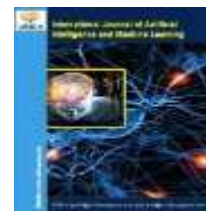




Svedberg Open
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

Lesion Focus-Lite: Knowledge Distillation For CPU-Deployable Skin Cancer Prediction With Multi-Level Validation

Meenakshi Dixit¹, Anwar Bari²

¹Department of Computer Application, Integral University, Lucknow, Uttar Pradesh, India. Email: mdixit@student.iul.ac.in

²Department of Computer Application, Integral University, Lucknow, Uttar Pradesh, India. Email: anwarbari@iul.ac.in

Abstract

Deep learning models for dermoscopic skin cancer prediction are usually reported on one test split and seldom built into software that can run outside a research GPU server. This study implements the Lesion Focus-Net framework, a segmentation-guided hybrid CNN-Transformer design with adaptive pre-processing, and validates it as a working system. The implementation has two tiers. Tier A runs the full framework on a GPU: trigger-based colour constancy, hair removal and contrast enhancement, an attention residual U-Net for lesion segmentation, and a dual-stream Swin-T/EfficientNet-B3 classifier with cross-attention, patient metadata, test-time augmentation and temperature scaling. Tier B, Lesion Focus-Lite, is a 4.2-million-parameter MobileNetV3 student distilled from the full framework, which needs no segmentation and runs through ONNX Runtime on a CPU. Both tiers were trained on a lesion-level split of HAM10000 and validated at five levels: an internal hold-out of 1,975 images, the independent ISIC 2018 Task 3 test set of 1,511 images, robustness to five image degradations, malignant triage with calibration, and a deployment profile. The full framework reached a balanced accuracy of 65.8% and a macro AUC of 0.962 internally, and 71.3% and 0.964 on the independent set, with a malignant-versus-benign AUC of 0.939 and an expected calibration error of 0.035. Distillation raised the light model's internal balanced accuracy from 57.7% to 60.4% and its macro-F1 from 54.2% to 59.4% (McNemar $p = 0.008$). Lesion Focus-Lite ran end to end in 127 ms per image on two CPU cores, about 23 times faster than the full pipeline, with a macro AUC above 0.94 on both test sets. Pre-processing made all models almost insensitive to colour casts, but Gaussian noise, blur and heavy JPEG compression reduced balanced accuracy by 10 to 36 points, and blur hurt the full framework more than the light model. The results show that the framework generalises to unseen patients, that a distilled model offers a practical CPU option, and that image quality at capture is now the main barrier to reliable use.

Keywords: skin cancer; dermoscopy; deep learning; implementation; knowledge distillation; ONNX deployment; external validation; robustness; calibration; HAM10000.

1. Introduction

Skin cancer is the most frequently diagnosed cancer in many countries, and melanoma, although it makes up a small share of cases, causes most skin cancer deaths; the global number of new melanoma cases is projected to keep rising [5]. Dermoscopy improves the accuracy of trained readers [28], and deep networks trained on dermoscopic images now reach or exceed specialist performance on curated test sets [15], [8], [43]. The open question is not anymore whether a neural network can detect skin lesions, but whether an entire system can be constructed, run on commodity hardware, and exhibit robust behaviour when tested on images that it has not been optimized for [25].

This paper marks the last stage of a sequence of related research works. A systematic review of the applications of deep learning for skin cancer detection pinpointed four common issues: poor ability to deal with artifacts during image acquisition, class imbalances, lack of external validation, and very little information about model implementation and inference speed. In a continuation of that study, adaptive pre-processing for accurate lesion segmentation and Lesion Focus-Net architecture were designed as a hybrid CNN-Transformer approach specifically designed to overcome these challenges. The task of this work was to put the design into action and evaluate it in a meaningful way. The present study does both.

On the implementation side, a framework that needs a GPU server is of limited use in the small clinics and tele dermatology settings where screening support is most needed. We therefore implement Lesion Focus-Net in two tiers: a reference pipeline that runs the full framework on a GPU, and a deployable pipeline in which a

compact student network, Lesion Focus-Lite, is trained by knowledge distillation [19] from the full framework and exported to ONNX for CPU inference. On the validation side, we go beyond a single accuracy figure and test both tiers at five levels: an internal lesion-level hold-out, an independent test set used once, robustness to six image-acquisition conditions, a clinical malignant-triage and calibration view, and a deployment check of size, speed and numerical fidelity.

The contributions are as follows:

- (a) a two-tier implementation of Lesion Focus-Net, with a GPU reference pipeline and a CPU-deployable distilled model, released as modular code with a documented execution environment;
- (b) a knowledge-distillation scheme that transfers the dual-stream teacher's behaviour to a 4.2-million-parameter MobileNetV3 student that needs no segmentation at inference, with an ablation against the same student trained without distillation;
- (c) a five-level validation protocol covering internal and independent test sets, image-degradation robustness, malignant triage with calibration, and deployment metrics;
- (d) evidence on where the full framework and its light version differ, including the conditions under which each one fails.

Section 2 reviews related work. Section 3 describes the framework and the two-tier implementation, Section 4 the validation design, and Section 5 the results. Section 6 discusses the findings, Section 7 lists the limitations, Section 8 covers ethics, and Section 9 concludes.

2. Related Work

2.1 Skin-Lesion Frameworks and Their Evaluation

Recent dermoscopy frameworks combine segmentation and classification, or fuse convolutional and transformer features. Abu Owida et al. joined an EfficientNet-B7 segmentation branch with a DenseNet-121 classifier across five benchmarks [2]; Karthik et al. fused a Swin Transformer with a custom CNN on HAM10000 [23]; Kaur et al. built a complete computer-aided pipeline with hair removal, contrast enhancement and segmentation before classification [24]. These studies report high accuracy, but almost all of it comes from a single split of a single dataset. Tschandl et al. demonstrated in an international reader study that algorithm performance is reduced on images that deviate from the training distribution [43], and subsequent work showed that skin tone and acquisition differences can lead to large drops in accuracy [13], [16]. Kelly et al. made the general argument that external testing, uncertainty reporting and consideration of deployment are what distinguish a research model from a clinically viable tool [25]. Comparative evaluation of various machine learning classifiers [35] CLAIM and TRIPOD+AI [11] ask for the same things.

2.2 Lightweight Models and Knowledge Distillation

MobileNetV3 was designed through hardware-aware architecture search to give good accuracy at low latency on mobile CPUs [20]. Knowledge distillation trains a small student network to match the softened output distribution of a larger teacher, transferring part of the teacher's learned structure between classes [19]. In skin imaging, lightweight networks have mostly been trained directly on labels; distillation from a segmentation-guided, multi-stream teacher to a student that sees only the whole image has received little attention. Federated approaches [46], [22] address a related deployment problem, namely that images often cannot leave a hospital, but they still assume a GPU-class model at each site.

2.3 Robustness to Acquisition Conditions

Hendrycks and Dietterich showed that image classifiers degrade under common corruptions such as blur, noise and compression, even when clean accuracy is high [18]. Rizwan et al. built a robust cross validated framework for disease prediction with XAI [6], hair removal [42] and contrast enhancement [47] are often proposed as defences, but their effect on robustness is rarely measured. We therefore test the implemented system under controlled degradations applied before pre-processing, so that the pre-processing stage is part of what is being tested.

2.4 Methodological Precedents from Integral University Research

Several design choices follow earlier work at our institution in other application areas. Alam and Ahamad used CNN transfer learning to detect rheumatoid arthritis [4], the same strategy of fine-tuning ImageNet-pretrained backbones on a modest medical image set that we use for every network here. Abdussami and Farooqui built an incremental deep neural network for intrusion detection in fog-based IoT environments, with an optimisation-assisted framework designed for constrained edge devices [1]; their concern with running deep models close to the data source is the motivation for our CPU tier. Ahamad and Bharti compared the fitness functions of K-means and kernel Fisher's discriminant analysis under a genetic algorithm [3], a reminder that the selection criterion shapes the final model; we select checkpoints by

validation macro-F1 rather than accuracy for that reason. Khan and Faisal predicted software risk with a Bayesian network [27], and Khan prioritised risks with a fuzzy approach [26]; both treat prediction as a graded risk, which is how we frame the malignant-triage score. Krishna and Ahamad benchmarked their Green Audit-GNN model across sectors and cross-checked two independent explanation methods before making any claim [29]; we follow the same discipline of testing each claim against more than one view of the data.

