Proceedings of CONF-APMM 2026 Symposium: Applied Mathematics and Quantitative Modeling
DOI: 10.54254/2753-8818/2026.PU37107

Handcrafted Features vs Frozen CNN Features vs Fine-Tuning: A Three-Route Comparison on Imbalanced Dermoscopic Classification

Yuchen Song

School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou, China
Yuchen.Song25@student.xjtlu.edu.cn

Abstract. Automated dermoscopic classification is often evaluated with a single modelling route and headline score, obscuring the effects of representation, partitioning and imbalance handling. This study compared three routes on the seven-class HAM10000 dataset, in which melanocytic nevi are 66.95% of the 10,015 images: handcrafted histogram-of-oriented-gradients (HOG) and colour features, frozen ImageNet-pretrained ResNet50 features, and end-to-end ResNet50 fine-tuning. Each route was crossed with image-level and lesion-grouped splits and unweighted and class-weighted training, giving 12 test configurations. In all four matched settings, fine-tuned ResNet50 achieved the highest macro-F1 (0.662–0.810), exceeding frozen features by 0.147–0.190, which in turn exceeded handcrafted features by 0.224–0.393. Class weighting raised melanoma recall and lowered melanocytic-nevus recall in all six matched comparisons, yet improved macro-F1 only once, because the additional rare-class predictions reduced precision. Lesion-grouped rather than image-level splitting reduced macro-F1 most strongly for fine-tuning, by 0.126–0.148, indicating sensitivity to repeated-lesion overlap and partition composition. Fixed splits, extracted features, result tables and epoch-level training records support transparent evaluation, reproducibility and extension to further pretrained backbones.

Keywords: dermoscopy, transfer learning, class imbalance, data leakage, medical image classification

1. Introduction

1.1. Research background and research gap

Automated dermoscopic classification may improve the consistency and scalability of skin-lesion assessment. Convolutional neural networks and transfer learning can reuse representations learned from large natural-image datasets for medical images [1-4]. However, studies on public datasets such as HAM10000 often evaluate a single pipeline and emphasize a headline score [1, 4], obscuring whether gains arise from representation, partitioning, optimization or imbalance handling. Melanocytic nevi constitute 66.95% of HAM10000's 10,015 images, so majority-class performance can dominate aggregate accuracy. The dataset contains 7,470 unique lesions, including repeated

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.PU37107

views, so image-level splitting can place the same lesion across partitions and yield optimistic estimates. Dataset audits have highlighted the effect of split structure on benchmark credibility [5-7]. A matched comparison of representation, class weighting and lesion-aware partitioning is therefore needed.

1.2. Main research questions and methods

This study addresses three related questions: how performance changes across handcrafted, frozen pretrained and end-to-end fine-tuned representations; whether class weighting improves recognition of minority diagnoses or mainly shifts predictions away from the dominant class; and how lesion-aware splitting affects performance estimates when repeated lesions are present. The modelling routes were evaluated in a matched design under both image-level and lesion-grouped partitions, with and without class weighting. Macro-F1 was the primary performance measure, with complementary class-specific and balanced metrics reported for interpretation.

1.3. Research objectives and significance

The benchmark is designed to separate gains from transferred representation and task-specific adaptation, identify the class-specific trade-offs of weighting, and assess sensitivity to repeated-lesion overlap while recognizing that grouping also changes test-set composition. Fixed splits, extracted features, result tables, training records and multi-seed fine-tuning results support audit and reuse. By reporting both aggregate and class-specific outcomes, the study treats the credibility of an estimate as part of its interpretation, particularly for rare diagnoses. A high score is informative only when the partition prevents unintended lesion sharing and when class-specific effects are not concealed by a majority-dominated aggregate.

2. Methodology

2.1. Dataset, licence, and variables

This study used HAM10000, comprising 10,015 non-synthetic dermoscopic images acquired over approximately two decades at two sites [8]. The CC BY-NC 4.0 dataset was downloaded from Harvard Dataverse (https://doi.org/10.7910/DVN/DBW86T) on 25 July 2026. Images were 600 × 450-pixel RGB JPEGs. The original image directories and HAM10000_metadata.csv were used, excluding downsampled pixel tables. The outcome was the seven-class diagnosis code (dx) [8]. image_id linked metadata to images; lesion_id identified repeated views for grouped splitting. No other metadata were predictors.

2.2. Class imbalance and repeated lesions

All 10,015 image identifiers were unique and mapped to decodable files, representing 7,470 lesions: 5,514 with one image and 1,956 with between two and six. The dataset was strongly imbalanced (Table 1): melanocytic nevi contributed 6,705 images and dermatofibroma 115 (58.30:1). Because those 1,956 lesions contribute more than one image, an image-level random split can place different views of the same lesion on both sides of a partition, whereas lesion-level grouping cannot. The rarest classes are also the ones whose test support shifts most between the two protocols (Table 1). Class weighting and lesion-level separation were therefore crossed in a 2 × 2 design [6, 7, 9].

