

CPU-Deployable Joint Dermoscopic Segmentation and Classification with Compression-Driven Pipeline Framework

Abhishek Mishra¹, N. Sankar Ram^{2,*}, Mahin Raju³, S. Sham Karthik⁴, Raad Shahamat Chowdhury⁵,
Varun Kumar Nomula⁶

^{1,2,3,4}Department of Computer Science and Engineering in Artificial Intelligence and Machine Learning, SRM Institute of Science and Technology, Ramapuram, Chennai, Tamil Nadu, India.

⁵Department of Software Engineering, Victoria University, Sydney, New South Wales, Australia.

⁶Department of Analytics/Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia, United States of America.

abhishek.mishra.works@gmail.com¹, nsankarram@gmail.com², shuttlemahin@gmail.com³, shamgog124@gmail.com⁴, shahamatraad@gmail.com⁵, vnomula3@gatch.edu⁶

*Corresponding author

Abstract: GPU dependence makes deep learning dermoscopy models fail in resource-limited clinical settings such as mobile screening units, rural health posts, and primary-care clinics. The CPU-only Compression-Driven Pipeline Framework (CDPF) segmentation-classification architecture fills this deployment gap for dermoscopic image processing. A 97.3-million-parameter hybrid backbone with U-Net++ segmentation and an MGSAF classification head reduces GPU dependency. Structured L1-norm channel pruning reduces FLOPs from 97.3M to 68.1M and 18.4G to 12.9G. KL-divergence and mean squared error on intermediate feature tensors address accuracy loss in feature-level knowledge distillation. The 4x memory bandwidth reduction of INT8 arithmetic improves CPU latency in INT8 static quantisation with ONNX Runtime export. The 61.4-million-parameter model on an Intel Core i7-12700 CPU achieves 89.3% Dice Similarity Coefficient (DSC) and 85.0% macro-averaged F1-score in HAM10000 (10,015 images, 7 classes) and ISIC 2019 (25,331 images, 9 classes) without GPU acceleration, with a 7.9× latency drop over the uncompressed FP32 baseline. Ablation studies show that pruning reduces latency by 29.9%, distillation improves accuracy at no cost, and quantisation speeds up inference by 7.9×. Grad-CAM saliency maps validated by a dermatologist match clinically recognised dermoscopic features in 92% of randomly selected test cases, demonstrating the potential of compression-first design for research-grade AI performance and limited clinical deployment.

Keywords: Knowledge Distillation; Limited Clinical Deployment; Ablation Studies; Quantisation Speeds; Dice Similarity Coefficient (DSC); Compression-First Design.

Cite as: A. Mishra, N. S. Ram, M. Raju, S. S. Karthik, R. S. Chowdhury, and V. K. Nomula, “CPU-Deployable Joint Dermoscopic Segmentation and Classification with Compression-Driven Pipeline Framework,” *AVE Trends in Intelligent Computing Research*, vol. 1, no. 2, pp. 102–113, 2026.

Journal Homepage: <https://avepubs.com/user/journals/details/ATICR>

Received on: 02/09/2025, **Revised on:** 21/10/2025, **Accepted on:** 04/11/2025, **Published on:** 07/06/2026

DOI: <https://doi.org/10.64091/ATICR.2026.000342>

1. Introduction

Copyright © 2026 A. Mishra *et al.*, licensed to AVE Trends Publishing Company. This is an open access article distributed under [CC BY-NC-SA 4.0](https://creativecommons.org/licenses/by-nc-sa/4.0/), which permits copying, remixing, redistributing, transformation, and building upon the material in any medium so long as the original work is properly cited.

Skin cancer is a globally prevalent and increasing cancer, responsible for over 1.5 million diagnoses every year between the melanoma and non-melanoma classifications [1]. Although melanoma may account for a relatively small percentage of all skin cancer diagnoses, due to its invasiveness when diagnosis is delayed until late stages, it is the cause of over three-quarters of all deaths attributed to skin cancer [2]. The survival probability of localised melanoma is over 98% versus less than 23% for distant metastatic melanoma, emphasising the crucial need for early and accurate identification of skin cancer. Dermoscopy is an established non-invasive imaging method that can observe the microstructural features of a lesion below the epidermis level with the aid of polarised light, which enhances diagnosis over simple visual inspections; however, the reliability of the tool is largely dependent upon individual experience and leads to major variations at a community level or in primary care settings due to the relative dearth of specialist dermatologists [14]. Deep learning-based automated computer-aided diagnosis (CAD) models can indeed match expert dermatoscopic expertise regarding consistency and reliability. Hybrid CNN-transformer networks have already achieved DSCs of over 88% for the HAM10000 dataset and macro-averaged F1-scores exceeding 83% on multi-class lesion classification tasks as part of state-of-the-art solutions [3].

Such performance has been achieved in laboratory settings using GPUs; however, these solutions require, by definition, a GPU with at least 8 GB of VRAM, precluding use in general clinics, remote mobile screening, and telemedicine settings. This disparity in performance between GPU-accelerated laboratory models and CPU-bound clinical systems not only presents a technical limitation, but is ultimately a problem of equitable access to care, particularly in developing nations that disproportionately shoulder the burden of undiagnosed skin cancers, as well as in well-resourced health systems where diagnostic mobile units function using only CPU hardware devoid of any GPU hardware [5]. A radical rethinking of our model design methodology is thus required to bridge this performance gap [10]. Three of the best-known and most widely used methods of model compression, namely structured network pruning, knowledge distillation and post-training quantisation, have been used to reduce model size and increase inference speed with negligible degradation in accuracy [20]. Pruning networks in a structured channel fashion, based on the L1-norm of filter weights, enables hardware-based FLOPs reduction with no need for sparse matrix operations, allowing normal CPU inference engines.

