STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 1–21.
Published online in International Academic Press (www.IAPress.org)

Lesion-Aware Evaluation of Frozen Swin Transformer and CNN Features for Binary Skin Lesion Classification: Model Comparison and Threshold Transfer

Esam Othman*

Information Systems Department, College of Computer and Information Sciences, Prince Sultan University, Riyadh, Saudi Arabia

Abstract Frozen pretrained features with a linear classifier are a low-cost way to adapt vision models to dermoscopy, but such evaluations rarely account for repeated images of the same lesion, sampling uncertainty, score calibration, or whether an operating point chosen on a small validation set holds on new data. We present a lesion-aware statistical evaluation of four frozen ImageNet backbones (Swin-T, ConvNeXt-T, MobileNetV2 and ResNet-50) on the official ISIC 2018 Task 3 partitions for a referral-oriented binary target (positive: melanoma, basal cell carcinoma, actinic keratosis/intraepithelial carcinoma). Each pipeline used an L2-regularised logistic classifier and a prespecified threshold (maximum recall at $\geq$80% validation specificity) applied once to the 1,512 test images; all inference used a paired lesion-cluster bootstrap, and archive lesion identifiers confirmed that the partitions are lesion-disjoint. Frozen Swin-T reached a test AUC of 0.893 (95% CI 0.870–0.914), above MobileNetV2 (0.834; difference 0.059, 0.037–0.084) and ResNet-50 (0.818), but there was no evidence of a difference from the size-matched ConvNeXt-T (difference −0.004, −0.022 to 0.013), and model-specific regularisation reduced the gap to MobileNetV2 to 0.019. At a matched development-specificity target, Swin-T detected 12.1 percentage points (6.7–17.7) more positive images than MobileNetV2. In decision-curve analysis, Swin-T had higher net benefit than MobileNetV2 and ResNet-50 at threshold probabilities of 0.10–0.30, with no evidence of a difference from ConvNeXt-T. Thresholds re-selected on resamples of the 193-image validation split moved test sensitivity between 70% and 90%, and even a stable development threshold missed its specificity target because the test negatives contained more benign keratoses, dermatofibromas and vascular lesions; 68% of the false positives were histopathology-confirmed benign lesions. Fine-tuning Swin-T (five seeds) raised the test AUC to 0.940 (range 0.923–0.950) without removing this threshold fragility. Validation-selected operating points should be reported together with their transfer uncertainty, calibration and case mix.

Keywords skin lesion classification; Swin Transformer; frozen features; cluster bootstrap; threshold selection; calibration; decision curve analysis; ISIC 2018

AMS 2010 subject classifications 62P10, 68T07, 62F40

DOI: 10.19139/soic-2310-5070-5019

1. Introduction

In automated classification of dermoscopic images, overall accuracy can hide missed positive lesions; a classifier intended to prioritise detection should therefore be judged by its sensitivity and specificity at a defined decision threshold. Diagnosis relies on clinical assessment, dermoscopy and, when needed, histopathology, and the accuracy of human readers varies with experience [1, 2]. Deep neural networks have been compared with dermatologists in several reader studies [3, 4], and reviews note persistent challenges, including the need for representative data and integration into clinical workflows [5, 6]. In a study of human–computer collaboration, good AI support improved diagnostic accuracy, whereas faulty AI misled clinicians of all experience levels [2]; the operating behaviour of a classifier, not only its ranking ability, therefore matters.

*Correspondence to: Esam Othman (Email: Eothman@psu.edu.sa). Prince Sultan University, Rafha Street, Riyadh 11586, Saudi Arabia.

ISSN 2310-5070 (online) ISSN 2311-004X (print)
Copyright © 202x International Academic Press

1.1. Motivation

A pretrained Swin Transformer [7] is a natural starting point: Swin-based dermoscopic classifiers have been reported under several training and evaluation designs [8, 9, 10], and using the transformer as a frozen feature extractor with a small regularised classifier (a linear probe) is cheap, reproducible and a common first step before fine-tuning [35]. Three statistical questions raised by this design are seldom addressed together. First, dermoscopy datasets contain several images of the same lesion, so images are not independent sampling units and image-level evaluation can overstate performance [11, 12]. Second, backbone comparisons are usually reported as point estimates on a single test set, without paired uncertainty and without asking whether the comparison isolates the architecture or the whole pretraining pipeline. Third, when a decision threshold is selected on a small validation subset, it is rarely checked whether the intended operating point is reached on new data, why it is missed, or whether the scores can be read as probabilities; a change in the mix of diagnoses between samples can by itself shift sensitivity and specificity [13], and a poorly calibrated score can keep its AUC while making threshold choice difficult to interpret [39]. For a clinical reader, these questions determine whether a reported "sensitivity of 90%" would hold in another clinic.

This study asks: *on the official ISIC 2018 Task 3 partitions, how large and how certain is the difference between frozen Swin-T features and frozen convolutional features for a binary positive-lesion target, and how reliably does a validation-selected operating point transfer to the test set?* We do not propose a new architecture; the comparison is between complete pretraining pipelines (architecture, ImageNet training recipe and preprocessing), not between isolated representations.

1.2. Contributions

The contributions are methodological and empirical:

- a lesion-level independence audit using archived lesion identifiers, showing that the official partitions are lesion-disjoint and quantifying the within-partition clustering that defines the resampling unit (with the caveat that lesion-disjoint is not necessarily patient-disjoint);
- lesion-clustered paired inference for the backbone comparison, at the level of ranking metrics, fixed operating points and a matched development-specificity target;
- an analysis of threshold transfer that separates the sampling variability of a 193-image validation split from a systematic case-mix shift between training and test negatives, and evaluates lesion-grouped out-of-fold thresholds as a more stable alternative;
- a clinical-utility layer: calibration diagnostics, decision curves and cost-sensitive operating points, and a case-level failure analysis linking false positives to the diagnostic confirmation pathway of the lesion;
- sensitivity analyses of classifier regularisation, an off-the-shelf linear probe, grouped versus image-level development resampling, convolutional comparators (MobileNetV2, ResNet-50 and the size-matched ConvNeXt-T), and a five-seed fine-tuned Swin-T reference.

MobileNetV2 [14] is the primary lightweight convolutional comparator, with ResNet-50 and ConvNeXt-T [15] as additional comparators; ConvNeXt-T was designed as a convolutional network of similar size and computational cost to Swin-T. All backbones use ImageNet weights [16]. Because the architectures, checkpoint training recipes and preprocessing differ, the comparisons concern pipelines and do not isolate the effect of shifted-window attention.

2. Related work

2.1. Deep learning for dermoscopic classification

Ajabani et al. [17] compared an EfficientNetB3 model with 69 healthcare practitioners on a 150-image test set. The ISIC 2018 challenge received 159 submissions for its disease-classification task [18]; one team ranked second with an ensemble of fine-tuned CNNs [19]. Several studies use Swin-based or hybrid transformers: Ayas [8] trained

a Swin Transformer with a class-weighted loss on ISIC 2019; Aruk et al. [9] compared 15 CNNs and 15 vision transformers on HAM10000 and ISIC 2019 and found the best transformers to be Swin-based; Nie et al. [20] combined part of a ResNet-50 with a transformer encoder for seven-class ISIC 2018 classification; Ramya and Jaisankar [21] used a Swin-T branch for melanoma versus non-melanoma classification; and Pathirana et al. [22] (preprint) fused a Swin backbone with clinical metadata. In this journal, El Idrissi El-Bouzaidi et al. [41] fine-tuned MobileNetV2 with a new pre-processing strategy for seven-class HAM10000 classification.

2.2. Frozen features, linear probing and dermatology foundation models

Fixed ImageNet features with a logistic-regression classifier are a standard way to measure how well a pretrained representation transfers; across 12 datasets, Kornblith et al. [35] found that fixed-feature transfer depends strongly on how the network was trained on ImageNet, and that regularisers which improve ImageNet accuracy can make penultimate-layer features worse for transfer. This supports reading our backbone comparison as one between training pipelines. Dermatology-specific foundation models now provide an alternative source of frozen features: PanDerm was pretrained with self-supervision on over two million images from 11 institutions and evaluated on 28 benchmarks [36]. Mahbod et al. [10] (preprint) applied non-linear probes (MLP, XGBoost, TabNet) to frozen PanDerm features, fully fine-tuned ViT-base and Swin V2-base models, and then fused the *predictions* of the PanDerm and Swin models.

2.3. Evaluation design, leakage, dataset shift and operating points

Two recent papers in this journal highlight the role of evaluation design. Muqtadir et al. [11] evaluated attention modules on a ConvNeXt-Tiny architecture under image-level and lesion-level grouped cross-validation and found that the validation protocol affected results more than the attention module; moving to lesion-level validation lowered macro-F1 by about 10.8 points on average in a seven-class setting. Belal et al. [23] used leakage-controlled evaluation for four-class feature fusion. HAM10000 contains several images of many lesions [12], which motivates lesion-grouped evaluation. Performance also changes under dataset shift: in the ISIC 2019 challenge, algorithms lost accuracy on images from other institutions and frequently labelled diagnoses outside the training categories as malignant [37], and dermatology models performed worse on a curated set with diverse skin tones [38]. When the positive class is a minority, precision–recall summaries are informative alongside ROC curves [24]. Studies of diagnostic test accuracy have long noted that sensitivity and specificity vary with the case mix of the tested population [13]; we use this idea for the transfer of a threshold between data partitions. Calibration—agreement between predicted and observed risk—is required before a score can be read as a probability [39, 28], and decision-curve analysis expresses clinical utility as net benefit across threshold probabilities that encode the relative harms of false-positive and false-negative decisions [40].

3. Materials and methods

3.1. Overview

Each pipeline (Figure 1) prepares an RGB image with the transforms specified for its checkpoint, encodes it with a frozen ImageNet backbone, standardises the features with training statistics, and scores them with an L2-regularised logistic classifier. The prespecified primary analysis selects one threshold per model on the official validation split and evaluates it once on the official test split, together with the default threshold of 0.5. The evaluation layer (Sections 3.6–3.9) quantifies uncertainty and threshold transfer without using test labels to choose any threshold or hyperparameter. The calibration, clinical-utility, linear-probe and failure analyses (Sections 3.10–3.12) and the additional fine-tuning seeds are post hoc additions to the prespecified protocol; they are labelled as such and follow the same rule that test labels are never used for selection.