DOI: 10.54254/2753-8818/2026.PU37107

Table 1. Class distribution, unique-lesion counts and fixed test-partition support under the image-level and lesion-grouped protocols

Code	Diagnosis	All images	Unique lesions	Image-level test	Lesion-grouped test
akiec	Actinic keratoses / intraepithelial carcinoma	327	228	49	48
bcc	Basal cell carcinoma	514	327	77	66
bkl	Benign keratosis-like lesions	1,099	727	165	172
df	Dermatofibroma	115	73	17	10
mel	Melanoma	1,113	614	167	186
nv	Melanocytic nevi	6,705	5,403	1,006	1,016
vasc	Vascular lesions	142	98	22	29

2.3. Image preprocessing and split protocols

Images were converted to RGB, bilinearly resized to a 256-pixel shorter edge, centre-cropped to 224 × 224 and cached as lossless PNGs. All routes used this cache; only Route 3 training batches received online stochastic augmentation.

Using seed 42, fixed 70/15/15 splits were generated and stored in splits.csv. Two-stage StratifiedShuffleSplit on dx produced image-level partitions of 7,009/1,503/1,503 images. Two-stage GroupShuffleSplit on lesion_id produced lesion-grouped partitions of 6,959/1,529/1,527 images. Whole-lesion assignment caused deviations from target proportions. Because GroupShuffleSplit is not stratified, class support differed between protocols (Table 1), potentially confounding lesion overlap with partition composition.

Overlap was flagged where a validation image shared its lesion_id with training, or a test image with training or validation. The image-level protocol contained 542/1,503 overlapping validation images (36.06%) and 606/1,503 overlapping test images (40.32%); lesion-grouped splitting produced none. Validation overlap could affect model selection, whereas test overlap could bias held-out estimates [6, 7].

2.4. Feature-representation routes

Route 1 combined 6,084 histogram-of-oriented-gradients (HOG) features, 96 hue-saturation-value (HSV) histogram features and nine HSV colour moments. HOG used greyscale 224 × 224 images, nine orientations, 16 × 16-pixel cells, 2 × 2-cell blocks and L2-Hys normalization [10]; HSV features used 32 normalized bins per channel plus channel-wise mean, standard deviation (s.d.) and skewness. The 6,189 features were standardized using training statistics and classified by L2-regularized multinomial logistic regression.

Route 2 used ImageNet-pretrained ResNet50 as a fixed feature extractor [11, 12], following prior transfer-learning work [1-4]. It was initialized with torchvision ResNet50_Weights.IMAGENET1K_V2, its classifier replaced with identity and gradients disabled. Routes 2 and 3 used ImageNet channel means (0.485, 0.456, 0.406) and s.d. values (0.229, 0.224,

DOI: 10.54254/2753-8818/2026.PU37107

0.225). Global average pooling yielded 2,048 features, standardized using training statistics and classified as in Route 1.

Route 3 replaced the classifier with a seven-class linear head and optimized all parameters end to end. Training used AdamW, learning rate 1×10^{-4}, weight decay 0.01, cosine annealing, mixed precision, batch size 32 and at most 20 epochs. Training-only augmentation comprised horizontal and vertical flips (each probability 0.5) and rotations within ±15°. Weighted runs applied class-weighted cross-entropy without weighted sampling. Validation macro-F1 determined checkpoint selection and early stopping with five-epoch patience. The selected checkpoint was evaluated without refitting.

2.5. Imbalance treatment and model selection

All routes were evaluated with and without class weighting. Equation (1) defines balanced weights, where N is training size, K = 7 and the denominator contains the class-c training count. Weights used training data only; validation/test labels and synthetic oversampling were excluded [9].

$$w_c = \frac{N}{Kn_c} \tag{1}$$

For Routes 1 and 2, the logistic regression used L-BFGS (maximum 2,000 iterations; tolerance 1×10^{-4}). Candidates $C \in \{0.01, 0.1, 1, 10\}$ were ranked by validation macro-F1, ties were resolved by balanced accuracy and the smaller C, and the selected model was refitted on training data independently for each weighting × protocol combination. Ordered image-level then lesion-grouped, unweighted before weighted within each, selected C values were 0.1, 0.01, 0.01 and 1 for Route 1 and 0.01, 0.1, 0.01 and 0.01 for Route 2. Route 3 used the prespecified Section 2.4 hyperparameters without search or tuning.