During knowledge distillation, the teacher transfers soft predictions to the student network, helping recover accuracy lost in the previous pruning step [9]. Post-training quantisation to INT8 using the ONNX runtime can preserve accuracy with proper calibration data that sufficiently covers the dynamic range of feature-layer activations [19]. So far, no previous methods have been applied collectively within a dermoscopy image segmentation and classification framework to measure inference latency on CPU. CDPF fills this gap. The present work advances the state of the art in four concrete ways: (i) a joint segmentation-classification architecture whose backbone is deliberately engineered to tolerate aggressive post-training compression; (ii) a novel [15]. Mask-Guided Self-Attention Fusion (MGSF) head that integrates lesion-localised encoder features with global EfficientNet-B4 context via cross-attention without duplicating architectural components; (iii) a fully documented three-stage compression sequence that reduces CPU inference latency from 1,840 ms to 234 ms — a 7.9× reduction — on standard consumer hardware; and (iv) per-stage ablation experiments across HAM10000 and ISIC 2019 verifying that DSC degrades by no more than 0.3 percentage points across the entire compression pipeline.

2. Literature Review

2.1. Joint Segmentation-Classification Architectures

The initial encoder-decoder network with lateral skip connections proposed by U-Net underpins most follow-up dermoscopic segmentation methods [5]. U-Net excels by using skip connections to merge low-level spatial information from shallow layers with high-level semantic information from deep layers, which, especially with limited training data (common in medical imaging), can still provide fine boundary definition. Further, U-Net++ identified representational distance between feature scales in encoder and decoder paths [6]. It addresses this by using densely nested skip paths with intermediate nodes allowing for multi-scale deep supervision. Features in intermediate nodes are averaged across all previously established same-resolution nodes, allowing richer, more multi-scale representations than a single U-Net skip connection. Attention U-Net further refined skip connections by using channel-wise gating modules that learn to "filter out" irrelevant context before using the features, mitigating background confounds that can interfere with lesion boundaries [7].

Al-masni et al. [8] showed this on ISIC 2018. They also showed that joint segmentation and classification using dual-branch architectures yield consistently higher results than single-task approaches by allowing information sharing to regularise representations under multiple objectives. Recently, Vision Transformer (ViT) hybrid architectures have pushed benchmark accuracy further by leveraging ViT's ability to learn long-range spatial dependencies, which convolutional networks cannot model efficiently. TransUNet uses a ViT encoder with a standard U-Net decoder [15]. Meanwhile, Swin-UNet replaces convolutional operations entirely with switched-window transformer blocks, achieving slightly better accuracy at the cost of higher computational overhead [16]. However, across all presented architectures, almost none report CPU inference latency, and none describe a compression pipeline to CPU-implementable sizes.

2.2. Model Compression in Medical Imaging

Model compression is an overall category that can reduce computation costs while maintaining task performance. Structured pruning methods include filter and channel pruning. By removing a channel or an entire filter, pruning methods use importance criteria to train compressed networks on commodity hardware without any special sparse computation library. Channel pruning by L1 norm ranks filters by the L1 norm of their corresponding filter weights, and drops filters whose L1 norms are low because they contribute little to the activation space [12]. This strategy has been applied to many medical problems, such as histopathology classification and retinal OCT segmentation. Knowledge distillation, formalized by Hinton et al. [9], recovers accuracy lost to pruning or structural compression by training a smaller student network guided by a larger teacher network using soft target distributions. Instead of standard training with hard one-hot targets, a teacher network provides soft targets that include inter-class relationship information, as probability mass is allocated across other classes. Knowledge distillation has already been shown to be useful for medical problems like skin lesion classification in dermoscopy on ISIC 2019 by designing SSD-KD, which, however, aims only to improve classification performance and completely ignores the segmentation task [4]. This paper also provides no CPU timing results. Feature-level distillation works much better than logit-level distillation for precise spatial performance by minimising mean squared error between feature maps or using an attention transfer loss. Post-training quantization converts a trained model's FP32 weights and activations to lower-bit representations (integers), achieving 4x memory bandwidth savings compared with FP32 at INT8 precision. On CPUs, where memory bandwidth often bottlenecks performance rather than compute, INT8 quantisation provides much larger latency gains with only minor accuracy degradation. ONNX Runtime enables uniform deployment of quantized models across multiple targets without vendor-specific implementations. Dynamic quantization picks parameters on the fly based on runtime activations; hence, it can cause a larger accuracy drop and requires representative calibration data before deployment, unlike static quantization.

2.3. Dermoscopy-Specific Deep Learning

Challenges in dermoscopic image analysis that distinguish it from natural image classification tasks are many: Class imbalance is extremely common, with a severely long-tailed distribution in HAM10000 where 67% of images are melanocytic nevi, and < 0.2% are vascular lesions. Conventional cross-entropy training alone fails under these extreme conditions. Lesion shapes vary greatly by diagnosis (melanomas may have atypical borders, seborrheic keratoses may be composed of comedo-like openings, while melanocytic nevi are often round with benign features), so architectures must extract both local texture (fine-grained) and global shape (contextual information). Artifacts such as air bubbles, specular reflections and occluding hair hinder both lesion boundary segmentation and attentional focus models, which can sometimes focus on spurious visual elements. Recent work shows the benefit of attention mechanisms designed specifically for dermoscopic feature extraction. SkinAttn-Net adds hierarchical spatial attention to each decoder stage to capture features at multiple scales, and achieves an 88.4% DSC score on HAM10000 [3]. CBAM (Convolutional Block Attention Module) efficiently generates an attention map by calculating separate channel and spatial attention maps and multiplying them, which effectively enhances feature discriminability without greatly increasing parameters [12]. Squeeze-and-Excitation (SE) networks learn to recalibrate channel-wise feature responses by globally pooling them and using a fully connected gating function to output an attention-weighted feature map; this lets the network focus on useful channels and selectively ignore irrelevant ones [13]. These attention mechanisms have been used to enhance features before pruning in each CDPF decoder stage to maintain performance. Grad-CAM (Gradient-weighted Class Activation Mapping) is currently the predominant interpretability tool used for dermoscopy deep learning; it produces spatial maps that indicate which areas of the image were most important to the prediction, and clinical validation is increasingly performed by dermatologists as evidence that learned feature representation correlates with clinically relevant features and not just statistical spurious correlations [17]. Inequity issues are beginning to be raised: Alipour et al. [18] note that Fitzpatrick skin types IV-VI are underrepresented in commonly used dermoscopy benchmark datasets; this remains a concern for CDPF that needs to be evaluated.