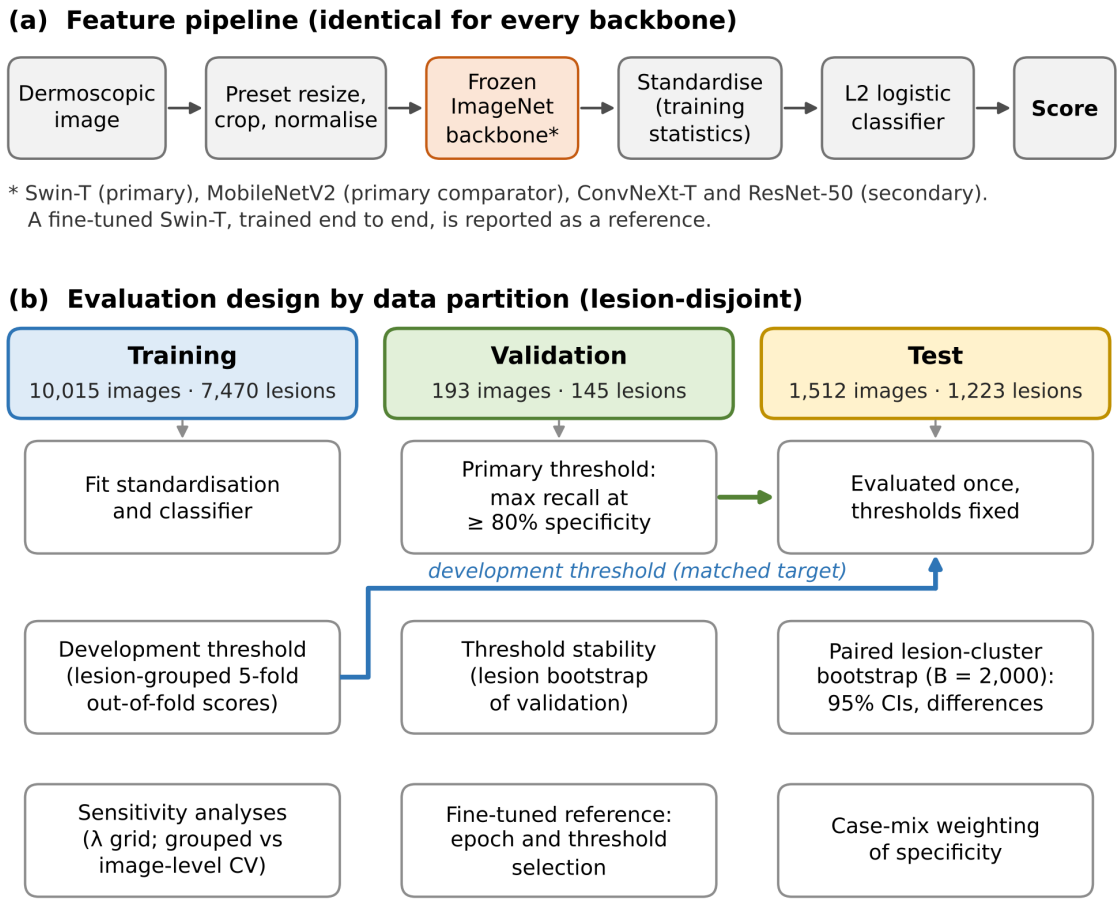

Figure 1. Study design. (a) Every backbone feeds the same standardisation, classifier and threshold rule. (b) Role of each official partition. The green arrow carries the primary validation-selected threshold to the test set; the blue arrow carries the threshold selected on lesion-grouped out-of-fold training scores, used for the matched-target comparison. Test labels are used only for evaluation, and all resampling is by lesion.

3.2. Data, binary target and lesion identifiers

We used the official ISIC 2018 Task 3 training, validation and test images and ground-truth files [18]; the training images are the HAM10000 dataset [12]. The seven categories are MEL (melanoma), NV (melanocytic nevus), BCC (basal cell carcinoma), AKIEC (actinic keratosis or Bowen disease/intraepithelial carcinoma), BKL (benign keratosis), DF (dermatofibroma) and VASC (vascular lesion). MEL, BCC and AKIEC are positive and NV, BKL, DF and VASC negative. The target is defined by management rather than by malignancy alone: the positive categories are those for which a dermatology referral is normally followed by an active intervention (excision, biopsy, destruction or field therapy) because the lesion is malignant (MEL, BCC) or is an intraepidermal carcinoma or recognised precursor of squamous cell carcinoma (AKIEC), whereas the negative categories are benign lesions usually managed by reassurance or monitoring. The binary task is therefore a referral-triage target, and its sensitivity should be read as the proportion of lesions needing treatment that would be referred. The positive class is clinically heterogeneous—its members differ in malignant potential and urgency—so detection rates are also reported separately for each diagnosis (Table 3). The target is specific to this study; results are not comparable with other binary tasks such as melanoma versus all others.

Lesion identifiers (`lesion_id`), diagnosis-confirmation type, approximate age, sex and anatomical site were retrieved for every image from the ISIC Archive API for the three Task 3 collections [25]. The archive holds no patient identifiers for these collections, so patient-level independence could not be assessed. We therefore describe the partitions as *lesion-disjoint*, which is not necessarily *patient-disjoint*: different lesions of the same patient could appear in different partitions, and any resulting dependence is not captured by the lesion-level analysis. We checked that every image had a lesion identifier, that diagnoses were consistent within lesions, and whether any lesion occurred in more than one partition. As a supplementary screen for re-uploaded or near-identical images, we computed, in each backbone's L2-normalised feature space, the cosine similarity of every validation and test image to its nearest training image, and compared it with similarities between different images of the same training lesion and between nearest neighbours from different training lesions.

3.3. Feature extraction and classifier

Swin-T used TorchVision's `Swin_T_Weights.IMAGENET1K_V1` preset (shorter side resized to 232 pixels with bicubic interpolation, 224×224 centre crop, ImageNet normalisation) [26]; the classification head was replaced by an identity layer, giving a 768-dimensional feature after the final normalisation and global pooling. MobileNetV2 used `MobileNet_V2_Weights.IMAGENET1K_V2` (bilinear resize to 232, 224×224 centre crop) and a 1,280-dimensional feature after global average pooling. The backbones were frozen in evaluation mode; features were extracted in batches of 64 with CUDA float16 autocasting and saved as float32. Two further backbones were extracted with the same code: ConvNeXt-T (`ConvNeXt_Tiny_Weights.IMAGENET1K_V1`; bilinear resize to 236, 224×224 centre crop; 768-dimensional feature after the final layer normalisation) and ResNet-50 (`ResNet50_Weights.IMAGENET1K_V2`; bilinear resize to 232, 224×224 centre crop; 2,048-dimensional pooled feature). Swin-T and MobileNetV2 form the prespecified primary comparison; ConvNeXt-T and ResNet-50 are secondary comparators assessed with the same classifier, threshold rules, cross-validation and bootstrap.

Each feature was standardised with its training mean and sample standard deviation (minimum 0.01). A single-layer logistic classifier with weights and bias initialised to zero minimised

$$\mathcal{L}(\mathbf{w}, b) = \frac{1}{n}\sum_{i=1}^{n}\Big[-y_i \log \sigma(\mathbf{w}^\top \mathbf{z}_i + b) - (1-y_i)\log\big(1-\sigma(\mathbf{w}^\top \mathbf{z}_i + b)\big)\Big] + \lambda\|\mathbf{w}\|_2^2, \tag{1}$$

where $\mathbf{z}_i$ is the standardised feature vector, σ the logistic function and $\lambda = 0.001$ in the main analysis (bias unpenalised), using L-BFGS [27] with a strong-Wolfe line search, at most 150 iterations and PyTorch's default tolerances (gradient 10^{-7}, change 10^{-9}). The value $\lambda = 0.001$ was fixed before the experiments rather than tuned; its effect is analysed in Section 3.9. This classifier is a linear probe in the usual sense; an off-the-shelf alternative is examined in Section 3.11. The score $s_i = \sigma(\mathbf{w}^\top \mathbf{z}_i + b)$ is used for ranking and thresholding; its calibration is assessed in Section 3.10.

3.4. Threshold rule

Candidate thresholds were the unique scores plus one value above the maximum. For a selection set S and specificity floor φ, the rule is

$$\tau^* = \arg\max_{\tau \in \mathcal{C}:\ \mathrm{Spec}_S(\tau) \geq \varphi} \mathrm{Sens}_S(\tau), \tag{2}$$

with ties broken by higher specificity and then higher threshold; an image is positive if $s_i \geq \tau^*$. In the primary analysis the selection set is the official validation split (193 images; 44 positive), giving thresholds of 0.1348 (Swin-T) and 0.1931 (MobileNetV2), which were applied unchanged to the test set. The floor $\varphi = 0.80$ was chosen to represent a sensitivity-oriented triage setting with a bounded false-positive load—no more than one in five benign lesions referred—and lies within the range of specificities reported for dermatologists in reader studies on dermoscopic images [1, 4]. It is an experimental constraint, not a clinical recommendation. The alternative floors of 70% and 90% were fixed in advance to bracket it by ±10 percentage points, representing a more detection-oriented and a more resource-constrained setting; they test whether the conclusions depend on the exact floor rather than define clinical targets. Clinically grounded alternatives, which weight false negatives and false positives explicitly, are given in Section 3.10.

3.5. Metrics

For the study-defined positive class we report the true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) and the following quantities:

$$\text{Sens} = \frac{TP}{TP+FN}, \quad \text{Spec} = \frac{TN}{TN+FP}, \quad \text{Prec} = \frac{TP}{TP+FP}, \quad \text{Acc} = \frac{TP+TN}{n}, \tag{3}$$

$$\text{BAcc} = \frac{\text{Sens}+\text{Spec}}{2}, \quad F_1 = \frac{2\,\text{Prec}\cdot\text{Sens}}{\text{Prec}+\text{Sens}}, \quad G = \sqrt{\text{Sens}\cdot\text{Spec}}. \tag{4}$$

With $\mathcal{P}$ and $\mathcal{N}$ the sets of positive and negative images (n_1 and n_0 images),

$$\text{AUC} = \frac{1}{n_1 n_0}\sum_{i\in\mathcal{P}}\sum_{j\in\mathcal{N}}\Big[\mathbb{1}(s_i > s_j) + \tfrac{1}{2}\mathbb{1}(s_i = s_j)\Big], \qquad \text{AP} = \sum_k (R_k - R_{k-1})\,P_k, \tag{5}$$

where R_k and P_k are the recall and precision at the k-th distinct score threshold in decreasing order ($R_0 = 0$). The reference AP for an uninformative score is the positive proportion (0.203 in the test set). We report the sensitivity on the empirical test ROC curve at 80% specificity as a threshold-free matched summary; it is a curve summary rather than a usable operating point. Detection rates within each positive diagnosis and false-positive rates within each negative diagnosis are reported at the selected thresholds.

3.6. Uncertainty and paired comparison

Uncertainty in test performance conditional on the fitted models and fixed thresholds was estimated with a lesion-cluster bootstrap [29, 30]. Let $\mathcal{L}^+$ (213 lesions) and $\mathcal{L}^-$ (1,010 lesions) be the positive and negative test lesions. In replicate b we draw $|\mathcal{L}^+|$ lesions from $\mathcal{L}^+$ and $|\mathcal{L}^-|$ from $\mathcal{L}^-$ with replacement, include all images of each drawn lesion, and recompute each statistic $\hat{\theta}_b^*$ ($b = 1,\ldots,B$; $B = 2{,}000$). The 95% interval is the percentile interval $[\hat{\theta}^*_{(0.025)}, \hat{\theta}^*_{(0.975)}]$. For a paired difference Δ the same resampled images are used for both models, and the two-sided bootstrap p-value is $2\min\{\Pr^*(\Delta^* \le 0), \Pr^*(\Delta^* \ge 0)\}$, capped at 1. P-values are reported only for a small set of prespecified differences (AUC, AP, sensitivity and specificity). As supplementary image-level checks we report DeLong's test for correlated AUCs [31] and exact McNemar tests [32] for paired classifications; these do not account for lesion clustering. This conditional analysis excludes variability from model fitting and threshold selection, which is addressed in Section 3.7. Bootstrap seeds were fixed in each script: 20261106 for the primary analyses (Tables 2–4), 606 for the four-backbone comparison (Table 6), 707 for the seed-42 fine-tuned run, 909 plus the training seed for the additional fine-tuned runs, 11 for Figure 4 and 20261118 for the post hoc analyses. L-BFGS iteration counts for all fits (Section 4.7) are recorded in the supplementary results.