2.6. Evaluation measures

Macro-F1, defined in equation (2), was the primary measure and balanced accuracy, defined in equation (3), was secondary [13]. Class-specific precision, recall, F1 and support, row-normalized 7 × 7 confusion matrices and macro one-versus-rest (OvR) area under the curve (AUC) were also reported. Accuracy was not primary because melanocytic nevi comprised 66.95% of images. Macro-F1 averages the per-class F1 scores with equal weight, so a rare diagnosis contributes as much as the dominant class, and balanced accuracy averages per-class recall in the same way. Test support was retained because the rarest lesion-grouped test class contained ten images.

$$\text{macro-F1} = \frac{1}{K} \sum_{c=1}^{K} \frac{2P_c R_c}{P_c + R_c} \tag{2}$$

$$\text{balanced accuracy} = \frac{1}{K} \sum_{c=1}^{K} R_c \tag{3}$$

2.7. Implementation and reproducibility

Experiments used Python 3.11.15 on Windows 10, CUDA 13.2 and an NVIDIA GeForce RTX 5060 Laptop GPU (8 GB). Packages were PyTorch 2.12.1+cu132, torchvision 0.27.1+cu132, NumPy 2.4.6, pandas 3.0.3, scikit-learn 1.9.0, scikit-image 0.26.0, Pillow 11.1.0, matplotlib 3.11.1 and joblib 1.5.3.

DOI: 10.54254/2753-8818/2026.PU37107

Route 3 was fitted with three independent seeds and reported as mean ± s.d. Routes 1 and 2 were deterministic for fixed splits and fitted once. Dispersion therefore represents Route 3 fitting variability, not a confidence or generalization interval.

3. Results: main comparison

3.1. Data overview

HAM10000 was strongly imbalanced (Table 1): in descending order of frequency, the seven classes were melanocytic nevi (nv), melanoma (mel), benign keratosis-like lesions (bkl), basal cell carcinoma (bcc), actinic keratoses or intraepithelial carcinoma (akiec), vascular lesions (vasc) and dermatofibroma (df). Representative examples showed within-class variation and overlapping colour, texture and border characteristics across diagnoses, together with visible hair and background skin (Figure 1).

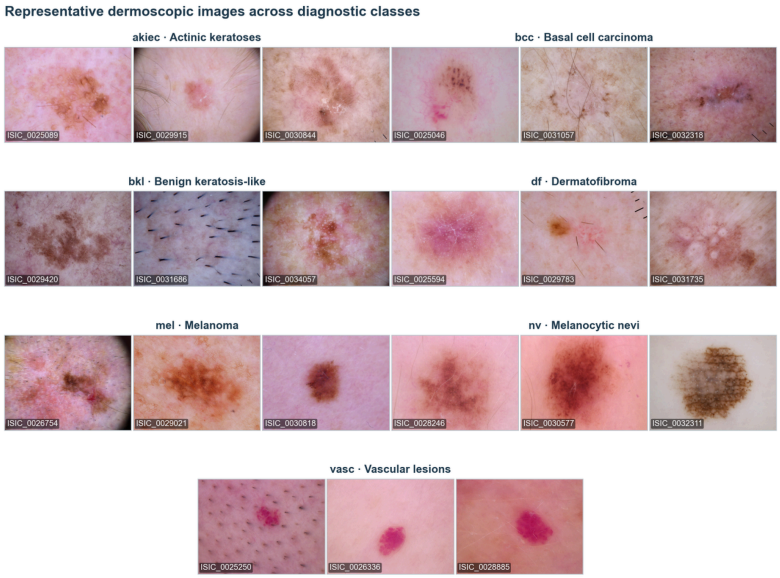


Figure 1. Representative dermoscopic images by diagnosis code [8]

(Data from: HAM10000, Tschandl et al. [8], licensed CC BY-NC 4.0)

3.2. Main results of the three routes

Route 3 achieved the highest macro-F1 in all four split-by-weighting settings, followed by Route 2 and Route 1 (Table 2). Macro-F1 was 0.244–0.292 for Route 1, 0.505–0.637 for Route 2 and 0.662–0.810 for Route 3, and balanced accuracy showed the same ordering. Route 3 values are means ± sample s.d. across seeds 42–44, whereas Routes 1 and 2 are fixed seed-42 results.

Across matched settings, frozen ResNet50 features exceeded handcrafted features by 0.224–0.393 in macro-F1 and fine-tuning added 0.147–0.190. Macro OvR AUC followed the same ordering, at 0.748–0.785, 0.897–0.929 and 0.935–0.979 for Routes 1–3. Differences were calculated from unrounded values.

DOI: 10.54254/2753-8818/2026.PU37107

Table 2. Test performance across the 12 combinations of route, split and weighting

Route	Split	Weighting	Macro-F1	Balanced accuracy	Macro OvR AUC	Test n
Route 1	Image-level	Unweighted	0.244	0.241	0.748	1,503
Route 1	Image-level	Class-weighted	0.276	0.287	0.752	1,503
Route 1	Lesion-grouped	Unweighted	0.292	0.272	0.785	1,527
Route 1	Lesion-grouped	Class-weighted	0.261	0.263	0.757	1,527
Route 2	Image-level	Unweighted	0.637	0.583	0.929	1,503
Route 2	Image-level	Class-weighted	0.597	0.605	0.912	1,503
Route 2	Lesion-grouped	Unweighted	0.515	0.491	0.910	1,527
Route 2	Lesion-grouped	Class-weighted	0.505	0.539	0.897	1,527
Route 3	Image-level	Unweighted	0.810 ± 0.023	0.795 ± 0.014	0.979 ± 0.002	1,503
Route 3	Image-level	Class-weighted	0.787 ± 0.051	0.804 ± 0.033	0.974 ± 0.009	1,503
Route 3	Lesion-grouped	Unweighted	0.663 ± 0.015	0.633 ± 0.012	0.953 ± 0.011	1,527
Route 3	Lesion-grouped	Class-weighted	0.662 ± 0.017	0.681 ± 0.004	0.935 ± 0.014	1,527

