Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

MelanoFusion: An Explainable and Uncertainty-Aware Hybrid CNN–Transformer Framework for Dermoscopic Melanoma Detection

V. Purushothama Raju
Shri Vishnu Engineering College for Women, Bhimavaram, A.P., India
Subhash R
Suvidha Foundation
T. Bharath Kumar
ICFAI Foundation for Higher Education
Smita Nirkhi
smita.singh@sitnagpur.siu.edu.in

Symbiosis International University
Manish Motghare
Suvidha Foundation

Article

Keywords: Melanoma Detection, Deep Learning, Skin Lesion Classification, Swin Transformer, Explainable Artificial Intelligence, Medical Image Analysis

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11208377/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Melanoma is an aggressive skin cancer requiring early detection, yet automated analysis remains affected by lesion variation, calibration errors, uncertainty, and subgroup disparities. This study presents MelanoFusion, a hybrid framework integrating explainability, uncertainty estimation, and fairness-oriented representation learning for melanoma detection. It combines EfficientNetV2-S and Swin Transformer encoders with cross-attention, lesion attention, feature-pyramid processing, multi-scale transformer layers, and an uncertainty head. Images were resized to 224 × 224 and processed using hair removal, median filtering, contrast enhancement, colour normalization, and watershed-based lesion priors; U-Net was used for lesion segmentation. MelanoFusion was evaluated against DenseNet121, ResNet101, EfficientNetV2-S, ConvNeXt-Tiny, and Swin Transformer on HAM10000, ISIC 2018, and PH². Segmentation Dice was 0.878 on HAM10000 and 0.830 on PH². MelanoFusion achieved ROC-AUC values of 0.687, 0.740, and 0.844 on HAM10000, ISIC 2018, and PH², respectively. Calibration was favourable on HAM10000, with an expected calibration error of 0.116. Uncertainty was significantly associated with classification error, while subgroup analysis showed significant sensitivity differences across age groups but not sex. Ablation results indicated that several proposed components did not improve discrimination. The findings support joint evaluation of discrimination, calibration, uncertainty, explainability, and subgroup behaviour, while indicating the need for external validation and clinical evaluation before deployment.

1 Introduction

Melanoma is a malignant tumour of melanocytes and an important target for early skin-cancer assessment. Dermoscopy provides visual information beyond conventional inspection, yet difficult lesions can share colour, texture, border and internal structural characteristics with benign lesions. Hairs, illumination, imaging devices and acquisition conditions can introduce additional variability. These factors make automated image-based assessment attractive but technically challenging [1], [2], [3], [4]. Recent reviews have highlighted the growing evidence base while also noting substantial heterogeneity in study design and evaluation [1], [5]. Recent work similarly emphasizes methodological, reliability and clinical-validation challenges in AI-based skin-cancer detection [1], [25], [26]. Deep learning has become a major computational paradigm for dermoscopic image analysis. CNNs learn hierarchical representations of local texture and shape, while transformer architectures provide mechanisms for modelling broader spatial relationships. DenseNet, ResNet and EfficientNetV2 represent established convolutional baselines; ConvNeXt modernizes convolutional design; and Swin Transformer provides hierarchical shifted-window attention [5], [8], [9], [10], [11]. Hybrid CNN–transformer approaches have subsequently attempted to combine local and contextual representations for skin-lesion analysis [6], [7], [8], [9], [12]. The evaluation problem is as important as the architecture problem. Results reported in different studies can be influenced by dataset selection, class balance, preprocessing, partitioning and decision thresholds. A model can have high accuracy while missing melanoma cases, or high specificity while sacrificing sensitivity. ROC-AUC measures ranking over thresholds but does not establish that a particular operating point is appropriate. Calibration, uncertainty, subgroup error and external validation

therefore provide complementary evidence [15], [16], [17], [18], [19], [20], [25], [26]. This study addresses that evaluation problem through a controlled comparison of six architectures across HAM10000, ISIC 2018 and PH². MelanoFusion combines EfficientNetV2-S and Swin-T representations through cross-attention, lesion-focused attention, multi-scale processing and additional uncertainty- and fairness-oriented components. The architecture is treated as an experimental construct: its value is determined by comparative evidence rather than by complexity or novelty alone.