3. Framework and Two-Tier Implementation

3.1 Lesion Focus-Net in Brief

Lesion Focus-Net has four stages. Stage 1 is adaptive pre-processing: dark lens corners are masked, Shades-of-Gray colour constancy ($p = 6$) [6] corrects the illuminant, hair is detected with a multi-orientation black-hat filter and removed by fast-marching in painting [42], and CLAHE [47] is applied to the lightness channel when contrast is low. Each step is triggered by a measured image property. Stage 2 is ARU-Net, an attention residual U-Net [40], [36] with an EfficientNet-B4 encoder [41], an ASPP bottleneck [9] and squeeze-and-excitation decoder blocks [21]; it outlines the lesion and defines a crop with a 15% margin. Stage 3 is a dual-stream classifier: a Swin-T transformer [31] reads the whole 224×224 image, an EfficientNet-B3 reads the 300×300 lesion crop, the crop features are pooled with the soft lesion mask and joined to the global tokens by eight-head cross-attention ($d = 256$), and an age-sex-site metadata branch is added before a 512-unit head. Stage 4 trains the classifier with class-balanced focal loss ($\beta = 0.999, \gamma = 2$) [12], [30] and calibrates it at inference with eight-fold dihedral test-time augmentation and temperature scaling [17].

3.2 Tier A: Reference Pipeline

Tier A runs all four stages on a GPU. It is the most accurate configuration and serves as the teacher for Tier B. Figure 1 shows both tiers.

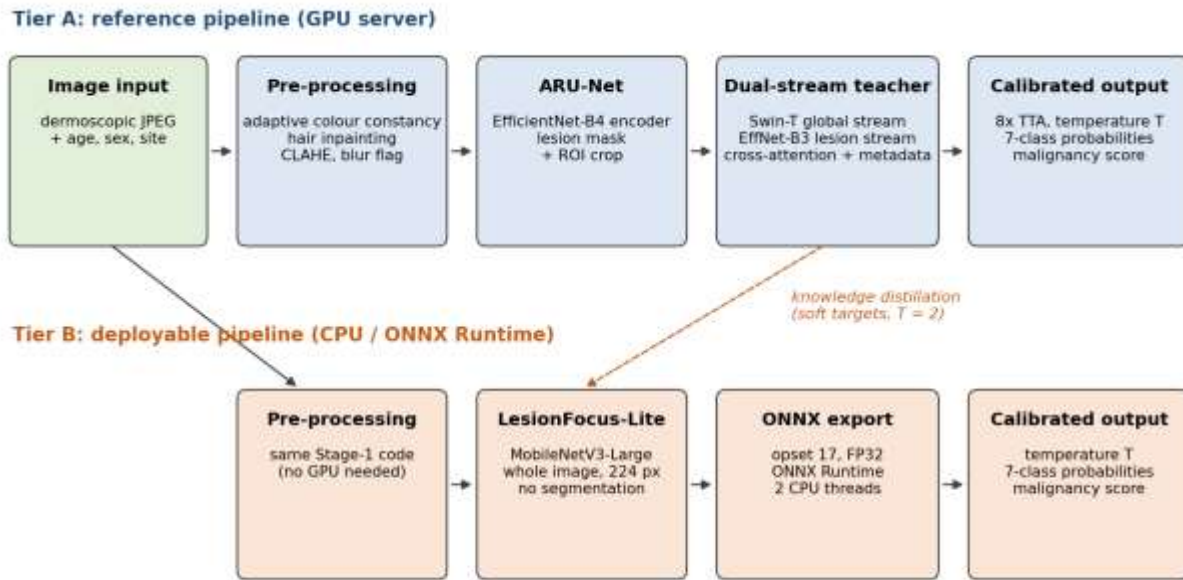


Figure 1. Two-tier implementation of Lesion Focus-Net. Tier A runs the complete framework on a GPU. Tier B reuses the same pre-processing code and replaces segmentation and the dual-stream classifier with a distilled MobileNetV3 student exported to ONNX for CPU inference.

3.3 Tier B: Lesion Focus-Lite by Knowledge Distillation

Tier B keeps Stage 1, which runs on a CPU in about 0.1 s, and replaces Stages 2 and 3 with a single MobileNetV3-Large network [20] that reads the whole pre-processed image at 224×224 pixels. No segmentation is needed at inference. The student is trained to match both the labels and the teacher's softened predictions:

$$L_{KD} = \alpha T^2 \cdot KL(\sigma(z_t / T) \parallel \sigma(z_s / T)) + (1 - \alpha) \cdot L_{CB-focal}(z_s, y) \quad (1)$$

where z_t and z_s are the teacher and student logits, σ is the softmax, $T = 2$ is the distillation temperature and $\alpha = 0.7$. The teacher logits are computed once on the training images after the teacher has been trained. To measure what distillation adds, the same student is also trained with the class-balanced focal loss alone. After training, the distilled student is calibrated by temperature scaling on the validation set and exported to ONNX (opset 17) for inference with ONNX Runtime on the CPU.

3.4 Software Organisation and Environment

The implementation is split into three Python modules: a data module that downloads the datasets from the Harvard Dataverse, runs Stage 1 in parallel CPU workers and stores the processed images and masks in memory-mapped arrays; a model module with the ARU-Net, dual-stream and student networks and the loss functions; and an experiment module that runs training, inference, robustness tests and profiling, and writes every result to a single JSON file and a time-stamped log. Table 1 lists the environment. Memory-mapped storage was introduced after the first trials showed that JPEG decoding on two CPU cores starved the GPU; it removes decoding from the training loop entirely.

Table 1. Execution environment of the implementation.

Component	Version / specification	Use
Platform	Google Colab, Ubuntu 24.04	All experiments
GPU	NVIDIA Tesla T4, 16 GB	Training and Tier A inference
CPU / memory	2 virtual cores, 12.7 GB RAM	Stage 1 pre-processing, Tier B inference
Language and DL library	Python 3.13, PyTorch 2.11 [38]	Networks, mixed-precision training [34]
Pretrained backbones	timm 1.0 [45]	EfficientNet-B3/B4, Swin-T, MobileNetV3-Large
Image processing	OpenCV 4 [7]	Colour constancy, black-hat, inpainting, CLAHE
Data handling and metrics	NumPy memory-mapped arrays, scikit-learn [39]	Image cache, StratifiedGroupKFold, metrics
Deployment runtime	ONNX opset 17, ONNX Runtime [37]	Tier B CPU inference

3.5 Training Protocol

HAM10000 [44] was split into five lesion-grouped, class-stratified folds with StratifiedGroupKFold (seed 2026), so that all photographs of a lesion fall in one fold. One fold served as the internal test set, the next as the validation set and the remaining three as training data (Table 2). ARU-Net was trained for three epochs on the training lesions with AdamW (learning rate 3×10^{-4}), a one-cycle schedule and a Dice, cross-entropy and boundary loss, and the epoch with the best validation Dice was kept. Ground-truth masks were used to crop training and validation images, while test images were cropped with ARU-Net's own predictions, as they would be in use. The teacher was trained for five epochs and each student for six, with AdamW, one-cycle schedules, batches of 32, mixed precision [34] and random dihedral augmentation; for every network the checkpoint with the best validation macro-F1 was kept. All evaluation ran in 32-bit precision.

Table 2. Data used for training and validation. HAM10000 was split by lesion; the ISIC 2018 Task 3 test set was used once, for V2 only.

Set	Images	Lesions	nv	mel	Bkl	bcc	akiec	vasc	df
HAM10000 training	6,014	4,482	4,044	644	676	304	184	81	81
HAM10000 validation	2,026	1,493	1,291	278	203	111	81	39	23
HAM10000 internal test	1,975	1,495	1,370	191	220	99	62	22	11
ISIC 2018 Task 3 test [10]	1,511	—	908	171	217	93	43	35	44

4. Validation Design

4.1 Five Validation Levels

Each model is checked at five levels (Figure 2). V1 measures discrimination on the internal lesion-level hold-out. V2 repeats the measurement on the official ISIC 2018 Task 3 test set [10], which was never used for training, model selection or calibration and was scored once. V3 measures robustness: the raw internal test images are degraded before pre-processing, and the whole pipeline, including pre-processing and segmentation, is rerun. V4 looks at the output the way a referral decision would, by grouping melanoma, basal cell carcinoma and actinic keratosis/intraepithelial carcinoma as malignant and the other four classes as benign, and by checking calibration. V5 measures model size, CPU and GPU latency, and the numerical agreement between the PyTorch and ONNX versions of the deployable model.