Table 3. Per-class precision, recall and test support for Route 1 (handcrafted features). U, unweighted; W, class-weighted; P/R, precision/recall

Class	n image	n grouped	Image U P/R	Image W P/R	Grouped U P/R	Grouped W P/R
akiec	49	48	0.12 / 0.06	0.15 / 0.14	0.14 / 0.08	0.09 / 0.08
bcc	77	66	0.21 / 0.18	0.23 / 0.32	0.23 / 0.20	0.18 / 0.18
bkl	165	172	0.34 / 0.33	0.30 / 0.38	0.34 / 0.32	0.29 / 0.33
df	17	10	0.00 / 0.00	0.12 / 0.06	0.20 / 0.10	0.06 / 0.10
mel	167	186	0.23 / 0.25	0.24 / 0.34	0.31 / 0.25	0.29 / 0.30
nv	1,006	1,016	0.77 / 0.81	0.82 / 0.72	0.77 / 0.85	0.80 / 0.78
vasc	22	29	0.12 / 0.05	0.06 / 0.05	0.60 / 0.10	0.17 / 0.07

DOI: 10.54254/2753-8818/2026.PU37107

Table 4. Per-class precision, recall and test support for Route 2 (frozen pretrained features). Abbreviations as in Table 3. Values are from seed 42

Class	n image	n grouped	Image U P/R	Image W P/R	Grouped U P/R	Grouped W P/R
akiec	49	48	0.61 / 0.45	0.45 / 0.49	0.42 / 0.46	0.36 / 0.48
bcc	77	66	0.63 / 0.57	0.50 / 0.65	0.51 / 0.55	0.43 / 0.61
bkl	165	172	0.59 / 0.59	0.51 / 0.59	0.56 / 0.58	0.49 / 0.61
df	17	10	0.88 / 0.41	0.56 / 0.53	1.00 / 0.10	0.07 / 0.10
mel	167	186	0.53 / 0.44	0.44 / 0.50	0.49 / 0.40	0.43 / 0.52
nv	1,006	1,016	0.87 / 0.93	0.90 / 0.84	0.87 / 0.90	0.92 / 0.80
vasc	22	29	1.00 / 0.68	0.82 / 0.64	0.76 / 0.45	0.68 / 0.66

Table 5. Per-class precision, recall and test support for Route 3 (end-to-end fine-tuning). Abbreviations as in Table 3

Class	n image	n grouped	Image U P/R	Image W P/R	Grouped U P/R	Grouped W P/R
akiec	49	48	0.89±0.08 / 0.68±0.10	0.94±0.07 / 0.69±0.04	0.76±0.08 / 0.42±0.06	0.62±0.02 / 0.37±0.03
bcc	77	66	0.86±0.03 / 0.81±0.01	0.83±0.05 / 0.78±0.06	0.73±0.10 / 0.67±0.03	0.69±0.05 / 0.77±0.07
bkl	165	172	0.72±0.05 / 0.85±0.05	0.65±0.12 / 0.85±0.04	0.59±0.01 / 0.81±0.02	0.57±0.01 / 0.79±0.03
df	17	10	0.84±0.14 / 0.75±0.09	0.74±0.18 / 0.80±0.07	0.78±0.20 / 0.30±0.00	0.52±0.09 / 0.47±0.06
mel	167	186	0.65±0.02 / 0.68±0.03	0.53±0.06 / 0.73±0.13	0.54±0.02 / 0.60±0.08	0.46±0.02 / 0.79±0.03
nv	1,006	1,016	0.95±0.02 / 0.93±0.02	0.96±0.00 / 0.86±0.04	0.93±0.01 / 0.89±0.01	0.96±0.01 / 0.79±0.02
vasc	22	29	0.95±0.05 / 0.86±0.05	0.91±0.06 / 0.91±0.08	0.91±0.06 / 0.75±0.02	0.89±0.08 / 0.79±0.09

3.3. Error distribution: confusion matrices

Under the unweighted image-level protocol, Route 1 predictions concentrated on nv: nv was the most frequent prediction for every non-nv class and non-nv diagonal proportions were 0.0–33.3% (Figure 2(a)). Route 2 gave the largest diagonal entry in every class, although nv remained the main error destination for bcc, bkl, df and mel (Figure 3(a)).

Route 3 showed stronger class separation: averaged across seeds, diagonal proportions ranged from 68.0% for akiec to 92.6% for nv, and mel-to-nv errors (22.4%) formed the largest off-diagonal

DOI: 10.54254/2753-8818/2026.PU37107

cell (Figure 4(a)). Image-level test support ranged from n = 17 for df to n = 1,006 for nv, so rare-class patterns should be interpreted cautiously.

