International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 10 | Oct - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Mitigating Skin Tone Bias in Lesion Screening via Confidence-Based Routing

M. Mukesh Kumar, A. Jayadithya, L. Varshitha, M. Bhagya Sri, Khasim Syed*
School of Computer Science and Engineering (SCOPE), Vellore Institute of Technology, Andhra Pradesh (VIT-AP University), Amaravati, Andhra Pradesh, India-522241
Corresponding author: Khasim Syed (syed.khasim@vitap.ac.in)

Abstract*—Deep-learning systems for skin-lesion screening are usually reported with one aggregate accuracy on one dataset, which says little about how they behave across skin tones, on images from other sources, or on photographs that are not lesions at all. We built DermaScan AI, a web-based pipeline that combines a DermNet-trained ResNet18 pre-filter, a three-tier (benign, monitor, urgent) ResNet18 risk classifier, Grad-CAM explanations, and a confidence gate whose threshold depends on the user's Fitzpatrick skin type. Our questions were how a baseline classifier behaves across skin types, whether a changed training recipe helps, and how far accuracy falls on data from another source. A baseline trained on HAM10000 reached 80.84% validation accuracy but only 44.2–51.9% on held-out Fitzpatrick17k images, depending on skin type. Retraining on HAM10000 plus half of the lesion-relevant Fitzpatrick17k images, with a lower learning rate, a scheduler and rotation augmentation (v3), raised validation accuracy to 88.69% and held-out accuracy to 54.4–74.8%. The gains for types I–V (14.2–25.5 points) exceeded run-to-run variation over three seeds (SD 1.1–2.2 points); the Type VI gain (2.5 points) did not. The gap between the best and worst skin type widened from 7.7 to 20.3 points. On ISIC 2019 images that are not in HAM10000 ($n = 2,403$), balanced accuracy was 55.70% against 80.50% on the HAM10000 validation split. The pre-filter reached 86.20% in-domain. Splits are image-level, labels are tiers we derived from dataset diagnoses, no clinician reviewed any output, and the routing rule was not evaluated by risk–coverage analysis, so these results say nothing about clinical safety.

***Index Terms*—skin lesion classification, dermatology image analysis, Fitzpatrick skin type, domain shift, Grad-CAM, selective prediction, explainable AI**

I. Introduction

Convolutional networks can classify skin lesions at a level comparable to dermatologists on curated test sets [1]. Deployment is harder than the headline numbers suggest. Performance has been reported to drop on darker skin tones and on rarer conditions [2], [3], and a classifier trained on dermoscopy images will still return a confident label when it is shown a rash, a bite or an unrelated photograph. A screening tool therefore has to answer three questions beyond "what is the accuracy": does it work comparably across skin types, what happens when images come from a different source, and what does it do with inputs it should not classify?

We started from these three questions after reading six studies on skin-lesion classification and explainability (Section II). Several of them state, in their own limitations, that they did not test under domain shift or could not analyse demographic bias, and none described an input check for non-lesion photographs. We then built a working system around those gaps and measured it. DermaScan AI sends every uploaded photograph through a pre-filter that separates common non-lesion conditions from images that need lesion assessment, classifies the remainder into three tiers, shows a Grad-CAM heatmap [4], and withholds the automatic result when the model's confidence is below a threshold chosen by skin type, in the spirit of selective classification [5].

Research questions. *RQ1:* How does a ResNet18 trained on HAM10000 perform across Fitzpatrick skin types, and does retraining with skin-tone-labelled data change this? *RQ2:* How much does accuracy fall on a source the model was not tuned for? *RQ3:* Can a pre-filter and a skin-type-dependent confidence gate be integrated into a usable screening workflow?

Contributions. (1) A matched evaluation protocol: the same held-out Fitzpatrick17k images are scored for every model version, giving per-skin-type results for a baseline (v1), a first retraining (v2) and a tuned retraining (v3). (2) A repeated-run check (three seeds) that separates real changes from training noise, which shows that the Type VI result is not distinguishable from noise. (3) An end-to-end implementation (React, Spring Boot, Flask/PyTorch, PostgreSQL) of the pre-filter, the confidence gate and Grad-CAM, with reports and a scan log. (4) A frank account of what did not work: v2 gave mixed results, the best-to-worst skin-type gap grew after retraining, accuracy fell sharply on another source, and the routing rule has not been validated on risk and coverage. We do not claim state-of-the-art accuracy, and our numbers cannot be compared with those of other papers because the label spaces and splits differ.

II. Related Work

A. Datasets and Classification

HAM10000 provides 10,015 dermoscopy images in seven diagnostic classes [6] and is part of the ISIC archive and challenges [7]; the ISIC 2019 release additionally contains BCN_20000 and other sources [8]. DermNet NZ supplies clinical photographs of many general skin conditions, which we use to teach the pre-filter what non-lesion images look like [9].

