Retrieval Familiarity for Selective Referral in Dermoscopic Classification: An External Evaluation Under Distribution Shift

Pronab Kumar Mondal[1]*, Subrata Kumar Das[2],
Sumon Kumar Debnath[3], Indrani Mandal[4], Md. Selim Al Mamun[5]
Department of Computer Science and Engineering,
Jatiya Kabi Kazi Nazrul Islam University, Mymensingh, Bangladesh[1,2,4,5]
Department of Electrical and Electronic Engineering, Begum Rokeya University, Rangpur, Bangladesh[3]

***Abstract*—This study audits retrieval familiarity as a post-hoc accept–refer signal for dermoscopic classification under dataset shift. Mean Top-5 familiarity was defined as the average cosine similarity between a query and its five nearest cases in a lesion-disjoint HAM10000 reference bank. Retrieval depth and the operating threshold were selected on HAM10000 validation data and then fixed. Evaluation used 14,885 ISIC 2019 images without matching development identifiers and 963 eligible Derm7pt images. The base classifier achieved 53.83% accuracy (macro-F1 0.4012) on ISIC 2019 and 64.38% accuracy (macro-F1 0.3926) on Derm7pt. Mean Top-5 familiarity obtained an area under the risk–coverage curve (AURC) of 0.3498 on ISIC 2019 and 0.2514 on Derm7pt, ranking fourth and ninth, respectively. A Combined Random Forest error score, constructed using consistent feature definitions across all cohorts, achieved lower AURCs of 0.3245 and 0.2096; the paired differences from Mean Top-5 were -0.0253 (95% CI -0.0324 to -0.0184; Holm-adjusted $p = 0.013$) and -0.0419 (-0.0670 to -0.0162; Holm-adjusted $p = 0.022$). A threshold designed for 50% validation coverage produced only 21.95% and 39.36% coverage for Mean Top-5 and 19.73% and 37.59% for the combined score. At 50% ISIC coverage, the familiarity rule accepted 1,034 of 2,153 misclassified melanomas. Predicted-class and cost-weighted analyses also showed that aggregate ranking did not establish clinical safety. The results support retrieval familiarity as an inspectable but weak signal whose value improved when combined with classifier information. They also show that score definitions, feature pipelines, thresholds, and class-level behaviour require careful external auditing before clinical use.**

Keywords*—Dermoscopy; selective classification; selective referral; distribution shift; image retrieval

I. Introduction

Deep learning has changed dermoscopic image analysis. HAM10000 and the ISIC collections enabled large multi-class comparisons [1], [2], while recent foundation models support several dermatological tasks within one representation [3]. Strong average performance, however, does not show whether one prediction is safe to accept. This matters because a missed melanoma is not comparable with an error between low-risk benign lesions. Devices, institutions, illumination, populations, and case mix can also shift after development, weakening otherwise useful uncertainty estimates [4]. Limited external validation and uncertain generalizability consequently remain barriers to clinical use [5], [6].

Selective classification addresses this problem by accepting predictions judged sufficiently reliable and referring the remainder [7], [8]. Its behaviour is described by *coverage*, the accepted proportion, and *selective risk*, the error rate among accepted predictions. A useful score places errors toward the high-risk end, allowing referral to remove them efficiently.

Post-hoc referral is useful when the classifier must remain unchanged: a separate score controls acceptance without retraining [9]. Softmax confidence, entropy, and probability margin are inexpensive but may remain confident on unfamiliar inputs. A frozen deterministic score also avoids repeated test-time sampling and limits the components requiring revalidation.

An alternative is to compare a query with observed cases. Nearest-neighbour similarity measures how well the query is represented in a learned feature space [10]. Dermoscopic retrieval has been studied as a diagnostic aid [11] and as a failure-detection signal [12]. The neighbours are inspectable, but similarity is not proof of correctness: a query may resemble the wrong class, an imbalanced bank may favour common lesions, and artefacts may drive visual matches. Moreover, a useful ranking does not guarantee that a threshold will preserve its intended workload after transfer.

This study evaluates Mean Top-5 retrieval familiarity as a post-hoc referral score under external dataset shift. A ConvNeXt-Tiny classifier [13] supplies both the prediction and image embedding. It is trained on one lesion-disjoint HAM10000 partition; another lesion-disjoint partition forms the fixed reference bank [1]. Familiarity is the mean cosine similarity between a query and its five nearest reference cases. Referral risk is one minus that value, so higher risk indicates a greater need for review.

Retrieval depth and threshold are selected on HAM10000 validation data, then transferred without recalibration to an ISIC 2019 subset after removing HAM10000 identifier matches and to an eligible Derm7pt cohort [2], [14]. The ISIC 2019 cohort is exact-image-ID-disjoint from development, although patient- or lesion-level independence could not be verified. The study concerns referral under severe shift, not

*Corresponding author.

development of a state-of-the-art diagnostic classifier.

The contribution is a multi-cohort external audit rather than a claim of a new or universally superior nearest-neighbour algorithm. The frozen rule was compared with 13 alternatives on ISIC 2019 and 12 on Derm7pt without using external labels for selection, fitting, or recalibration. The analysis separates overall error ranking, fixed-threshold transfer, and retrospective true-diagnosis-stratified selection. Mean Top-5 familiarity ranked fourth on ISIC 2019 and ninth by point estimate on Derm7pt; its difference from predictive entropy on ISIC 2019 did not remain statistically significant after Holm correction. The Combined Random Forest, an error model using classifier and retrieval features, ranked first in both external cohorts. A threshold designed for 50% validation coverage produced 21.95% coverage on ISIC 2019 and 39.36% on Derm7pt. Four ISIC 2019 true-diagnosis strata, including melanoma, also showed an unfavourable direction. Thus, an aggregate ranking signal can coexist with unstable workload and unfavourable retrospective subgroup selection. Familiarity is examined as an inspectable but limited complementary signal rather than as a replacement for predictive uncertainty or a validated clinical safety mechanism.

II. Related Work

A. Dermoscopic Classification Under Dataset Shift

Dermoscopic classifiers are often developed on a small number of curated collections. Foundation models broaden representation learning but do not remove the need for independent evaluation [3]. HAM10000 is multi-source yet dominated by nevi [1]; BCN20000 contains less constrained clinical images [15]; and Derm7pt combines images, metadata, and checklist annotations [14]. Their different devices, acquisition practices, and diagnostic mixes make transfer both informative and difficult.

Confidence may not remain reliable after this change. Networks can assign high probability to errors [16], and uncertainty often deteriorates under distribution shift [4]. Dermatology reviews likewise report limited external validation and uncertain population transfer [5], [6]. The present study therefore examines the referral signal under severe shift rather than another diagnostic backbone.

B. Selective Classification and Post-Hoc Referral

Selective classification balances coverage against error among accepted cases [7], [8]. SelectiveNet learns the selection function during training [17]; post-hoc methods instead change only the accept–refer score, which suits an existing frozen classifier and permits fair comparison on identical predictions [9].

Maximum softmax probability, entropy, and probability margin are inexpensive but remain tied to calibration and logit scale [18]. SIRC adds an auxiliary confidence signal [19]; generalized margins target label and covariate shift [20]; and likelihood-ratio methods contrast evidence associated with correct and incorrect predictions [21].

Risk–coverage curves evaluate the whole score ordering. Lower AURC means that errors are removed earlier [8], [9]. AURC alone, however, does not show whether a development threshold preserves its coverage or risk after transfer [22]. Ranking and fixed-threshold transfer are therefore assessed separately.

C. Representation-Based Confidence and Dermoscopic Retrieval

Representation-based confidence compares a query with development examples or feature distributions. Mahalanobis confidence uses class-conditional distance, with shrinkage stabilizing high-dimensional covariance estimation [23], [24]. Nearest-neighbour methods avoid an explicit distributional model and perform well on normalized deep embeddings [10].

Dermoscopic retrieval has been evaluated as a diagnostic aid [11] and in human–computer collaboration [25]. A retrieved neighbourhood is more inspectable than one confidence value, but it is neither a causal explanation nor diagnostic proof.

Super-TrustScore is the closest failure-detection method: it combines global and class-conditional neighbourhood information [12]. The simpler score evaluated here averages the five nearest cases without class restriction. The unresolved issue is whether this global signal transfers across cohorts. Three aspects are examined together: aggregate error ranking, transfer of a fixed threshold, and the direction of selection within individual diagnoses.

III. Materials and Methods

A. Study Design and Datasets

This retrospective computational study used HAM10000 for all development decisions and two external datasets for frozen evaluation. No external label was used to fit a score, select a parameter, or recalibrate a threshold. Mean Top-5 familiarity was designated as the primary score before external comparison, and AURC was the primary endpoint.

The evaluation examined whether familiarity ranked external errors, whether its fixed threshold preserved the intended operating point, and whether aggregate improvement held within each diagnosis. Class-restricted retrieval and a learned combination were secondary analyses.