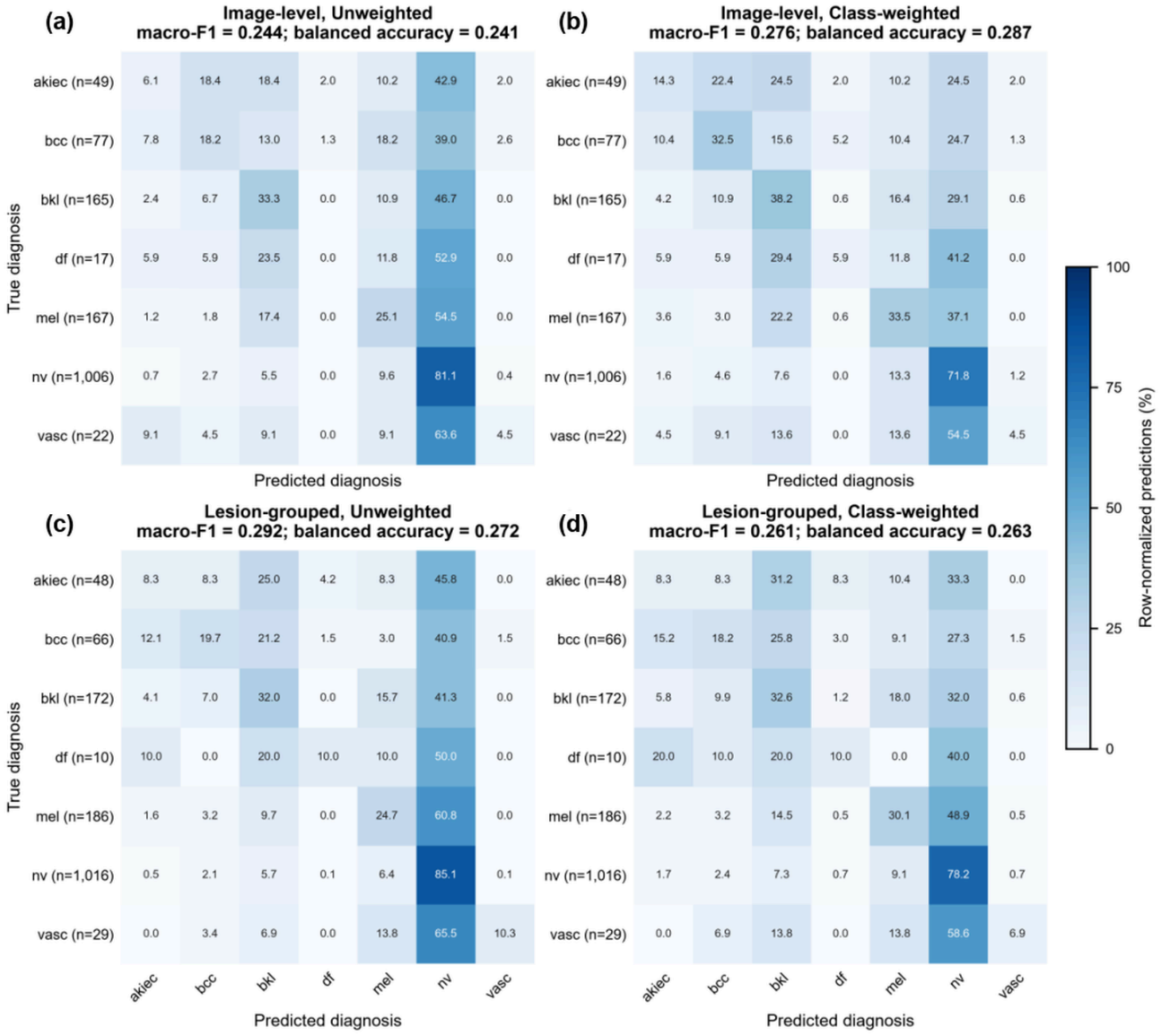


Figure 2. Confusion matrices for Route 1 (handcrafted features). Row-normalized matrices are shown for the image-level and lesion-grouped splits after unweighted and class-weighted training. Rows indicate true diagnoses and columns indicate predicted diagnoses; each row label reports test support

(Picture credit: Original, Data from: HAM10000 [8])