B. Reviewed Studies

We reviewed six studies in detail (Table I). A multi-condition EfficientNetB0 system with a decision engine reports 87.2% over nine classes [10]. A custom CNN reaches 98% on HAM10000, against 83% for ResNet50V2 and 67% for VGG19, from a single dataset [11]. A multi-task Grad-CAM system for basal cell carcinoma reports 90% for BCC versus non-BCC, validated against dermatologist segmentations at one centre [12]. A reader study found 81% accuracy for an explainable classifier against 66.2% balanced accuracy for clinicians, and its authors state that the system was tested under artificial conditions and that domain shift was not addressed [13]. DUNEScan is a web server for uncertainty estimation whose authors note that skin-lesion datasets are less standardized than typical image datasets (different doctors, cameras and lighting) [14]. A benchmark of Grad-CAM across 11 CNNs on DermXDB states that the dataset carries little information on gender, age and ethnicity, which prevented an in-depth bias analysis [15].

C. Skin-Tone Fairness and Evaluation

Skin colour is commonly described with the Fitzpatrick scale (types I–VI) [16]. Fitzpatrick17k attaches these labels to about 16,500 clinical images and was used to show accuracy differences between skin types [2]. A curated, pathology-confirmed set of diverse images reported lower algorithm performance on dark skin tones and uncommon conditions, and found that fine-tuning on diverse images reduced the gap [3]. Our contribution is not to rediscover this effect but to measure it on our own pipeline, change the training, and report what changed, including the unfavourable parts.

D. Abstention, Calibration and Explanation

Selective classification lets a model refuse low-confidence inputs and evaluates the resulting risk and coverage [5]. Softmax confidences of modern networks are often overconfident [17], so a threshold on them should be treated cautiously. Grad-CAM highlights regions that increased a class score [4]; it does not prove how the model reasons.

Gap. Among the studies we reviewed, we did not find one that (a) filters non-lesion inputs before risk classification, (b) reports per-skin-type results for successive training changes on matched images, and (c) ties the acceptance threshold to measured per-skin-type reliability, in one working system. This is a narrow statement from a limited review, not a novelty claim; related ideas exist [3], [5].

III. Proposed Method

A. System Overview

Fig. 1 shows the pipeline. A React front end sends the photograph, and optionally the user's Fitzpatrick type, to a Spring Boot service. The service forwards the image to a Python Flask service that hosts both models and the Grad-CAM module. The Flask service returns predictions, the three-class probability breakdown and the heatmap; Spring Boot applies the routing rule, stores the record in PostgreSQL and returns it to the browser.

B. Preprocessing

The Flask service converts any image mode to RGB and downsizes photographs whose longer side exceeds 1,600 pixels (Lanczos resampling), because phone images of 3,000–4,000 pixels slowed Grad-CAM and overloaded memory. For the networks, the image is resized to 224×224 and normalized with ImageNet mean and standard deviation.

C. Stage 1: Pre-Filter

A ResNet18 [18] initialised from ImageNet weights, with its last layer replaced by two outputs, is fine-tuned on DermNet. The 23 DermNet folders are merged into two classes by the team: `needs_lesion_check` (malignant lesions, melanoma and nevi, seborrheic keratoses and benign tumours, vascular tumours) and `harmless_common` (the other 19 folders, such as eczema, psoriasis, fungal and viral infections). If the pre-filter returns `harmless_common`, the pipeline stops and the user sees an informational message; the core model is not run.

D. Stage 2: Core Risk Model

A second ResNet18 outputs $K = 3$ logits that are turned into probabilities $p_k = \mathrm{softmax}(z)_k$, with confidence $c = \max_k p_k$. The tiers are derived from HAM10000 diagnoses: *benign* (nv, bkl, vasc, df), *monitor* (akiec) and *urgent* (mel, bcc). To counter class imbalance we use a cross-entropy loss with class weights

$$w_k = \frac{N}{K\, n_k}, \tag{1}$$

where n_k is the number of training images in tier k and N the total. Three versions were trained (Section IV); v3 is deployed.

E. Skin-Type-Dependent Confidence Gate

Let $g \in \{\mathrm{I}, \ldots, \mathrm{VI}\}$ be the selected skin type and a_g the held-out accuracy of the deployed model for that type. We map lower measured accuracy to a stricter threshold,

$$\tau_g = \tau_{\max} - \frac{a_g - a_{\min}}{a_{\max} - a_{\min}} (\tau_{\max} - \tau_{\min}), \tag{2}$$

with $\tau_{\min} = 0.60$, $\tau_{\max} = 0.80$, and $a_{\min}$, $a_{\max}$ the lowest and highest a_g. The gate is

$$\text{decision}(x, g) = \begin{cases} \text{show tier} & c(x) \geq \tau_g, \\ \text{uncertain} & c(x) < \tau_g, \end{cases} \tag{3}$$

TABLE I
REVIEWED STUDIES. HEADLINE FIGURES ARE AS REPORTED BY THE AUTHORS AND ARE NOT COMPARABLE WITH EACH OTHER OR WITH OUR RESULTS.