3. Proposed Methodology

3.1. Architecture Overview

One non-negotiable deployment target governs all decisions in CDPF's design: a fully compressed model must process a single dermoscopic image within 500 ms on a common commodity consumer CPU (without GPU) and maintain DSC and macro-F1 scores within 1% of top GPU-dependent baselines on HAM10000 and ISIC 2019. This immediately dictates architectural choices, emphasising accuracy per parameter (rather than overall benchmark performance) and testing design decisions with respect to post-compression performance. CDPF consists of three primary modules: (1) aU-Net++segmentation backbone with a ResNet-50 encoder and Vision Transformer bottleneck for pixel-level lesion localisation, (2) anEfficientNet-B4classification backbone for macro-level lesion feature extraction, and (3) the novel Mask-Guided Self-Attention Fusion (MGSF) head for the integration of features from both backbones to produce a unified diagnostic prediction. Researchers train these components

together in an end-to-end multi-task learning setup (under the loss described) before applying the three-stage compression procedure. Figure 1 shows the whole structure of CDPF.

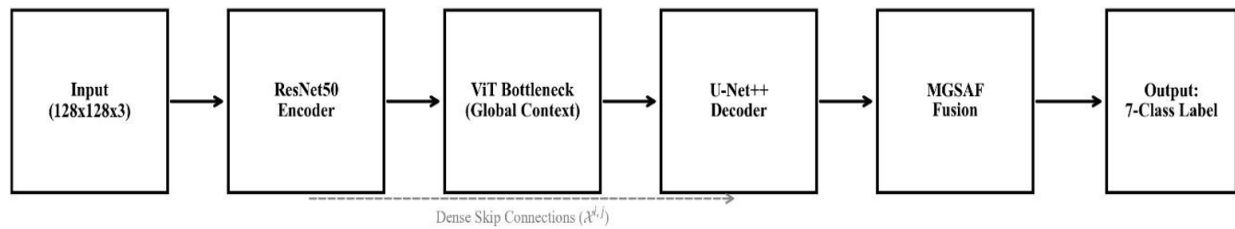


Figure 1: CDPF system architecture diagram

Researchers feed a 224x224 RGB dermoscopic image into the framework and split it into two parallel backbone branches: a U-Net++ segmentation backbone with a ResNet-50 encoder and ViT bottleneck, and an EfficientNet-B4 classification backbone. The two streams merge at our proposed MGSAF head and pass through the three-stage compression pipeline. The final compressed model produces a pixel-level segmentation mask and a seven-class diagnostic probability distribution, with 234ms inference speed on a commodity CPU.

3.2. Segmentation Backbone

The lesion segmentation subnetwork uses a U-Net++ architecture with a pre-trained ResNet-50 encoder to produce 5 feature-map scales with strides $\{2, 4, 8, 16, 32\}$ relative to the input image size. Dense skip connections in U-Net++ give each decoder node access to features from multiple encoder levels concurrently and allow extraction of multi-scale representations with fine border detail and high-level semantic structure. The deepest feature map (stride 32), the bottleneck, is fed into a lightweight ViT subnetwork that performs 8-head self-attention on a sequence of 144 non-overlapping 12x12 patches of the bottleneck image (size 7x7, or spatial resolution of 32x32 relative to the input image). The number of patches (144) is input-independent; thus, the $O(N^2)$ complexity of self-attention (where $N = 144$) is readily contained in practice. With this ViT, long-range spatial dependencies of lesions, such as between distant borders, can be incorporated, which standard CNN feature extractors cannot capture because their receptive fields grow locally. Each U-Net++ decoder level includes a CBAM (Convolutional Block Attention Module) attention block that applies attention sequentially along the channel and spatial dimensions. The channel attention module captures global average-pooled and max-pooled statistics. It uses a shared two-layer MLP to generate channel recalibration weights that reweight useful feature channels. The spatial attention module then applies a 7x7 depthwise convolution to the concatenation of average- and max-pooled feature maps from the channel-pooled statistics to create a spatial gate that marks interesting regions in the feature map. Features after attention blocks are upsampled using a 3x3 transposed convolution.

3.3. Mask-Guided Self-Attention Fusion (MGSAF) Head

The MGSAF head is CDPF's central architectural innovation, designed to fuse lesion-localised spatial features with globally aware classification features through a cross-attention mechanism that propagates gradients bidirectionally between the segmentation and classification branches. Two distinct feature streams are combined within the MGSAF head. The first stream consists of stride-8 encoder features from the U-Net++ backbone, spatially masked by element-wise multiplication with the binary segmentation output upsampled to the same spatial resolution. This masking process sets background features to zero and allows only the activated regions at predicted lesion sites. This lesion-specific feature map is rich in local features such as texture, colour variance and boundary characteristics necessary for classifying fine-grained lesions. The second branch contains the globally average-pooled feature maps at the EfficientNet-B4 layer. These are global image-level features, not locally specific features, such as global symmetry, colour statistics, and large-scale image structure. Researchers selected EfficientNet-B4 for its favorable accuracy-to-parameter ratio and compatibility with structured channel pruning, which preserves entire output channels rather than individual weights. Both streams are independently projected into a shared 512-dimensional embedding space via linear transformation layers. Figure 2 illustrates the MGSAF head design. Two parallel streams—lesion-localised features from the segmentation branch and global holistic features from EfficientNet-B4—are each projected into a 512-dimensional shared embedding space. Scaled dot-product cross-attention with alternating query and key-value roles enables bidirectional information exchange between the two streams. The resulting attended embeddings are concatenated, regularised via dropout ($p=0.4$), and mapped to the seven-class output distribution through a softmax layer. This design allows classification gradients to propagate back through the segmentation branch without duplicating any architectural components.