HAM10000 contains 10,015 dermoscopic images from 7,470 lesions across seven diagnoses [1]: actinic keratoses and intraepithelial carcinoma (`akiec`), basal cell carcinoma (`bcc`), benign keratosis-like lesions (`bkl`), dermatofibroma (`df`), melanoma (`mel`), melanocytic nevi (`nv`), and vascular lesions (`vasc`). Ten folds were generated with `StratifiedGroupKFold`, using diagnosis for stratification and `lesion_id` for grouping (seed 42). Folds 3–9 supplied 7,009 training images. Folds 0, 1, and 2 supplied 1,002 images each for validation, the reference bank, and the internal test set, respectively. Pairwise comparison of lesion identifiers confirmed that these partitions were lesion-disjoint. The internal test set was not consulted during checkpoint, retrieval, threshold, or comparator selection, and reference-bank images were not used for classifier optimization.

ISIC 2019 was the primary external cohort [2], [15]. Starting from 25,331 labelled training images, 628 squamous-cell-carcinoma images outside the classifier's output space

TABLE I. DEVELOPMENT PARTITIONS AND EXTERNAL EVALUATION COHORTS

Cohort or partition	Images	Classes	Purpose
HAM10000 training	7,009	7	Classifier optimization
HAM10000 validation	1,002	7	Model and policy selection
HAM10000 reference bank	1,002	7	Retrieval cases
HAM10000 internal test	1,002	7	In-domain evaluation
ISIC 2019	14,885	7	Primary external evaluation
Derm7pt	963	6	Additional external evaluation

TABLE II. CLASS DISTRIBUTION OF THE FIXED HAM10000 REFERENCE BANK

Class	Images	Percentage
AKIEC	33	3.29
BCC	52	5.19
BKL	110	10.98
DF	11	1.10
MEL	111	11.08
NV	671	66.97
VASC	14	1.40

were excluded, followed by 9,818 images whose normalized ISIC identifiers occurred in HAM10000. A second identifier audit found no residual match. The resulting exact-image-ID-disjoint cohort contained 14,885 images: 6,170 `nv`, 3,409 `mel`, 2,809 `bcc`, 1,525 `bkl`, 737 `akiec`, 124 `df`, and 111 `vasc`. Patient- or lesion-level independence could not be verified from the released metadata.

Derm7pt provided a second transfer setting [14]. A total of 963 dermoscopic images with diagnoses that could be mapped unambiguously to the target taxonomy were retained: 42 `bcc`, 45 `bkl`, 20 `df`, 252 `mel`, 575 `nv`, and 29 `vasc`. No eligible case mapped to `akiec`; 48 records were excluded. Four melanoma-metastasis records were included in the primary analysis and removed in a sensitivity analysis. Table I summarizes the development partitions and external cohorts.

The reference bank preserved the natural HAM10000 imbalance. Of its 1,002 images, 671 (66.97%) were nevi, whereas DF and VASC contributed only 11 and 14 images, respectively (Table II).

B. Preprocessing and Classifier

Images were resized to 224×224 pixels and normalized using ImageNet channel statistics. Training augmentation comprised horizontal and vertical flips, rotations within $\pm 20^\circ$, and modest colour jitter. No random augmentation was applied to validation, reference-bank, reserved, or external images.

An ImageNet-pretrained ConvNeXt-Tiny model was fitted with a seven-output linear head [13]. The complete backbone and classification head were fine-tuned jointly; no layer was frozen during optimization. For image x, trainable parameters θ, and class probability $p_\theta(c \mid x)$, the prediction was

$$\hat{y}(x) = \arg \max_{c \in \{1,\ldots,C\}} p_\theta(c \mid x), \qquad C = 7. \tag{1}$$

Class imbalance was handled with weighted cross-entropy, using $w_c = N_{\text{train}}/(Cn_c)$, where N_{train} is the number of training images and n_c is the number from class c. AdamW optimization was run for at most 15 epochs with learning rate 10^{-4}, weight decay 10^{-5}, batch size 16, and cosine annealing with $T_{\max} = 15$. Checkpoint selection used validation macro-F1 only. Epoch 14 was retained (macro-F1=0.6678). Python, NumPy, and PyTorch used seed 42. Training and inference were performed on a CPU-based Windows system without CUDA acceleration or mixed-precision computation.

Only one training run, data split, architecture, checkpoint, random seed, and embedding layer were evaluated. The objective was to compare referral scores while holding the underlying predictions fixed, rather than to establish stability across classifier-training conditions. The reported rankings and class-specific behaviour are therefore conditional on this classifier and should not be generalized to stronger classifiers, foundation models, alternative splits, or independently trained seeds.

C. Reference Bank and Primary Referral Score

The frozen encoder produced a 768-dimensional pre-classification embedding $f_\theta(x)$. Query and reference embeddings were normalized as

$$z(x) = \frac{f_\theta(x)}{\|f_\theta(x)\|_2}, \qquad s(x, r_j) = z(x)^\top z(r_j), \tag{2}$$

where, r_j is reference image j and $s(x, r_j)$ is cosine similarity. Let $r_{(1)}, \ldots, r_{(k)}$ denote the k reference cases ordered from greatest to smallest similarity. Familiarity and referral risk were

$$F_k(x) = \frac{1}{k} \sum_{j=1}^{k} s\big(x, r_{(j)}\big), \qquad R_{\text{fam}}(x) = 1 - F_k(x). \tag{3}$$

Candidate depths $k \in \{1,2,3,4,5\}$ were compared on HAM10000 validation AURC. Mean Top-5 familiarity achieved the lowest value (0.1162) and was fixed at $k = 5$. The risk threshold was then selected on HAM10000 validation data to produce 50% coverage. This procedure yielded $\tau = 0.397352$:

$$a_\tau(x) = \begin{cases} 1, & R_{\text{fam}}(x) \le \tau \quad \text{(accept)}, \\ 0, & R_{\text{fam}}(x) > \tau \quad \text{(refer)}. \end{cases} \tag{4}$$

The numerical separation among the best validation candidates was small. In particular, the AURC difference between $k = 1$ and the selected boundary value $k = 5$ was approximately 0.0003. Thus, $k = 5$ was retained because it followed the prespecified selection rule, not because it was shown to be a uniquely optimal or generally transferable neighbourhood depth. The exploratory results for broader neighbourhoods require independent confirmation.

The threshold was calculated from the unrounded validation scores and retained to six decimal places so that the accept–refer assignments could be reproduced exactly. This numerical precision is an implementation detail and does not imply comparable clinical precision. The encoder, reference

Algorithm 1 Frozen Mean Top-5 familiarity accept–refer procedure

Require: Query image x; frozen embedding function f_θ; frozen class-probability function $p_\theta(c \mid x)$; normalized reference embeddings $\mathscr{Z}_{\text{ref}}$; $k = 5$; $\tau = 0.397352$.

Ensure: Predicted class $\hat{y}(x)$, familiarity risk $R_{\text{fam}}(x)$, and decision $d_\tau(x) \in \{\text{ACCEPT}, \text{REFER}\}$.

1: Predict $\hat{y}(x) = \arg\max_{c\in\{1,\dots,7\}} p_\theta(c \mid x)$ and calculate $z(x) = f_\theta(x)/\|f_\theta(x)\|_2$.
2: Compute the cosine similarity between $z(x)$ and each embedding in $\mathscr{Z}_{\text{ref}}$, then retrieve the five greatest similarities, $s_{(1)}(x), \dots, s_{(5)}(x)$.
3: Calculate $F_5(x) = 5^{-1}\sum_{j=1}^{5} s_{(j)}(x)$ and set $R_{\text{fam}}(x) = 1 - F_5(x)$.
4: Set $d_\tau(x) = \text{ACCEPT}$ when $R_{\text{fam}}(x) \le \tau$; otherwise set $d_\tau(x) = \text{REFER}$ and withhold the automated prediction for expert assessment.
5: **return** $\hat{y}(x)$, $R_{\text{fam}}(x)$, and $d_\tau(x)$.

bank, k, and τ were then locked. Fig. 1 summarizes the workflow, and Algorithm 1 gives the inference procedure.

D. Comparators and Evaluation

Thirteen alternatives were evaluated on the same predictions and oriented so that larger values indicated greater referral risk. Twelve were evaluated on Derm7pt because the stored Derm7pt generalized geometric margin used a different formula from the validation and ISIC 2019 implementation and was therefore excluded. Classifier-derived scores were maximum-softmax risk, normalized predictive entropy, and probability-margin risk [18]. Retrieval ablations used Top-1 similarity, fifth-neighbour similarity, and the fraction of the five retrieved labels that disagreed with the prediction. Other comparators were a class-conditional Mahalanobis score with Ledoit–Wolf covariance shrinkage [23], [24], generalized confidence and geometric margins [20], an SIRC-style combination [19], adapted Δ-KNN-style risk [21], a validation-trained Combined Random Forest error model, and expected random referral. Every fitted parameter came from HAM10000 development data.