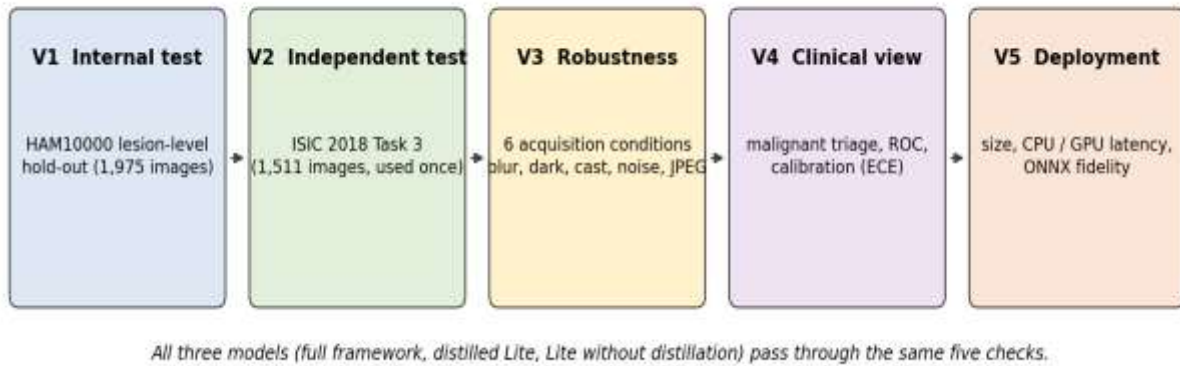


Figure 2. Five-level validation protocol applied to the full framework, the distilled Lesion Focus-Lite model and the same student trained without distillation.

4.2 Degradation Conditions

Five degradations represent common problems in clinical photographs: Gaussian blur ($\sigma = 2$ px) for focus errors, a 30% reduction in brightness for under-exposure, a warm colour cast (blue channel $\times 0.85$, red channel $\times 1.2$) for uncorrected illumination, additive Gaussian noise ($\sigma = 12$ grey levels) for sensor noise, and JPEG compression at quality 15 for images sent through messaging applications. Each condition is applied to all images of the internal test set, and the unmodified images are run through the same code as a reference

4.3 Metrics

Balanced accuracy (BA), the mean of per-class recall, and macro-F1 are the main metrics, because overall accuracy on HAM10000 is dominated by nevi. We also report the one-vs-rest macro AUC, the expected calibration error (ECE) over 15 confidence bins, and for the triage view the area under the ROC curve of the summed malignant probability, the sensitivity and specificity of the predicted class, and the sensitivity at 90% specificity:

$$BA = (1/K) \sum_{k=1}^K TP_k / (TP_k + FN_k), \quad ECE = \sum_{m=1}^M (|B_m| / N) \cdot |acc(B_m) - conf(B_m)| \quad (2)$$

$$S_{mal}(x) = \hat{p}_{mel}(x) + \hat{p}_{bcc}(x) + \hat{p}_{akiec}(x) \quad (3)$$

Ninety-five per cent confidence intervals for BA and macro-F1 come from 200 bootstrap resamples of the test images [14], and paired comparisons between models on the same images use McNemar's test [33].

5. Results

5.1 Implementation Run and Training Behaviour

The full experiment, from raw images to all reported numbers, ran on one T4 GPU. Stage 1 pre-processing took 119.5 ± 44.5 ms per HAM10000 image on two CPU cores; colour correction fired on 99.1% of images, hair in painting on 75.8% and CLAHE on 80.1%, and 8.7% were flagged as blurred. Training took 11.3 minutes for ARU-Net, 17.2 minutes for the teacher and about 9.5 minutes for each student (Table 3). Figure 3 shows the training curves. ARU-Net reached a validation Dice of 0.926 after three epochs and scored 0.933 ± 0.078 (IoU 0.883) on the 1,975 internal test images. The teacher's validation macro-F1 rose steadily from 56.5% to 71.2% over five epochs. The distilled student learned faster than the undistilled one: after one

epoch its validation macro-F1 was 54.3% against 51.6%, and its best value was 66.1% against 63.9%. Its training loss stays higher because it includes the distillation term, which keeps the student close to the teacher's softer output distribution rather than letting it fit the hard labels completely.

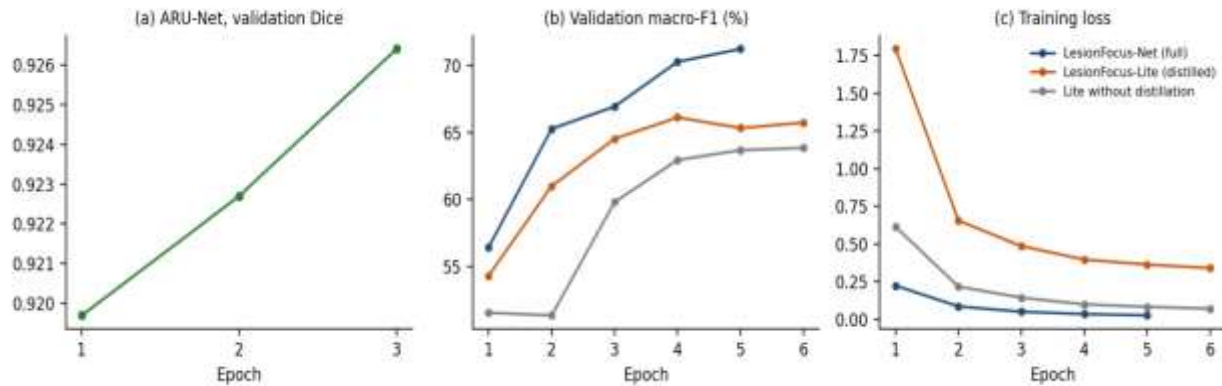


Figure 3. Training behaviour. (a) Validation Dice of ARU-Net. (b) Validation macro-F1 of the teacher and the two students. (c) Training loss; the distilled student's loss includes the distillation term and is therefore not on the same scale.

Table 3. Training cost of each component on one NVIDIA T4 GPU.

Component	Epochs	Time per epoch (s)	Total time (min)	Selected epoch
ARU-Net segmentation	3	193	11.3	3 (best val. Dice)
Dual-stream teacher	5	199	17.2	5 (best val. macro-F1)
LesionFocus-Lite, distilled	6	90	9.5	4 (best val. macro-F1)
Lite without distillation	6	88	9.2	6 (best val. macro-F1)
Complete run incl. inference	—	—	58.0	—

5.2 V1 and V2: Internal and Independent Test Performance

Table 4 and Figure 4 give the main results. On the internal hold-out, the full framework reached a balanced accuracy of 65.8% (95% CI 60.9–71.2), a macro-F1 of 65.9% and a macro AUC of 0.962. On the independent ISIC 2018 test set, which none of the models had seen, it did better: 71.3% balanced accuracy, 74.1% macro-F1 and an AUC of 0.964. The framework therefore did not lose performance when moved to new patients; the independent set contains more of the minority classes (44 dermatofibromas against 11 in the internal fold), on which it performs well, and that lifts its balanced accuracy.

Lesion Focus-Lite kept most of the teacher's ranking ability. Its macro AUC was 0.947 internally and 0.943 externally, about 0.02 below the teacher, but its balanced accuracy was lower by 5.4 points internally and 11.1 points externally. Distillation clearly helped on the internal set, raising balanced accuracy from 57.7% to 60.4% and macro-F1 from 54.2% to 59.4% compared with the same network trained without the teacher; McNemar's test confirmed the difference in per-image correctness ($p = 0.008$). On the independent set the two students were close (balanced accuracy 60.3% against 61.5%, macro-F1 63.0% against 62.1%, $p = 0.16$), although the distilled student kept a higher AUC (0.943 against 0.933) and melanoma AUC (0.869 against 0.849). The teacher was significantly more often correct than either student on both test sets ($p < 0.001$).