Study	Task and data	Headline result	Gap relevant to this work
IJCRT 2025 [10]	EfficientNetB0, 9 classes, decision engine	87.2% accuracy	Single aggregate figure; skin-tone claim without a protocol we could verify
Agustin & Nurhaida [11]	Custom CNN vs. ResNet50V2, VGG19; HAM10000	98% / 83% / 67%	One dataset; no skin-tone or external evaluation; "bias" discussed only as class imbalance
BCC dual-explanation [12]	BCC vs. non-BCC; multi-task + Grad-CAM	90% accuracy	One condition; single-centre validation
Multimodal XAI reader study [13]	Explainable classifier vs. clinicians	81% vs. 66.2% (balanced)	Authors: artificial test conditions; domain shift not addressed
DUNEScan [14]	Uncertainty-estimation web server	No single accuracy	Authors: small, non-standardized training images
DermXDB Grad-CAM benchmark [15]	Grad-CAM on 11 CNNs	Mean explainability score 0.46	Authors: no demographic metadata for bias analysis

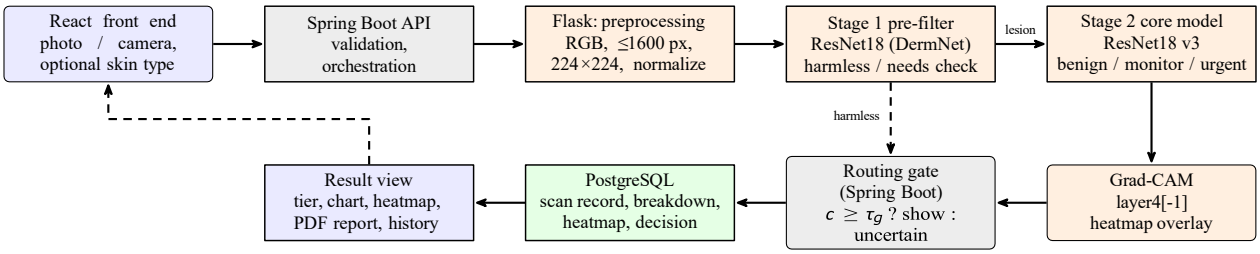


Fig. 1. DermaScan AI architecture and request flow. Solid arrows: lesion path; dashed arrow: images the pre-filter calls harmless skip Stage 2 and Grad-CAM.

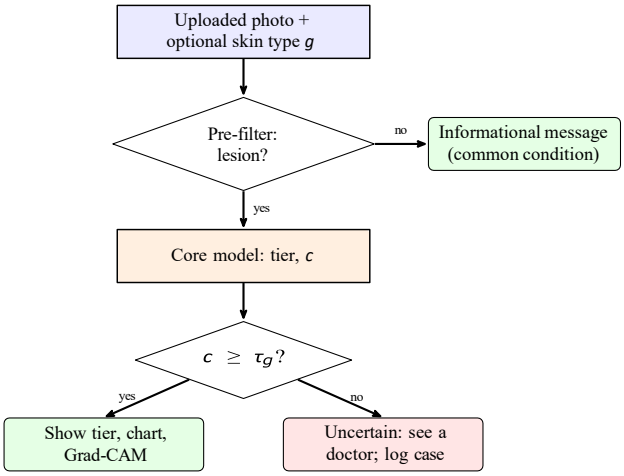


Fig. 2. Decision logic from upload to result.

and $\tau = 0.65$ is used when no skin type is given. The thresholds are computed offline, stored in a JSON file and loaded by the backend at start-up. An "uncertain" outcome shows no tier and advises the user to see a doctor; the case is also logged. Fig. 2 summarises the decision logic. The bounds 0.60 and 0.80 are design choices of ours, not fitted values.

F. Explainability and Output

Grad-CAM weights the feature maps A^k of the last convolutional block (`layer4[-1]`) by the pooled gradient of the predicted class score y^c,

$$L^c = \mathrm{ReLU}\left(\sum_k \alpha_k A^k\right), \quad \alpha_k = \frac{1}{Z}\sum_{i,j} \frac{\partial y^c}{\partial A^k_{ij}}, \tag{4}$$

computed with the `pytorch-grad-cam` library. The map is resized to the original photograph's resolution and overlaid on it. The interface shows the original and the overlay side by side, a bar chart of the three class probabilities, and a two-step indicator of which stages ran. A report is produced in the browser (jsPDF) from the same data. A session history and a log of uncertain cases are kept; the log is not editable by users because it is meant as an audit record.