For clarity, let $p_c(x) = p_\theta(c \mid x)$ denote the predicted probability of class c. Maximum-softmax risk was $1 - p_{(1)}(x)$, where $p_{(1)}(x)$ is the largest predicted probability. Normalized entropy was

$$R_{\text{ent}}(x) = -\frac{1}{\log C}\sum_{c=1}^{C} p_c(x)\log p_c(x), \tag{5}$$

and probability-margin risk was $1 - [p_{(1)}(x) - p_{(2)}(x)]$, with $p_{(2)}(x)$ denoting the second-largest probability. Top-1 and fifth-neighbour risks were respectively $1 - s(x, r_{(1)})$ and $1 - s(x, r_{(5)})$. Retrieval-label support risk was one minus the proportion of five retrieved labels that agreed with $\hat{y}(x)$.

Mahalanobis risk was the minimum squared distance to the seven class-conditional reference-bank means. A shared within-class covariance matrix was estimated with Ledoit–Wolf shrinkage [23], [24]. Generalized confidence margin was the negative log ratio of the two leading probabilities; the geometric variant normalized the two leading logits using the norms of their classification-head weight vectors [20].

The SIRC-style comparator combined maximum softmax probability with Mean Top-5 familiarity [19]. Its location and scale parameters were estimated from validation familiarity values only. It is termed *SIRC-style* because retrieval familiarity replaced the auxiliary score used in the original formulation. Adapted Δ-KNN contrasted the fifth-neighbour distances from a query to validation embeddings associated with correct and incorrect predictions [21]. This was also an adaptation: the implementation used correctness groups and $k = 5$ rather than reproducing every element of the original average Top-k configuration. The Combined Random Forest estimated error probability from 13 classifier and retrieval features: confidence, uncertainty, entropy, probability margin, retrieval confidence, ambiguity and margin, mean, maximum and minimum neighbour similarity, classifier–retrieval agreement, total-variation disagreement, and retrieval support for the predicted class. The same feature definitions and the final reference-bank retrieval were used for validation, ISIC 2019, and Derm7pt. A 500-tree model with maximum depth 8, minimum leaf size 5, balanced class weights, and seed 42 was trained on the 1,002 validation cases. Five-fold stratified out-of-fold probabilities supplied its validation scores; the model was then refitted on all validation cases and frozen for external evaluation. This reconstruction corrected an earlier analysis in which development and external retrieval features had been generated by inconsistent pipelines.

Risk scores were ordered from lowest to highest. When several observations shared a score, the error contribution at each position in that tied block was replaced by the block's mean error rate. The resulting curve is the exact expectation under random ordering within tied groups and is independent of evaluation-table order. This treatment is particularly important for retrieval-label support, which takes only the values $0, 0.2, \dots, 1.0$. For that score, the exact ISIC 2019 AURC was 0.41221; 500 random within-tie permutations produced a mean of 0.41219 (SD 0.0018). The trapezoidal integration step was separately checked using `sklearn.metrics.auc` and agreed to machine precision; the permutation experiment provided the independent check of tie handling.

Consider an evaluation cohort $\mathscr{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where N is the number of images, x_i is image i, and y_i is its true class. Let $\hat{y}_i = \hat{y}(x_i)$ denote the corresponding classifier prediction and $R_i = R(x_i)$ its referral-risk score. The error and acceptance indicators are defined as $e_i = \mathbb{I}(\hat{y}_i \neq y_i)$ and $a_i(t) = \mathbb{I}(R_i \le t)$, respectively, where t is the risk threshold. The function $\mathbb{I}(\cdot)$ equals one when its argument is true and zero otherwise. Coverage and selective risk were

$$\text{Coverage}(t) = \frac{1}{N}\sum_{i=1}^{N} a_i(t), \qquad \text{Risk}(t) = \frac{\sum_{i=1}^{N} a_i(t)e_i}{\sum_{i=1}^{N} a_i(t)}. \tag{6}$$

At threshold t, error capture measured the fraction of all classifier errors placed in the referred set:

$$\text{ErrorCapture}(t) = \frac{\sum_{i=1}^{N}[1 - a_i(t)]e_i}{\sum_{i=1}^{N} e_i}. \tag{7}$$

Referral precision was the error rate among referred images. Accepted-case accuracy was $1 - \text{Risk}(t)$. The evaluation

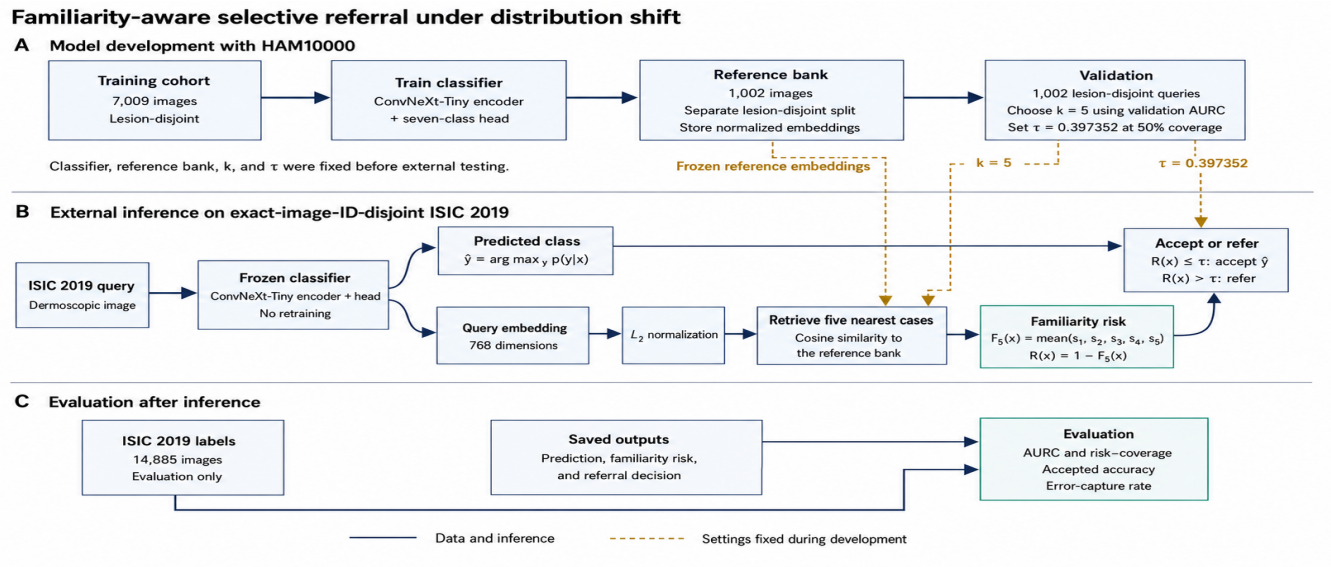

Fig. 1. Familiarity-aware selective referral. A classifier and a separate HAM10000 reference bank were developed from lesion-disjoint partitions. Mean similarity to the five nearest reference embeddings was converted to referral risk. The complete configuration was transferred unchanged to ISIC 2019 and Derm7pt; external labels were introduced only for evaluation.

covered the complete ranking, the descriptive 50% coverage point, and direct transfer of the fixed validation threshold. The 50% point permits methods to be compared at the same workload; the fixed threshold tests whether a development-selected policy transfers without adjustment. These operating conditions were kept distinct throughout the analysis.

For each score, images were ordered from lowest to highest referral risk. At accepted-set size $m \in \{1,\ldots,N\}$, coverage was m/N, and selective risk was the cumulative number of errors among the first m images divided by m. The zero-coverage point was excluded because selective risk is undefined when no image is accepted. AURC was calculated by trapezoidal integration over the observed nonzero coverage–risk coordinates, including the full-coverage endpoint; lower values indicate better error ranking [8], [9]. The same procedure was applied to every comparator. Error-detection AUROC, area under the precision–recall curve (AUPRC), error capture among referred images, and accepted-case accuracy are also reported. Because AURC can obscure aspects of selective behaviour [22], fixed-coverage, fixed-threshold, and class-stratified outcomes were examined alongside it.

Validation thresholds targeting 50% coverage were transferred unchanged for comparators whose validation and external scores used compatible definitions and scales. Adapted Δ-KNN was excluded from this analysis because its available validation score was calculated in-sample. The Derm7pt generalized geometric margin was excluded because its stored value used a different formula from the validation implementation. Retrieval-label support could not attain exactly 50% with a deterministic threshold because many cases lay on the boundary; strict and inclusive coverage bounds were therefore reported.

Two additional safety analyses were conducted at 50% ISIC 2019 coverage. First, results were stratified by the classifier's predicted class, which is available when a referral decision is made. The number of true melanomas that were misclassified and accepted was estimated with a 10,000-resample image-level bootstrap. Second, an illustrative cost-weighted analysis assigned error weights of 5 to melanoma, 3 to AKIEC, 2 to BCC, and 1 to the remaining diagnoses. These weights were chosen for sensitivity analysis and were not clinically elicited.

The matched-risk analysis was secondary and descriptive. For a specified selective-risk target, matched-risk coverage was defined as the greatest observed coverage at which selective risk did not exceed that target. Predictive entropy was used as the comparator because it is a conventional, classifier-derived uncertainty score that requires no additional model.

The measures were kept distinct: AUROC evaluates error discrimination, AURC the retained predictions across coverage, error capture the referred set, and accepted accuracy the automated set. Under random referral, referring a proportion q is expected to capture approximately q of all errors.