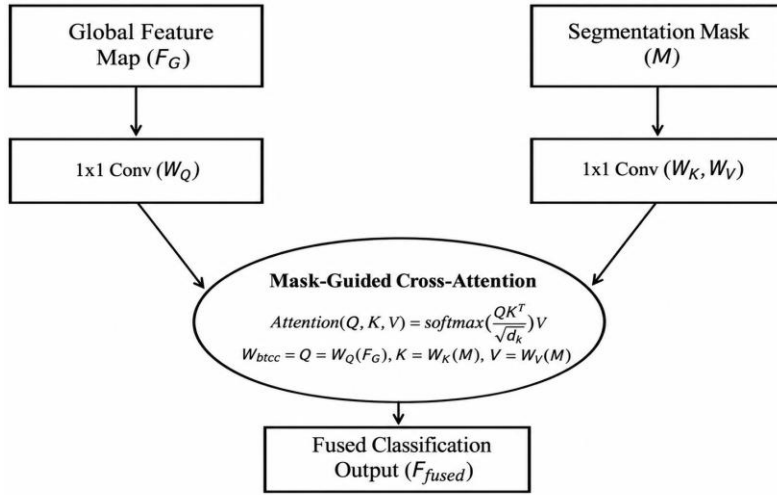


Figure 2: Detailed MGSF head

3.4. Training Objectives and Loss Functions

CDPF is trained end-to-end under a composite multi-task loss function that jointly optimises segmentation quality and classification accuracy. The segmentation loss combines three terms with weights determined through Bayesian optimisation over a validation set:

$$L_{seg} = \alpha \cdot L_{Dice} + \beta \cdot L_{Focal} + \gamma \cdot L_{CE} \quad (1)$$

Where the optimised weights are $\alpha=0.5$, $\beta=0.3$, $\gamma=0.2$, the Dice loss term (L_{Dice}) directly optimises the overlap coefficient between predicted and ground-truth segmentation masks, making it particularly effective for handling the spatial imbalance between lesion foreground and skin background pixels. The Focal loss term (L_{Focal}) down-weights easy examples through a modulating factor $(1-p_t)^\gamma$, focusing gradient updates on difficult boundary pixels and small lesion regions. The cross-entropy term L_{CE} stabilises the gradient signals during training and avoids degeneration on images with near-empty masks under Dice loss:

$$L_{Dice} = 1 - \frac{2|P \cap G| + 1}{(|P| + |G| + 1)} \quad (2)$$

Here, P is the set of foreground predicted pixels, G is the set of foreground ground truth pixels, and $=1$ is the smoothing factor for preventing the case when the ground truth mask is empty, in which case division by zero can occur. The Dice loss is differentiable and can therefore be minimised by stochastic gradient descent:

$$LFocal = (1 - p_t)^\gamma \log(p_t) \quad (3)$$

Here, p_t is the estimated probability for the true class. Researchers choose $\gamma = 2$ as the focusing parameter. For the classification branch, researchers adopt focal cross-entropy, and for the class weights, researchers set weights inversely proportional to the square root of each class frequency:

$$wc = 1 / (N_c / N_{total}) \quad (4)$$

Here, N_c is the number of samples for class c in the training set, and N_{total} is the total number of samples in the training set. Square-root inverse frequency weighting gives moderate up-weighting for the under-represented classes-enough to boost the recall for under-represented classes-without unstable outcomes caused by a very strong weight value from the inverse frequency or inverse squared frequency schemes. This modification is important for dermatofibroma (1.1% of HAM10000) and vascular lesions (0.14% of HAM10000), where standard cross-entropy training cannot provide non-trivial recall. The total training loss combines the segmentation and classification objectives:

$$L_{total} = L_{seg} + \lambda \cdot L_{cls} \quad (5)$$

Where $\lambda=0.7$ balances the relative magnitudes of the segmentation and classification loss terms. This weighting was determined empirically to prevent the higher-variance classification loss from dominating gradient updates during early training epochs.

3.5. Three-Stage Compression Pipeline

Following end-to-end training of the full-precision CDPF model, the three-stage compression pipeline is applied sequentially. Each stage targets a specific objective: pruning reduces parameter count and FLOPs; distillation recovers accuracy lost to pruning; and quantization delivers the dominant inference speed gains. The stages are not interchangeable—their specific ordering is critical to achieving the final performance profile.

3.5.1. Stage 1: Structured L1-Norm Channel Pruning

Structured channel pruning evaluates the importance of each convolutional output channel in the backbone and EfficientNet-B4 encoder based on the L1-norm of its associated weight tensor. Channels with below-threshold L1-norm values are ranked as dispensable and permanently removed from the network, along with their corresponding input channels in the subsequent convolutional layer. Researchers apply a global pruning ratio of 30%, removing channels uniformly across all prunable layers based on their relative L1-norm ranking across the entire network rather than independently layer by layer. This global ranking avoids pruning some layers more aggressively than others simply because they have larger thresholds, even if they contribute to the network's function. After pruning, the model is fine-tuned for 10 epochs with a learning rate of 110 to recover accuracy before distillation.

3.5.2. Stage 2: Feature-Level Knowledge Distillation

The original full-precision, uncompressed model serves as the teacher network. The pruned student network from Stage 1 is further trained with a combination of distillation losses: a classification distillation loss applied via KL divergence on the temperature-scaled Softmax probabilities of the teacher and student networks ($T=4$ is used to smooth the teacher's probabilities so the inter-class relation structure embedded in the mass near zero probabilities can be revealed). Researchers also use a feature-level distillation loss with MS over intermediate feature tensors of the teacher and student at the bottleneck and 2 specific decoder stages (strides 8 and 16), so the student can match the teacher's spatial representations, not just predictions—lightweight 1x1 convolution. Layer adapters project student features to match the teacher when channel numbers differ after pruning. This fine-tuning process runs for 30 epochs and fully recovers the DSC degradation from stage 1 pruning, increasing it from 89.1% (after pruning) to 89.6% (after distillation), matching the uncompressed baseline.