IV. EXPERIMENTAL SETUP

A. Datasets and Splits

Table II lists the data. Fitzpatrick17k images without a skin-type label (565 of 16,577) were dropped, leaving 16,012. For model development we kept only the 40 of 114 Fitzpatrick17k conditions that we judged lesion-like (e.g., melanoma, carcinomas, nevi, seborrheic keratosis, keloid), which gives 4,817 images, and mapped them to the same three tiers (8 labels to urgent, 5 to monitor, 27 to benign). This subset was split 50/50, stratified by tier and skin type: 2,408 images were added to the v2/v3 training set and 2,409 were held out and never used for training or checkpoint selection. The held-out counts per skin type are 484, 753, 507, 380, 206 and 79 (types I–VI). The combined training set has 10,920 images (benign 8,095, monitor 565, urgent 2,260). Only 288 of them (2.6%) are type V or VI, and HAM10000 itself carries no skin-type labels; it was collected in Austria and Australia, so lighter skin probably dominates.

ISIC 2019 was never used for training or checkpoint selection. Its training release aggregates several sources including HAM10000 [7], [8], so we removed every image whose identifier appears in the HAM10000 metadata (10,015 of 25,331

TABLE II
DATASETS AND THEIR ROLES. "VAL" IS A STRATIFIED 15% SPLIT OF THE TRAINING FOLDER (SEED 42).

Dataset	Role	Images used	Classes (ours)	Notes
HAM10000	Train and validate core model (v1–v3)	10,015: 8,512 train / 1,503 val	benign 8,061; monitor 327; urgent 1,627	Tiers from 7 diagnoses; multiple images per lesion, split at image level
DermNet (Kaggle train folder)	Train and validate pre-filter	15,557: 13,223 train / 2,334 val	harmless 12,092; needs check 3,465	23 folders merged by us; provided test folder not used
Fitzpatrick17k	Train (half of lesion-like images) and held-out skin-type evaluation; pre-filter stress test	4,817 lesion-like: 2,408 train / 2,409 held out; 16,012 for pre-filter test	benign 2,490; monitor 573; urgent 1,754 (lesion-like subset)	Skin types I–VI; 114 original labels mapped by us
ISIC 2019	Cross-source evaluation only	2,403 HAM10000-free images (cap 833 per tier) of 15,316 available, plus 1,796 HAM10000-derived images as a contamination check	benign, monitor (AK), urgent (MEL, BCC, SCC)	HAM10000 image IDs removed from the main evaluation set

TABLE III
TRAINING CONFIGURATIONS.

	Pre-filter	v1	v2 / v3
Training images	13,223	8,512	10,920
Output classes	2	3	3
Epochs	8	10	10 / 16
Learning rate	3×10^{-4}	3×10^{-4}	3×10^{-4} / 1×10^{-4}
Scheduler	none	none	none / step (6, 0.5)
Rotation aug.	no	no	no / $\pm15^{\circ}$

labelled images) before sampling; the remaining images come mainly from other collections such as BCN_20000. To show how much the overlap matters, we also evaluated a sample of the removed, HAM10000-derived images; roughly 85% of these were in v3's training split and the rest in its validation split.

B. Models and Training

Table III gives the settings. All models are ImageNet-pretrained ResNet18 networks with a replaced final layer, trained with Adam, batch size 16, input 224×224, class-weighted cross-entropy (Eq. 1; weights 0.414/10.21/2.05 for v1 and 0.450/6.44/1.61 for v2/v3, in benign/monitor/urgent order), random flips, and checkpoint selection by best validation accuracy on the HAM10000 validation split. v1 uses HAM10000 only; v2 adds the 2,408 Fitzpatrick17k training images with the v1 recipe; v3 uses the v2 data with a lower learning rate, a step scheduler (halved every 6 epochs), 16 epochs and $\pm15^{\circ}$ rotation augmentation. Two further v3 runs used seeds 123 and 456; the original v3 run was not seeded explicitly. Training used a laptop NVIDIA RTX 3050 (4 GB), PyTorch 2.5.1 (CUDA 12.1), Python 3.11; an epoch took roughly 2–4 minutes.

C. Evaluation Protocol

We report (i) accuracy, per-tier recall and balanced accuracy (the mean of the three per-tier recalls) of v3 on the HAM10000 validation split (natural class frequencies); (ii) accuracy on the 2,409 held-out Fitzpatrick17k images, overall and per skin type, for v1, v2, v3 and the two extra seeds, all on the same images, plus balanced accuracy for v3; (iii) balanced accuracy and per-tier recall of v3 on two ISIC 2019 subsets, images whose identifiers are not in HAM10000 and images that are (the core model is evaluated directly, without the pre-filter); and (iv) pre-filter accuracy on its validation split and on the 16,012 Fitzpatrick17k images, using our own mapping of the 114 conditions as ground truth. Balanced accuracy is given with 95% intervals from 1,000 bootstrap resamples within tiers, per-tier recall with Wilson intervals, and per-skin-type accuracies with approximate binomial half-widths $1.96\sqrt{\hat{p}(1-\hat{p})/n}$ for v3, which assume independent images. No significance tests were run. Seeds and splits use fixed values, and the evaluation scripts are kept with the code.