DOI: 10.54254/2753-8818/2026.PU37107

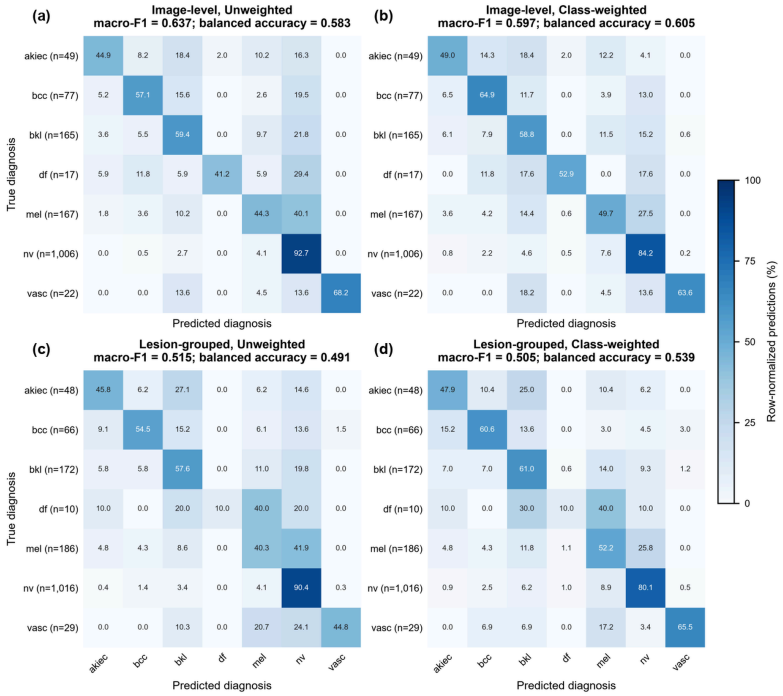


Figure 3. Confusion matrices for Route 2 (frozen pretrained features). Row-normalized matrices are shown for the image-level and lesion-grouped splits after unweighted and class-weighted training. Row and column conventions and support labels are as in Figure 2

(Picture credit: Original, Data from: HAM10000 [8])

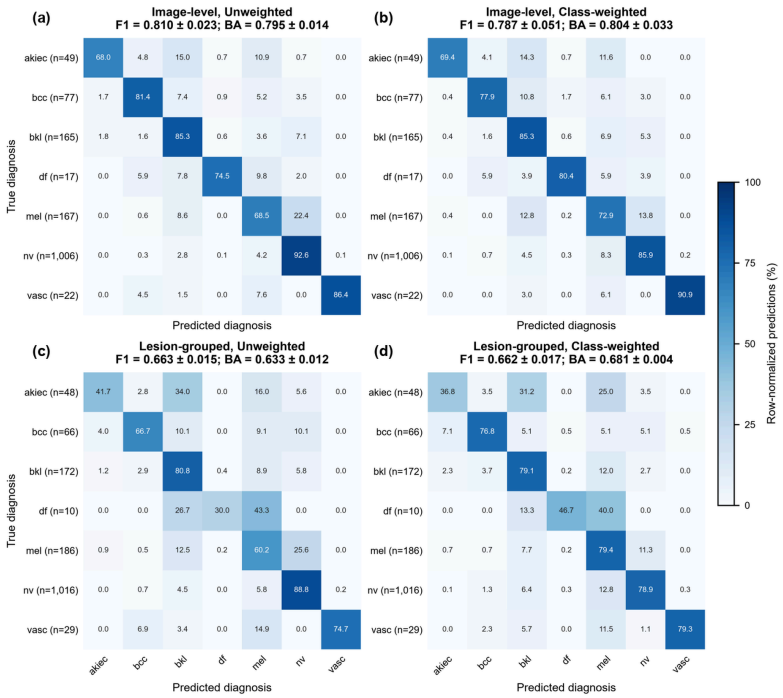


Figure 4. Confusion matrices for Route 3 (end-to-end fine-tuning). Cells show mean row-normalized proportions across seeds 42–44 for image-level and lesion-grouped splits after unweighted and class-weighted training. Panel titles report mean macro-F1 ± sample s.d.

DOI: 10.54254/2753-8818/2026.PU37107

(Picture credit: Original, Data from: HAM10000 [8])

4. Results: weighting and split protocol

4.1. Effect of class weighting

Class weighting increased mel recall and reduced nv recall in all six route-by-protocol comparisons. Under the image-level split, mel recall changed by +8.4, +5.4 and +4.4 percentage points for Routes 1–3 and nv recall by −9.3, −8.5 and −6.7 points; under lesion grouping, the mel changes were +5.4, +11.8 and +19.2 points and the nv changes were −6.9, −10.2 and −9.9 points. Class-specific results appear in Tables 3–5 and confusion matrices in Figures 2–4.

These shifts did not produce consistent aggregate gains: weighting improved macro-F1 only for Route 1 under the image-level protocol and reduced it in the other five comparisons (Table 2). Balanced accuracy increased in five comparisons and decreased only for lesion-grouped Route 1. For Routes 2 and 3, higher mel recall coincided with lower precision under both protocols (Tables 4 and 5), illustrating the main precision–recall trade-off. Effects for df and vasc should be interpreted cautiously because their test support was only 10–29 images, compared with more than 1,000 for nv.