This study has six main aims. First, it evaluates the preprocessing and lesion-segmentation pipeline using median filtering, CLAHE, colour normalization, watershed localization and U-Net segmentation. Second, it benchmarks DenseNet121, ResNet101, EfficientNetV2-S, ConvNeXt-Tiny and Swin Transformer against MelanoFusion under the same evaluation framework. Third, it examines model decisions using Grad-CAM, LIME, SHAP, attention and occlusion analyses. Fourth, it evaluates the fairness-oriented representation component using the available subgroup information without making unsupported claims about skin-tone fairness. Fifth, it assesses uncertainty and probability quality using uncertainty/error analysis, Brier score and ECE. Finally, it evaluates external computational generalization and uses retrained ablation and statistical analyses to determine whether the additional MelanoFusion components provide measurable benefit.

Although the literature contains strong CNNs, Transformers, segmentation methods, explainability techniques, uncertainty-aware approaches and fairness-oriented studies, these capabilities are often evaluated separately. The specific gap addressed by this study is therefore an evaluation gap: MelanoFusion is assessed against strong standalone reference models within one controlled framework that examines segmentation, discrimination, operating-point behaviour, calibration, uncertainty, subgroup error, external datasets and component-level ablation. The contribution is methodological and empirical rather than a claim of universal model superiority.

2 Related Work

Recent literature reviews and contemporary benchmark studies document rapid growth of artificial intelligence for skin-cancer detection while emphasizing heterogeneity across datasets, image sources and evaluation protocols [1], [5], [13]. This motivates controlled benchmarking rather than treating isolated published scores as directly comparable.

2.1 Datasets and Dataset Shift

HAM10000 is a major public multi-source dermoscopic resource containing 10,015 images. PH² provides 200 dermoscopic cases with diagnostic and segmentation information in recent benchmark studies, while ISIC 2018 challenge data provide a separate benchmark environment. Studies of diverse skin-image datasets have shown that population representation can affect model behaviour. The present work therefore treats ISIC 2018 and PH² as independent computational tests.

2.2 CNNs, Transformers and Hybrid Models

The benchmark set was selected to span distinct and historically important representation strategies rather than to favour a single model family. DenseNet121 is represented by recent skin-cancer benchmarking work that evaluates DenseNet121 directly against other CNN families [11], [30]. ResNet101 is represented through recent melanoma/skin-lesion studies using deep residual CNNs, including ResNet-family comparisons [3], [14]. EfficientNetV2-S represents a recent convolutional family whose EfficientNetV2 variants have been evaluated in skin-cancer classification [5], [10]. ConvNeXt-Tiny provides a modernized pure-convolutional comparator evaluated in recent skin-disease classification and fusion studies [9]. Swin Transformer provides a hierarchical Transformer with shifted-window self-attention and multi-scale representations, with recent dermoscopic applications supporting its use as a contemporary benchmark [2], [8], [27]. This selection therefore covers dense connectivity, residual learning, efficient scaling, modernized convolution and hierarchical attention.

Skin-lesion literature has increasingly examined Transformer and hybrid CNN–Transformer designs. Swin-based approaches have been evaluated for dermoscopic lesion classification, including settings motivated by the need to capture long-range and multi-scale visual relationships [13], [14], [16], [17]. Other work has explicitly combined convolutional feature extraction with Transformer encoding for skin-lesion classification [12], [8]. These studies provide literature support for testing complementary convolutional and Transformer representations, but they do not establish that combining additional modules will improve a particular dataset or training regime. The present study therefore evaluates the proposed combination empirically rather than assuming an advantage from architectural complexity.