D. Implementation

Backend: Java 21, Spring Boot 4.2.0-M1 and PostgreSQL 18, with three JUnit persistence tests plus the default context test. Model service: Flask with PyTorch and `pytorch-grad-cam` 1.5.7. Front end: React with Vite, Recharts and jsPDF. The system runs as four local processes; it is not containerised and has no authentication.

V. RESULTS AND DISCUSSION

A. Validation Performance

Table IV and Fig. 3 give HAM10000 validation accuracy. v2 improved on v1 by 1.8 points, and v3 by a further 6.1. The two extra seeds gave 88.76% and 88.16%; the three v3-recipe runs have a mean of 88.54% and a range of 0.6 points. Checkpoints were chosen on this same split, and HAM10000 contains several images of the same lesion that our image-level split can place on both sides, so these figures are optimistic. For v3, per-tier recall on the validation split was 92.64% (benign), 77.55% (monitor) and 71.31% (urgent), giving a balanced accuracy of 80.50% (95% interval 76.3–84.6).

B. Accuracy Across Skin Types

Table V and Fig. 4 show the matched held-out comparison. Overall, v1 scored 46.62%, v2 46.82% and v3 64.51%.

Gains. v3 improved every skin type over v1, by 14.2 to 25.5 points for types I–V and 2.5 points for type VI. For types I–V,

TABLE IV
VALIDATION ACCURACY ON THE HAM10000 SPLIT (1,503 IMAGES).

Model	Training data	Val. acc. (%)
v1	HAM10000	80.84
v2	+ Fitzpatrick17k (half)	82.63
v3	+ new recipe	88.69
v3, seed 123	same	88.76
v3, seed 456	same	88.16
Pre-filter (DermNet, 2 classes)	—	86.20

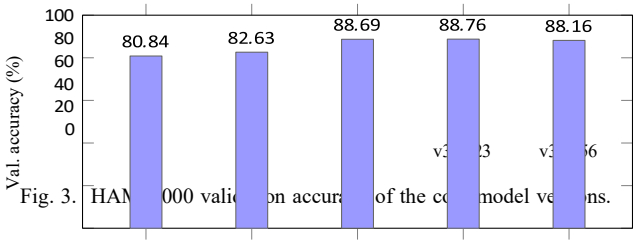


Fig. 3. HAM 000 vali on accura of the c model ve ns.

the gain is well beyond the approximate interval half-widths (3.5–5.9 points for v3). For type VI (n = 79) the half-width is 11.0 points, so the 2.5-point gain cannot be separated from zero.

The gap grew. The spread between the best and worst skin type was 7.7 points for v1 (44.2%–51.9%, type VI best) and 20.3 points for v3 (54.4%–74.8%, type VI worst). Retraining raised absolute accuracy for all groups but did not make the pipeline more even; the group with the fewest training images (80 of type VI) benefited least. This is the main negative finding for RQ1.

v2 was mixed. Against v1, v2 changed accuracy by +1.5, −0.7, +1.2, +7.1, −10.7 and −10.1 points for types I–VI. We ran v2 only once, and the large drops fall on the two smallest groups, so we cannot say how much is noise. One explanation we did not test is that the v1 recipe was not suited to the larger, more varied training set; v3's changes (lower rate, scheduler, more epochs, rotation) addressed that, but we did not ablate them individually, so we cannot say which change mattered.

Attribution. The held-out images come from the same collection as the 2,408 Fitzpatrick17k images added to training, whereas v1 never saw that collection. Part of the v1-to-v3 gain may therefore be adaptation to Fitzpatrick17k's photographs and our label mapping, not to skin-tone diversity as such. A control that adds the same number of images without skin-type balance was not run.

An initial evaluation of v1 through the full pipeline on all 16,012 Fitzpatrick17k images gave 35.1–56.6% for the 1,628 images routed to Stage 2. Because that depended on the pre-filter and used different images, we rely on the matched comparison above.

C. Run-to-Run Consistency

Across the three v3-recipe runs, per-type accuracy had a standard deviation of 1.65, 1.34, 2.17, 1.60, 1.12 and 1.93 points for types I–VI, and overall accuracy was 64.09± 0.58% (mean ± SD). The v1-to-v3 gains for types I–V are about 7 to 20 times these deviations, so they are unlikely to be seed effects. The type VI gain (2.5 points) is comparable to its deviation (1.93), which supports the interval-based reading above. Three runs give only a rough estimate of the spread.