Table 4. Seven-class performance on the internal ($n = 1,975$) and independent ($n = 1,511$) test sets. 95% bootstrap confidence intervals in brackets for the internal set.

Test set / model	Acc. (%)	Balanced acc. (%)	Macro-F1 (%)	AUC	ECE
Internal: Lesion Focus-Net (full)	84.7	65.8 [60.9–71.2]	65.9 [60.8–70.7]	0.962	0.022

Test set / model	Acc. (%)	Balanced acc. (%)	Macro-F1 (%)	AUC	ECE
Internal: full, without TTA and T	82.4	64.7	63.2	0.950	0.064
Internal: Lesion Focus-Lite (distilled)	79.4	60.4 [55.8–65.4]	59.4 [54.8–63.9]	0.947	0.027
Internal: Lite without distillation	77.1	57.7 [53.1–63.4]	54.2 [50.6–58.1]	0.929	0.022
Independent: Lesion Focus-Net (full)	83.3	71.3	74.1	0.964	0.035
Independent: Lesion Focus-Lite (distilled)	75.6	60.3	63.0	0.943	0.040
Independent: Lite without distillation	74.1	61.5	62.1	0.933	0.028

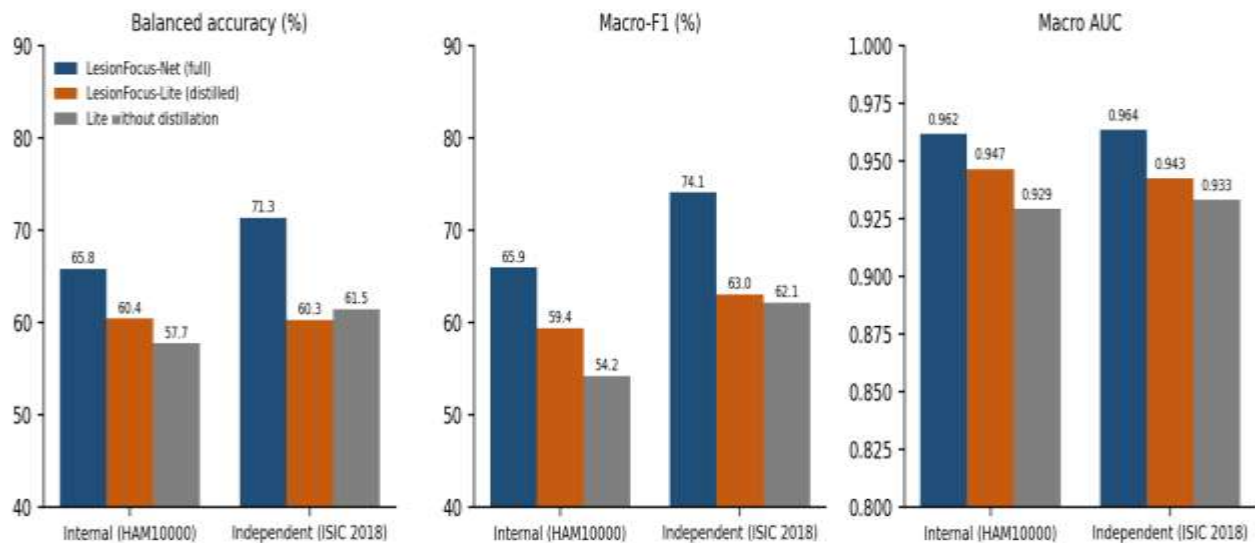


Figure 4. Balanced accuracy, macro-F1 and macro AUC of the three models on the internal HAM10000 test set and the independent ISIC 2018 Task 3 test set.

Test-time augmentation and temperature scaling both earned their place in the teacher. Without them, internal balanced accuracy was 64.7%, macro-F1 63.2% and ECE 0.064; with them, 65.8%, 65.9% and 0.022. Per class (Table 5), the teacher recalled 93.0% of nevi, 76.8% of basal cell carcinomas, 70.9% of benign keratoses, 64.5% of actinic keratoses and 55.5% of melanomas internally, and 60.8% of melanomas externally. Dermatofibroma was the weakest internal class (2 of 11 found) but reached 68.2% recall on the larger external sample, which shows how unstable per-class figures are when a class has only a handful of test images.

Table 5. Per-class recall and precision of the full framework and Lesion Focus-Lite (distilled) on both test sets.

Class	n (int.)	Recall full / Lite	Precision full / Lite	n (ind.)	Recall full / Lite	Precision full / Lite
nv	1370	0.93 / 0.89	0.93 / 0.90	908	0.94 / 0.90	0.88 / 0.85
mel	191	0.56 / 0.53	0.54 / 0.43	171	0.61 / 0.47	0.68 / 0.45
bkl	220	0.71 / 0.57	0.77 / 0.64	217	0.70 / 0.57	0.80 / 0.66
bcc	99	0.77 / 0.65	0.71 / 0.68	93	0.69 / 0.62	0.74 / 0.68

Class	n (int.)	Recall full / Lite	Precision full / Lite	n (ind.)	Recall full / Lite	Precision full / Lite
akiec	62	0.65 / 0.50	0.62 / 0.57	43	0.72 / 0.60	0.58 / 0.58
vasc	22	0.82 / 0.91	0.78 / 0.69	35	0.66 / 0.60	0.92 / 0.72
df	11	0.18 / 0.18	0.33 / 0.22	44	0.68 / 0.45	0.86 / 0.77

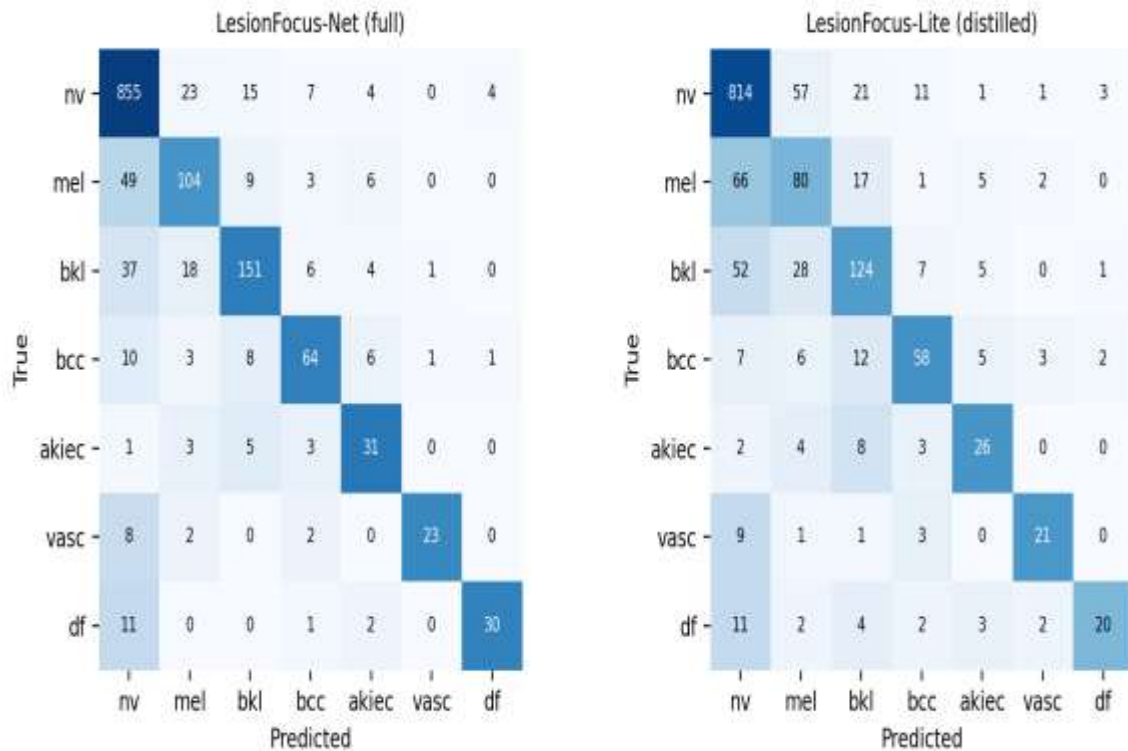


Figure 5. Confusion matrices on the ISIC 2018 independent test set for the full framework and Lesion Focus-Lite. Cell values are image counts; shading is row-normalised recall.