3.7. Threshold stability and transfer

Four analyses were prespecified, none of which used test labels for selection. (i) Validation resampling: the 145 validation lesions were bootstrap-resampled (2,000 replicates), stratified by the binary lesion label so that every resample contains 23 positive and 122 negative lesions with all their images; the threshold rule was reapplied to each resample and each re-selected threshold was evaluated on the fixed test scores. (ii) Out-of-fold thresholds: lesion-grouped stratified five-fold cross-validation within the training partition gave out-of-fold (OOF) scores for all 10,015 images; the same rule was applied to these scores, the resulting threshold was used with the model trained on all training data, and its stability was assessed by resampling the 7,470 training lesions (1,000 replicates, stratified in the same way). (iii) Refit bootstrap: training lesions were resampled, the classifier was refitted and the threshold re-selected on the official validation split (200 replicates), capturing variability from both model fitting and threshold selection. (iv) Alternative floors: the rule was repeated with floors of 70% and 90% on both the official validation split and the OOF scores; these are a prespecified sensitivity analysis and do not replace the primary threshold. The OOF threshold at the 80% floor gives each model an operating point aimed at the same development specificity, which we use for a matched comparison.

3.8. Case-mix standardisation of specificity

Specificity is a weighted mean of the specificities within the four negative diagnoses, weighted by their shares among negatives. To test whether the gap between intended and achieved specificity reflected a different negative case mix, we recomputed the test specificity as $\mathrm{Spec}_{\mathrm{std}} = \sum_d w_d \,\mathrm{Spec}_d$, where Spec_d is the test specificity within diagnosis $d \in \{\mathrm{NV}, \mathrm{BKL}, \mathrm{DF}, \mathrm{VASC}\}$ and w_d is the share of d among training negatives [13]. A lesion-cluster bootstrap interval and leave-one-diagnosis-out weights were added post hoc to show how much the estimate depends on the small DF and VASC groups.

3.9. Development-data sensitivity analyses and numerical checks

Using the training partition only, we compared lesion-grouped and image-level stratified five-fold cross-validation (three repetitions each) for $\lambda \in \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$. For each model the λ with the highest mean grouped-CV AUC was taken as the tuned value, and the tuned models were evaluated on the test set with validation and OOF thresholds as a secondary analysis. For numerical auditability we recorded the number of L-BFGS iterations and function evaluations, the final objective and the gradient norm, and refitted each primary classifier in float64 with up to 2,000 iterations and tolerances of 10^{-10} (gradient) and 10^{-14} (change) to check that predictions did not depend on the optimisation budget. All re-analyses used the cached features and the original fitting algorithm on CPU.

3.10. Calibration and clinical utility (post hoc)

Calibration of the test scores was summarised by the Brier score $\frac{1}{n}\sum_i (s_i - y_i)^2$, the expected calibration error over ten equal-width bins, $\mathrm{ECE} = \sum_k \frac{n_k}{n} |\bar{y}_k - \bar{s}_k|$, the calibration slope β from the logistic model $\mathrm{logit}\,\Pr(y = 1) = \alpha + \beta\,\mathrm{logit}(s)$, and the calibration-in-the-large intercept (the same model with β fixed at 1) [39]. A slope below 1 indicates scores that are too extreme. For the frozen models we also fitted a Platt recalibration map (α, β) on the lesion-grouped OOF training scores and applied it unchanged to the test scores. Because the threshold rule is based on ranks within the selection set, a monotone recalibration does not change which images are flagged at a selected threshold; it changes only how the score, and thresholds such as 0.5, can be interpreted.

Clinical utility was assessed with decision curves [40]. For a threshold probability p_t, at which a clinician would be indifferent between referring and not referring, the net benefit is

$$\mathrm{NB}(p_t) = \frac{TP(p_t)}{n} - \frac{FP(p_t)}{n} \cdot \frac{p_t}{1 - p_t}, \tag{6}$$

so that $p_t/(1 - p_t)$ is the number of unnecessary referrals one is willing to accept per correctly referred positive lesion divided by one; for example $p_t = 0.10$ accepts up to nine unnecessary referrals per positive lesion found. We report NB for p_t from 0.05 to 0.50 and paired lesion-bootstrap differences at $p_t \in \{0.10, 0.15, 0.20, 0.30\}$, using both the raw and the OOF-recalibrated scores. As a direct cost-sensitive alternative to the specificity floor, for false-negative to false-positive cost ratios $r \in \{3, 5, 10, 20\}$ we chose on the OOF training scores the threshold minimising $r \cdot FN(\tau) + FP(\tau)$ and applied it once to the test set; under calibration, ratio r corresponds to $p_t = 1/(1 + r)$.

3.11. Off-the-shelf linear probe (post hoc)

To check that conclusions did not depend on our classifier implementation, we fitted scikit-learn's `LogisticRegression` with default settings (L2 penalty, $C = 1$, L-BFGS, up to 5,000 iterations) to the same standardised features. On the mean-loss scale of Eq. (1) this corresponds to $\lambda = 1/(2Cn) \approx 5 \times 10^{-5}$, a weaker penalty than the primary $\lambda = 10^{-3}$. Its test AUC, validation-selected operating point and paired differences were estimated with the same lesion-cluster bootstrap.

3.12. Case-level failure analysis (post hoc)

At the primary thresholds, false-positive rates among negative images were tabulated by diagnosis and by the archived confirmation type (histopathology, single-image expert consensus, serial imaging showing no change, or

confocal microscopy), approximate age band and anatomical site; detection among positive images was tabulated by age band and, for melanoma, by site. The confirmation type is informative because benign lesions are usually excised and sent for histopathology only when they were clinically suspicious, whereas lesions followed by serial imaging were considered stable enough to monitor. We also inspected the eight highest-scoring false positives among BKL images, four each among DF and VASC, the eight lowest-scoring missed melanomas, and correctly classified contrasts (Figure 6). This inspection is descriptive and was not performed by a dermatologist.

3.13. Scope

The experiments use dermoscopic images only. The fine-tuned Swin-T runs serve only as a reference for the cost of freezing the backbone (Section 3.15); external validation, explainability, patient metadata and prospective clinical use are outside the scope of this study.

3.14. Software and hardware

Feature extraction, the primary fits and fine-tuning used Python 3.12.14, PyTorch 2.7.1+cu128 and TorchVision 0.22.1+cu128 on an NVIDIA GeForce RTX 5070 Ti (16 GB). ConvNeXt-T and ResNet-50 features were extracted in the same environment. The re-analyses used Python 3.13.18, PyTorch 2.15.0 (CPU), NumPy 2.5.3, SciPy 1.18.1, scikit-learn 1.9.1 (`StratifiedGroupKFold`, `LogisticRegression`), pandas 3.0.5 and Matplotlib 3.11.2. Version numbers were read from the run logs and the analysis environment.

3.15. Fine-tuned reference

To quantify the cost of freezing, Swin-T was fine-tuned end to end with a fixed protocol and no hyperparameter search: initialisation with `Swin_T_Weights.IMAGENET1K_V1`, a new single-output head, random horizontal and vertical flips and 90° rotations after the preset transform, AdamW (backbone learning rate 10^{-4}, head 10^{-3}, weight decay 0.05) with a cosine schedule, 12 epochs, batch size 32, unweighted binary cross-entropy and float16 autocasting. The official validation split was used to choose the epoch with the highest validation AUC and then to select the threshold with the same rule as the frozen pipelines; test predictions were generated once at the chosen epoch. Validation performance is optimistic because the epoch and threshold were chosen on the same 193 images, but the test evaluation is not affected. The first run used seed 42; the identical protocol was then repeated with seeds 1–4 (five runs in total) to show variability across training runs. Each run took about 4 minutes on the same GPU.

4. Results

4.1. Data partitions and lesion independence

Every image in the three partitions had an archived lesion identifier (Table 1). No lesion occurred in more than one partition, and diagnoses were consistent within every lesion. Clustering within partitions was substantial: 4,501 training images (44.9%) and 515 test images (34.1%) belonged to lesions with two or more images, and the 307 positive test images came from 213 lesions. Every positive image in every partition was confirmed by histopathology; negatives were confirmed by histopathology, serial imaging, expert consensus or confocal microscopy.

In the supplementary feature-space screen, no validation or test image had a training image with cosine similarity of 0.99 or more in any feature space; the highest test-to-training similarities were 0.958 (Swin-T) and 0.954 (MobileNetV2), below the highest similarity between nearest neighbours from different training lesions (0.978 and 0.962). The screen has limited sensitivity: images of the same training lesion had a median similarity of 0.832 (Swin-T), and 9,599 of 10,015 training images had a nearest neighbour from a different lesion that was more similar than a quarter of same-lesion pairs. It therefore rules out near-identical copies but cannot by itself establish independence; the lesion identifiers are the main evidence.

Table 1. Official ISIC 2018 Task 3 partitions under the study-defined binary target, with lesion structure from the ISIC archive.

Partition	Images	Positive images	Lesions	Positive lesions	Lesions ≥2 images	Images in multi-image lesions
Training (HAM10000)	10,015	1,954 (19.5%)	7,470	1,169	1,956	4,501
Validation	193	44 (22.8%)	145	23	39	87
Test	1,512	307 (20.3%)	1,223	213	226	515

Lesions shared between partitions: 0 for every pair. Test positives: 171 MEL, 93 BCC, 43 AKIEC; test negatives: 909 NV, 217 BKL, 44 DF, 35 VASC. Patient identifiers are not available, so lesion-disjoint partitions are not necessarily patient-disjoint.

4.2. Test performance with uncertainty

Table 2 gives ranking and threshold-dependent metrics with lesion-cluster bootstrap intervals, and Figures 2 and 3 show the curves. Swin-T reached a test ROC AUC of 0.893 (95% CI 0.870–0.914) and an AP of 0.676 (0.618–0.733), against 0.834 (0.804–0.859) and 0.551 (0.495–0.610) for MobileNetV2. At its validation-selected threshold, Swin-T detected 276 of 307 positive images (sensitivity 0.899, 0.859–0.936) with specificity 0.738 (0.710–0.765) and precision 0.466 (0.434–0.501); it flagged 592 of 1,512 images, 316 of them negative. At the threshold of 0.5, sensitivity was 0.638 (0.580–0.698) and specificity 0.905 (0.887–0.923).

Table 2. Test-set performance (1,512 images; 307 positive) with 95% lesion-cluster bootstrap intervals ($B = 2{,}000$; models and thresholds fixed).

Metric	Swin-T	MobileNetV2	Difference (Swin-T − MobileNetV2)
ROC AUC	0.893 (0.870–0.914)	0.834 (0.804–0.859)	0.059 (0.037–0.084)
Average precision	0.676 (0.618–0.733)	0.551 (0.495–0.610)	0.125 (0.064–0.183)
Sensitivity on test ROC at 80% specificity	0.840 (0.793–0.894)	0.717 (0.640–0.773)	0.124 (0.068–0.204)
Validation-selected threshold (Swin-T 0.1348; MobileNetV2 0.1931)			
Sensitivity	0.899 (0.859–0.936)	0.746 (0.691–0.798)	0.153 (0.099–0.211)
Specificity	0.738 (0.710–0.765)	0.778 (0.752–0.803)	−0.041 (−0.068 to −0.014)
Precision	0.466 (0.434–0.501)	0.462 (0.424–0.498)	–
Balanced accuracy	0.818 (0.795–0.841)	0.762 (0.732–0.790)	0.056 (0.027–0.090)
TP / FP / TN / FN	276 / 316 / 889 / 31	229 / 267 / 938 / 78	–
Threshold 0.5			
Sensitivity	0.638 (0.580–0.698)	0.528 (0.466–0.587)	0.111 (0.046–0.180)
Specificity	0.905 (0.887–0.923)	0.890 (0.872–0.909)	0.015 (−0.007 to 0.036)
Accuracy	0.851 (0.832–0.870)	0.817 (0.797–0.835)	–
TP / FP / TN / FN	196 / 114 / 1,091 / 111	162 / 132 / 1,073 / 145	–

Positive class: MEL, BCC, AKIEC. Accuracy at the selected thresholds was 0.771 (Swin-T) and 0.772 (MobileNetV2), below the 0.797 of an always-negative rule; F1 was 0.614 and 0.570 and the G-score 0.814 and 0.762. The ROC sensitivity at 80% specificity is a curve summary, not a selectable operating point.