E. Secondary Robustness and Computational Analyses

Several secondary analyses examined the limits of the primary configuration. Reference-bank sensitivity was evaluated with 20 naturally distributed subsamples at sizes 70, 125, 250, and 500, alongside the complete 1,002-image bank (seed 42). A balanced unique bank could contain at most 77 images because DF, the rarest reference class, contributed 11 cases; therefore, a 70-image balanced bank was evaluated.

The aggregation search was then extended to $k \in \{1,2,3,4,5,10,20\}$ and five summaries of the retrieved similarities: mean, minimum (the kth neighbour), median, lower quantile, and distance-weighted mean. The best configuration was selected on HAM10000 validation AURC and transferred unchanged. This broader search followed the original Mean Top-5 analysis and was therefore exploratory; its paired bootstrap comparison with the primary score was not adjusted for the 35 configurations searched.

Temperature scaling was fitted on HAM10000 validation probabilities by minimizing negative log-likelihood and transferred unchanged. The comparison included negative log-likelihood, 15-bin expected calibration error, and the AURCs of maximum-softmax and entropy risks before and after scaling. A separate class-wise analysis compared Mean Top-5 familiarity with entropy, SIRC-style risk, and adapted Δ-KNN at 50% overall ISIC 2019 coverage.

CPU cost was measured for 200 images after 10 warm-up passes. Online retrieval latency was measured for one query over 1,000 repetitions with banks of 125, 250, 500, and 1,002 FP32 embeddings; 500-query batches were also profiled. The profiling computer used an Intel Core i7-1165G7 CPU (four physical cores, eight logical processors), Python 3.14.3, PyTorch 2.12.0 CPU, torchvision 0.27.0, `timm` 1.0.25, scikit-learn 1.8.0, NumPy 2.4.2, and pandas 3.0.0.

F. Statistical Analysis and Leakage Control

Confidence intervals for every score comparison, including the Combined Random Forest, were calculated with 2,000 paired image-level bootstrap resamples (seed 42) using the percentile method. All scores in a cohort were evaluated in one bootstrap run, so each paired replicate used the same sampled indices for every compared score. The accepted-missed-melanoma count used 10,000 resamples (seed 20261017). For comparison with Mean Top-5 familiarity, a two-sided bootstrap p value was calculated from the paired AURC differences, using a finite-bootstrap $+1$ correction. Holm's step-down procedure controlled the family-wise error rate across the 13 ISIC 2019 and 12 Derm7pt comparator tests; adjusted $p < 0.05$ was considered statistically significant. Class-wise AUROCs were exploratory and were not included in the multiplicity-adjusted comparisons.

HAM10000 partitions were lesion-disjoint, and the reference bank was separate from classifier training and validation. ISIC identifiers matching HAM10000 were removed before evaluation. Neither external dataset informed checkpoint selection, retrieval depth, threshold estimation, covariance fitting, comparator fitting, or score direction.

Exact identifier matching excluded known matching image identifiers but could not establish complete patient- or lesion-level independence across source collections. Unrecognized source, patient, or lesion relationships between HAM10000 and ISIC 2019 may therefore remain. Furthermore, image-level bootstrap resampling may underestimate uncertainty when multiple images originate from the same lesion. The reported intervals should therefore be interpreted as image-level rather than lesion-cluster-adjusted intervals.

G. Reproducibility and Ethical Scope

Split assignment, checkpoint, normalization, score directions, seeds, bootstrap settings, and fitted comparator parameters were recorded before external evaluation. Validation queries searched only the reference bank; the internal test fold was used after development choices were fixed. External labels were introduced only for evaluation.

HAM10000, ISIC 2019, and Derm7pt are public, de-identified datasets whose source publications describe collection and governance [1], [2], [14]. No participant was recruited or contacted. Here, referral means withholding an automated output for expert assessment, not correcting a diagnosis or replacing dermatologist review.

TABLE III. PERFORMANCE OF THE FROZEN CLASSIFIER BEFORE AND AFTER EXTERNAL TRANSFER

Cohort	Images	Classes	Accuracy	Macro-F1	Balanced accuracy
HAM10000 internal test	1,002	7	78.64%	0.6857	0.7460
ISIC 2019	14,885	7	53.83%	0.4012	0.4118
Derm7pt	963	6	64.38%	0.3926	0.4628

TABLE IV. SENSITIVITY TO RETRIEVAL DEPTH. ONLY VALIDATION AURC WAS USED FOR SELECTION AMONG THE PRESPECIFIED VALUES $k \in \{1,\ldots,5\}$; LOWER AURC IS BETTER

k	Validation AURC	ISIC 2019 AURC	ISIC accuracy at 50%
1	0.1165	0.3569	65.61%
2	0.1194	0.3536	65.61%
3	0.1191	0.3522	65.80%
4	0.1197	0.3505	65.90%
5	**0.1162**	**0.3498**	**66.06%**

IV. RESULTS

A. Classifier and Cohort Performance

On the held-out HAM10000 test fold, the frozen classifier achieved 78.64% accuracy, macro-F1 0.6857, and balanced accuracy 0.7460. Performance fell substantially after external transfer. On ISIC 2019, 8,012 of 14,885 predictions were correct (53.83% accuracy), while Derm7pt accuracy was 64.38% (620/963). The study therefore examines referral under a marked shift; these figures are not offered as competitive diagnostic benchmarks. Full cohort-level results are reported in Table III.

On ISIC 2019, support-weighted F1 was 0.5387 and recall ranged from 15.88% for AKIEC to 72.97% for NV. All referral methods used identical predictions; only their risk ordering differed.

B. Validation Selection of Retrieval Depth

Mean Top-5 produced the lowest validation AURC (0.1162) and was selected without external data (Table IV; Fig. 2). On ISIC 2019, AURC decreased from 0.3569 at $k = 1$ to 0.3498 at $k = 5$; these external results did not affect selection.

C. Post-Hoc Robustness Analyses

The extension to larger neighbourhoods was conducted after the primary analysis and is therefore exploratory. Across 35 combinations of retrieval depth and aggregation, the validation-selected rule was the 20th-neighbour similarity (the minimum similarity among the 20 nearest references). Its AURC was 0.0817 on validation, 0.3275 on ISIC 2019, and 0.2166 on Derm7pt, compared with 0.1162, 0.3498, and 0.2514 for the original Mean Top-5 score. Paired bootstrap differences favoured the exploratory rule on ISIC 2019 (-0.0224, 95% CI -0.0260 to -0.0189) and Derm7pt (-0.0349, 95% CI -0.0506 to -0.0198; unadjusted $p = 0.001$ for both). Because 35 configurations were searched on one validation set,

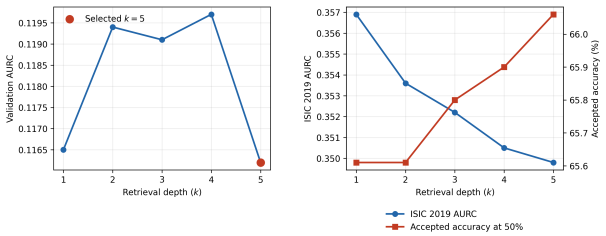

Fig. 2. Sensitivity to retrieval depth. HAM10000 validation AURC determined $k = 5$ before external evaluation; ISIC 2019 values are shown only to describe transfer of the selected family of scores.

TABLE V. POST-HOC ROBUSTNESS ANALYSES. LOWER AURC IS BETTER. THE 20TH-NEIGHBOUR RULE WAS SELECTED AFTER SEARCHING 35 CONFIGURATIONS AND IS NOT A NEW PRIMARY ENDPOINT

Analysis	Validation	ISIC 2019	Derm7pt
Original Mean Top-5 AURC	0.1162	0.3498	0.2514
Exploratory 20th-neighbour AURC	0.0817	0.3275	0.2166
Raw maximum-softmax AURC	0.0686	0.3649	0.2378
Temperature-scaled maximum-softmax AURC	0.0674	0.3616	0.2330

this finding is exploratory and does not replace the primary analysis.

Reference-bank size also mattered, but not in a simple "larger is better" direction. Twenty natural-distribution subsamples produced mean AURCs of 0.3375, 0.3360, 0.3420, and 0.3459 on ISIC 2019 for bank sizes 70, 125, 250, and 500, respectively; the full bank yielded 0.3498. On Derm7pt the corresponding values were 0.2251, 0.2267, 0.2375, 0.2489, and 0.2514. A class-balanced bank without replacement was limited to 70 images because the reference fold contained only 11 DF cases. Its mean AURC was 0.3644 on ISIC 2019 and 0.2556 on Derm7pt. Neither enlarging the imbalanced bank nor balancing a small bank guaranteed better ranking.

Temperature scaling fitted on validation data reduced expected calibration error from 0.0754 to 0.0352 on validation, from 0.2542 to 0.1422 on ISIC 2019, and from 0.1668 to 0.0517 on Derm7pt. Its effect on referral ranking was much smaller: maximum-softmax AURC changed from 0.3649 to 0.3616 on ISIC 2019 and from 0.2378 to 0.2330 on Derm7pt. Calibration improved probability estimates but changed referral ranking only slightly. Table V summarizes these robustness checks.