5.3 V4: Malignant Triage and Calibration

For referral, the summed malignant probability matters more than the exact class. Figure 6 shows the malignant-versus-benign ROC curves and Table 6 the operating points. The teacher's triage AUC was 0.929 (95% CI 0.917–0.943) internally and 0.939 externally. At the default decision rule it detected 71.0% of malignant lesions internally and 72.6% externally, with specificities of 92.7% and 94.3%; at a threshold fixed for 90% specificity its sensitivity was 77.3% and 80.8%. Lesion Focus-Lite reached triage AUCs of 0.892 and 0.887 and sensitivities at 90% specificity of 67.0% and 65.8%. In practical terms, on the independent set the teacher missed 84 of 307 malignant lesions and the distilled student missed 119.

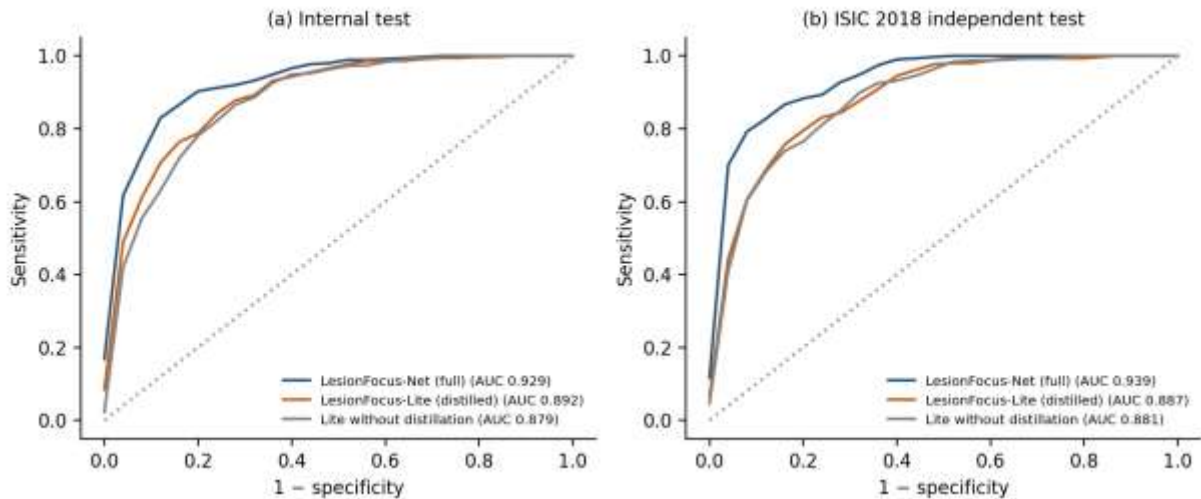


Figure 6. Malignant (mel, bcc, akiec) against benign ROC curves of the three models on (a) the internal and (b) the independent test set.

Table 6. Malignant (mel, bcc, akiec) against benign triage. TP / FN / FP at the default decision rule; Sens@90 = sensitivity at 90% specificity; Spec@95 = specificity at 95% sensitivity.

Test set / model	AUC	TP / FN / FP	Sens. (%)	Spec. (%)	Sens@90 (%)	Spec@95 (%)	Mel. AUC
Internal: full	0.929	250 / 102 / 119	71.0	92.7	77.3	64.0	0.911
Internal: Lite (distilled)	0.892	232 / 120 / 152	65.9	90.6	67.0	57.2	0.865
Internal: Lite (no distillation)	0.879	215 / 137 / 189	61.1	88.3	60.2	59.2	0.849
Independent: full	0.939	223 / 84 / 69	72.6	94.3	80.8	67.9	0.928
Independent: Lite (distilled)	0.887	188 / 119 / 120	61.2	90.0	65.8	59.1	0.869
Independent: Lite (no distillation)	0.881	205 / 102 / 131	66.8	89.1	64.8	54.6	0.849

All three models were well calibrated when temperature scaled. On the internal set, the teacher's ECE decreased from 0.093 to 0.022 with a fitted temperature of 0.67, thus sharpening its outputs: test-time averaging had made the raw teacher under-confident. The distilled student required almost no correction ($T = 0.93$; ECE 0.038 to 0.027), whereas the undistilled student was over-confident and required a temperature of 1.32 (ECE 0.041 to 0.023). Figure 7 shows the reliability diagrams. On the independent set ECE remained low for all models (0.035, 0.040 and 0.028).

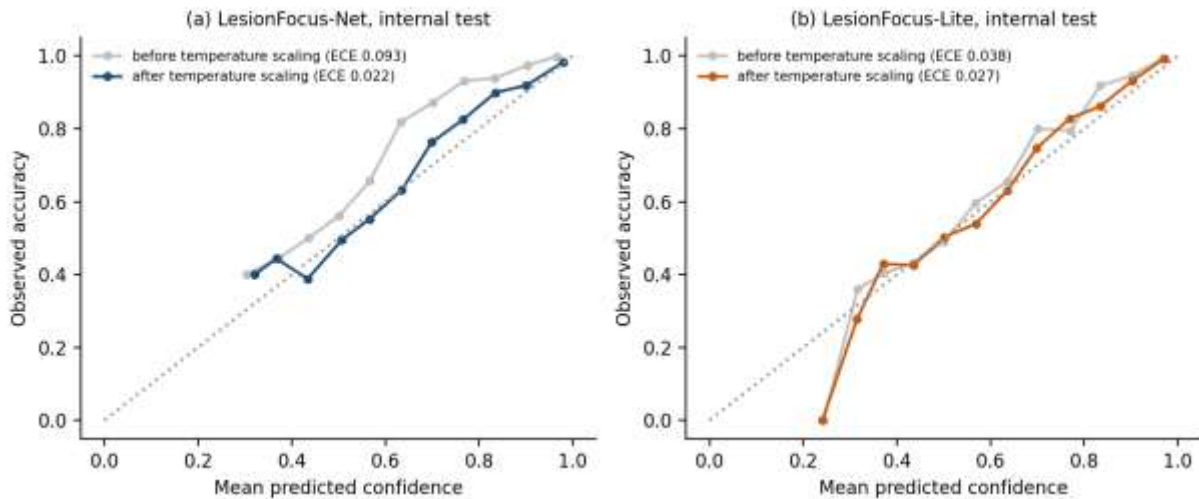


Figure 7. Reliability diagrams on the internal test set before and after temperature scaling for (a) the full framework and (b) Lesion Focus-Lite. Points on the diagonal indicate perfect calibration.

5.4 V3: Robustness to Image Degradation

Table 7 and Fig. 8 show how the three models responded to the raw images being degraded before entering the pipeline, using results from the same 1,000 images processed without degradation as references. Two conditions had almost no effect: under a warm colour cast, balanced accuracy changed by less than 1.2 points for any model (because most of the cast was removed by the Shades-of-Gray step in Stage 1 before any network saw it), and a 30% loss of brightness cost 2.0-4.9 points. ARU-Net was barely affected by either (Dice 0.932 and 0.928 compared with 0.934).

The other three conditions were harmful and highlight particular weaknesses in the system. Gaussian noise was worst of all: ARU-Net's Dice dropped to 0.768 and every classifier had less than 32% balanced accuracy. Heavy JPEG compression cost the teacher 21.3 points and the students 25-29 points; distillation helped here, since the distilled student retained 35.8% against 29.2% for the undistilled one. Blur behaved differently from the others: it hurt the full framework more than the light models (-25.3 points compared with -12.1 and -10.4) and after blurring the distilled student was more accurate than the teacher (49.1% versus 42.6%). The teacher's local stream reads fine texture from a high-resolution lesion crop, which is exactly what blur removes, while the student, which sees the whole image at 224 pixels, is less reliant on fine detail. Nevertheless, the teacher retained the best triage AUC under blur (0.889 versus 0.837).

Table 7. Robustness on 1,000 internal test images degraded before pre-processing. Balanced accuracy (%) and triage AUC; change from the clean reference in brackets.