4.3. Paired comparison

The paired differences in AUC (0.059, 0.037–0.084), AP (0.125, 0.064–0.183) and ROC sensitivity at 80% specificity (0.124, 0.068–0.204) all favoured Swin-T (bootstrap $p < 0.001$ for each). DeLong's image-level test gave the same AUC difference (95% CI 0.036–0.083; $p < 0.001$). At the model-specific selected thresholds, Swin-T had higher sensitivity (difference 0.153, 0.099–0.211) but lower specificity (-0.041, -0.068 to -0.014), and the two pipelines had similar accuracy (McNemar: 172 images correct only for Swin-T and 174 only for MobileNetV2; $p = 0.96$). Among positives, 61 images were detected only by Swin-T and 14 only by MobileNetV2 ($p < 0.001$); among negatives, 111 were correctly cleared only by Swin-T and 160 only by MobileNetV2 ($p = 0.003$). Because the selected thresholds place the two models at different operating points, these threshold-specific differences should be read together; the matched comparison in Section 4.5 addresses this directly.

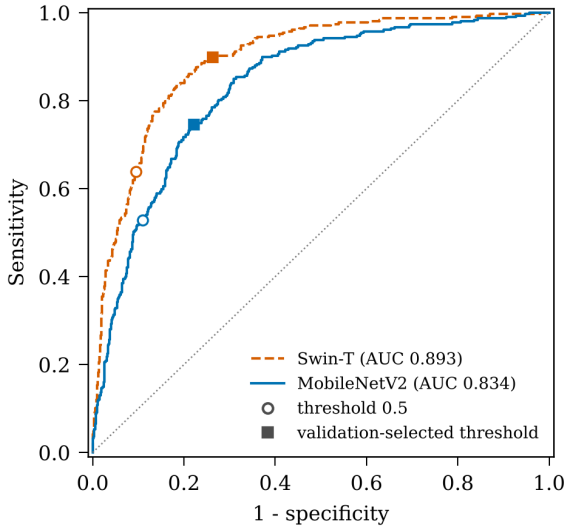

Figure 2. Test ROC curves for frozen Swin-T (dashed) and MobileNetV2 (solid). Open circles: threshold 0.5; filled squares: validation-selected thresholds (0.1348 and 0.1931). Dotted diagonal: chance.

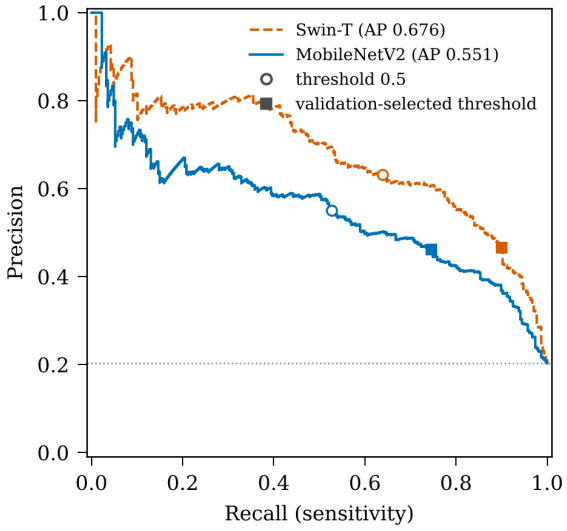

Figure 3. Test precision–recall curves with the same operating points as Figure 2. Dotted line: positive proportion (0.203).

4.4. Errors by diagnosis

Table 3 shows detection rates within positive diagnoses and false-positive rates within negative diagnoses at the selected thresholds. Swin-T detected 88.3% of melanoma images (0.821–0.939) and MobileNetV2 69.6% (0.617–0.770). False positives came mostly from outside the dominant nevus category: Swin-T flagged 15.8% of NV images but 59.9% of BKL, 47.7% of DF and 60.0% of VASC images, so BKL alone produced 130 of the 316 false positives. MobileNetV2 showed the same pattern with lower rates for BKL and VASC. Aggregate specificity therefore hides diagnosis-specific error concentrations, especially for benign keratoses.

Table 3. Detection and false-positive rates by diagnosis at the validation-selected thresholds (95% lesion-cluster bootstrap intervals).

Diagnosis (n)	Swin-T flagged	Swin-T rate (95% CI)	MobileNetV2 flagged	MobileNetV2 rate (95% CI)
Positive diagnoses – detection				
MEL (171)	151	0.883 (0.821–0.939)	119	0.696 (0.617–0.770)
BCC (93)	85	0.914 (0.854–0.967)	72	0.774 (0.677–0.861)
AKIEC (43)	40	0.930 (0.826–1.000)	38	0.884 (0.788–0.971)
Negative diagnoses – false positives				
NV (909)	144	0.158 (0.134–0.184)	137	0.151 (0.127–0.177)
BKL (217)	130	0.599 (0.525–0.672)	102	0.470 (0.401–0.541)
DF (44)	21	0.477 (0.302–0.630)	17	0.386 (0.219–0.538)
VASC (35)	21	0.600 (0.417–0.760)	11	0.314 (0.154–0.487)

Intervals for small categories (AKIEC, DF, VASC) are wide. Flagged counts sum to the TP and FP totals in Table 2.

4.5. Threshold stability and transfer

On the validation split used for selection, Swin-T detected all 44 positives with specificity 0.812, and MobileNetV2 had sensitivity 0.818 and specificity 0.826; both figures are optimistic by construction. Thresholds re-selected on lesion-bootstrap resamples of the validation split varied widely (Table 4; Figure 4): for Swin-T the 95% range of

thresholds was 0.135–0.400, corresponding to test sensitivities of 0.700–0.899 and specificities of 0.737–0.886, and only 44.6% of resamples reached a test specificity of at least 80% (10.0% for MobileNetV2). The refit bootstrap, which also varies the fitted classifier, gave a similar picture (Swin-T thresholds 0.101–0.253; test sensitivity 0.772–0.893; specificity 0.709–0.824).

Thresholds selected on the 10,015 lesion-grouped OOF training scores were far more stable (Swin-T 0.173, 95% resampling range 0.161–0.186; MobileNetV2 0.188, 0.169–0.209). Both models reached 80.0% OOF specificity at these thresholds by construction; on the test set Swin-T had sensitivity 0.866 and specificity 0.778, and MobileNetV2 0.746 and 0.773. At this matched development target, the paired sensitivity difference was 0.121 (0.067–0.177) and the specificity difference 0.005 (−0.022 to 0.032).

Table 4. Threshold transfer: thresholds selected on the official validation split or on lesion-grouped OOF training scores, and the resulting test operating points ($\lambda = 0.001$).

Selection data and floor	Swin-T threshold	Swin-T test sens / spec	MobileNetV2 threshold	MobileNetV2 test sens / spec
Official validation, 80% (primary)	0.135	0.899 / 0.738	0.193	0.746 / 0.778
validation bootstrap, 95% range	0.135–0.400	0.700–0.899 / 0.737–0.886	0.095–0.289	0.664–0.840 / 0.695–0.828
refit bootstrap, 95% range	0.101–0.253	0.772–0.893 / 0.709–0.824	0.092–0.268	0.619–0.779 / 0.723–0.823
Official validation, 70%	0.135	0.899 / 0.737	0.095	0.840 / 0.695
Official validation, 90%	0.435	0.671 / 0.894	0.436	0.567 / 0.872
Grouped OOF, 70%	0.082	0.932 / 0.652	0.069	0.860 / 0.661
Grouped OOF, 80%	0.173	0.866 / 0.778	0.188	0.746 / 0.773
OOF bootstrap, 95% range	0.161–0.186	0.860–0.876 / 0.771–0.785	0.169–0.209	0.733–0.756 / 0.760–0.789
Grouped OOF, 90%	0.373	0.739 / 0.877	0.489	0.537 / 0.888

Floors other than 80% are a prespecified sensitivity analysis. For Swin-T the 70% and 80% validation floors select the same threshold because validation recall is already 1.0. Rows other than the primary row come from the CPU refit of the original code (Section 4.7); at the primary thresholds the refit reproduces the original confusion counts for MobileNetV2 and differs by one false positive for Swin-T.

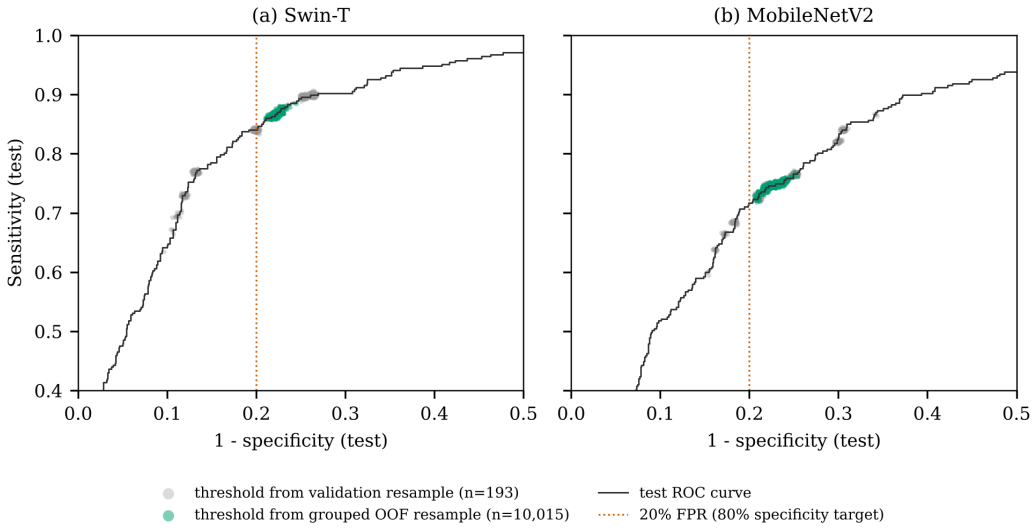

Figure 4. Test operating points obtained when the threshold rule (maximum recall at ≥80% specificity) is reapplied to 1,000 lesion-bootstrap resamples of the 193-image validation split (grey) or of the 10,015 lesion-grouped out-of-fold training scores (green). The fitted models and the test set are fixed. The dotted line marks a 20% false-positive rate.