D. External Error Ranking

On ISIC 2019, Mean Top-5 familiarity achieved an AURC of 0.3498 (95% CI 0.3390–0.3603) and ranked fourth among 14 scores (Table VI). The Combined Random Forest ranked first (AURC 0.3245, 95% CI 0.3143–0.3344), followed by fifth-neighbour similarity (0.3466) and adapted Δ-KNN (0.3490). The Combined Random Forest–Mean Top-5 difference was -0.0253 (95% CI -0.0324 to -0.0184; unadjusted $p = 0.001$, $p_{\text{Holm}} = 0.013$). The fifth-neighbour difference was statistically significant ($\Delta\text{AURC} = 0.0031$, $p_{\text{Holm}} = 0.013$), whereas adapted Δ-KNN did not differ from the primary score ($p_{\text{Holm}} = 0.870$). Mean Top-5 familiarity had significantly lower AURC than Top-1 similarity, Mahalanobis risk, maximum-softmax risk, and probability-margin risk after Holm correction. Its difference from entropy was significant before adjustment ($p = 0.019$), but not afterward ($p_{\text{Holm}} = 0.057$).

The primary score's error-detection AUROC and AUPRC were 0.6537 and 0.5881. At 50% coverage, selective risk was 33.94%, and the referred half contained 63.25% of all errors.

The frozen comparison on Derm7pt showed a similar advantage for the combined score. Mean Top-5 familiarity achieved an AURC of 0.2514 (95% CI 0.2152–0.2878) and ranked ninth numerically (Table VII). The Combined Random Forest ranked first (AURC 0.2096, 95% CI 0.1785–0.2423). Its difference from Mean Top-5 was -0.0419 (95% CI -0.0670 to -0.0162; unadjusted $p = 0.002$, $p_{\text{Holm}} = 0.022$ across 12 comparisons). Random referral was the only other comparator that differed significantly from Mean Top-5 ($p_{\text{Holm}} = 0.012$). Removing four melanoma-metastasis records changed its AURC only from 0.2514 to 0.2505.

On Derm7pt, error-detection AUROC was 0.6724 and AUPRC 0.5049. At 50% coverage, accepted accuracy was 77.75%, and the referred half contained 68.80% of all errors. Confidence intervals were wide and overlapping.

E. Threshold Transfer and Class-Specific Behaviour

The threshold designed for 50% validation coverage accepted 21.95% of ISIC 2019 and 39.36% of Derm7pt (Table VIII). Accepted predictions were more accurate than the unselective classifier in both cohorts, but the coverage difference shows that useful ranking did not preserve the operating point.

A matched-risk comparison provided a complementary view. At a selective-risk target of 0.35 on ISIC 2019, Mean Top-5 familiarity retained 53.75% coverage, whereas predictive entropy retained 39.81%. The difference was 13.94 percentage points, indicating a smaller referral workload for familiarity at that risk level. This advantage was not consistent across the complete operating range. It narrowed at a risk target of 0.40 and reversed at 0.45.

On ISIC 2019, 3,267 images were accepted and 11,618 referred; the referred set captured 86.22% of 6,873 errors, 8.17 percentage points above random referral at the same workload. On Derm7pt, 379 images were accepted and 584 referred; error capture was 75.80%, a 15.16-point lift over random referral. The rule therefore enriched both referred sets despite different acceptance rates.

Applying the same validation-threshold rule to compatible comparators showed that the shortfall was not unique to familiarity. ISIC 2019/Derm7pt coverages were 21.55%/37.80% for fifth-neighbour similarity, 17.71%/14.95% for SIRC-style risk, 22.18%/40.60% for Top-1 similarity, 18.40%/14.02% for entropy, 14.33%/29.39% for Mahalanobis risk, 18.37%/14.12% for maximum-softmax risk, 18.16%/14.43% for probability margin, and 19.13%/15.58% for generalized confidence margin. The Combined Random Forest achieved 19.73% and 37.59%. Thus, transferred thresholds generally undershot the intended workload, but the magnitude depended on the score and cohort. For retrieval-label support, the tied boundary at risk

TABLE VI. REFERRAL-SCORE COMPARISON ON 14,885 ISIC 2019 IMAGES. LOWER AURC AND RISK ARE BETTER; HIGHER ERROR CAPTURE IS BETTER

Rank	Referral score	AURC (95% CI)	Risk at 50%	Capture at 50%	p_{Holm}
1	Combined Random Forest	0.3245 (0.3143–0.3344)	0.3332	0.6392	0.013
2	Fifth-neighbour similarity†	0.3466 (0.3359–0.3571)	0.3369	0.6352	0.013
3	Adapted Δ-KNN-style	0.3490 (0.3386–0.3591)	0.3561	0.6144	0.870
4	**Mean Top-5 familiarity***	**0.3498 (0.3390–0.3603)**	**0.3394**	**0.6325**	Reference
5	SIRC-style	0.3565 (0.3460–0.3670)	0.3566	0.6139	0.368
6	Top-1 similarity	0.3569 (0.3460–0.3676)	0.3439	0.6277	0.013
7	Predictive entropy	0.3632 (0.3526–0.3736)	0.3667	0.6029	0.057
8	Mahalanobis risk	0.3633 (0.3522–0.3742)	0.3550	0.6156	0.013
9	Maximum-softmax risk	0.3649 (0.3543–0.3752)	0.3675	0.6021	0.044
10	Probability-margin risk	0.3664 (0.3559–0.3768)	0.3675	0.6021	0.025
11	Generalized confidence margin	0.3681 (0.3577–0.3785)	0.3705	0.5989	0.013
12	Generalized geometric margin	0.3752 (0.3651–0.3856)	0.3835	0.5848	0.013
13	Retrieval-label support‡	0.4122 (0.4013–0.4233)	0.3846	0.5836	0.013
14	Random referral	0.4617 (0.4537–0.4694)	0.4617	0.5000	0.013

*Primary score; †exploratory ablation. Holm-adjusted p values (13 comparisons, 2,000 paired resamples) test the paired-bootstrap null hypothesis that each method and Mean Top-5 familiarity have equal AURC. With the finite-bootstrap $+1$ correction, the smallest attainable two-sided unadjusted p value was approximately 0.001.
‡Tie-averaged evaluation was used because retrieval-label support has only six distinct score values.

TABLE VII. FROZEN REFERRAL-SCORE COMPARISON ON 963 DERM7PT IMAGES. HOLM-ADJUSTED p VALUES COMPARE EACH METHOD WITH MEAN TOP-5 FAMILIARITY

Rank	Referral score	AURC (95% CI)	p_{Holm}
1	Combined Random Forest	0.2096 (0.1785–0.2423)	0.022
2	Adapted Δ-KNN-style	0.2319 (0.1999–0.2661)	1.000
3	SIRC-style	0.2322 (0.1981–0.2679)	1.000
4	Predictive entropy	0.2354 (0.2010–0.2707)	1.000
5	Maximum-softmax risk	0.2378 (0.2029–0.2739)	1.000
6	Probability-margin risk	0.2397 (0.2040–0.2754)	1.000
7	Generalized confidence margin	0.2424 (0.2065–0.2783)	1.000
8	Fifth-neighbour similarity†	0.2498 (0.2135–0.2871)	1.000
9	**Mean Top-5 familiarity***	**0.2514 (0.2152–0.2878)**	Reference
10	Top-1 similarity	0.2574 (0.2200–0.2943)	1.000
11	Mahalanobis risk	0.2686 (0.2300–0.3063)	0.837
12	Retrieval-label support‡	0.2847 (0.2470–0.3201)	0.550
13	Random referral	0.3558 (0.3267–0.3849)	0.012

*Primary score; †exploratory ablation; ‡tie-averaged evaluation. Holm-adjusted p values (12 comparisons, 2,000 paired resamples) test the paired-bootstrap null hypothesis of equal AURC relative to Mean Top-5 familiarity. The generalized geometric margin was excluded because its stored Derm7pt implementation was not comparable with validation.

TABLE VIII. PERFORMANCE OF MEAN TOP-5 FAMILIARITY AFTER DIRECT TRANSFER OF THE VALIDATION-DERIVED THRESHOLD. ERROR CAPTURE IS THE PROPORTION OF ALL CLASSIFIER ERRORS PLACED IN THE REFERRED SET

Cohort	Base accuracy	Coverage	Accepted accuracy	Error capture
ISIC 2019	53.83%	21.95%	71.01%	86.22%
Derm7pt	64.38%	39.36%	78.10%	75.80%

0.2 gave strict/inclusive coverage bounds of 47.31%/62.28% on validation, 18.32%/46.04% on ISIC 2019, and 34.68%/64.90% on Derm7pt.

At 50% overall coverage, accepted accuracy improved for BCC, NV, and VASC but decreased for AKIEC, BKL, DF, and MEL (Table IX; Fig. 3). Within-class error AUROC was also below 0.5 for those four diagnoses, meaning that errors tended to receive lower risk—and higher familiarity—than correct cases.

Error capture was 77.26% for AKIEC, 92.83% for BCC, 56.69% for BKL, 79.79% for DF, 51.97% for MEL, 50.42% for NV, and 79.66% for VASC. AKIEC shows why capture cannot stand alone: 77.26% of its errors were referred, yet accepted accuracy was only 1.40%.