Condition	ARU-Net Dice	Full: BA (Δ)	Lite: BA (Δ)	Lite no KD: BA (Δ)	Full / Lite triage AUC
Clean (reference)	0.934	68.0	61.2	58.0	0.936 / 0.894
Gaussian blur, $\sigma = 2$ px	0.928	42.6 (-25.3)	49.1 (-12.1)	47.6 (-10.4)	0.889 / 0.837
Brightness $\times 0.7$	0.928	63.5 (-4.5)	59.2 (-2.0)	53.0 (-4.9)	0.914 / 0.855
Warm colour cast	0.932	67.8 (-0.1)	62.3 (+1.1)	58.7 (+0.7)	0.936 / 0.897
Gaussian noise, $\sigma = 12$	0.768	31.9 (-36.0)	26.2 (-35.0)	26.6 (-31.3)	0.764 / 0.649
JPEG, quality 15	0.909	46.6 (-21.3)	35.8 (-25.4)	29.2 (-28.8)	0.799 / 0.800

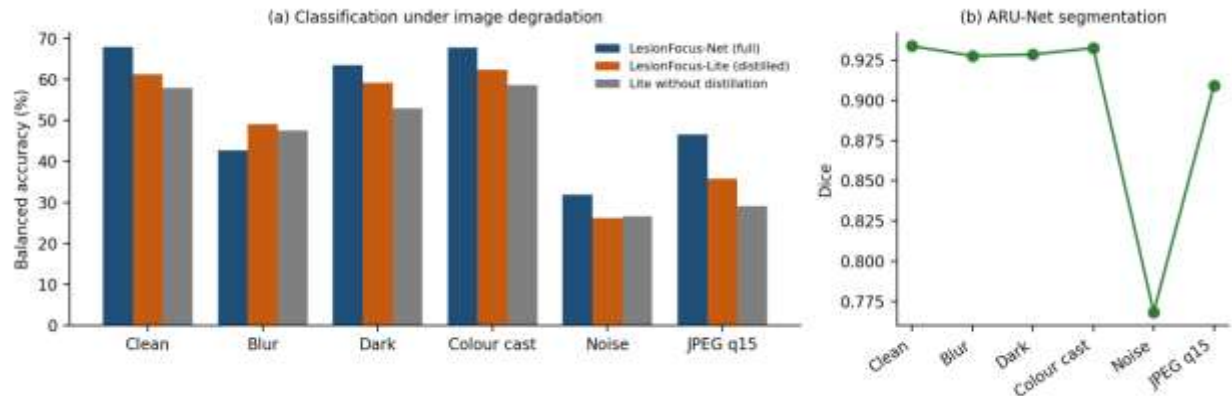


Figure 8. Robustness of the implemented pipeline. (a) Balanced accuracy of the three classifiers when the raw images are degraded before pre-processing. (b) Dice of the ARU-Net segmentation under the same conditions.

5.5 V5: Deployment Profile

Table 8 and Figure 9 summarise the cost of running each tier. On two CPU cores, the complete Tier A pipeline needed about 2.9 s per image: 119 ms for pre-processing, 235 ms for segmentation and 2.53 s for the teacher with eight-fold test-time augmentation. On a T4 GPU the same networks took 25 ms and 31 ms per pass. Tier B needed 127 ms per image on the same two CPU cores, 119 ms of which is pre-processing; the ONNX student itself ran in 7.2 ms, 4.6 times faster than the same network in PyTorch on the CPU (32.9 ms). The Tier B model has 4.2 million parameters and a 16.8 MB file, against 63.8 million parameters for the segmentation and teacher networks together, so the CPU tier is about 23 times faster end to end with 15 times fewer parameters. The ONNX export was numerically faithful: on 300 test images, the largest difference from the PyTorch logits was 1.4×10^{-5} and all 300 predicted classes agreed. We also tried dynamic INT8 quantisation of the ONNX student. It shrank the file to 4.4 MB but was slower on this CPU (56 ms) and changed 23% of the predicted classes, so it was rejected.

Table 8. Deployment profile. CPU times use two threads; GPU times use a T4 with mixed precision; batch size 1.

Component	Params (M)	Size (MB)	CPU time (ms)	GPU time (ms)
Stage 1 pre-processing (OpenCV)	—	—	119	—
ARU-Net segmentation	22.8	91	235	24.9
Dual-stream teacher, single pass	41.0	164	316	31.3
Dual-stream teacher, 8× TTA	41.0	164	2529	251
LesionFocus-Lite, PyTorch	4.2	17	32.9	8.3
LesionFocus-Lite, ONNX Runtime FP32	4.2	16.8	7.2	—
Tier A end to end (CPU-only fallback)	63.8	—	2883	—
Tier B end to end	4.2	16.8	127	—

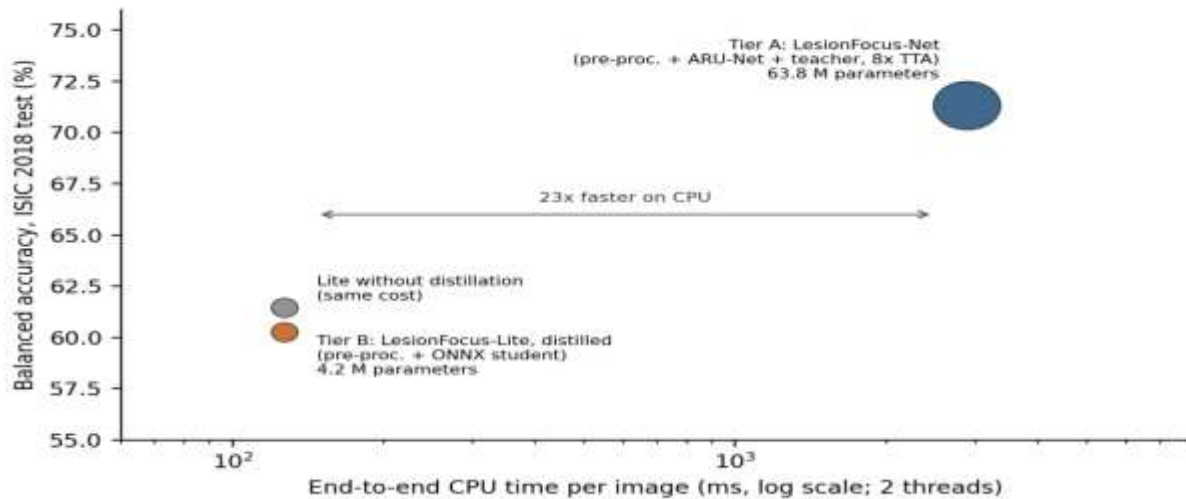


Figure 9. Accuracy–cost trade-off on two CPU cores. Balanced accuracy on the independent ISIC 2018 test set against end-to-end CPU time per image, with bubble area proportional to the number of parameters.

5.6 Validation Summary

Table 9 brings the five levels together. The full framework met every criterion except robustness to noise, blur and heavy compression. Lesion Focus-Lite met the deployment criteria by a wide margin and kept a macro AUC above 0.94 on both test sets, but paid for its speed with 5 to 11 points of balanced accuracy and a lower malignant sensitivity. This digital transformation enhances the convenience of computing, storing, and retrieving medical images records, ultimately leading to improved treatment experiences for patients[48].

Table 9. Outcome of the five-level validation for the full framework (Tier A) and Lesion Focus-Lite (Tier B).

Level	Criterion	Tier A: Lesion Focus-Net	Tier B: Lesion Focus-Lite
V1 Internal test	Balanced accuracy and macro AUC on unseen lesions	Met (BA 65.8%, AUC 0.962)	Partly met (BA 60.4%, AUC 0.947)
V2 Independent test	No loss of performance on ISIC 2018	Met (BA 71.3%, AUC 0.964)	Partly met (BA 60.3%, AUC 0.943)
V3 Robustness	Balanced-accuracy loss under 5 points	Met for colour cast and darkness; not met for blur, noise, JPEG	Met for colour cast and darkness; not met for blur, noise, JPEG
V4 Clinical view	Triage AUC ≥ 0.90 and ECE < 0.05	Met (AUC 0.929 / 0.939; ECE 0.022 / 0.035)	Partly met (AUC 0.892 / 0.887; ECE 0.027 / 0.040)
V5 Deployment	Under 1 s per image on 2 CPU cores; faithful export	Not met on CPU (2.9 s); met on GPU	Met (127 ms; 100% ONNX agreement)

6. Discussion

The first point is that the full framework succeeded upon application to unseen data: the balanced accuracy on the independent ISIC 2018 set (71.3%) was higher than on the internal hold-out (65.8%), the AUC was the same (0.964 against 0.962), and it was well calibrated. This is important, since the framework was built around particular choices, namely trigger-based pre-processing, lesion cropping and class-balanced training, and a design that only worked on the split it was tuned on would be of little value. The gain on the independent set is mostly due to the minority classes, which are better represented there; with only 11 dermatoid fibromas in the internal fold, internal per-class figures for that class say very little.