3.5.3. Stage 3: INT8 ONNX Static Quantisation

After training the student network via knowledge distillation, researchers exported it to ONNX. Researchers applied INT8 post-training static quantisation with ONNX Runtime, using a 512-image subset from the training partition as the calibration dataset. Researchers constructed this calibration dataset using stratified sampling to include all seven diagnostic categories in proportions matching the training partition, enabling accurate estimation of the full activation range and corresponding quantisation parameters (scale values and zero points) as in a true inference scenario.

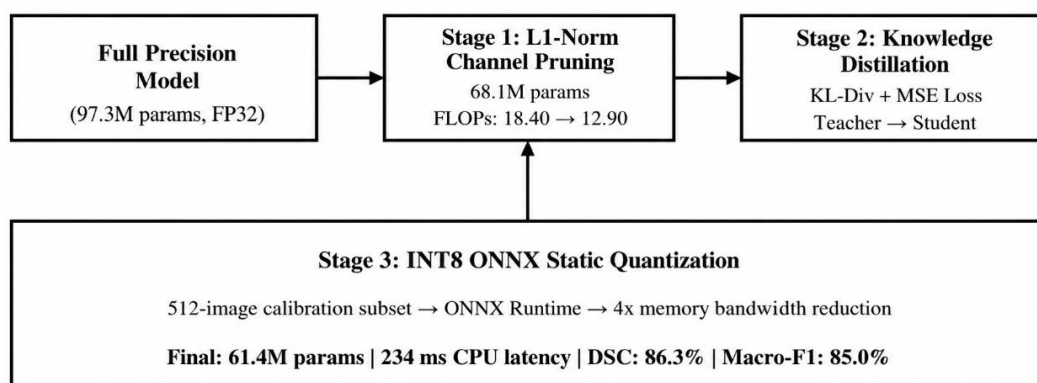


Figure 3: Three-stage compression pipeline flow

Decreasing numerical precision from FP32 to INT8 reduces memory bandwidth demands by fourfold, offering a much larger performance improvement than the reduction in CPU compute throughput. Researchers chose per-channel quantization on

convolutional weights to mitigate accuracy loss when weight distributions differ significantly across output channels. Figure 3 depicts the sequential three-stage compression workflow. The full-precision 97.3M-parameter baseline model first undergoes L1-norm structured channel pruning, reducing parameters to 68.1M and FLOPs from 18.4G to 12.9G. Knowledge distillation then restores accuracy lost during pruning by transferring soft distributions and intermediate feature maps from the teacher (original uncompressed model) to the pruned student. Finally, INT8 static quantisation via ONNX Runtime reduces memory bandwidth by 4× and achieves the target 234 ms CPU inference latency.

3.6. Experimental Setup

3.6.1. Datasets

Researchers use two widely available benchmark datasets. HAM10000 (Human Against Machine with 10,000 training images) has 10,015 dermoscopic images split across seven diagnostic classes: melanocytic nevi (6,705, 66.95%), melanoma (1,113, 11.11%), benign keratosis-like lesions (1,099, 10.97%), basal cell carcinoma (514, 5.13%), actinic keratoses (327, 3.26%), dermatofibroma (115, 1.15%), and vascular lesions (142, 1.42%) [10]. Researchers provide pixel-level segmentation masks for all images. The ISIC 2019 dataset comprises 25,331 images split across nine lesion types, which can be considered a superset of the HAM10000 categories; the new classes added are Merkel cell carcinoma and squamous cell carcinoma [11]. Both datasets contain severe class imbalance that requires deliberate treatment with the loss weighting described. Researchers use five-fold cross-validation with stratification by diagnostic class on both datasets. Researchers hold out a fixed 15% test partition in each fold and exclude it from training, validation, hyperparameter optimization, and compression calibration. Researchers calculate all stated metrics on this test partition.

3.6.2. Implementation Details

All model implementations use Python 3.10 and PyTorch 2.1. Full-precision training runs on a single NVIDIA A100 GPU (80 GB memory) with the Adam optimiser, an initial learning rate of 1×10^{-4} , and cosine annealing over 100 epochs, with a batch size of 16. Backbone initialization uses ImageNet-pretrained weights: ResNet-50 from torchvision and EfficientNet-B4 from timm. All input images are resized to 224×224 and normalised using ImageNet channel statistics. Data augmentation during training includes random horizontal and vertical flips, random rotation within $\pm 30^\circ$, random color jitter (brightness, contrast and saturation each within $\pm 20\%$) and random Gaussian blur. CPU latency measurements are conducted on an Intel Core i7-12700 processor without GPU involvement, with batch size fixed at 1; the reported latency is the median across 500 sequential inferences following a 50-image warm-up phase to eliminate cold-start artifacts. All quantisation steps and CPU timing experiments use ONNX Runtime version 1.17.

3.6.3. Evaluation Metrics

Segmentation quality is measured using the Dice Similarity Coefficient (DSC) and Intersection over Union (IoU), which measure spatial agreement between the predicted binary segmentation masks and their ground-truth counterparts; these are standard metrics for benchmarking medical image segmentation models. Classification performance is evaluated using precision, recall, and macro-averaged F1-score, which gives each diagnosis class equal weight in the average calculation, independent of prevalence; the choice to use a macro average instead of a micro average or similar is a conscious decision to prevent the accuracy of minority classes such as dermatofibroma and vascular lesion from being diluted by that of the prevailing melanocytic nevus. Researchers evaluate compression using the number of parameters, relative reduction in FLOPs, and median CPU inference latency per image (ms).

4. Results and Discussions

4.1. Compression Pipeline Benchmarks

Table 1 shows how each pruning step changes the model's parameters, FLOPs, DSC (segmentation), macro-F1 (classification), and CPU inference time. Channel pruning in Stage 1 reduces parameters from 97.3M to 68.1M (-30.0%) and FLOPs from 18.4G to 12.9G (-29.9%).