D. Cross-Source Evaluation

We removed from ISIC 2019 every image whose identifier appears in the HAM10000 metadata (10,015 of 25,331 labelled images). Of the remaining 15,316 images, we evaluated 2,403 (up to 833 per tier). v3 reached a balanced accuracy of 55.70% (95% interval 54.0–57.4) and an overall accuracy of 56.60%. Per-tier recall was 84.03% for benign (n = 833; 95% CI 81.4–86.4), 33.24% for monitor (n = 737; 29.9–36.7) and 49.82% for urgent (n = 833; 46.4–53.2).

Fig. 5 compares balanced accuracy on four sets: the HAM10000 validation split (80.50%), the held-out Fitzpatrick17k images (65.97%; overall accuracy 64.51%), the HAM10000-derived ISIC images (90.97%, n = 1, 796) and the HAM10000-free ISIC images (55.70%). Balanced accu-racy on the HAM10000-free images is 24.8 points lower than on the HAM10000 validation split. The HAM10000-derived images scored 35.3 points higher than the HAM10000-free images; because most of them were training images, they are not an estimate for unseen data.

All four values are balanced accuracies, so they share the same class mix. Two caveats remain. The ISIC images outside HAM10000 come largely from other collections, so this is one cross-source test and not a general measure of robustness. Fitzpatrick17k also differs from the training label space (e.g., keloid or xanthoma were never seen as tiers), so part of its low score reflects our mapping rather than image appearance.

E. Pre-Filter

The pre-filter reached 86.20% on its DermNet validation split (best of 8 epochs; accuracy fluctuated between 77.1% and 86.2% across epochs). On the 16,012 Fitzpatrick17k images it scored 67.84% overall (64.77–70.08% across skin types) against our own mapping of 114 conditions into lesion-like and common. This drop shows that the pre-filter is itself sensitive to the source of the images, and its errors are not independent of the core model's: an image wrongly called harmless is never risk-classified. Pre-filter accuracy was fairly even across skin types in this test.

F. Routing Thresholds

Applying Eq. 2 to the v3 accuracies gives the thresholds in Table VI and Fig. 6: 0.80 for type VI and 0.60 for type V. As a functional check on the deployed system, an image from the held-out set with a core-model confidence of 0.693 was shown directly when type V was selected and returned as "uncertain" when type VI was selected. This verifies that the

TABLE V
ACCURACY (%) ON THE 2,409 HELD-OUT FITZPATRICK17K IMAGES. Δ = V3 MINUS V1 (PERCENTAGE POINTS). ± = APPROXIMATE 95% BINOMIAL HALF-WIDTH FOR V3.

Skin type	*n*	v1	v2	v3	Δ (v3−v1)	v3 ±	v3 seed 123	v3 seed 456
I	484	44.83	46.28	60.54	+15.7	4.4	58.88	57.23
II	753	47.01	46.35	61.22	+14.2	3.5	62.28	59.63
III	507	46.75	47.93	66.67	+19.9	4.1	66.27	70.22
IV	380	44.21	51.32	69.74	+25.5	4.6	70.00	67.11
V	206	51.46	40.78	74.76	+23.3	5.9	72.82	72.82
VI	79	51.90	41.77	54.43	+2.5	11.0	55.70	51.90
All	2,409	46.62	46.82	64.51	+17.9	1.9	64.34	63.43

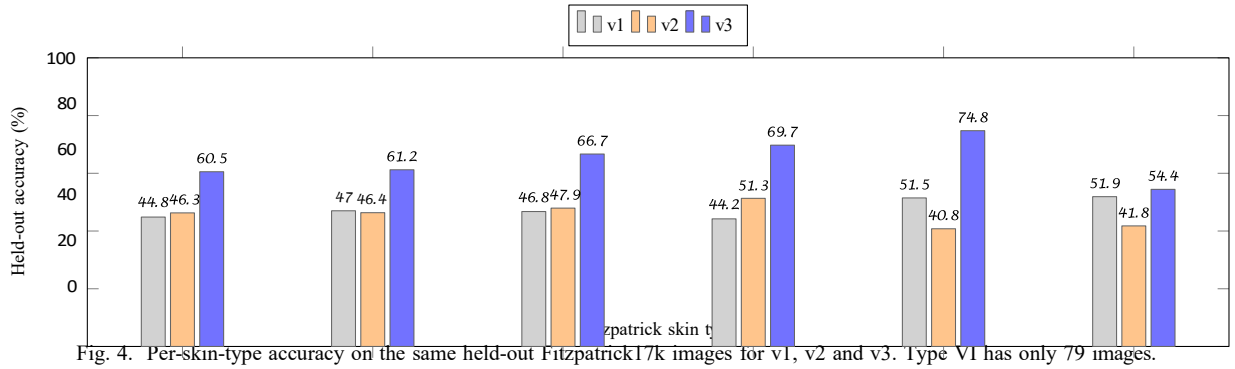


Fig. 4. Per-skin-type accuracy on the same held-out Fitzpatrick17k images for v1, v2 and v3. Type VI has only 79 images.