The second point concerns the deployable tier. Distillation did what it was supposed to on the internal set: a student that sees only the whole image, with no segmentation, gained 2.7 points of balanced accuracy and 5.2

points of macro-F1 over the same network trained on labels alone, learned faster, and arrived better calibrated (fitted temperature 0.93 against 1.32). On the independent set the advantage shrank to a higher AUC and melanoma AUC, with no gain in balanced accuracy. We read this as a sign that, with only six epochs, the student absorbed the teacher's ranking of classes better than its decision boundaries. The gap between teacher and student (5 to 11 points of balanced accuracy) is the price of a 23-fold speed-up on a CPU, and whether that price is acceptable depends on the setting. For a screening aid on a low-cost computer in a primary-care clinic, a model that runs in about a tenth of a second without a GPU may be useful as a first filter even if it is less accurate than the full framework, provided suspicious cases are sent on for a full assessment. The third point is the most practical. The robustness tests show that the weakest link is not any network but the image itself. Pre-processing handled the problems it was designed for: colour casts were almost entirely neutralised, and darkness cost little. Noise, blur and heavy compression, which pre-processing does not address, cut balance accuracy by 10 to 36 points. Blur was especially instructive, because it hurt the full framework more than the light student; the high-resolution lesion stream that makes the full framework strong on clean images makes it fragile when fine texture is lost. Robustness benchmarks in dermatology have reported similar sensitivity to small acquisition changes [32]. Two simple measures follow. A capture-time quality check, using the variance-of-Laplacian measure that Stage 1 already computes, could reject blurred or noisy photographs before they reach the classifier. And photographs sent through messaging applications should be transmitted without heavy compression.

These results also show how the stages of the work fit together. The review that preceded this study found that external testing, robustness and implementation detail were rarely reported; the pre-processing analysis and framework design addressed artefacts, imbalance and lesion focus; and the present implementation and validation show which of those design choices hold up outside the development data and which do not. Colour constancy and class-balanced training held up. Reliance on fine lesion texture is a strength on clean images and a weakness on degraded ones.

7. Limitations

The study has several limitations. Firstly, in order to fit within a single working session on a free cloud GPU, each network was trained just once, on one lesion-level split, and for a small number of epochs (three for segmentation, five for the teacher and six for the students). Longer training and full five-fold cross-validation would probably increase all scores and provide between-fold variability, which a single split cannot. Secondly, the independent test set is official and was used only once, but it was collected by the same contributing centres as HAM10000, testing new patients and lesions rather than new devices or hospitals. Third, the robustness tests use synthetic degradations on a random subset of 1,000 internal test images; real clinical variation is more varied and may interact in ways that single degradations do not capture. Fourth, the internal test set contains only 11 dermatofibromas and 22 vascular lesions, so per-class figures for those classes are unstable. Fifth, latency was measured on two cloud CPU cores and one T4 GPU and will differ on other hardware; a naive dynamic INT8 quantisation of the student broke its predictions and was not pursued further. Finally, the data come mostly from lighter-skinned patients [13], [16], and no dermatologist has yet used the system.

8. Ethical Considerations

All images and metadata come from public, de-identified research datasets; no patient was contacted or imaged for this work. Lesion Focus-Net and Lesion Focus-Lite are research prototypes. They have not been approved for clinical use, have not been tested prospectively, and must not be used to diagnose a lesion or to reassure a patient. The robustness results show that a poorly focused photograph can mislead even the strongest model, so any future use would need an image-quality check at capture time. Before any clinical study, the system would need ethical approval, testing on darker skin tones and on images from other devices, and a design in which a clinician always makes the final decision.

9. Conclusion and Future Work

This study implemented Lesion Focus-Net as working software in two tiers and validated both tiers at five levels. The full framework reached a balanced accuracy of 65.8% and a macro AUC of 0.962 on unseen HAM10000 lesions and 71.3% and 0.964 on the independent ISIC 2018 test set, with a malignant-triage AUC of 0.939 and an expected calibration error below 0.04 on both sets. Its distilled counterpart, Lesion Focus-Lite, needs no segmentation, has 4.2 million parameters and runs end to end in 127 ms on two CPU cores, about 23 times faster than the full pipeline, while keeping a macro AUC above 0.94. Distillation improved the light model on the internal test set and gave a better-calibrated model. The robustness tests showed that

colour and illumination problems are handled well by the pre-processing stage, but that blur, noise and heavy compression remain serious threats to every model.

Future work will follow directly from these findings: full five-fold cross-validation with longer training; testing on images from other devices and hospitals and on darker skin tones; a capture-time image-quality gate; noise- and blur-aware augmentation during training; quantisation-aware training for an integer version of Lesion Focus-Lite; and a prospective reader study with practising dermatologists in which the system supports, rather than replaces, the clinician.

Data Availability

The HAM10000 images, metadata and lesion masks and the ISIC 2018 Task 3 test images and diagnoses analysed in this study are publicly available from the Harvard Data verse.

(<https://doi.org/10.7910/DVN/DBW86T>). The source code of the implementation, the lesion-level split, the trained model weights, the exported ONNX model and the results file are available from the corresponding author on reasonable request.

Conflicts of Interest

The authors confirm that there are no conflicts of interest to disclose.

Funding

This research received no external funding.

Acknowledgements

Statement of Approval. This study has been approved by Department of Computer Application, Integral University, Lucknow, India with Reference No. IU/ R&D/2026-MCN0005118

References

- [1] Abdussami, A. A., & Farooqui, M. F. (2021). Incremental deep neural network intrusion detection in fog based IoT environment: An optimization assisted framework. *Indian Journal of Computer Science and Engineering*, 12(6). <https://doi.org/10.21817/indjcse/2021/v12i6/211206191>
- [2] Abu Owida, H., Abd El-Fattah, I., Abuowaida, S., Alshdaifat, N., Mashagba, H. A., Abd Aziz, A. B., Alzoubi, A., Laguech, S., & Al-Bawri, S. S. (2025). A deep learning-based dual-branch framework for automated skin lesion segmentation and classification via dermoscopic images. *Scientific Reports*, 15, Article 37823. <https://doi.org/10.1038/s41598-025-21783-z>
- [3] Ahamad, M. K., & Bharti, A. K. (2020). Comparative analysis the fitness function of K-means and kernel Fisher's discriminant analysis (KFDA) with genetic algorithm. *European Journal of Molecular & Clinical Medicine*, 7(11).
- [4] Alam, A., & Ahamad, M. K. (2024). Detection of rheumatoid arthritis using CNN by transfer learning (pp. 99–112).
- [5] Arnold, M., Singh, D., Laversanne, M., Vignat, J., Vaccarella, S., Meheus, F., Cust, A. E., de Vries, E., Whiteman, D. C., & Bray, F. (2022). Global burden of cutaneous melanoma in 2020 and projections to 2040. *JAMA Dermatology*, 158(5), 495–503. <https://doi.org/10.1001/jamadermatol.2022.0160>
- [6] Akhtar, R., & Ahamad, M. K. (2026b). Explainable Cross-Validated Greedy Stacking Ensemble with RFECV-Based Feature Selection for T2DM Prediction. *International Journal of Artificial Intelligence and Machine Learning*, 6(10s), 2079–2099. <https://doi.org/10.51483/ijaiml.6.10s.2026.2079-2099>
- [7] Bradski, G. (2000). The OpenCV library. *Dr. Dobbs's Journal of Software Tools*, 25(11), 120–125.
- [8] Brinker, T. J., Hekler, A., Enk, A. H., et al. (2019). Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. *European Journal of Cancer*, 113, 47–54. <https://doi.org/10.1016/j.ejca.2019.04.001>
- [9] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4), 834–848. <https://doi.org/10.1109/TPAMI.2017.2699184>
- [10] Codella, N., Rotemberg, V., Tschandl, P., Celebi, M. E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., Kittler, H., & Halpern, A. (2019). Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the International Skin Imaging Collaboration (ISIC). *arXiv preprint arXiv:1902.03368*. <https://arxiv.org/abs/1902.03368>
- [11] Collins, G. S., Moons, K. G. M., Dhiman, P., et al. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385, e078378. <https://doi.org/10.1136/bmj-2023-078378>