TABLE IX. CLASS-SPECIFIC PERFORMANCE OF MEAN TOP-5 FAMILIARITY ON ISIC 2019 AT 50% OVERALL COVERAGE. ERROR AUROC TREATS CLASSIFIER ERROR AS THE POSITIVE OUTCOME; VALUES BELOW 0.5 INDICATE AN UNFAVOURABLE WITHIN-CLASS RANKING

Class	Images	Accepted-n	Base accuracy	Coverage	Accepted accuracy	Error AUROC
AKIEC	737	143	15.88%	19.40%	1.40%	0.3637
BCC	2,809	1,102	51.33%	39.23%	91.11%	0.8790
BKL	1,525	550	40.20%	36.07%	28.18%	0.4278
DF	124	19	24.19%	15.32%	0.00%	0.3248
MEL	3,409	1,440	36.84%	42.24%	28.19%	0.3568
NV	6,170	4,142	72.97%	67.13%	80.03%	0.6906
VASC	111	46	46.85%	41.44%	73.91%	0.8080

The reversal was not shared by every score. Melanoma accepted accuracy was 28.19% for Mean Top-5 and 18.53% for adapted Δ-KNN, versus 47.93% for entropy, 46.80% for SIRC-style scoring, and 36.84% without selection. Entropy reversed in no class; adapted Δ-KNN reversed in AKIEC and MEL, and SIRC-style scoring reversed marginally in AKIEC (-0.17 percentage points). The melanoma finding is therefore particularly relevant to neighbourhood-based selection.

Predicted-class stratification reached a similar conclusion without using the unknown reference diagnosis at decision time. Among 2,363 cases predicted as melanoma, 35.42% were accepted and accuracy decreased from 53.15% overall to 48.51% in the accepted subset. Only 15 of 673 cases predicted as AKIEC were accepted (2.23% coverage; 13.33% accepted accuracy), and no case predicted as DF was accepted. Of 3,409 true melanomas, the classifier misclassified 2,153; the familiarity rule accepted 1,034 of these missed melanomas (48.03%; bootstrap 95% CI for the count, 974–1,091).

The illustrative cost-weighted analysis was also unfavourable to familiarity as a stand-alone safety score. On ISIC 2019, weighted AURC was 1.0069 for Mean Top-5, 0.9407 for the Combined Random Forest, 0.8968 for entropy, and 0.8952 for SIRC-style risk. On Derm7pt, the corresponding Mean

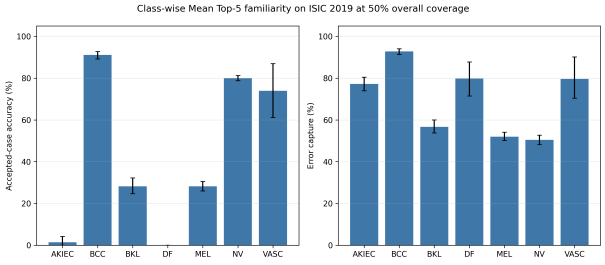

Fig. 3. Class-specific performance of Mean Top-5 familiarity on ISIC 2019 at 50% overall coverage. Accepted-case accuracy and error capture varied substantially across diagnoses. Error bars show 95% bootstrap confidence intervals.

Top-5 and Combined Random Forest values were 0.9198 and 0.8047. The weights were illustrative; they do not constitute a clinically validated utility function.

The secondary experiments did not show that a more complicated formulation would resolve these weaknesses. The primary Mean Top-5 score searched the complete reference bank, regardless of the classifier's predicted category. A second formulation restricted the search to reference images belonging to the predicted class. This restriction increased validation AURC from 0.1162 to 0.1222 and produced an ISIC 2019 AURC of 0.3752, compared with 0.3498 for the complete-bank score. Restricting the neighbourhood to the predicted class therefore worsened error ranking rather than improving it.

A feature-pipeline audit showed that the earlier Random Forest analysis had used inconsistent feature definitions across development and external cohorts. After rebuilding the features consistently, the Combined Random Forest achieved a five-fold out-of-fold validation AURC of 0.0656. Its external AURCs were 0.3245 on ISIC 2019 and 0.2096 on Derm7pt. A binary disagreement sensitivity analysis gave similar external AURCs of 0.3217 and 0.2099. The corrected result shows that combining familiarity with classifier and retrieval-agreement information was more effective than familiarity alone. The three design analyses are summarized in Table X.

F. Computational Cost

Profiling on an Intel Core i7-1165G7 CPU gave a mean preprocessing time of 29.89 ms and a mean ConvNeXt feature-extraction time of 86.80 ms per image. Exact cosine search over all 1,002 reference embeddings took 3.02 ms per query (1,000 repetitions). The raw FP32 bank occupied 2.94 MiB. The resulting online time was about 119.7 ms per image, of which retrieval accounted for roughly 2.5%. Search time rose from 0.20 ms for 125 references to 3.02 ms for 1,002 references. These measurements describe the present CPU implementation; larger banks may require indexed approximate search and separate validation of any accuracy–latency trade-off.

These measurements describe single-process execution on the specified research computer and do not constitute a complete deployment benchmark. Concurrent-request throughput, p50/p95/p99 latency, peak memory under load, queue behaviour, and performance with substantially larger or dynamically updated banks were not evaluated. No claim regarding production scalability is made.

G. Representative Retrieval Outcomes

Fig. 4 illustrates both the value and the limit of the score. A single example was selected descriptively from each of the four combinations of prediction correctness and accept–refer decision. The cases were chosen to illustrate contrasting outcomes; they were not randomly sampled and should not be interpreted as estimates of how frequently those outcomes occurred. The accepted correct NV query retrieved visually similar pigmented lesions, although the neighbour labels were mixed. The referred AKIEC prediction was pale and low contrast, with circular vignetting and comparatively weak similarities. The most revealing counterexample was an NV misclassified as melanoma and accepted: its close neighbours shared dark pigmentation and shape, while hair and a peripheral red marking were also visible. Conversely, a correct NV was referred despite visible ruler markings and hair. These observations cannot establish that an artefact caused a retrieval decision, but they show why visual inspection of the neighbours remains necessary.

V. Discussion

A. Principal Findings

Mean Top-5 familiarity retained error information in both external cohorts but was not consistently strongest. It ranked fourth on ISIC 2019 and ninth on Derm7pt. The Combined Random Forest ranked first in both cohorts, showing that familiarity became more useful when combined with classifier confidence and retrieval-agreement information. Mean Top-5 still outperformed several single-score baselines on ISIC 2019, but its relative value changed with the cohort and evaluation criterion.

Top-1 similarity performed worse than Mean Top-5 on ISIC 2019, whereas the fifth neighbour performed better. The boundary neighbour may represent local support more clearly than the mean in this embedding space. The broader search also favoured 20th-neighbour similarity in validation and both external cohorts. Both findings are exploratory: the latter was selected from 35 configurations on one validation set and requires independent confirmation.

The feature audit was consequential. The earlier Random Forest result had combined validation similarities from an obsolete retrieval pipeline with external similarities from the final pipeline, and disagreement had been defined continuously during development but as a binary flag externally. Harmonizing these definitions reversed the conclusion: the combined score outperformed Mean Top-5 in both cohorts. A referral model can therefore appear to fail under shift when its feature pipeline has shifted internally before the external cohort is considered.

The primary familiarity score requires one deterministic forward pass and no separately learned risk model. Its neighbours are inspectable, although they neither explain causality nor prove a diagnosis. Super-TrustScore instead includes a class-conditional local component [12], which may matter when common diagnoses dominate a global bank.

TABLE X. SECONDARY ANALYSES OF REFERENCE-BANK AND REFERRAL-SCORE DESIGN. MEAN TOP-5 FAMILIARITY USING THE COMPLETE REFERENCE BANK WAS SELECTED AS THE PRIMARY SCORE BEFORE EXTERNAL EVALUATION. LOWER AURC IS BETTER

Analysis	Validation AURC	ISIC 2019 AURC	Interpretation
Mean Top-5 using the complete reference bank	0.1162	0.3498	Primary score selected on validation
Mean Top-5 using predicted-class references only	0.1222	0.3752	No improvement
Combined Random Forest	0.0656†	0.3245	Consistent combined features improved ranking

†Five-fold out-of-fold validation AURC; the model was refitted on all validation cases and frozen before external evaluation. Its Derm7pt AURC was 0.2096.