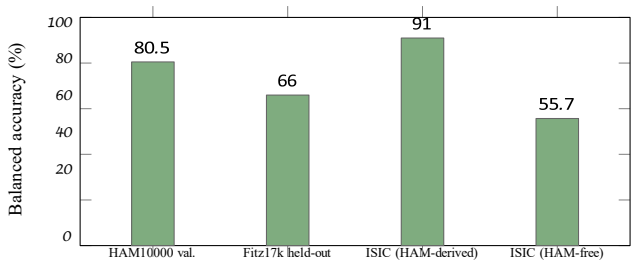


Fig. 5. Balanced accuracy (mean per-tier recall) of v3 on four evaluation sets. The HAM10000-derived ISIC images were mostly training images.

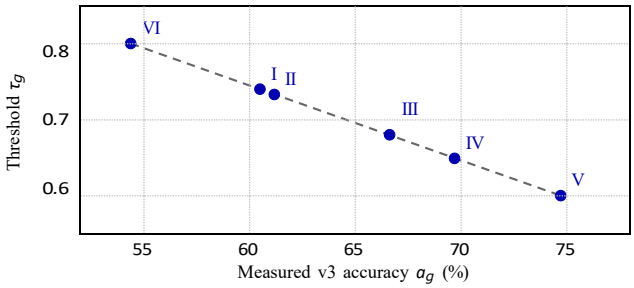


Fig. 6. Threshold as a function of measured accuracy per skin type.

TABLE VI
MEASURED V3 ACCURACY AND RESULTING THRESHOLD (EQ. 2).

Skin type	I	II	III	IV	V	VI
a_g (%)	60.54	61.22	66.67	69.74	74.76	54.43
τ_g	0.740	0.733	0.680	0.649	0.600	0.800

rule is applied; it is a single example and not an evaluation. We did not compute risk–coverage curves, so we cannot say whether the gate lowers errors among accepted results, nor how many correct predictions it withholds. The thresholds also inherit the uncertainty of the accuracies behind them (type VI: ±11 points), and softmax confidence was not calibrated.

G. Explanations

Fig. 7 shows the intended display: the original photograph beside its Grad-CAM overlay. In the examples we inspected, the strongest activation sat on or near the central lesion. We did not measure localization against annotations, so the heatmaps are explanatory aids and not evidence that the model attends to clinically relevant features.

H. What the Results Do and Do Not Show

They show that a standard ResNet18 baseline varied across skin types on matched held-out images; that a retraining recipe raised accuracy for every group, with gains for types I–V that exceed seed noise; that the best-to-worst gap nevertheless widened; that balanced accuracy on ISIC 2019 images outside HAM10000 was 24.8 points lower than on the HAM10000

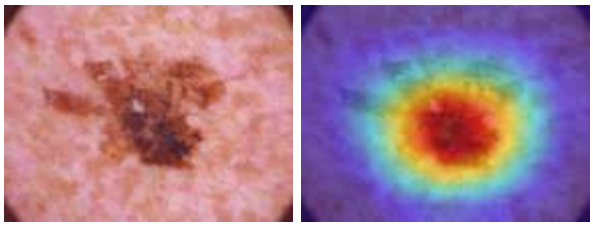


Fig. 7. Example result view: original image (left) and Grad-CAM overlay (right).

validation split (a single cross-source set, with the caveats above); and that the gate, pre-filter and Grad-CAM work together in a running application. They do not show that the gains come from skin-tone diversity rather than from adding any Fitzpatrick17k data, that the gate improves safety, that the pre-filter is reliable on real phone photographs, or anything about clinical performance. We also cannot rank our accuracies against those of the reviewed studies.

VI. Limitations and Ethical Considerations

Labels. All tiers were assigned by us from dataset diagnoses; we cross-checked key assignments (actinic keratosis as premalignant, melanoma and basal and squamous cell carcinoma as malignant, seborrheic keratosis as benign) against standard clinical references, but no dermatologist reviewed the mapping. The "urgent" tier groups conditions of different severity. **Splits.** Splits are at image level, not lesion level, and checkpoints were selected on the evaluation split used for HAM10000, so in-source figures are optimistic. **Sample sizes.** Type VI has 79 held-out images and 80 training images; many per-type conclusions are indicative only. **Domain shift.** Only one cross-source family was used (ISIC 2019 images outside HAM10000), and its monitor tier has 737 images. **Confidence.** Softmax confidence was not calibrated and the gate was not evaluated by risk and coverage. **Phone photographs.** The pre-filter and core model were trained on dermoscopy and atlas photographs; behaviour on phone images was only tested informally. **Bias.** Skin-type input is optional and self-reported; users who skip it receive a single default threshold, and darker skin remains under-represented in training.