2.3 Model Design Rationale from the Literature

The MelanoFusion design was assembled as a modular response to limitations identified across the literature rather than copied from a single prior architecture. The first design choice was to retain complementary convolutional and hierarchical Transformer encoders: EfficientNetV2-S supplies a modern convolutional representation, while Swin-T supplies hierarchical attention features [9], [10]. The second choice was feature fusion through cross-attention so that information from the two representations can interact rather than being concatenated without learned correspondence; related hybrid CNN–Transformer studies in dermoscopy motivate this general fusion strategy [12], [8]. The subsequent lesion-attention and feature-pyramid stages were included to preserve spatially informative and multi-resolution lesion structure, consistent with the broader literature on lesion segmentation and multi-scale representation [21], [22], [23], [24]. A Transformer encoder was then used to model relationships among the fused spatial tokens, extending the contextual modelling already provided by the Swin branch. Finally, uncertainty, calibration, explainability and subgroup analysis were treated as separate evaluation dimensions because discrimination alone does not establish probability quality, reliability or subgroup robustness [15], [16], [17], [18], [19], [20], [25]. The resulting architecture is therefore a literature-grounded composition of established mechanisms; its scientific contribution in this study is the controlled evaluation of whether that composition provides measurable benefit under the executed training protocol.

2.4 Segmentation and Trustworthy AI

U-Net and related encoder–decoder architectures remain common for lesion segmentation because skip connections preserve spatial information [21], [22], [23], [24]. Segmentation quality should not be treated as an automatic proxy for classification quality and is therefore reported as a separate endpoint. Grad-CAM, LIME and SHAP provide complementary post-hoc explanations, while attention and occlusion provide additional evidence about image regions [19], [20], [25], [29], [30]. Recent explainability literature also emphasizes the need to distinguish visual plausibility from quantitative faithfulness and clinical usefulness [19], [20], [29]. Uncertainty and calibration address prediction difficulty and probability quality [15], [16], [17]. Fairness studies show that subgroup representation and patient demographics can influence dermatological model performance [17], [18]. Recent segmentation research has also emphasized multi-scale and efficient feature representations for dermoscopic lesions [23], [24]. The current study reports only evidence supported by the available metadata and executed analyses.

2.5 External and Clinical Validation

External computational testing and prospective clinical validation represent different evidence levels. External datasets test distributional robustness, whereas prospective studies address workflow, clinician interaction and patient-facing consequences [1], [2], [26]. Accordingly, ISIC 2018 and PH² are interpreted as independent computational evaluations and not as substitutes for dermatologist reader studies or prospective clinical validation.

3 Methods

3.1 Study Design

The executed workflow comprises six functional modules: image preprocessing and watershed localization; U-Net lesion segmentation; benchmark classification with five reference models; MelanoFusion hybrid classification; explainability, uncertainty and calibration analysis; and fairness-related and external comparative analysis. The experimental design was organized according to these scientific functions rather than according to notebook phase labels.

3.2 Datasets and Experimental Environment

HAM10000 was partitioned using lesion-level grouping. The executed audit identified 7,470 unique lesions and found no intentional lesion overlap across the final development partitions. The classification split contained 7,000 training images, 1,505 validation images and 1,510 test images. External evaluation used 193 ISIC 2018 validation images, 1,512 ISIC 2018 test images and all 200 PH² cases available to the workflow. The notebook also includes explicit image-ID and lesion-level leakage checks for the HAM10000 development partitions and a frozen-test contamination guard. Dataset sources are cited in the reference list for HAM10000 [31], PH² [32] and ISIC 2018 [33]; restricted datasets should not be redistributed with the code.

Computational environment and training configuration: Experiments were executed in a Kaggle Linux environment using Python 3.12.13, PyTorch 2.10.0 + cu128 and CUDA 12.8, with two NVIDIA Tesla T4

GPUs available and CUDA execution enabled. A fixed random seed of 42 was used for Python, NumPy and PyTorch; deterministic execution was enabled and benchmarking was disabled. Images were processed at 224 × 224 pixels. For the controlled MelanoFusion training run, the physical batch size was 4 with gradient accumulation of 4 (effective batch size 16). AdamW optimization was used with a learning rate of 5×10^{-6} and weight decay of 1×10^{-4}. Training was limited to 10 epochs with early-stopping patience of 3 epochs. Two data-loader workers were used, automatic mixed precision was disabled, and the pretrained EfficientNetV2-S and Swin-T encoders were frozen.