The stable OOF thresholds gave a test specificity about 2 percentage points below the 80% development target. The negative case mix differed between partitions: NV made up 83.2% of training negatives but only 75.4% of test negatives, whereas BKL, DF and VASC made up 13.6%, 1.4% and 1.8% of training negatives against 18.0%, 3.7% and 2.9% of test negatives. Re-weighting the diagnosis-specific test specificities to the training mix raised the Swin-T test specificity from 0.778 to 0.806 (95% CI 0.782–0.828) at the OOF threshold (MobileNetV2: 0.773 to

0.795, 0.772–0.818), and from 0.738 to 0.769 (0.743–0.794) at the primary validation threshold. The adjustment is driven mainly by BKL: leaving DF or VASC out of the re-weighting changed the standardised Swin-T specificity at the OOF threshold by at most 0.006 (0.808 and 0.812), whereas leaving BKL out raised it to 0.861. The shortfall at the OOF threshold is therefore mostly explained by a harder negative case mix in the test set, whereas the much larger variability of validation-selected thresholds reflects the small validation sample.

4.6. Grouped versus image-level development evaluation and regularisation

Within the training partition, image-level cross-validation gave higher AUC and AP than lesion-grouped cross-validation at every λ (Table 5); at $\lambda = 0.001$ the inflation was 0.016 in AUC and 0.039 in AP for Swin-T, and 0.018 and 0.043 for MobileNetV2 (repetition SD $\leq$0.004 in AUC). Grouped-CV AUC peaked at $\lambda = 0.01$ for Swin-T (0.894) and at $\lambda = 0.1$ for MobileNetV2 (0.870), so the common $\lambda = 0.001$ was closer to optimal for Swin-T than for the higher-dimensional MobileNetV2 features.

Table 5. Lesion-grouped and image-level five-fold cross-validation within the training partition (mean of three repetitions) and single test evaluation by λ.

	Swin-T			MobileNetV2		
λ	Grouped CV AUC / AP	Image CV AUC / AP	Test AUC	Grouped CV AUC / AP	Image CV AUC / AP	Test AUC
10^{-5}	0.868 / 0.623	0.887 / 0.667	0.880	0.812 / 0.484	0.831 / 0.520	0.823
10^{-4}	0.872 / 0.631	0.891 / 0.674	0.883	0.818 / 0.496	0.837 / 0.534	0.824
10^{-3} (primary)	0.885 / 0.656	0.901 / 0.695	0.893	0.838 / 0.533	0.856 / 0.576	0.834
10^{-2}	0.894 / 0.672	0.904 / 0.698	0.897	0.863 / 0.589	0.878 / 0.628	0.860
10^{-1}	0.885 / 0.639	0.889 / 0.651	0.883	0.870 / 0.598	0.878 / 0.618	0.878

λ was chosen by grouped-CV AUC only; test AUCs document sensitivity and were not used for selection. At $\lambda \geq 10^{-3}$ all fits converged before the 150-iteration cap; at $\lambda \leq 10^{-4}$ some fits reached it.

At each model's grouped-CV λ (secondary analysis), the test AUC was 0.897 for Swin-T and 0.878 for MobileNetV2: a paired AUC difference of 0.019 (0.003–0.035), an AP difference of 0.040 ($-$0.012 to 0.091) and a ROC sensitivity difference at 80% specificity of 0.072 (0.017–0.125). At OOF thresholds aimed at 80% development specificity, the tuned Swin-T detected 0.883 of positives at 0.778 specificity and the tuned MobileNetV2 0.775 at 0.798 (sensitivity difference 0.107, 0.065–0.153; specificity difference $-$0.019, $-$0.042 to 0.003). The direction of the comparison was unchanged, but its size in ranking metrics depended materially on classifier regularisation.

4.7. Reproducibility and convergence

Re-running the original code on CPU with the cached features reproduced the saved test scores to within 0.0019 (Swin-T) and 0.0006 (MobileNetV2), and the reported AUC and AP to three decimals; the main analyses use the original saved scores. With the original budget, L-BFGS stopped on its tolerance criteria after 81 (Swin-T) and 79 (MobileNetV2) iterations, below the 150-iteration cap, with final objectives of 0.24458 and 0.22325 and maximum gradient norms of 6.8×10^{-6} and 7.9×10^{-6}. The float64 extended-budget refit lowered the objectives by less than 2×10^{-7}, changed no test score by more than 0.004, and changed the classification of one Swin-T test image and no MobileNetV2 test image at the original thresholds.

4.8. Additional convolutional comparators

Table 6 places Swin-T alongside the two additional frozen backbones. ConvNeXt-T performed like Swin-T, with a test AUC of 0.897 (0.874–0.917) and an AP of 0.692; there was no evidence of a difference between Swin-T and ConvNeXt-T: the differences were $-$0.004 ($-$0.022 to 0.013) in AUC, $-$0.016 ($-$0.066 to 0.034) in AP, and, at the matched OOF threshold, 0.020 ($-$0.030 to 0.067) in sensitivity and 0.006 ($-$0.019 to 0.030) in specificity. With tuned regularisation the AUC difference was $-$0.004 ($-$0.019 to 0.011). Because no equivalence margin was prespecified, these results do not establish equivalence; they show that differences larger than about 0.02 AUC in either direction are not supported by the data. ResNet-50 behaved like MobileNetV2, with an AUC of 0.818

(0.787–0.846) at $\lambda = 0.001$ that rose to 0.869 with a grouped-CV λ of 0.1; Swin-T exceeded it by 0.075 (0.048–0.102) in AUC at $\lambda = 0.001$ and by 0.028 (0.011–0.046) after tuning, and by 0.150 (0.094–0.209) in sensitivity at the matched OOF threshold. As with MobileNetV2, grouped CV selected the strongest penalty for the high-dimensional ResNet-50 features and $\lambda = 0.01$ for the two 768-dimensional feature sets.

Table 6. Four frozen backbones under the same protocol, with the fine-tuned Swin-T reference (test set; 95% lesion-cluster bootstrap intervals; $\lambda = 0.001$ unless stated).

Backbone (dim.)	Test AUC (95% CI)	AP	Val. threshold: sens / spec	OOF threshold (80%): sens / spec	Grouped-CV λ; tuned AUC	Swin-T − backbone, AUC (95% CI)
Swin-T (768)	0.893 (0.871–0.914)	0.676	0.899 / 0.738	0.866 / 0.778	10^{-2}; 0.897	–
ConvNeXt-T (768)	0.897 (0.874–0.917)	0.692	0.879 / 0.747	0.847 / 0.773	10^{-2}; 0.901	−0.004 (−0.022 to 0.013)
MobileNetV2 (1,280)	0.834 (0.805–0.860)	0.551	0.746 / 0.778	0.746 / 0.773	10^{-1}; 0.878	0.059 (0.037 to 0.084)
ResNet-50 (2,048)	0.818 (0.787–0.846)	0.501	0.717 / 0.797	0.717 / 0.797	10^{-1}; 0.869	0.075 (0.048 to 0.102)
Swin-T fine-tuned, seed 42	0.945 (0.925–0.962)	0.842	0.850 / 0.886	n/a	n/a	−0.052 (−0.072 to −0.032)
Swin-T fine-tuned, 5 seeds	0.940 ± 0.011	0.829	0.902 / 0.799	n/a	n/a	−0.047 (range 0.030–0.057)

Intervals for the frozen backbones come from one paired bootstrap of all four models and can differ from Table 2 in the third decimal. Selected validation thresholds: ConvNeXt-T 0.1174, ResNet-50 0.1450; OOF thresholds 0.1481 and 0.1461. The five-seed row gives the mean ± SD across runs (sens / spec: mean at each run's validation-selected threshold); per-run results are in Table 7. OOF thresholds and λ do not apply to the fine-tuned model.

4.9. Fine-tuned Swin-T reference

The seed-42 fine-tuned run reached its highest validation AUC (0.971) at epoch 11. Its test AUC was 0.945 (0.925–0.962) and AP 0.842 (0.796–0.883), exceeding frozen Swin-T by 0.052 (0.032–0.072) and 0.166 (0.114–0.216), with a ROC sensitivity at 80% specificity of 0.922 (0.879–0.962), 0.081 (0.032–0.128) higher than the frozen model. At its validation-selected threshold (0.0334) it detected 261 of 307 positives (sensitivity 0.850, 0.800–0.898) with specificity 0.886 (0.865–0.907) and precision 0.656, flagging 137 negatives (79 NV, 49 BKL, 7 DF, 2 VASC) instead of 316. Compared with frozen Swin-T at its own selected threshold, specificity rose by 0.149 (0.119–0.175) and sensitivity fell by 0.049 (−0.099 to 0.003). Threshold transfer remained fragile: thresholds re-selected on lesion-bootstrap resamples of the validation split ranged from below 0.0001 to 0.231 (95% range), giving test sensitivity of 0.785–0.971 and specificity of 0.645–0.930, because the fine-tuned scores are concentrated near 0 and 1 and the 44 validation positives determine where the threshold falls.

Across the five training seeds (Table 7), test AUC was 0.940 ± 0.011 (range 0.923–0.950) and the gain over frozen Swin-T was 0.047 (range 0.030–0.057), so the cost of freezing is consistent across runs. The operating point varied more than the ranking: at each run's validation-selected threshold, test sensitivity ranged from 0.850 to 0.958 and specificity from 0.724 to 0.886. The epoch selected on the 193 validation images ranged from 1 to 11 across seeds and the selected thresholds from 0.0008 to 0.220, so the same model-selection fragility that affects the frozen thresholds also affects epoch selection.

Table 7. Fine-tuned Swin-T reference repeated with five training seeds (identical protocol; epoch and threshold selected on the official validation split; test evaluated once per run).

Seed	Epoch	Val. AUC	Threshold	Test AUC (95% CI)	AP	Test sens / spec	AUC gain vs frozen	Val.-bootstrap sens / spec range
42	11	0.971	0.0334	0.945 (0.926–0.962)	0.842	0.850 / 0.886	0.052	0.77–0.97 / 0.64–0.94
1	8	0.969	0.1327	0.950 (0.935–0.964)	0.854	0.889 / 0.860	0.057	0.74–0.89 / 0.86–0.96
2	7	0.970	0.0008	0.946 (0.931–0.960)	0.840	0.958 / 0.724	0.053	0.68–0.96 / 0.70–0.96
3	1	0.969	0.2200	0.923 (0.903–0.942)	0.784	0.902 / 0.766	0.030	0.80–0.95 / 0.64–0.88
4	2	0.973	0.1213	0.935 (0.917–0.951)	0.824	0.909 / 0.758	0.042	0.79–0.91 / 0.76–0.91
Mean ± SD				0.940 ± 0.011	0.829 ± 0.028	0.902 / 0.799	0.047 ± 0.011	

Test AUC intervals: 95% lesion-cluster bootstrap. Validation-bootstrap ranges: 95% range of test sensitivity and specificity when the threshold is re-selected on 2,000 lesion-bootstrap resamples of the validation split.

4.10. Calibration and clinical utility

Calibration differed between backbones more than discrimination did (Table 8; Figure 5a). Frozen Swin-T and ConvNeXt-T were close to calibrated in the large (mean score 0.227 and 0.214 against a test prevalence of 0.203)

but their scores were somewhat too extreme (calibration slopes 0.73 and 0.68; ECE 0.036 and 0.042). The higher-dimensional MobileNetV2 and ResNet-50 features gave much more extreme scores (slopes 0.41 and 0.27; ECE 0.084 and 0.126), consistent with the weak penalty relative to their dimension noted in Section 3.9. The fine-tuned model had the lowest Brier score (0.079) but the most extreme scores (slope 0.28), as expected from scores concentrated near 0 and 1. Platt recalibration fitted on OOF training scores brought the slopes of all frozen models to 0.97–1.09 and their ECE to about 0.03. Because recalibration is monotone, it does not change the images flagged at a validation- or OOF-selected threshold; it does change the meaning of the default threshold of 0.5, at which the uncalibrated MobileNetV2 and ResNet-50 scores are not comparable with those of Swin-T.