Ethics. The system could miss a malignant lesion or reassure a user wrongly, and fluent colour-coded output can be over-trusted. It is a research prototype for screening assistance, not a diagnostic device; every result screen states this and recommends consulting a doctor. We removed an originally planned "human review queue" from the user interface because no clinical reviewers exist in this project, and presenting one would be misleading; uncertain cases are only logged. Uploaded images are processed in memory; the database stores the heatmap and results but there is no access control, and the system must not be used with real patient data in its current form.

VII. Conclusion and Future Work

We built and measured a two-stage skin-lesion screening pipeline with a skin-type-dependent confidence gate and Grad-CAM explanations. Retraining raised validation accuracy from 80.84% to 88.69% and held-out accuracy for every skin type, with gains for types I–V that exceeded seed variation, but the gap between best and worst skin type grew from 7.7 to 20.3 points, the type VI gain was not distinguishable from noise, and balanced accuracy on ISIC 2019 images outside HAM10000 was 55.70% against 80.50% on the HAM10000 validation split. The gate, pre-filter and explanations work end to end, but their value for safety is unproven.

Future work (not completed): split HAM10000 by lesion; compute calibration error for all sets; evaluate the gate with risk–coverage curves; ablate the v3 changes and add a control with skin-type-unbalanced Fitzpatrick17k data; collect more type V–VI images; test on phone photographs; and have dermatologists review the tier mapping and the system's outputs.

Acknowledgment

This work was carried out as a final-year B.Tech project at the School of Computer Science and Engineering (SCOPE), VIT-AP University, under the guidance of Dr. Khasim Syed. We thank the guide for supervision. We also thank the providers of HAM10000, DermNet NZ, Fitzpatrick17k and ISIC for making their data available.

Data and Code Availability

HAM10000, DermNet, Fitzpatrick17k and ISIC 2019 are available from their providers under their respective licence terms (some are non-commercial). The code, trained weights and logs are kept private and are not publicly released.

Funding

This work received no external funding.

Ethics Statement

Only existing, publicly available, de-identified datasets were used. No new images were collected and no patients, clinicians or other human participants took part. The system is a research prototype and is not intended for clinical use.

References

[1] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, pp. 115–118, 2017.

[2] M. Groh *et al.*, "Evaluating deep neural networks trained on clinical images in dermatology with the Fitzpatrick 17k dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2021, pp. 1820–1828.

[3] R. Daneshjou *et al.*, "Disparities in dermatology AI performance on a diverse, curated clinical image set," *Sci. Adv.*, vol. 8, no. 32, Art. no. eabq6147, 2022.

[4] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 618–626.

[5] Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017.

[6] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci. Data*, vol. 5, Art. no. 180161, 2018.

[7] N. Codella *et al.*, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the International Skin Imaging Collaboration (ISIC)," arXiv:1902.03368, 2019.
[8] M. Combalia *et al.*, "BCN20000: Dermoscopic lesions in the wild," arXiv:1908.02288, 2019.
[9] DermNet NZ, "Dermatology image dataset," https://dermnetnz.org/dermatology-image-dataset. Accessed: Aug. 18, 2026.
[10] C. Vadivelraju, K. E. S., V. M. B. Kumar, P. Suraj, and M. H. Reddy, "DermaScan AI: Deep learning system for preliminary diagnosis of dermatological manifestations," *Int. J. Creative Research Thoughts (IJCRT)*, 2025, paper IJCRT2505389.
[11] A. Agustin and I. Nurhaida, "Dermascan: Convolutional neural network-based skin cancer early detection system," *Electron. J. Educ., Soc. Econ. Technol.*, vol. 6, no. 2, 2025.
[12] I. Matas, C. Serrano, F. Silva-Claver´ıa, A. Serrano, T. Toledo-Pastrana, and B. Acha, "MultiTask Learning AI system to assist BCC diagnosis with dual explanation," arXiv:2407.00104, 2024.
[13] T. Chanda *et al.*, "Dermatologist-like explainable AI enhances trust and confidence in diagnosing melanoma," *Nature Commun.*, article s41467-023-43095-4, 2023.
[14] B. Mazoure, A. Mazoure, J. Be´dard, and V. Makarenkov, "DUNEScan: a web server for uncertainty estimation in skin cancer detection with deep neural networks," *Sci. Rep.*, article s41598-021-03889-2, 2022.
[15] R. Jalaboi, O. Winther, and A. Galimzianova, "Dermatological di-agnosis explainability benchmark for convolutional neural networks," arXiv:2302.12084, 2023.
[16] T. B. Fitzpatrick, "The validity and practicality of sun-reactive skin types I through VI," *Arch. Dermatol.*, vol. 124, no. 6, pp. 869–871, 1988.
[17] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 1321–1330.
[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.