3.3 Preprocessing and Lesion Localization

Images were resized to 224 × 224 pixels and passed through hair removal, 3 × 3 median filtering, contrast-limited adaptive histogram equalization (CLAHE), reference-based colour normalization and centre-aware watershed localization; recent skin-lesion studies likewise emphasize preprocessing, augmentation and artifact handling as important sources of variation in model performance [5], [10], [28]. The watershed output was an auxiliary lesion-region prior rather than a ground-truth mask. RGB data and the prior were concatenated to form the four-channel U-Net input.

3.4 U-Net Lesion Segmentation

The U-Net encoder–decoder receives four channels and produces a single-channel lesion probability map. Skip connections transfer spatial information from encoder to decoder stages. Predicted probabilities are thresholded to obtain binary masks. The executed segmentation evaluation reported Dice, IoU, boundary accuracy and Hausdorff distance.

3.5 Benchmark Models

The five benchmark models were DenseNet121, ResNet101, EfficientNetV2-S, ConvNeXt-Tiny and Swin Transformer. They were selected to represent complementary architectural families and were evaluated alongside MelanoFusion under the same dataset definitions and outcome metrics. DenseNet121 provides a dense-connectivity CNN reference with strong feature reuse and feature propagation [11], [30]; ResNet101 provides a deep residual-learning reference whose shortcut connections facilitate optimization of deeper networks [3], [14]; EfficientNetV2-S provides a modern convolutional reference designed around training speed and parameter efficiency [5], [10]; ConvNeXt-Tiny provides a contemporary pure-CNN comparator designed in the Transformer era [9]; and Swin Transformer provides a hierarchical shifted-window Transformer reference with multi-scale contextual representations [2], [8], [27]. The set therefore spans dense connectivity, residual learning, efficient convolutional scaling, modernized convolution and hierarchical attention. Importantly, EfficientNetV2-S and Swin-T are also the two encoders retained in MelanoFusion, allowing the study to test whether learned cross-family fusion and additional reliability-oriented modules provide measurable value beyond either encoder alone. The comparison is not intended to imply parameter-count equivalence; rather, it asks whether the added complexity of MelanoFusion is justified by the observed multi-dimensional evidence.

3.6 Proposed Model: MelanoFusion

MelanoFusion was proposed as a modular hybrid CNN–Transformer architecture. The model was developed by retaining EfficientNetV2-S as the convolutional branch and Swin-T as the hierarchical Transformer branch, then connecting these complementary representations through cross-attention. The fused representation was further refined with lesion attention and a three-scale feature pyramid, followed by a two-layer Transformer encoder. A fairness-oriented projection was included in the representation stage, and separate classification and uncertainty heads were added at the output. These components were selected from mechanisms supported by the related literature and were combined to test whether their joint use provides measurable benefits under a common training and evaluation protocol. Each major added component was subsequently evaluated through retrained ablation, so the architecture was treated as an experimental hypothesis rather than as an assumption that greater complexity would improve performance.

Figure 2 summarizes the implemented computational pathway, from the 224 × 224 RGB input through feature extraction, feature fusion, lesion-aware enhancement, multi-scale processing, representation learning, and the two output heads. MelanoFusion uses parallel pretrained EfficientNetV2-S and Swin-T encoders. For the 224 × 224 input, their final feature maps contain 1280 and 768 channels, respectively, and are projected by 1 × 1 convolutions to 512 channels. The projected maps are converted to token sequences and fused by an eight-head cross-attention module, with EfficientNetV2-S providing the queries and Swin-T providing the keys and values. The fused features are then refined by a learned spatial lesion-attention block and processed through a three-branch feature pyramid. The original 7 × 7 scale is processed directly, while 3 × 3 and 1 × 1 branches are generated by average pooling, processed, and upsampled to 7 × 7 before concatenation and 1 × 1 fusion to 512 channels. A two-layer, eight-head Transformer encoder operates on 49 tokens of dimension 512. Mean pooling produces a 512-dimensional representation, which is passed through the fairness-oriented projection layer and then to a 512-to-1 melanoma classifier and a separate uncertainty head consisting of 512-to-256 and 256-to-1 linear layers with ReLU and dropout. The executed MelanoFusion model contains 65,287,373 parameters. The comparison with the five standalone backbones is not intended to imply parameter-count equivalence; instead, it tests whether the additional representation and reliability mechanisms provide measurable value under the same evaluation framework. The executed results show a mixed outcome: MelanoFusion did not achieve the highest mean ROC-AUC, although it showed favourable HAM10000 accuracy, specificity and calibration. The retrained ablation study further showed that every tested component-removal condition improved validation AUC, indicating that the complete configuration was not optimal under the evaluated training regime.