4.2. Effect of the split protocol

Route 3 was most sensitive to the protocol change (Table 6): macro-F1 decreased by 0.148 without weighting and 0.126 with weighting, with the largest declines in balanced accuracy and macro OvR AUC. Route 2 also declined on all three metrics under both weighting conditions. Route 1 was mixed: macro-F1 and balanced accuracy rose without weighting but fell with weighting, whereas macro OvR AUC rose under both settings. Because lesion grouping also changed test-set composition, these differences should not be attributed solely to removal of repeated-lesion overlap.

Table 6. Changes in test metrics after replacing the image-level random split with the lesion-grouped split. Values are lesion-grouped minus image-level results; positive values indicate higher performance under lesion grouping. Differences were calculated from the underlying unrounded metric values

Route	Weighting	Δ Macro-F1	Δ Balanced accuracy	Δ Macro OvR AUC
Route 1	Unweighted	+0.047	+0.031	+0.036
Route 1	Class-weighted	−0.015	−0.023	+0.005
Route 2	Unweighted	−0.122	−0.092	−0.019
Route 2	Class-weighted	−0.092	−0.065	−0.015
Route 3	Unweighted	−0.148	−0.162	−0.026
Route 3	Class-weighted	−0.126	−0.123	−0.039

4.3. Fine-tuning process observations

Across all four Route 3 conditions, mean training loss declined markedly, while validation macro-F1 (Figure 5) improved overall but remained non-monotonic with visible across-seed dispersion. Four class-weighted runs reached the 20-epoch limit and no unweighted run did. For seeds 42, 43 and 44 respectively, training ended at epochs 17, 13 and 18 under image-level unweighted training and 20,

DOI: 10.54254/2753-8818/2026.PU37107

10 and 20 under image-level weighted training; under lesion-grouped training the epochs were 10, 16 and 18 without weighting and 20, 20 and 17 with weighting.

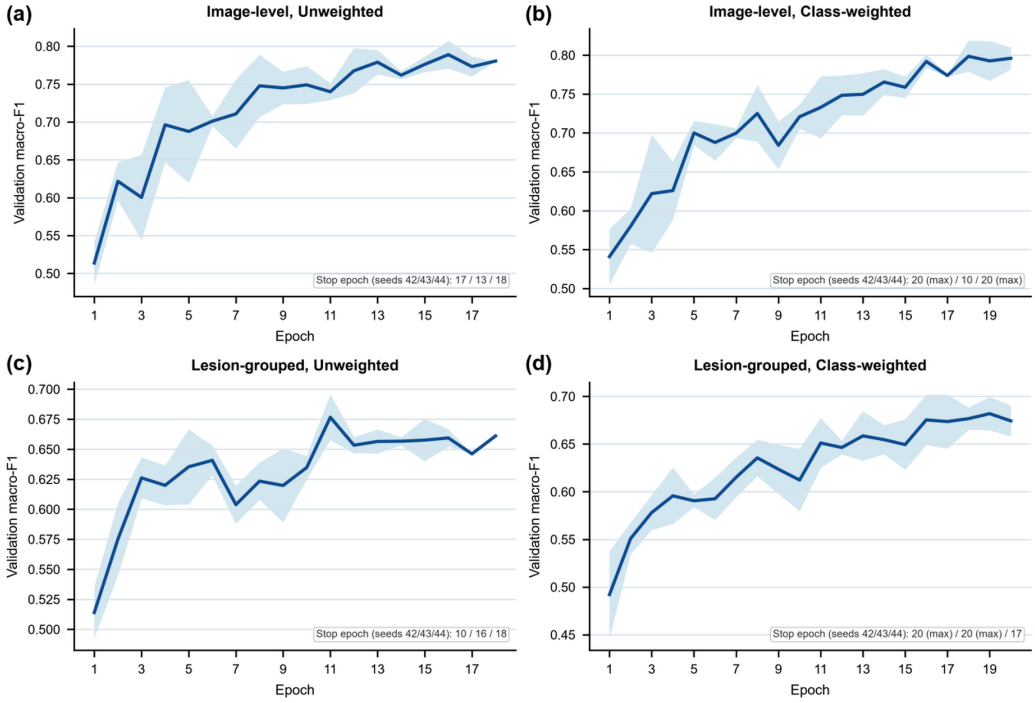


Figure 5. Route 3 validation macro-F1 during fine-tuning. Lines show means across seeds and shading indicates ±1 s.d.; observations after a run stopped are omitted. Annotations give the stop epoch of each seed in the order 42/43/44, and (max) marks runs that reached the 20-epoch limit

(Picture credit: Original)

5. Discussion and limitations

5.1. Interpretation of results

Fine-tuning ranked above frozen ResNet50 features and both above handcrafted features under macro-F1 and macro OvR AUC. Class weighting increased minority-class recall but often reduced precision rather than improving both, most clearly in the shift from nv to mel: under lesion-grouped Route 3, mean mel recall rose from 0.60 to 0.79 while mean precision fell from 0.54 to 0.46, and 325 of 443 newly assigned mel predictions had previously been labelled nv (Table 5). Lesion grouping reduced performance most for fine-tuning, suggesting sensitivity to repeated-lesion overlap and changes in test-set composition, so these effects should not be attributed to leakage alone.