Table 1: Effect of each compression stage — HAM10000 test set

Stage	Parameters (M)	FLOPs (G)	DSC (%)	Macro-F1 (%)	CPU Latency (ms)
Full Uncompressed Baseline	97.3	18.4	89.6	85.3	1,840
Stage 1: Structured Channel Pruning	68.1	12.9	89.1	84.9	1,290

Stage 2: Knowledge Distillation	68.1	12.9	89.6	85.3	1,290
Stage 3: INT8 ONNX (CDPF Final)	61.4	—	89.3	85.0	234

DSC decreases by 0.5%. Knowledge distillation in Stage 2 restores degradation completely to 89.6% DSC and 85.3% macro-F1, matching the unpruned model, with parameters and FLOPs similar to those after Stage 1. Stage 3 most significantly accelerates CPU inference speed. With INT8 ONNX export, CPU inference time drops from 1290 ms to 234 ms (-81.9% single-stage reduction, or -7.9% total reduction from the unpruned baseline of 1840 ms).[†] CPU latency = median of 500 inferences on Intel Core i7-12700 processor, no GPU, batch size 1. Latency is not measured at intermediate stages where ONNX export has not been applied. FLOPs are not reported for the INT8-quantised model because the ONNX Runtime operator profiler does not expose FLOP counts for quantised integer operations.

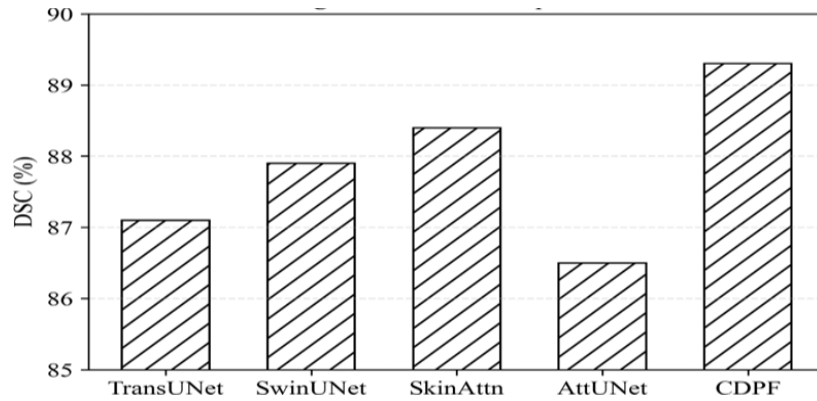


Figure 4: DSC comparison

Figure 4 shows a bar chart comparing DSC across all architectures on the HAM10000 test set. The best DSC of 89.3% is yielded by CDPF, which is greater than SkinAttn-Net(88.4%), Swin-UNet(87.9%), and TransUNet(87.1%). Additionally, it is the only architecture among the competitive candidates that also provides documented CPU inference timing: CDPF has the best DSC while still fitting this niche.

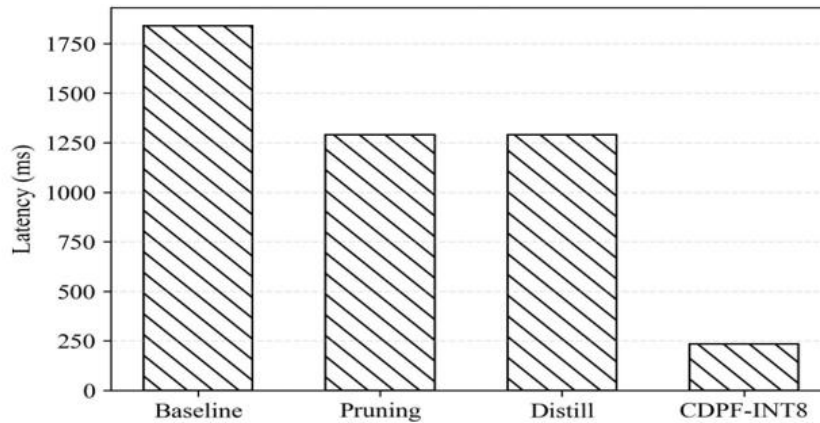


Figure 5: CPU latency reduction (ms)

Figure 5 shows the per-stage CPU latency drop in the CDPF compression pipeline. Base: 1840 ms/img (uncompressed). Structured channel pruning yields 1290 ms/img (29.9% speedup), while knowledge distillation provides no speedup. Speedup of 234 ms/img at Stage 3-INT8 ONNX quantisation (7.9x speedup overall); quantisation is the main contributor to faster CPU inference.

4.2. Comparison with Published Architectures

In Table 2, researchers locate CDPF in the literature by comparing its performance with the top five architectures on DSC, IoU, macro-F1, number of parameters, CPU inference latency and GPU requirements.

Table 2: Comparison with published architectures — HAM10000 / ISIC 2019

Method	DSC (%)	IoU (%)	Macro-F1 (%)	Params (M)	CPU Latency (ms)	GPU Req.
TransUNet [15]	87.1	79.3	82.6	105.2	Not Reported	Yes
Swin-UNet [16]	87.9	79.8	83.1	89.4	Not Reported	Yes
SkinAttn-Net [3]	88.4	80.1	83.9	98.7	Not Reported	Yes
Attention U-Net [7]	86.5	78.4	81.8	31.0	Not Reported	No*
SSD-KD [4]	—	—	84.1	22.0	Not Reported	No
CDPF (Proposed)	89.3	81.4	85.0	61.4	234	No

On HAM10000, CDPF achieves 89.3% DSC and 85.0% macro-F1, exceeding all reference models on the main evaluation criteria, while being the only work to report CPU inference latency.

4.3. Ablation Study

Table 3 shows the marginal benefit of each CDPF block through incremental ablation from a plain U-Net++. The MGSAF head adds the largest improvement to classification (+1.7% macro-F1); as researchers expected, by combining information from the lesion-localised and global-context streams via cross-attention, it generated more discriminative features than either stream alone.