3.6.1 Dual Encoder Feature Extraction

The preprocessed RGB image is supplied to two parallel pretrained encoders: EfficientNetV2-S and Swin-T. For the evaluated 224 × 224 input, their final feature maps contain 1280 and 768 channels, respectively. Each feature map is projected by a 1 × 1 convolution to 512 channels, giving compatible 7 × 7 × 512 representations. The dual pathway is intended to retain complementary convolutional and hierarchical transformer features rather than relying on a single feature-extraction family.

3.6.2 Cross-Attention and Lesion-Aware Feature Enhancement

The two projected representations are converted into token sequences and integrated with an eight-head cross-attention module. In the implemented configuration, the EfficientNetV2-S representation supplies queries (Q), while the Swin-T representation supplies keys (K) and values (V). For attention head h, the attention operation is expressed as:

$A_h = \text{Softmax}((Q_h K_h^T) / \sqrt{d_h})$ (1)

$Z_h = A_h V_h$ (2)

Here, Q_h, K_h, and V_h denote the query, key, and value matrices for attention head h, respectively; d_h denotes the dimensionality of attention head h; A_h denotes the resulting attention-weight matrix; and Z_h denotes the attended output representation for head h. The head outputs are combined to form the fused representation. A learned spatial lesion-attention block then assigns spatial weighting to the fused feature map. This mechanism is used to enhance spatially informative lesion features without treating the watershed prior as a direct input to the lesion-attention block.

$\text{Fused} = \text{LayerNorm}(\text{Attention}(Q, K, V) + F_E)$ (3)

$\text{Attention}(Q, K, V) = \text{Softmax}(QK^T / \sqrt{d})V$ (4)

$Q = F_E W_Q, K = F_S W_K, V = F_S W_V$ (5)

$F_E = \text{ReLU}(\text{BN}(\text{Conv}_{1\times1}(E))), F_S = \text{ReLU}(\text{BN}(\text{Conv}_{1\times1}(S)))$ (6)

Here, F_E and F_S denote the projected EfficientNetV2-S and Swin-T feature maps, respectively; W_Q, W_K, and W_V are the learned query, key, and value projection matrices; Q, K, and V are the resulting query, key, and value token representations; and d is the per-head attention dimension. The fused representation is obtained by adding the attention output to F_E followed by LayerNorm.

3.6.3 Three-Scale Feature Pyramid

The lesion-attended 7 × 7 × 512 representation is processed through three spatial branches. The original 7 × 7 branch is retained, while average pooling produces 3 × 3 and 1 × 1 branches. These branches are processed and then upsampled to the common 7 × 7 spatial resolution. The resulting representations are concatenated and fused by a 1 × 1 convolution to recover 512 output channels. This design provides multi-resolution context while preserving a common representation for subsequent token-based processing.

$P = \text{ReLU}(\text{BN}(\text{Conv}_{1\times1}([P_1, P_2, P_3])))$ (7)

$P_1 = Conv_{3\times3}(F_L)$; $P_2 = Up(Conv_{3\times3}(AvgPool_2(F_L)))$; $P_3 = Up(Conv_{3\times3}(AvgPool_4(F_L)))$ (8)

Here, F_L denotes the lesion-attended feature map; P_1 is the directly processed 7 × 7 branch; P_2 is the branch obtained by 2 × 2 average pooling, 3 × 3 convolution, and upsampling to 7 × 7; P_3 is the branch obtained by 4 × 4 average pooling, 3 × 3 convolution, and upsampling to 7 × 7; and P is the fused 7 × 7 × 512 feature representation.