Table 8. Calibration on the test set and net benefit at selected threshold probabilities (95% lesion-cluster bootstrap intervals).

Model	Brier (95% CI)	ECE	Slope	Intercept	NB at $p_t = 0.10$	NB at $p_t = 0.20$	NB at $p_t = 0.30$
Swin-T	0.103 (0.091–0.114)	0.036	0.73	−0.29	0.156	0.131	0.107
ConvNeXt-T	0.100 (0.089–0.112)	0.042	0.68	−0.15	0.159	0.128	0.103
MobileNetV2	0.135 (0.122–0.149)	0.084	0.41	−0.20	0.142	0.107	0.075
ResNet-50	0.154 (0.137–0.171)	0.126	0.27	−0.35	0.131	0.100	0.070
Swin-T fine-tuned (seed 42)	0.079 (0.064–0.093)	0.068	0.28	0.39	0.160	0.147	0.136
Treat all	–	–	–	–	0.115	0.004	−0.139
Paired NB difference, Swin-T minus comparator (raw scores)							
vs ConvNeXt-T					−0.003 (−0.011 to 0.006)	0.004 (−0.007 to 0.016)	0.004 (−0.010 to 0.017)
vs MobileNetV2					0.014 (0.005 to 0.024)	0.025 (0.013 to 0.037)	0.032 (0.017 to 0.047)
vs ResNet-50					0.025 (0.014 to 0.036)	0.032 (0.019 to 0.045)	0.037 (0.022 to 0.051)
vs fine-tuned					−0.003 (−0.015 to 0.008)	−0.015 (−0.028 to −0.003)	−0.029 (−0.043 to −0.015)

ECE: ten equal-width bins. Slope and intercept from logistic recalibration of the test scores (intercept with slope fixed at 1). NB: net benefit per test image; p_t is the threshold probability. After OOF Platt recalibration, the Swin-T minus MobileNetV2 differences were 0.011 (0.005 to 0.017), 0.023 (0.012 to 0.034) and 0.032 (0.017 to 0.047), and the Swin-T minus ConvNeXt-T differences remained compatible with zero.

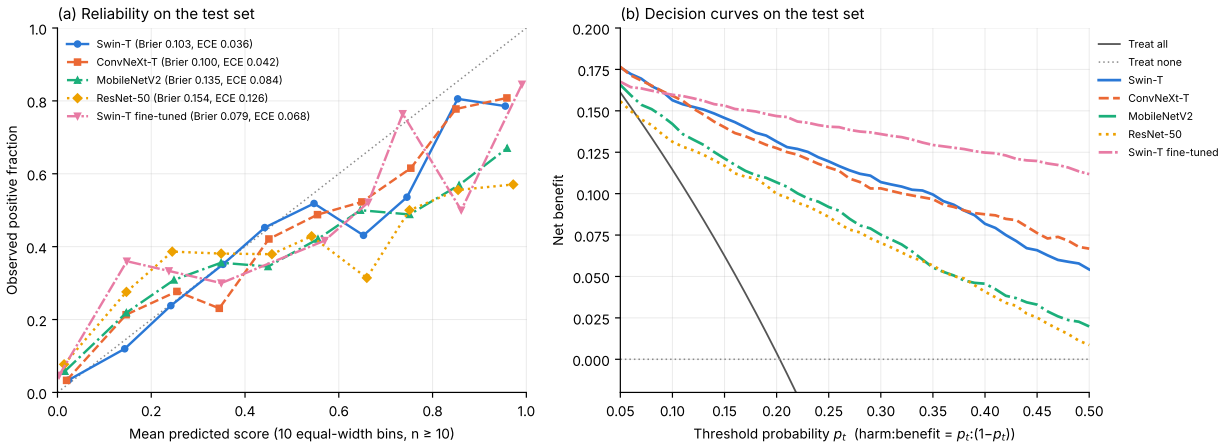

Figure 5. (a) Reliability diagram on the test set (ten equal-width bins with at least 10 images). (b) Decision curves: net benefit of referring images whose score is at least the threshold probability p_t, compared with referring all images (treat all) or none.

In decision-curve analysis (Figure 5b), all models had higher net benefit than referring every lesion across p_t of 0.05–0.50. Swin-T had higher net benefit than MobileNetV2 and ResNet-50 at p_t of 0.10–0.30, by 0.014–0.037 (equivalent to 1.4–3.7 additional correctly referred positive lesions per 100 images at the same harm weighting), and the differences persisted after recalibration; there was no evidence of a difference from ConvNeXt-T. The fine-tuned model's advantage grew with p_t, being clear from $p_t = 0.20$, that is, in settings where unnecessary referrals are weighted more heavily.

Cost-sensitive thresholds chosen on the OOF scores (Table 9) make the clinical trade-off explicit. With a false-negative cost three times that of a false positive, the Swin-T threshold (0.197) referred 339 test images per 1,000 and missed 29.8 positive images per 1,000; at a ratio of 10, it referred 532 per 1,000 and missed 8.6. At every

cost ratio, Swin-T and ConvNeXt-T either missed fewer positives than MobileNetV2 and ResNet-50 at a similar referral burden or missed a similar number with fewer referrals. Test specificities were again up to 7 percentage points below the OOF values, the same case-mix effect as in Section 4.5.

Table 9. Cost-sensitive operating points: threshold minimising $r \cdot FN + FP$ on lesion-grouped OOF training scores, applied once to the test set.

Cost ratio r (implied p_t)	Model	Threshold	OOF sens / spec	Test sens / spec	Referred per 1,000	Missed per 1,000
3 (0.25)	Swin-T	0.197	0.778 / 0.819	0.853 / 0.792	339	29.8
	ConvNeXt-T	0.148	0.848 / 0.800	0.847 / 0.773	353	31.1
	MobileNetV2	0.178	0.735 / 0.795	0.749 / 0.764	340	50.9
	ResNet-50	0.152	0.703 / 0.804	0.713 / 0.797	307	58.2
5 (0.17)	Swin-T	0.081	0.896 / 0.699	0.932 / 0.651	468	13.9
	ConvNeXt-T	0.145	0.851 / 0.797	0.857 / 0.769	358	29.1
	MobileNetV2	0.069	0.829 / 0.700	0.860 / 0.661	445	28.4
	ResNet-50	0.035	0.824 / 0.698	0.850 / 0.665	440	30.4
10 (0.09)	Swin-T	0.049	0.934 / 0.624	0.958 / 0.576	532	8.6
	ConvNeXt-T	0.056	0.931 / 0.675	0.935 / 0.659	462	13.2
	MobileNetV2	0.013	0.919 / 0.542	0.938 / 0.495	593	12.6
	ResNet-50	0.006	0.908 / 0.561	0.919 / 0.500	585	16.5
20 (0.05)	Swin-T	0.016	0.976 / 0.463	0.977 / 0.406	672	4.6
	ConvNeXt-T	0.024	0.967 / 0.560	0.964 / 0.524	575	7.3
	MobileNetV2	0.003	0.962 / 0.385	0.967 / 0.328	732	6.6
	ResNet-50	<0.001	0.966 / 0.371	0.948 / 0.301	749	10.6

Implied $p_t = 1/(1 + r)$ under calibration. Referred and missed counts are per 1,000 test images (prevalence 20.3%).

4.11. Off-the-shelf linear probe

The scikit-learn probe converged in 165–199 iterations and gave test AUCs of 0.882 (0.857–0.906) for Swin-T, 0.886 (0.861–0.909) for ConvNeXt-T, 0.824 (0.796–0.851) for MobileNetV2 and 0.783 (0.751–0.817) for ResNet-50, 0.010–0.035 below the primary classifier (paired differences all excluding zero), as expected from its weaker penalty. The comparison between backbones was unchanged: Swin-T minus MobileNetV2 0.058 (0.033–0.085) and Swin-T minus ConvNeXt-T −0.004 (−0.026 to 0.017). At its validation-selected threshold the Swin-T probe had test sensitivity 0.873 and specificity 0.744. The conclusions therefore do not depend on the particular implementation of the linear head, although the absolute AUC does depend on its regularisation.

4.12. Case-level failure analysis

False positives were concentrated in lesions that had been clinically suspicious (Table 10). At the primary threshold, Swin-T flagged 45.1% of histopathology-confirmed negative images but 2.4% of nevi confirmed by serial imaging; histopathology-confirmed lesions accounted for 216 of the 316 false positives (68%). Within nevi, the false-positive rate was 33.4% for excised (histopathology-confirmed) nevi and 2.4% for monitored nevi (difference 0.276, 95% CI 0.218–0.336); within BKL it was 72.5% for histopathology-confirmed and 42.2% for consensus-confirmed lesions. MobileNetV2 and the fine-tuned model showed the same gradient at lower or similar levels. False-positive rates were also higher among negatives from patients aged 60 or more (42.5% vs 13.7% at 40–59 and 22.5% under 40) and from the head and neck (51.9%), the age and site profile of benign keratoses. Among positives, detection of melanoma was lowest on the lower extremity (20 of 28, 71%; trunk 55 of 59, head and neck 26 of 27); 8 of the 20 missed melanomas were on the lower extremity.

Visual inspection (Figure 6) was consistent with these patterns. The highest-scoring benign keratoses and vascular lesions were dark, multicoloured or irregular lesions that share dermoscopic features with melanoma or basal cell carcinoma, whereas correctly cleared benign keratoses were light, uniform lesions. Several of the lowest-scoring missed melanomas were small, low-contrast or out-of-focus images, and some showed parallel line patterns suggestive of acral skin, a site that is uncommon in the training data. These observations are descriptive

Table 10. False-positive rates among negative test images by diagnosis and confirmation type at the validation-selected thresholds.

Diagnosis	Confirmation	n	Frozen Swin-T	Frozen MobileNetV2	Fine-tuned Swin-T (seed 42)
NV	histopathology	329	0.334	0.325	0.207
NV	expert consensus	122	0.189	0.139	0.082
NV	serial imaging, no change	458	0.024	0.028	0.002
BKL	histopathology	120	0.725	0.525	0.308
BKL	expert consensus	90	0.422	0.400	0.133
BKL	confocal microscopy	7	0.714	0.429	0.000
DF	histopathology / consensus	16 / 28	0.500 / 0.464	0.375 / 0.393	0.375 / 0.036
VASC	histopathology / consensus	14 / 21	0.786 / 0.476	0.214 / 0.381	0.071 / 0.048
All negatives	histopathology	479	0.451	0.374	0.234
All negatives	other confirmation	726	0.138	0.121	0.034

All positive test images were histopathology-confirmed. Small DF, VASC and confocal groups give imprecise rates.

and were not verified by a dermatologist; they suggest that part of the error reflects genuinely ambiguous lesions—those clinicians themselves chose to excise—rather than failures on easy cases, and that acral and image-quality subgroups deserve targeted evaluation.