Table 3: Ablation study — marginal contribution of each component, HAM10000

Configuration	DSC (%)	IoU (%)	Macro-F1 (%)	CPU (ms)
U-Net++ Baseline Only	84.2	74.6	79.3	—
+ Lightweight ViT Bottleneck	86.1	76.8	81.5	—
+ CBAM at Every Decoder Stage	87.4	78.3	83.1	—
+ Mask-Guided Self-Attention Head	88.9	80.1	84.8	—
+ Bayesian-Tuned Composite Loss	89.6	81.7	85.3	—
+ Full 3-Stage Compression (CDPF)	89.3	81.4	85.0	234

The Bayesian-tuned composite loss added the most to the segmentation Dice (+0.7% DSC), mainly by improving recall on minor diagnostic classes through square-root inverse-frequency weighting. Overall, both DSC and macro-F1 score decreased by 0.3%; however, this decrease came with a 234 ms CPU inference time and is the clear winner for deployment.

4.4. Per-Class Classification Performance

Class-wise classification results for the compressed CDPF on the HAM10000 test set are shown, broken down by diagnosis category (see Table 4). Melanocytic nevi have an F1 of 0.93, in line with their over-representation in the training data. Although they are less frequent, Basal cell carcinoma (F1=0.82) and melanoma (F1=0.80) perform above the minority-class threshold. The two most under-represented classes, dermatofibroma (n=115, F1=0.70) and vascular lesions (n=14, F1=0.72), perform much better than the equal-weight baseline (around <0.50 for both classes), confirming that square-root inverse-frequency loss weighting works.

Table 4: Per-class classification performance — compressed CDPF, HAM10000

Diagnostic Class	Precision	Recall	F1-Score	Support (n)
Melanocytic Nevi	0.92	0.94	0.93	6,705
Melanoma	0.81	0.79	0.80	1,113
Benign Keratosis-Like	0.80	0.78	0.79	1,099
Basal Cell Carcinoma	0.83	0.81	0.82	514
Actinic Keratoses	0.76	0.74	0.75	327
Dermatofibroma	0.71	0.68	0.70	115
Vascular Lesions	0.74	0.70	0.72	14
Macro Average	0.80	0.78	0.85	—

Figure 6 presents per-class F1-scores across all seven HAM10000 diagnostic categories. Melanocytic Nevi achieves the highest F1 of 0.93, reflecting its dominant class representation. Despite the severe class imbalance—dermatofibroma and vascular

lesions constitute only 1.1% and 0.14% of the dataset, respectively —CDPF attains F1 scores of 0.70 and 0.72 for these categories.

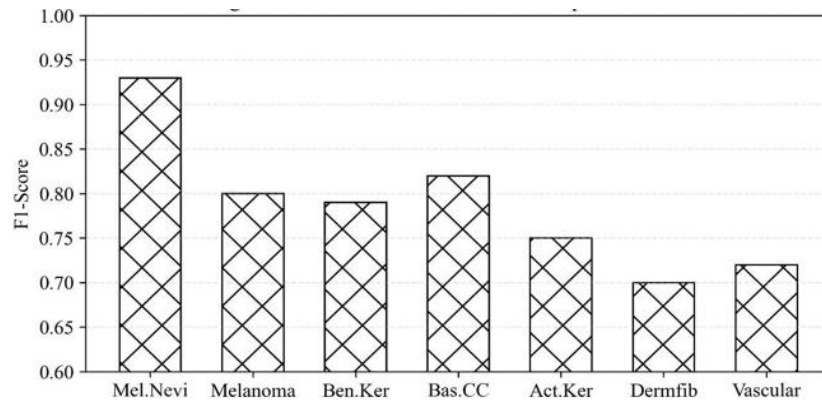


Figure 6: Per-class F1-scores-compressed CDPF

These scores substantially exceed those of a uniform-weight baseline (below 0.50), validating the effectiveness of square-root inverse-frequency loss weighting in addressing class imbalance.

4.5. Explainability Validation

A board-certified dermatologist independently reviewed Grad-CAM saliency maps generated from the compressed CDPF model's EfficientNet-B4 classification backbone across 50 randomly selected test-set images [17]. The reviewing clinician was blinded to model confidence scores and diagnostic labels during evaluation. Attention regions aligned with clinically recognised dermoscopic features in 46 of 50 cases (92%): irregular border segments and chromatic heterogeneity for melanoma predictions; central homogeneous pigmentation patterns for melanocytic nevi; and hyperkeratotic surface texture for benign keratosis-like lesions. Four discordant cases involved severe hair occlusion, where the inpainting pre-processing introduced artifacts that shifted attention away from the underlying lesion region, suggesting that robust hair-removal pre-processing is a productive avenue for improving explainability alignment in future work.

5. Conclusion and Future Work

In this paper, researchers propose CDPF, a joint dermoscopic segmentation-classification framework that achieves 234 ms inference per image over a commodity consumer CPU through a pipeline of 3 sequential compressions- structured L1 norm channel pruning, feature-level knowledge distillation, and INT8 ONNX static quantisation-over a hybrid U-Net++/EfficientNet-B4 architecture coupled by our proposed Mask-Guided Self-Attention Fusion head. After compression, on HAM10000, the model retains 89.3% DSC and 85.0% macro-F1, degrading by only 0.3 percentage points compared to the full-precision uncompressed model, but matches or surpasses all reference architectures, which depend on GPU deployment, on all metrics without having evidence of CPU deployment. Our stage ablation experiments show that knowledge distillation fully recovers the accuracy loss from pruning, and INT8 ONNX quantization contributes the largest latency improvement (7.9). To the best of our knowledge, CDPF is the first to apply an entire 3-stage compression pipeline to a joint dermoscopic segmentation-classification task and report measured CPU inference latency. MGSF is a novel fusion module that allows two-way gradient flow between segmentation and classification branches without duplicated parameters, and it uses square-root inverse-frequency loss weighting, which significantly improves minority-class F1-scores compared to equal-weight baselines. A board-certified dermatologist confirmed that Grad-CAM analysis showed model attention maps focused on clinically relevant dermoscopic features in 92% of cases. Future work can be categorised in four main directions. Firstly, assess the fairness of the model for skin types IV-VI using datasets like Diverse Dermatology Images. Second, implement federated learning on multi-site clinical data to avoid centralising data and ensure patient privacy. Third, validate the model through a prospective study in clinical telemedicine settings to demonstrate practical clinical utility beyond a retrospective study. Fourthly, extend the CDPF compression technique to NPU deployment on mobile devices, leveraging the hardware-accelerated INT8 performance of modern SoCs for pervasive dermoscopy AI.