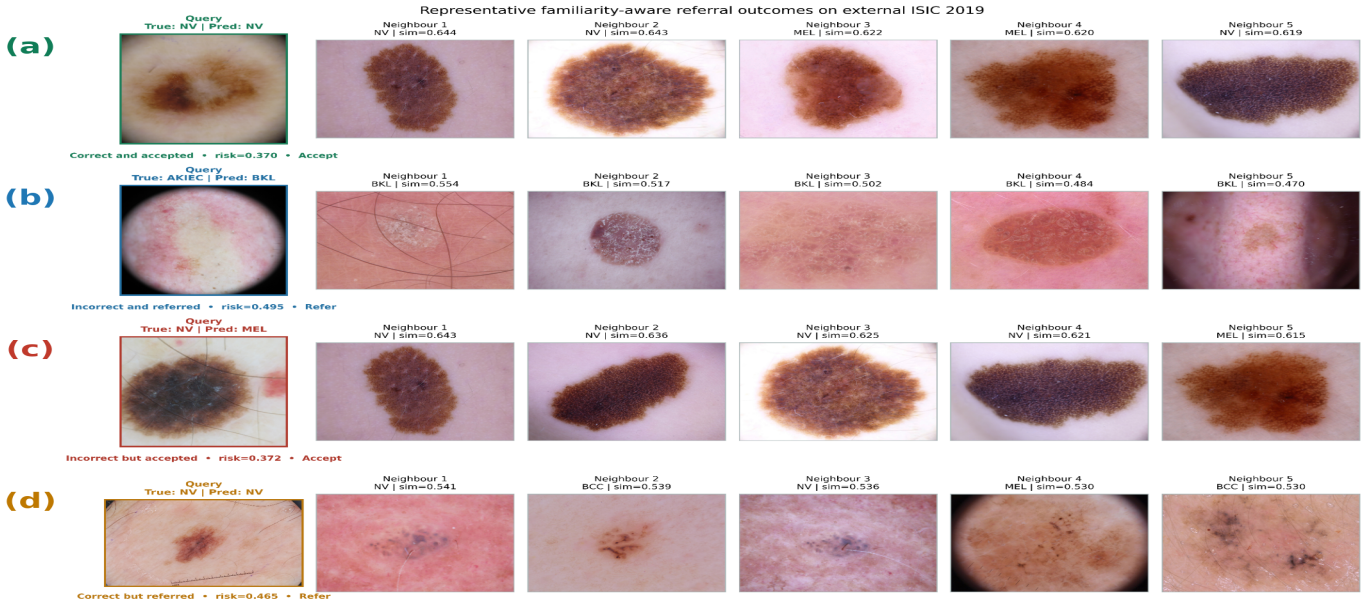

Fig. 4. Descriptively selected ISIC 2019 queries and their five nearest HAM10000 reference cases: (a) A correct prediction that was accepted; (b) An incorrect prediction that was referred; (c) An incorrect prediction that was accepted; and (d) A correct prediction that was referred. Hair, vignetting, ruler markings, colour variation, and lesion scale were examined as possible influences on visual similarity. The cases were selected to illustrate the four possible referral outcomes rather than to estimate their frequency. Retrieved similarity reflects representation-space proximity and should not be interpreted as diagnostic evidence.

Adapted Δ-KNN was similar to or numerically better than Mean Top-5, but it builds neighbourhoods using development correctness labels and is therefore not training-free in the same sense.

B. Ranking Did Not Guarantee Threshold Transfer

The clearest result was the separation between ranking and operating-point transfer. Mean Top-5 ranked errors better than random referral in both cohorts, yet its threshold—designed for 50% validation coverage—produced 21.95% coverage on ISIC 2019 and 39.36% on Derm7pt. Ranking travelled better than workload.

Equivalent validation-derived thresholds were transferred for all comparators with compatible development and external score definitions. Most produced coverage well below 50% in both external cohorts. The Combined Random Forest reached 19.73% on ISIC 2019 and 37.59% on Derm7pt; entropy reached 18.40% and 14.02%, respectively. Threshold instability was therefore not unique to retrieval familiarity, and its severity depended on both score and cohort.

A service with fixed review capacity cannot assume that an imported threshold will preserve referral volume. Local calibration may stabilize workload, but it requires suitable data and becomes an adapted, not fully transferred, policy. AURC alone would hide this instability.

A possible alternative is to estimate a target-domain score quantile using unlabelled local cases and adjust the threshold to a specified referral capacity. Such a procedure could stabilize workload but would constitute target-domain workload recalibration rather than frozen threshold transfer. It would also provide no guarantee regarding accepted-set error without labelled outcomes. This alternative was not evaluated in the present study.

At a selective-risk target of 0.35, familiarity retained almost 14 percentage points more coverage than entropy, but the advantage nearly disappeared at 0.40 and reversed at 0.45. Its workload benefit was therefore confined to the lower-risk region examined.

C. Implications for Deployment and System Design

Deployment should be staged. Familiarity should first be audited on local data; if its ranking remains useful, a threshold can be chosen for local referral capacity. This is adaptation,

not transfer of the original operating point, and changes in acquisition or case mix require continued monitoring.

Accepted outputs would remain decision support. Referred cases could enter a dermatologist or teledermatology queue, with urgency determined from clinical information beyond this image-only model; referral alone would not recommend biopsy. This workflow was not tested here.

Neighbours may reveal artefacts, duplicates, or unexpected matches but do not explain the classifier or establish a diagnosis. An interface should separate the prediction, referral recommendation, and retrieved examples.

Because bank updates change referral scores without changing classifier weights, a deployed bank requires version control, provenance and duplicate checks, label verification, and repeated class-wise evaluation. These controls are especially important for rare diagnoses.

Familiarity is not a replacement for probability calibration or formal coverage guarantees. It asks whether a query lies near observed cases in the chosen representation. Any combination of these signals still requires local evaluation.

D. Proposed Human-Review Evaluation

No dermatologist reader study was conducted. Consequently, the computational results do not establish whether retrieved neighbours improve diagnostic accuracy, reduce review time, or cause automation bias. A prospective evaluation should first record each dermatologist's unaided diagnosis, confidence, and review time while withholding the classifier output and retrieved labels. The referral recommendation and retrieved images could then be displayed, followed by recording the revised diagnosis, confidence, and additional review time. Retrieved labels should initially remain hidden or be evaluated under randomized display conditions to measure anchoring and automation bias. Diagnostic disagreements should be resolved by an independent expert panel using the clinical reference standard. Relevant outcomes include diagnostic accuracy, malignancy sensitivity, decision-change frequency, review time, inter-reader agreement, and harmful changes from an initially correct to an incorrect decision. This protocol is proposed for future work and was not evaluated in the present computational study.

E. Class-Wise Reversal

This analysis was stratified by reference-standard diagnosis and was conducted only as a retrospective failure audit. True diagnosis was unavailable to the referral mechanism and could not be used to make a class-conditional decision at inference time. These findings therefore identify possible failure patterns but do not define a deployable diagnosis-specific referral policy.

Aggregate results concealed a potentially important failure pattern. At 50% overall coverage, accepted accuracy fell for AKIEC, BKL, DF, and MEL, and their within-class error AUROCs were below 0.5. The score ranked these classes in the wrong direction rather than merely weakly.

Bank composition is one plausible explanation. Because HAM10000 is dominated by nevi, melanoma resemblance to nevi can increase familiarity while also causing an NV prediction. Rare diagnoses also have little local support. This mechanism remains a hypothesis: subsampling and small-bank balancing changed AURC but did not isolate its cause. Entropy did not reverse in any class, whereas adapted Δ-KNN also reversed for AKIEC and melanoma and SIRC-style scoring reversed marginally for AKIEC, so the problem was not universal but was not unique to Mean Top-5.

More generally, proximity can look useful overall while selecting a minority group unfavourably. Class-stratified analysis is therefore necessary when error costs differ. Restricting retrieval to the predicted class did not solve the problem and may make an incorrect prediction appear self-consistent.

Rebalanced banks or class-specific thresholds deserve study, although rare classes provide little calibration support. SIRC-style scoring already combined familiarity with softmax confidence but did not solve cross-cohort transfer. A better design must address class imbalance and within-class selection directly, be fixed without external labels, and be evaluated with class-wise curves.

F. Limitations

All results come from one ConvNeXt-Tiny model, trained once on a single split and seed, with one embedding layer and one HAM10000 reference bank. Stronger classifiers, other representations, repeated training runs, or a separately collected bank could change the rankings. Exact HAM10000 identifiers were removed from ISIC 2019, but overlap at the patient or lesion level could not be ruled out. Because the bootstrap resampled images rather than lesions, its intervals, including the one for accepted misclassified melanomas, may be too narrow.

Similarity between neighbours may reflect hair, rulers, background, or illumination rather than the lesion itself; no masking or perturbation experiments were run to separate these effects. The class-specific, predicted-class, and cost-weighted analyses were exploratory, some class subsets were small, and the cost weights were chosen for illustration, not derived clinically.

Two comparators were left out of threshold transfer: adapted Δ-KNN, whose validation score was computed in-sample, and the Derm7pt geometric margin, whose stored values were on an incompatible scale. The Combined Random Forest needs validation correctness labels and is therefore not training-free. Finally, no reader study, artefact-controlled analysis, multi-architecture comparison, or load test was carried out. The results therefore do not establish clinical safety, benefit to human reviewers, stability across architectures, or production scalability.

VI. Conclusion

This external audit found that Mean Top-5 retrieval familiarity retained some error-ranking information across two shifted dermoscopic cohorts but was not consistently superior to alternative referral scores. It ranked fourth on ISIC 2019 and ninth on Derm7pt. The Combined Random Forest ranked first, with AURCs of 0.3245 and 0.2096 compared with 0.3498 and 0.2514 for Mean Top-5. The result supports familiarity

as complementary information rather than a sufficient safety score. The selected value $k = 5$ was only marginally favoured on validation data and should not be regarded as a generally optimal retrieval depth.

