Life*, vol. 14, no. 8, p. 1009, 2024. https://doi.org/10.3390/life14081009

[15] J. Ye, "Enhancing skin lesion classification generalization with active domain adaptation," *arXiv preprint* arXiv:2412.00702, 2024. https://doi.org/10.48550/arXiv.2412.00702

[16] N. Sultana, W. Lu, X. Fan, and M. H. Yap, "Selective alignment transfer for domain adaptation in skin lesion analysis," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, LNCS, vol. 15965, pp. 596–606, Springer Nature Switzerland, 2025.

[17] S. Chamarthi, K. Fogelberg, R. C. Maron, T. J. Brinker, and J. Nölke, "Benchmarking unsupervised domain adaptation methods for skin lesion classification," 2024.

[18] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11976–11986.

[19] R. Müller, S. Kornblith, and G. Hinton, "When does label smoothing help?," in *Advances in Neural Information Processing Systems*, vol. 32, 2019, pp. 4694–4703.

[20] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. International Conference on Learning Representations (ICLR)*, 2019.

[21] T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. S. Marcal, and J. Rozeira, "PH² – A dermoscopic image database for research and benchmarking," in *Proc. 35th Annu. Int. Conf. IEEE Engineering in Medicine and Biology Society (EMBC)*, 2013, pp. 5437–5440.

[22] L. Ballerini, R. B. Fisher, B. Aldridge, and J. Rees, "A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions," in *Color Medical Image Analysis*, M. E. Celebi and G. Schaefer, Eds. Springer, 2013, pp. 63–86.

[23] Y. Liu, A. Jain, C. Eng, D. H. Way, K. Lee, P. Bui, K. Kanada, et al., "A deep learning system for differential diagnosis of skin diseases," *Nature Medicine*, vol. 26, pp. 900–908, 2020.