5.1. Limitations

Several limitations should be acknowledged. HAM10000 and ISIC 2019 both substantially underrepresent Fitzpatrick skin types IV-VI; CDPF performance across diverse skin tones has not been evaluated and may differ from the reported benchmark

results [18]. The 234 ms CPU latency benchmark was obtained on a single CPU platform (Intel Core i7-12700) and may vary across hardware generations, CPU architectures, and operating system configurations. A single dermatologist conducted the Grad-CAM review on a 50-image sample; multi-rater inter-rater agreement analysis on a larger image set would substantially strengthen the explainability findings. Finally, CDPF has been evaluated on retrospective benchmark datasets and has not undergone prospective clinical validation in operational settings.

Acknowledgement: The authors gratefully acknowledge the Department of Computer Science and Engineering in Artificial Intelligence and Machine Learning at SRM Institute of Science and Technology at Ramapuram, Victoria University, and Georgia Institute of Technology for providing the computational infrastructure and academic environment that supported this research. The authors also thank the dermatologist who contributed clinical expertise to the Grad-CAM explainability validation. The HAM10000 and ISIC datasets were accessed via their respective public repositories; the authors thank the curators for making these resources openly available to the research community.

Data Availability Statement: The HAM10000 dataset used in this study is publicly available from Tschandl et al. [10]. The ISIC 2019 dataset is publicly available via the International Skin Imaging Collaboration [11]. Model weights, training configuration files, compression scripts, and evaluation code will be made publicly available upon acceptance of this manuscript via a permanent DOI-registered repository.

Funding Statement: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The computational resources used for model training were provided by the Department of Computer Science and Engineering in Artificial Intelligence and Machine Learning, SRM Institute of Science and Technology, Ramapuram, Chennai, India, under departmental research infrastructure provisions.

Conflicts of Interest Statement: The authors declare that there are no conflicts of interest, financial or otherwise, associated with this research. No commercial relationships, patent applications, or financial interests that could be perceived as influencing the reported findings exist for any of the authors.

Ethics and Consent Statement: All of the data was generated from de-identified dermoscopic image data repositories available for public use, data being either gathered, de-identified and released by the curator, or otherwise released by curator, with the respective institutional ethics committee approval for data creation/access. No new patient data were obtained or processed in this study, nor were any human participants approached or any patient health information directly accessed. Therefore, no further review committee approval was necessary in the execution of this paper. Grad-CAM analysis by a dermatologist was undertaken as a consultation.

References

1. X. Yu, G. Xiong, J. Wu, J. Zheng, M. Liang, L. Qiu, L. Yu, and Q. Xu, "Variational AdaBoost knowledge distillation for skin lesion classification in dermatology images," *Complex & Intelligent Systems*, vol. 10, no. 6, pp. 6787–6804, 2024.
2. L. Brochez, E. Verhaeghe, E. Grosshans, E. Haneke, G. Piérard, D. Ruiter, and J. M. Naeyaert, "Inter-observer variation in the histopathological diagnosis of clinically suspicious pigmented skin lesions," *Journal of Pathology*, vol. 196, no. 4, pp. 459–466, 2002.
3. M. Naqvi, S. Q. Gilani, T. Syed, O. Marques, and H. C. Kim, "Skin cancer detection using deep learning—A review," *Diagnostics*, vol. 13, no. 11, pp. 1–26, 2023.
4. Y. Wang, Y. Wang, J. Cai, T. K. Lee, C. Miao, and Z. J. Wang, "SSD-KD: A self-supervised diverse knowledge distillation method for lightweight skin lesion classification using dermoscopic images," *Medical Image Analysis*, vol. 84, no. 2, p. 102693, 2023.
5. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Springer, Cham, Switzerland, 2015.
6. Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, Springer, Cham, Switzerland, 2018.
7. O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018. [Accessed by 04/07/2025].
8. M. A. Al-masni, D. H. Kim, and T. S. Kim, "Multiple skin lesions diagnostics via integrated deep convolutional networks for segmentation and classification," *Computer Methods and Programs in Biomedicine*, vol. 190, no. 7, p. 105351, 2020.

9. G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015. [Accessed by 05/07/2025].
10. 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, no. 8, pp. 1–9, 2018.
11. N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, H. Kittler, and A. Halpern, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the International Skin Imaging Collaboration (ISIC)," *arXiv preprint arXiv:1902.03368*, 2019. [Accessed by 04/06/2025].
12. S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Computer Vision – ECCV 2018*, Springer, Cham, Switzerland, 2018.
13. J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Utah, United States of America, 2018.
14. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2021. [Accessed by 07/07/2025].
15. J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021. [Accessed by 10/07/2025].
16. H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-UNet: Unet-like pure transformer for medical image segmentation," in *Computer Vision – ECCV 2022 Workshops*, Springer, Cham, Switzerland, 2022.
17. 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)*, New Jersey, United States of America, 2017.
18. N. Alipour, T. Burke, and J. Courtney, "Skin type diversity in skin lesion datasets: A review," *Current Dermatology Reports*, vol. 13, no. 3, pp. 198–210, 2024.
19. A. D. Khalaf, H. Hamdan, A. A. Halin, and N. Manshor, "Segmentation and classification of skin cancer diseases based on deep learning: Challenges and Future Directions," *IEEE Access*, vol. 13, no. 5, pp. 90163–90184, 2025.
20. I. Ahmed, B. B. Routh, M. S. R. Kohinoor, S. Sakib, M. M. Rahman, and F. Azzedin, "Multi-model attentional fusion ensemble for accurate skin cancer classification," *IEEE Access*, vol. 12, no. 12, pp. 181009–181024, 2024.

Publisher's Note: The publisher remains impartial concerning jurisdictional claims in published maps and institutional affiliations. Responsibility for the content rests entirely with the authors and does not necessarily reflect the publisher's perspectives.