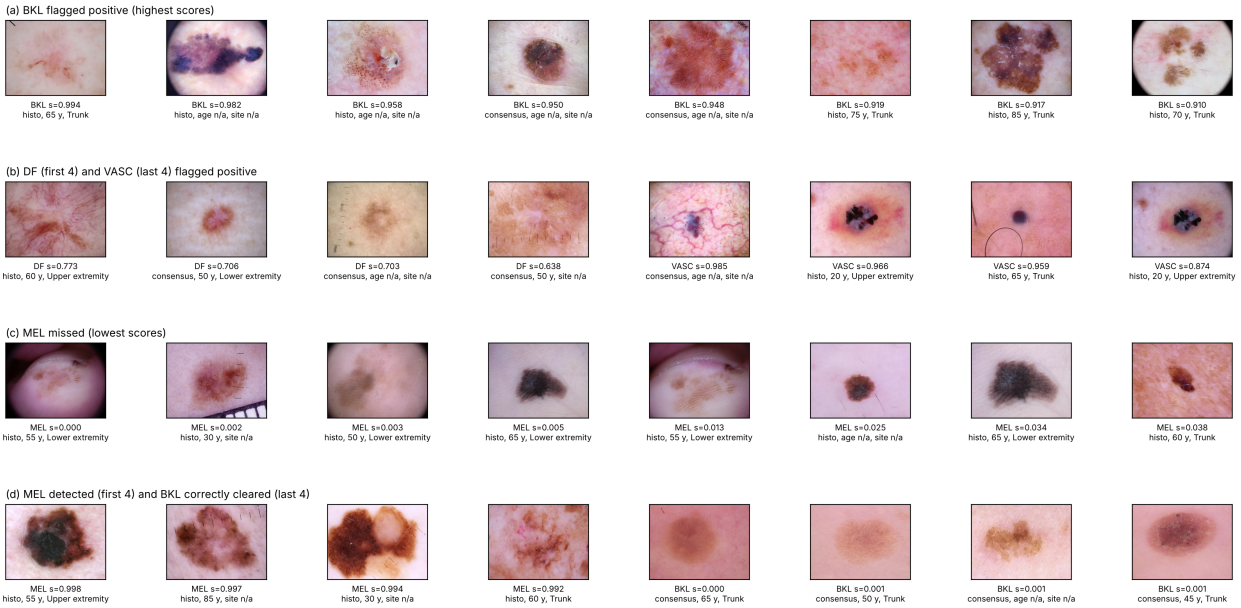

Figure 6. Examples of errors by frozen Swin-T at the validation-selected threshold (0.1348). (a) Highest-scoring false positives among BKL images; (b) highest-scoring false positives among DF (first four) and VASC (last four) images; (c) lowest-scoring missed melanomas; (d) correctly detected melanomas and correctly cleared BKL images for contrast. Labels: diagnosis, score s, confirmation type, approximate age and site. Images: ISIC 2018 Task 3 test set [18, 12], reproduced under the CC BY-NC 4.0 licence.

5. Comparison with published work

Because we could not confirm that any published study used the official ISIC 2018 Task 3 test partition with the MEL/BCC/AKIEC binary target and a validation-selected threshold, no published results are directly comparable.

Table 11 summarises the protocol differences for a set of relevant studies identified during the work; it is not a ranking or the result of a systematic search, and values are given as reported. The "lesion-disjoint" column records whether the evaluation split was reported to keep all images of a lesion in one partition; studies using random image-level splits may report inflated results because images of the same lesion can appear in both training and test data [11]. Aruk et al. report the highest Swin-based accuracies, 0.9212 on HAM10000 and 0.9187 on ISIC 2019 [9].

Table 11. Selected related studies and their evaluation protocols (values as published; not a ranking).

Study	Data / partition	Label setting	Lesion-disjoint test	Reported result
Nie et al. 2023 [20]	ISIC 2018 Task 3 training images; random 0.7/0.1/0.2 split	7-class	No (random image-level)	Accuracy 89.48%
Ramya & Jaisankar 2026 [21]	ISIC 2019; image-level stratified 70/15/15	Melanoma vs non-melanoma	No (image-level)	Accuracy 98.55%; sensitivity 95.73%; specificity 99.17%
Mahbod et al. 2025 [10] (preprint)	HAM10000; predefined 8,207 / 575 / 1,232 split	7-class	Not reported	Swin V2: accuracy 91.80%, balanced accuracy 81.43%
Pathirana et al. 2026 [22] (preprint)	HAM10000; group-aware split as reported	7-class	Yes (as reported)	Accuracy 0.9255; macro F1 0.9133 (with metadata)
Kitada & Iyatomi 2018 [33]	ISIC 2018 Task 3; official validation (193 images)	7-class	Yes (official partitions)	Balanced accuracy 87.2%
This study	ISIC 2018 Task 3; official validation and test	Binary (MEL, BCC, AKIEC positive)	Yes (verified)	Test AUC 0.893 (0.870–0.914); sensitivity 0.899 at specificity 0.738; accuracy 0.771 at the selected threshold and 0.851 at threshold 0.5

"Yes (official partitions)": the ISIC 2018 Task 3 partitions share no lesions (Table 1). Accuracy for this study is computed from the confusion counts in Table 2; at the selected threshold it is below the 0.797 of an always-negative rule, which illustrates why accuracy is not a suitable primary metric for this target.

6. Discussion

6.1. Principal findings

On lesion-disjoint official partitions, frozen Swin-T features ranked study-defined positive images better than frozen MobileNetV2 features, with a difference well outside sampling uncertainty under the prespecified protocol (AUC difference 0.059, 0.037–0.084). The most decision-relevant evidence is the matched development target: when both models were given a threshold aimed at 80% specificity on 10,015 OOF training scores, Swin-T detected about 12 percentage points more positive test images at almost identical test specificity, and decision curves showed a corresponding net-benefit gain of 1.4–3.2 correctly referred positives per 100 images over MobileNetV2 across plausible threshold probabilities. By contrast, comparison at the model-specific validation thresholds mixes a sensitivity gain with a specificity loss, which is why the two pipelines had similar accuracy. The additional comparators refine the picture: there was no evidence of a difference between frozen ConvNeXt-T, a convolutional network of similar size, and Swin-T on any summary, whereas ResNet-50 resembled MobileNetV2. The advantage of Swin-T is therefore an advantage over older or lighter convolutional backbones, consistent with differences in model generation and pretraining recipe [35], and not evidence that shifted-window attention itself gives better frozen features for this task; strictly, the comparison is between pipelines rather than representations. Freezing has a visible cost: fine-tuning Swin-T improved test AUC by about 0.05 (0.030–0.057 across five seeds) and, at its own selected threshold in the seed-42 run, more than halved false positives; across seeds, however, specificity at the validation-selected threshold ranged from 0.724 to 0.886, so this operating-point gain was not reproducible. The frozen pipelines should be read as cheap, reproducible baselines rather than as competing classifiers.

6.2. Clinical interpretation

For a referral-triage reading of the binary target, three results matter most. First, a reported sensitivity is only as good as the threshold behind it: with 44 validation positives, the same rule could have produced a test sensitivity anywhere between about 70% and 90%, so a single validation-selected operating point should not be quoted as the expected clinical performance. Second, the false-positive burden falls on a specific, recognisable group—benign keratoses and benign lesions that clinicians had themselves found suspicious enough to excise—so a deployment in a population with more such lesions (older patients, head and neck sites) will see lower specificity than the development data suggest, as the case-mix analysis showed. Third, the choice of operating point is a cost judgement: at a false-negative to false-positive cost ratio of 3 the frozen Swin-T pipeline would refer about a third of lesions and miss about 15% of positive images (30 per 1,000 images); at a ratio of 10 it would refer about half and miss about 4% (9 per 1,000). These numbers depend on the case mix and prevalence of this benchmark and are not a screening recommendation, but they show the form in which a clinically useful report should present such a classifier.

6.3. What limits the transfer of an operating point

Two distinct mechanisms affected threshold transfer. The first is sampling variability: with 44 positive images from 23 lesions, the validation split cannot locate a threshold precisely, and resampling moved the Swin-T test operating point along most of the ROC curve between 70% and 90% sensitivity (Figure 4). The original selected threshold lay at the low end of this distribution, giving high test sensitivity but specificity well below the target. The second is a systematic shift in case mix: test negatives contained more benign keratoses, dermatofibromas and vascular lesions, which both models flag far more often than nevi. A development threshold, even when precisely located, missed its specificity target on the test data, and standardising to the training case mix closed most of that gap; the adjustment rests mainly on the large BKL group and is not sensitive to the small DF and VASC groups. This has a practical consequence: enlarging the selection sample (here, with lesion-grouped OOF scores) addresses the first problem but not the second, which requires knowledge of the case mix in the target setting [13], and which is likely to be larger in multi-centre data [37, 38]. The fine-tuned reference shows that better ranking does not remove the first problem: its validation-selected threshold moved test specificity from about 65% to 93% across validation resamples, and its operating point varied between training seeds.

6.4. Calibration

Calibration and discrimination told different stories. MobileNetV2 and ResNet-50 had much more extreme scores than Swin-T and ConvNeXt-T, so at the default threshold of 0.5 their scores carry different meanings, and comparisons at 0.5 confound representation quality with calibration. The fine-tuned model had the best Brier score but the most extreme probabilities. Simple Platt recalibration on development data corrected the frozen models, and because the threshold rule is rank-based it did not alter the selected operating points; calibration matters for interpreting scores and for decision-analytic thresholds, not for the ranking results reported above [39]. Recalibration on the target population would be needed before scores are presented to clinicians as probabilities.

6.5. Dependence on evaluation and tuning choices

Lesion grouping mattered even for a linear classifier on frozen features: image-level cross-validation raised AUC by about 0.016–0.018 and AP by about 0.04, in line with the larger effects reported for fine-tuned seven-class models [11]. The official partitions are lesion-disjoint, so the reported test estimates are not affected by this leakage, but within-partition clustering still had to be respected in the bootstrap. The backbone comparison depended on regularisation: a fixed λ favoured the lower-dimensional Swin-T features, and per-model tuning reduced the AUC gap by two-thirds while preserving a clear sensitivity advantage at a matched development target. Evaluations of frozen backbones that fix classifier hyperparameters across feature spaces should therefore report this sensitivity; ResNet-50, whose 2,048-dimensional features also favoured the strongest penalty, showed the same pattern. The conclusion of a backbone comparison also depends on the comparator: the transformer appears better against MobileNetV2 or ResNet-50 but not against ConvNeXt-T.

6.6. Errors by diagnosis

The most frequent missed positives were melanomas (20 of 171 images for Swin-T at the selected threshold), whereas false positives were concentrated in benign keratoses, dermatofibromas and vascular lesions. Because the positive class combines clinically different diagnoses, aggregate sensitivity and specificity should be read together with Tables 3 and 10; the intervals for the small categories are wide.

6.7. Limitations

Patient identifiers are unavailable, so the partitions are lesion-disjoint but not demonstrably patient-disjoint; multiple lesions of one patient could appear in different partitions, and patient-level dependence is not reflected in the lesion-cluster intervals. All data come from one benchmark collected at two clinical sources [12]; there was no external validation, and no evaluation across skin types, devices or populations, so the generalisation claims are bounded by this benchmark's case mix, and threshold instability may be more severe in multi-centre data. The binary grouping follows a management-based rationale (Section 3.2) but combines malignant and pre-malignant lesions, and the 80% specificity floor is a design choice rather than a clinical parameter; the cost ratios and threshold probabilities explored in Section 4.10 are illustrative and were not elicited from clinicians. Only four frozen ImageNet backbones with a linear classifier were evaluated; they differ in architecture, checkpoint training, interpolation and feature dimension, so the effect of shifted-window attention is not isolated even in the size-matched ConvNeXt-T comparison. The fine-tuned reference uses five runs of a fixed, untuned recipe with epoch and threshold chosen on a small validation split; it indicates the approximate cost of freezing, not the best achievable fine-tuned performance, and dermatology-specific models such as PanDerm [36] were not evaluated. The test-conditional bootstrap excludes fitting variability; the refit bootstrap includes it, but resampling training lesions reduces the effective training data and degrades performance, so it is used only to describe variability. The calibration, decision-curve, linear-probe and failure analyses were added post hoc, and the visual failure review was not performed by a dermatologist.