5.2. Limitations and future work

HAM10000 was collected at two centres with limited histopathological confirmation, while rare-class test support was only 10–29 images. Each split protocol was sampled once; Routes 1 and 2 used one fit, whereas Route 3 used three seeds, limiting estimates of partition and fitting variability. The study evaluated one dataset, three modelling routes, one weighting scheme and image-level classification without lesion segmentation, so the results should not be generalized to clinical

DOI: 10.54254/2753-8818/2026.PU37107

performance. Future work should use multicentre cohorts, additional pretrained backbones and lesion-aware or segmentation-based models.

6. Conclusion

This study compared three modelling routes under matched split and weighting conditions, giving 12 test configurations. Frozen ResNet50 features outperformed handcrafted HOG and colour features by 0.224–0.393 in macro-F1, and end-to-end fine-tuning added a further 0.147–0.190, ranking first in all four matched settings with macro-F1 between 0.662 and 0.810. Macro OvR AUC preserved the same ordering, so the advantage was not confined to a single decision threshold. The comparison therefore separates the contribution of a transferred representation from that of task-specific adaptation. Class weighting increased minority-class recall but generally reduced precision. It raised melanoma recall and lowered melanocytic-nevus recall in all six matched comparisons, yet improved macro-F1 only once, which shows that weighting mainly broadened rare-class prediction rather than sharpening it. Evaluation design mattered as much as model choice. Lesion grouping caused the largest performance losses for fine-tuning, at 0.126–0.148 in macro-F1, showing that repeated-lesion overlap and partition composition materially affect reported performance. Because all three routes started from the same cached images, the same fixed split files and the same validation-based selection rule, the reported differences are attributable to the feature representation rather than to preprocessing or partitioning. Fixed splits, extracted features, result tables and epoch-level training records make every reported value auditable and allow the comparison to be extended to further backbones. These findings support evaluating representation, imbalance handling and lesion-level partitioning as one decision rather than as separate choices, although interpretation remains limited by one dataset, small rare-class support and a restricted seed design, and the results do not establish clinical utility or support diagnostic use.

References

[1] Wu, Y., Chen, B., Zeng, A., Pan, D., Wang, R., & Zhao, S. (2022). Skin cancer classification with deep learning: A systematic review. *Frontiers in Oncology, 12*, 893972. https://doi.org/10.3389/fonc.2022.893972

[2] Debelee, T. G. (2023). Skin lesion classification and detection using machine learning techniques: A systematic review. *Diagnostics, 13*(19), 3147. https://doi.org/10.3390/diagnostics13193147

[3] Naqvi, M., Gilani, S. Q., Syed, T., Marques, O., & Kim, H.-C. (2023). Skin cancer detection using deep learning: A review. *Diagnostics, 13*(11), 1911. https://doi.org/10.3390/diagnostics13111911

[4] Mehmood, A., Gulzar, Y., Ilyas, Q. M., Jabbari, A., Ahmad, M., & Iqbal, S. (2023). SBXception: A shallower and broader Xception architecture for efficient classification of skin lesions. *Cancers, 15*(14), 3604. https://doi.org/10.3390/cancers15143604

[5] Fogelberg, K., Chamarthi, S., Maron, R. C., Niebling, J., & Brinker, T. J. (2023). Domain shifts in dermoscopic skin cancer datasets: Evaluation of essential limitations for clinical translation. *New Biotechnology, 76*, 106-117. https://doi.org/10.1016/j.nbt.2023.04.006

[6] Varoquaux, G., & Cheplygina, V. (2022). Machine learning for medical imaging: Methodological failures and recommendations for the future. *npj Digital Medicine, 5*, 48. https://doi.org/10.1038/s41746-022-00592-y

[7] Bradshaw, T. J., Huemann, Z., Hu, J., & Rahmim, A. (2023). A guide to cross-validation for artificial intelligence in medical imaging. *Radiology: Artificial Intelligence, 5*(4), e220232. https://doi.org/10.1148/ryai.220232

[8] 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. https://doi.org/10.1038/sdata.2018.161

[9] Salmi, M., Atif, D., Oliva, D., Abraham, A., & Ventura, S. (2024). Handling imbalanced medical datasets: Review of a decade of research. *Artificial Intelligence Review, 57*, 259. https://doi.org/10.1007/s10462-024-10884-2

[10] Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 886-893).

DOI: 10.54254/2753-8818/2026.PU37107

https://doi.org/10.1109/CVPR.2005.177

[11] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770-778). https://doi.org/10.1109/CVPR.2016.90

[12] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Li, F.-F. (2009). ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248-255). https://doi.org/10.1109/CVPR.2009.5206848

[13] Farhadpour, S., Warner, T. A., & Maxwell, A. E. (2024). Selecting and interpreting multiclass loss and accuracy assessment metrics for classifications with class imbalance: Guidance and best practices. *Remote Sensing, 16*(3), 533. https://doi.org/10.3390/rs16030533