- [12] Cui, Y., Jia, M., Lin, T.-Y., Song, Y., & Belongie, S. (2019). Class-balanced loss based on effective number of samples. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9268–9277). IEEE.
- [13] Daneshjou, R., Vodrahalli, K., Novoa, R. A., et al. (2022). Disparities in dermatology AI performance on a diverse, curated clinical image set. *Science Advances*, 8(31), Article eabq6147.
<https://doi.org/10.1126/sciadv.abq6147>
- [14] Efron, B., & Tibshirani, R. J. (1993). An introduction to the bootstrap. Chapman & Hall/CRC.
- [15] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115–118.
<https://doi.org/10.1038/nature21056>
- [16] Groh, M., Harris, C., Soenksen, L., Lau, F., Han, R., Kim, A., Koochek, A., & Badri, O. (2021). Evaluating deep neural networks trained on clinical images in dermatology with the Fitzpatrick 17k dataset. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 1820–1828). IEEE.
- [17] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (PMLR 70, pp. 1321–1330).
- [18] Hendrycks, D., & Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. In International Conference on Learning Representations (ICLR).
<https://arxiv.org/abs/1903.12261>
- [19] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531. <https://arxiv.org/abs/1503.02531>
- [20] Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q. V., & Adam, H. (2019). Searching for MobileNetV3. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1314–1324). IEEE. <https://doi.org/10.1109/ICCV.2019.00140>
- [21] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7132–7141). IEEE.
- [22] Junejo, R. H., Abbas, Q., Awais, M., Akram, T., & Aldajani, M. B. (2025). Federated ViT: A distributed deep learning framework for skin cancer classification. *IEEE Access*, 13, 166326–166341.
<https://doi.org/10.1109/ACCESS.2025.3612477>
- [23] Karthik, R., Menaka, R., Atre, S., Cho, J., & Easwaramoorthy, S. V. (2024). A hybrid deep learning approach for skin cancer classification using Swin Transformer and dense group shuffle non-local attention network. *IEEE Access*, 12, 158040–158051. <https://doi.org/10.1109/ACCESS.2024.3485507>
- [24] Kaur, R., GholamHosseini, H., & Lindén, M. (2025). Advanced deep learning models for melanoma diagnosis in computer-aided skin cancer detection. *Sensors*, 25(3), Article 594.
<https://doi.org/10.3390/s25030594>
- [25] Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17, Article 195.
<https://doi.org/10.1186/s12916-019-1426-2>
- [26] Khan, E. A. T. (2023). Risk prioritization using a fuzzy based approach in software development design phase. *International Journal on Recent and Innovation Trends in Computing and Communication*, 11(9), 1674–1687. <https://doi.org/10.17762/ijritcc.v11i9.9153>
- [27] Khan, T., & Faisal, M. (2023). An efficient Bayesian network model (BNM) for software risk prediction in design phase development. *International Journal of Information Technology*, 15(4), 2147–2160.
<https://doi.org/10.1007/s41870-023-01244-4>
- [28] Kittler, H., Pehamberger, H., Wolff, K., & Binder, M. (2002). Diagnostic accuracy of dermoscopy. *The Lancet Oncology*, 3(3), 159–165. [https://doi.org/10.1016/S1470-2045\(02\)00679-4](https://doi.org/10.1016/S1470-2045(02)00679-4)
- [29] Krishna, R., & Ahamad, M. K. (2026). GreenAudit-GNN: Sector-stratified benchmarking and dual-method explainability validation for ESG anomaly detection. *International Journal of Artificial Intelligence and Machine Learning*, 6(11s), 1082–1094. <https://doi.org/10.51483/IJAIML.6.11s.2026.1082-1094>
- [30] Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2980–2988). IEEE.
- [31] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 10012–10022). IEEE.
- [32] Maron, R. C., Schlager, J. G., Haggemüller, S., von Kalle, C., Utikal, J. S., Meier, F., Gellrich, F. F., Hobelsberger, S., Hauschild, A., French, L., Heinzerling, L., Schlaak, M., Ghoreschi, K., Hilke, F. J., Poch, G., Heppt, M. V., Berking, C., Haferkamp, S., Sondermann, W., ... Brinker, T. J. (2021). A benchmark for neural network

robustness in skin cancer classification. *European Journal of Cancer*, 155, 191–199.

<https://doi.org/10.1016/j.ejca.2021.06.047>

[33] McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2), 153–157. <https://doi.org/10.1007/BF02295996>

[34] Micikevicius, P., Narang, S., Alben, J., et al. (2018). Mixed precision training. In *International Conference on Learning Representations (ICLR)*. <https://arxiv.org/abs/1710.03740>

[35] Akhtar, R., & Ahamad, M. K. (2026). A comparative evaluation of various machine learning techniques for prediction of Type 2 diabetes mellitus. *ITEGAM- Journal of Engineering and Technology for Industrial Applications (ITEGAM-JETIA)*, 12(60), 1215–1231. <https://doi.org/10.5935/jetia.v12i60.4052>

[36] Oktay, O., Schlemper, J., Le Folgoc, L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., Glocker, B., & Rueckert, D. (2018). Attention U-Net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*. <https://arxiv.org/abs/1804.03999>

[37] ONNX Runtime developers. (2021). ONNX Runtime [Computer software]. <https://onnxruntime.ai>

[38] Paszke, A., Gross, S., Massa, F., et al. (2019). PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32* (pp. 8024–8035).

[39] Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

[40] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 (LNCS 9351, pp. 234–241)*. Springer. https://doi.org/10.1007/978-3-319-24574-4_28

[41] Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning (PMLR 97, pp. 6105–6114)*.

[42] Telea, A. (2004). An image inpainting technique based on the fast marching method. *Journal of Graphics Tools*, 9(1), 23–34. <https://doi.org/10.1080/10867651.2004.10487596>

[43] Tschandl, P., Codella, N., Akay, B. N., et al. (2019). Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: An open, web-based, international, diagnostic study. *The Lancet Oncology*, 20(7), 938–947. [https://doi.org/10.1016/S1470-2045\(19\)30333-X](https://doi.org/10.1016/S1470-2045(19)30333-X)

[44] Tschandl, P., Rosendahl, C., & Kittler, H. (2018). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5, Article 180161. <https://doi.org/10.1038/sdata.2018.161>

[45] Wightman, R. (2019). PyTorch image models [Computer software]. GitHub. <https://doi.org/10.5281/zenodo.4414861>

[46] Yaqoob, M. M., Alsulami, M., Khan, M. A., Alsadie, D., Saudagar, A. K. J., & AlKhathami, M. (2023). Federated machine learning for skin lesion diagnosis: An asynchronous and weighted approach. *Diagnostics*, 13(11), Article 1964. <https://doi.org/10.3390/diagnostics13111964>

[47] Zuiderveld, K. (1994). Contrast limited adaptive histogram equalization. In P. S. Heckbert (Ed.), *Graphics Gems IV* (pp. 474–485). Academic Press.

[48] Abida Khanam, Mohd Faizan Farooqui (2023) Enhancing the Security and Privacy of eHealth Records through Blockchain-based Management: A Comprehensive Framework. *International Journal on Recent and Innovation Trends in Computing and Communication* ISSN: 2321-8169 Volume: 11 Issue: 9 Article Received:25July 2023

[49] Zaiba Haroon, Mohammed Faizan Farooqui, Design of Efficient Framework to Enhance Data Privacy and Data Security, *J. Electrical Systems* 20-11s(2024): 3769-3783.