6.8. Future directions

Repeated fine-tuning with lesion-grouped development splits for epoch and threshold selection, external validation with a documented case mix, recalibration on the target population, threshold selection that accounts for an expected case mix, and comparison with frozen dermatology foundation-model features [36] are natural next steps. Evaluating patient metadata [34] requires linked records and comparison with the image-only pipeline.

7. Conclusions

In a lesion-aware evaluation of the official ISIC 2018 Task 3 partitions, frozen Swin-T features outperformed frozen MobileNetV2 and ResNet-50 features for a referral-oriented binary target—in ranking (AUC 0.893 vs 0.834 and 0.818), in sensitivity at a matched development-specificity target (+12.1 percentage points over MobileNetV2) and in net benefit—but there was no evidence of a difference from the size-matched convolutional ConvNeXt-T, and the size of the advantage over older backbones depended on classifier regularisation. Fine-tuning improved AUC consistently across five seeds. For clinical interpretation, the more important findings concern the operating point: a threshold chosen on the 193-image validation split transferred poorly for both frozen and fine-tuned models, a stable development threshold still missed its specificity target because test negatives were harder, and false positives concentrated in benign lesions that clinicians had themselves considered suspicious. Evaluations of frozen-feature classifiers on this dataset should therefore report lesion-clustered uncertainty, matched operating points, calibration and the transfer behaviour of any validation-selected threshold, rather than a single sensitivity–specificity pair.

Declarations

Data and code availability. ISIC 2018 Task 3 images and ground truth are available from the ISIC Challenge archive; lesion identifiers and confirmation types were retrieved from the ISIC Archive API [25]. The benchmark code, re-analysis scripts, the lesion metadata table, saved test predictions and all numerical results (including the post hoc analyses and the additional fine-tuning runs) are provided as supplementary material.
Conflict of interest. The author declares no conflict of interest.
Ethics statement. This study used only de-identified, retrospective dermoscopic images and labels released publicly through the ISIC Archive; no participants were recruited and no new data were collected. Ethical approval was therefore not sought.
Use of AI-assisted tools. AI-assisted tools (ChatGPT, Codex, Claude and QuillBot) were used for code development, statistical re-analysis and language editing; the author verified all code, results and text and takes full responsibility for the content.

Acknowledgement

The author would like to acknowledge the support of Prince Sultan University for paying the Article Processing Charges (APC) of this publication.

REFERENCES

1. P. Tschandl, N. Codella, B. N. Akay, et al., *Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study*, The Lancet Oncology, vol. 20, no. 7, pp. 938–947, 2019.
2. P. Tschandl, C. Rinner, Z. Apalla, et al., *Human–computer collaboration for skin cancer recognition*, Nature Medicine, vol. 26, no. 8, pp. 1229–1234, 2020. doi:10.1038/s41591-020-0942-0.
3. A. Esteva, B. Kuprel, R. A. Novoa, et al., *Dermatologist-level classification of skin cancer with deep neural networks*, Nature, vol. 542, no. 7639, pp. 115–118, 2017.
4. H. A. Haenssle, C. Fink, R. Schneiderbauer, et al., *Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists*, Annals of Oncology, vol. 29, no. 8, pp. 1836–1842, 2018.
5. K. Behara, E. Bhero, and J. T. Agee, *AI in dermatology: a comprehensive review into skin cancer detection*, PeerJ Computer Science, vol. 10, e2530, 2024. doi:10.7717/peerj-cs.2530.
6. B. C. R. S. Furriel, B. D. Oliveira, R. Prôa, et al., *Artificial intelligence for skin cancer detection and classification for clinical environment: a systematic review*, Frontiers in Medicine, vol. 10, 1305954, 2024. doi:10.3389/fmed.2023.1305954.
7. Z. Liu, Y. Lin, Y. Cao, et al., *Swin Transformer: hierarchical vision transformer using shifted windows*, Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10012–10022, 2021.
8. S. Ayas, *Multiclass skin lesion classification in dermoscopic images using swin transformer model*, Neural Computing and Applications, vol. 35, no. 9, pp. 6713–6722, 2023. doi:10.1007/s00521-022-08053-z.
9. I. Aruk, I. Pacal, and A. N. Toprak, *A comprehensive comparison of convolutional neural network and visual transformer models on skin cancer classification*, Computational Biology and Chemistry, vol. 120, pt. 2, 108713, 2026. doi:10.1016/j.compbiolchem.2025.108713.
10. A. Mahbod, R. Ecker, and R. Woitek, *Fusion of foundation and vision transformer model features for dermatoscopic image classification*, preprint, arXiv:2505.16338, 2025.
11. A. Muqtadir, K. Adi, and V. Gunawan, *Benchmarking attention mechanisms in ConvNeXt-Tiny for dermoscopic skin lesion classification under stratified and lesion-level leakage-safe cross-validation*, Statistics, Optimization & Information Computing, vol. 16, no. 4, pp. 3252–3271, 2026. doi:10.19139/soic-2310-5070-4134.
12. 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, 180161, 2018. doi:10.1038/sdata.2018.161.
13. M. M. G. Leeflang, A. W. S. Rutjes, J. B. Reitsma, L. Hooft, and P. M. M. Bossuyt, *Variation of a test's sensitivity and specificity with disease prevalence*, CMAJ, vol. 185, no. 11, pp. E537–E544, 2013. doi:10.1503/cmaj.121286.
14. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, *MobileNetV2: inverted residuals and linear bottlenecks*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4510–4520, 2018.
15. Z. Liu, H. Mao, C. Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, *A ConvNet for the 2020s*, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11976–11986, 2022.
16. J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, *ImageNet: a large-scale hierarchical image database*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 248–255, 2009.

17. D. Ajabani, Z. A. Shaikh, A. Yousef, K. Ali, and M. A. Albahar, *Enhancing skin lesion classification: a CNN approach with human baseline comparison*, PeerJ Computer Science, vol. 11, e2795, 2025. doi:10.7717/peerj-cs.2795.
18. N. Codella, V. Rotemberg, P. Tschandl, et al., *Skin lesion analysis toward melanoma detection 2018: a challenge hosted by the International Skin Imaging Collaboration (ISIC)*, preprint, arXiv:1902.03368, 2019.
19. N. Gessert, T. Sentker, F. Madesta, et al., *Skin lesion diagnosis using ensembles, unscaled multi-crop evaluation and loss weighting*, preprint, arXiv:1808.01694, 2018.
20. Y. Nie, P. Sommella, M. Carratù, M. O'Nils, and J. Lundgren, *A deep CNN transformer hybrid model for skin lesion classification of dermoscopic images using focal loss*, Diagnostics, vol. 13, no. 1, 72, 2023. doi:10.3390/diagnostics13010072.
21. R. Ramya and N. Jaisankar, *Design of a feature-optimized hybrid CNN–Transformer model for melanoma-vs.-non-melanoma classification*, Frontiers in Artificial Intelligence, vol. 9, 1936676, 2026. doi:10.3389/frai.2026.1936676.
22. N. Pathirana, I. Munasinghe, and D. Alwis, *Multimodal skin lesion classification with Swin Transformer and clinical metadata fusion*, preprint, arXiv:2608.07574, 2026.
23. M. A. Belal, M. M. El-Gazzar, B. B. S. Tawfik, and S. I. Ibrahim, *Leakage-controlled evaluation of handcrafted-deep feature fusion and confidence stacking for four-class skin lesion classification*, Statistics, Optimization & Information Computing, vol. 16, no. 4, pp. 3089–3102, 2026. doi:10.19139/soic-2310-5070-4022.
24. T. Saito and M. Rehmsmeier, *The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets*, PLoS ONE, vol. 10, no. 3, e0118432, 2015.
25. International Skin Imaging Collaboration, *ISIC Archive API v2, collections 66 (Challenge 2018: Task 3: Training), 73 (Validation) and 67 (Test): image metadata including lesion_id and diagnosis_confirm_type*, https://api.isic-archive.com/, accessed 6 November 2026.
26. PyTorch, *TorchVision 0.22 documentation: torchvision.models (weights metadata and inference transforms)*, https://docs.pytorch.org/vision/0.22/models.html, accessed 5 November 2026.
27. D. C. Liu and J. Nocedal, *On the limited memory BFGS method for large scale optimization*, Mathematical Programming, vol. 45, no. 1–3, pp. 503–528, 1989.
28. C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, *On calibration of modern neural networks*, Proceedings of the 34th International Conference on Machine Learning (PMLR 70), pp. 1321–1330, 2017.
29. B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall, New York, 1993.
30. C. A. Field and A. H. Welsh, *Bootstrapping clustered data*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 69, no. 3, pp. 369–390, 2007. doi:10.1111/j.1467-9868.2007.00593.x.
31. E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, *Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach*, Biometrics, vol. 44, no. 3, pp. 837–845, 1988. doi:10.2307/2531595.
32. Q. McNemar, *Note on the sampling error of the difference between correlated proportions or percentages*, Psychometrika, vol. 12, no. 2, pp. 153–157, 1947.
33. S. Kitada and H. Iyatomi, *Skin lesion classification with ensemble of squeeze-and-excitation networks and semi-supervised learning*, preprint, arXiv:1809.02568, 2018.
34. J. Höhn, A. Hekler, E. Krieghoff-Henning, et al., *Integrating patient data into skin cancer classification using convolutional neural networks: systematic review*, Journal of Medical Internet Research, vol. 23, no. 7, e20708, 2021. doi:10.2196/20708.
35. S. Kornblith, J. Shlens, and Q. V. Le, *Do better ImageNet models transfer better?*, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2661–2671, 2019.
36. S. Yan, Z. Yu, C. Primiero, et al., *A multimodal vision foundation model for clinical dermatology*, Nature Medicine, vol. 31, no. 8, pp. 2691–2702, 2025. doi:10.1038/s41591-025-03747-y.
37. M. Combalia, N. Codella, V. Rotemberg, et al., *Validation of artificial intelligence prediction models for skin cancer diagnosis using dermoscopy images: the 2019 International Skin Imaging Collaboration Grand Challenge*, The Lancet Digital Health, vol. 4, no. 5, pp. e330–e339, 2022. doi:10.1016/S2589-7500(22)00021-8.
38. R. Daneshjou, K. Vodrahalli, R. A. Novoa, et al., *Disparities in dermatology AI performance on a diverse, curated clinical image set*, Science Advances, vol. 8, no. 32, eabq6147, 2022. doi:10.1126/sciadv.abq6147.
39. B. Van Calster, D. J. McLernon, M. van Smeden, L. Wynants, and E. W. Steyerberg, *Calibration: the Achilles heel of predictive analytics*, BMC Medicine, vol. 17, 230, 2019. doi:10.1186/s12916-019-1466-7.
40. A. J. Vickers and E. B. Elkin, *Decision curve analysis: a novel method for evaluating prediction models*, Medical Decision Making, vol. 26, no. 6, pp. 565–574, 2006. doi:10.1177/0272989X06295361.
41. Y. El Idrissi El-Bouzaidi, S. El Khamlichi, and O. Abdoun, *Diagnosis with deep learning and a novel pre-processing strategy on a large-scale dermoscopic image dataset*, Statistics, Optimization & Information Computing, vol. 15, no. 3, pp. 2069–2085, 2025. doi:10.19139/soic-2310-5070-3122.