The results also distinguished ranking performance from operating-point transfer. Validation thresholds targeting 50% coverage produced markedly lower coverage externally for familiarity, the combined model, and most compatible comparators. True- and predicted-class analyses identified unfavourable selection patterns, and the familiarity policy accepted 1,034 of 2,153 misclassified melanomas at 50% coverage. The illustrative cost-weighted analysis likewise did not support familiarity as a stand-alone safety rule.

The findings are conditional on one classifier, representation, training run, reference bank, and image-level evaluation procedure. They do not establish clinical safety, human-review benefit, architecture-independent stability, or production deployment performance. Retrieval familiarity is therefore best viewed as an inspectable complementary signal whose operating threshold, reference-bank composition, neighbourhood definition, feature consistency, and class-level behaviour require independent validation. Future studies should evaluate multiple representations and training seeds, clinically elicited cost-sensitive outcomes, artefact-controlled retrieval, independent reference banks, and prospective clinician interaction.

Statements and Declarations

Funding

The authors received no specific funding for this study.

Competing Interests

The authors declare that they have no competing interests.

Ethics Approval

Not applicable. This retrospective computational study used publicly available, de-identified dermoscopic datasets and involved no prospective recruitment, participant contact, intervention, or access to direct personal identifiers. The source publications describe the ethical approval and consent procedures for the original dataset collections.

Consent to Participate

Not applicable to this secondary analysis of publicly available, de-identified data.

Consent for Publication

Not applicable.

Data Availability

HAM10000, ISIC 2019, and Derm7pt are publicly available from their respective repositories under their original licences. Partition identifiers and derived evaluation results are available from the corresponding authors upon reasonable request.

Code Availability

Analysis code and configuration files are available from the corresponding authors upon reasonable request.

Declaration of Generative AI and AI-Assisted Technologies

During manuscript preparation, the authors used ChatGPT (OpenAI) to assist with language editing and organization. The authors reviewed and approved the final text and remain responsible for its content.

References

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

[2] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, and A. Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC)," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018, pp. 168–172. [Online]. Available: https://doi.org/10.1109/ISBI.2018.8363547

[3] S. Yan, Z. Yu, C. Primiero, C. Vico-Alonso, Z. Wang, L. Yang, P. Tschandl, M. Hu, L. Ju, G. Tan, V. Tang, A. B. Ng, D. Powell, P. Bonnington, S. See, E. Magnaterra, P. Ferguson, J. Nguyen, P. Guitera, J. Bañuls, M. Janda, V. Mar, H. Kittler, H. P. Soyer, and Z. Ge, "A multimodal vision foundation model for clinical dermatology," *Nature Medicine*, vol. 31, no. 8, pp. 2691–2702, 2025. [Online]. Available: https://doi.org/10.1038/s41591-025-03747-y

[4] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. V. Dillon, B. Lakshminarayanan, and J. Snoek, "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift," in *Advances in Neural Information Processing Systems*, vol. 32, 2019, pp. 13 991–14 002. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/8558cb408c1d76621371888657d2eb1d-Abstract.html

[5] N. Jairath, V. Pahalyants, R. Shah, J. Weed, J. A. Carucci, and M. C. Criscito, "Artificial intelligence in dermatology: A systematic review of its applications in melanoma and keratinocyte carcinoma diagnosis," *Dermatologic Surgery*, vol. 50, no. 9, pp. 791–798, 2024. [Online]. Available: https://doi.org/10.1097/DSS.0000000000004223

[6] J.-W. Tjiu and C.-F. Lu, "Equity and generalizability of artificial intelligence for skin-lesion diagnosis using clinical, dermoscopic, and smartphone images: A systematic review and meta-analysis," *Medicina*, vol. 61, no. 12, p. 2186, 2025. [Online]. Available: https://doi.org/10.3390/medicina61122186

[7] R. El-Yaniv and Y. Wiener, "On the foundations of noise-free selective classification," *Journal of Machine Learning Research*, vol. 11, pp. 1605–1641, 2010. [Online]. Available: https://www.jmlr.org/papers/v11/el-yaniv10a.html

[8] Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," in *Advances in Neural Information Processing Systems*, vol. 30, 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html

[9] L. F. P. Cattelan and D. Silva, "How to fix a broken confidence estimator: Evaluating post-hoc methods for selective classification with deep neural networks," in *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, ser. Proceedings of Machine Learning Research, vol. 244. PMLR, 2024, pp. 547–584. [Online]. Available: https://proceedings.mlr.press/v244/cattelan24a.html

[10] Y. Sun, Y. Ming, X. Zhu, and Y. Li, "Out-of-distribution detection with deep nearest neighbors," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 162. PMLR, 2022, pp. 20 827–20 840. [Online]. Available: https://proceedings.mlr.press/v162/sun22d.html

[11] P. Tschandl, G. Argenziano, M. Razmara, and J. Yap, "Diagnostic accuracy of content-based dermatoscopic image retrieval with deep classification features," *British Journal of Dermatology*, vol. 181, no. 1, pp. 155–165, 2019. [Online]. Available: https://doi.org/10.1111/bjd.17189

[12] J. Naushad and I. Voiculescu, "Super-TrustScore: Reliable failure detection for automated skin lesion diagnosis," in *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*. Athens, Greece: IEEE, 2024, pp. 1–4. [Online]. Available: https://doi.org/10.1109/ISBI56570.2024.10635815

[13] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 976–11 986. [Online]. Available: https://doi.org/10.1109/CVPR52688.2022.01167

[14] J. Kawahara, S. Daneshvar, G. Argenziano, and G. Hamarneh, "Seven-point checklist and skin lesion classification using multi-task multi-modal neural nets," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 2, pp. 538–546, 2019. [Online]. Available: https://doi.org/10.1109/JBHI.2018.2824327

[15] C. Hernández-Pérez, M. Combalia, S. Podlipnik, N. C. F. Codella, V. Rotemberg, A. C. Halpern, O. Reiter, C. Carrera, A. Barreiro, B. Helba, S. Puig, V. Vilaplana, and J. Malvehy, "BCN20000: Dermoscopic lesions in the wild," *Scientific Data*, vol. 11, no. 1, p. 641, 2024. [Online]. Available: https://doi.org/10.1038/s41597-024-03387-w

[16] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 70. PMLR, 2017, pp. 1321–1330. [Online]. Available: https://proceedings.mlr.press/v70/guo17a.html

[17] Y. Geifman and R. El-Yaniv, "SelectiveNet: A deep neural network with an integrated reject option," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 97. PMLR, 2019, pp. 2151–2159. [Online]. Available: https://proceedings.mlr.press/v97/geifman19a.html

[18] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=Hkg4TI9xl

[19] G. Xia and C.-S. Bouganis, "Augmenting the softmax with additional confidence scores for improved selective classification with out-of-distribution data," *International Journal of Computer Vision*, vol. 132, pp. 3714–3752, 2024. [Online]. Available: https://doi.org/10.1007/s11263-024-02029-3

[20] H. Liang, L. Peng, and J. Sun, "Selective classification under distribution shifts," *Transactions on Machine Learning Research*, 2024. [Online]. Available: https://openreview.net/forum?id=dmxMGW6J7N

[21] A. Heng and H. Soh, "Know when to abstain: Optimal selective classification with likelihood ratios," in *International Conference on Learning Representations*, 2026. [Online]. Available: https://proceedings.iclr.cc/paper_files/paper/2026/hash/3fe2a777282299ecb4f9e7ebb531f0ab-Abstract-Conference.html

[22] J. Traub, T. J. Bungert, C. T. Lüth, M. Baumgartner, K. H. Maier-Hein, L. Maier-Hein, and P. F. Jaeger, "Overcoming common flaws in the evaluation of selective classification systems," in *Advances in Neural Information Processing Systems*, vol. 37, 2024, pp. 2323–2347. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/hash/047c84ec50bd8ea29349b996fc64af4b-Abstract-Conference.html

[23] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," in *Advances in Neural Information Processing Systems*, vol. 31, 2018, pp. 7167–7177. [Online]. Available: https://proceedings.neurips.cc/paper/2018/hash/abdeb6f575ac5c6676b747bca8d09cc2-Abstract.html

[24] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of Multivariate Analysis*, vol. 88, no. 2, pp. 365–411, 2004. [Online]. Available: https://doi.org/10.1016/S0047-259X(03)00096-4

[25] P. Tschandl, C. Rinner, Z. Apalla, G. Argenziano, N. Codella, A. Halpern, M. Janda, A. Lallas, C. Longo, J. Malvehy, J. Paoli, S. Puig, C. Rosendahl, H. P. Soyer, I. Zalaudek, and H. Kittler, "Human–computer collaboration for skin cancer recognition," *Nature Medicine*, vol. 26, no. 8, pp. 1229–1234, 2020. [Online]. Available: https://doi.org/10.1038/s41591-020-0942-0