3.6.4 Multi-Scale Transformer and Fairness-Oriented Representation

The fused 7 × 7 × 512 features are represented as 49 tokens of dimension 512 and passed through two Transformer encoder layers, each using eight attention heads. Mean pooling over the token sequence produces a 512-dimensional global representation. A fairness-oriented projection is then applied during the representation-learning stage. A fairness regularization term was implemented in the composite loss, but it was inactive during the main MelanoFusion training because validated protected-group labels were unavailable in the executed metadata audit. The architecture is therefore evaluated as a fairness-oriented design component rather than as evidence of fairness optimization.

$r = (1/49) \Sigma_{i=1}^{49} T_i$, $r_f = ReLU(LayerNorm(Linear(r)))$ (9)

$T = TransformerEncoder(F_P)$, $T \in \mathbb{R}^{49\times512}$ (10)

Here, F_P denotes the fused feature-pyramid output, T_i denotes the 512-dimensional token at spatial position i after Transformer encoding, T denotes the complete 49 × 512 encoded token sequence, r is the mean-pooled 512-dimensional representation, and r_f is the fairness-oriented representation produced by the linear, LayerNorm, and ReLU projection.

3.6.5 Classification and Uncertainty Heads

The 512-dimensional representation feeds the binary melanoma classifier and the separate uncertainty head. The classifier uses a 512-to-1 linear mapping to produce the melanoma logit, which is converted to a probability for the binary decision. The uncertainty head uses a 512-to-256 linear layer followed by ReLU and dropout, followed by a 256-to-1 linear output. The two heads therefore provide complementary prediction and predictive-uncertainty information. The executed MelanoFusion contains 65,287,373 parameters.

$u = W_2\ Dropout(ReLU(W_1 r_f + b_1)) + b_2$ (11)

$z = W_c r_f + b_c$, $p = \sigma(z) = 1/(1 + e^{-z})$ (12)

Here, r_f is the 512-dimensional representation supplied to both heads; u is the scalar output of the uncertainty head; W_1, W_2, b_1, and b_2 are the learned parameters of its two linear layers; z is the

melanoma classification logit; W_c and b_c are the classifier parameters; and p is the predicted melanoma probability obtained with the sigmoid function σ(·).

$L_{att} = \text{mean}|\Delta_x A| + \text{mean}|\Delta_y A|$ (13)

$L_{cal} = \text{mean}_i(p_i - y_i)^2$ (14)

$L_{fair} = \text{Var}_g(\text{mean}_{i \in g}\, p_i)$ (15)

$L_{Dice} = 1 - (2\Sigma_i p_i y_i + 1)/(\Sigma_i p_i + \Sigma_i y_i + 1)$ (16)

$L_{cls} = \text{BCEWithLogitsLoss}(z, y)$ (17)

$L_{total} = L_{cls} + 0.5L_{Dice} + 0.10L_{fair} + 0.10L_{cal} + 0.10L_{att}$ (18)

For the loss terms below, A denotes the spatial lesion-attention map (distinct from the cross-attention weights A_h); $\Delta_x A$ and $\Delta_y A$ denote its spatial differences along the horizontal and vertical directions. In the calibration and Dice terms, p_i denotes the predicted melanoma probability and y_i the corresponding binary target for sample i. In the fairness term, g indexes the available subgroup and the variance is computed across subgroup mean predicted probabilities. L_{cls}, L_{Dice}, L_{fair}, L_{cal}, and L_{att} denote the classification, Dice, fairness, calibration, and attention-consistency losses, respectively. During the main MelanoFusion training run, L_{fair} evaluated to zero because validated subgroup labels were unavailable.

3.6.6 Melanoma-Specific Design Rationale and Verification

The melanoma-specific design preserves information that may be relevant to heterogeneous dermoscopic appearance while explicitly evaluating reliability and subgroup behaviour. Preprocessing reduces common image artefacts and localizes the lesion region. The CNN and Swin branches provide complementary feature representations; cross-attention integrates these representations; lesion attention and the feature pyramid enhance spatial and multi-resolution information; and the Transformer models relationships among the resulting spatial tokens. Uncertainty and explainability are evaluated as separate reliability and inspection mechanisms. The retrained ablation study tested removal of cross-attention, lesion attention, the feature pyramid, the multi-scale Transformer, the uncertainty head and the fairness-oriented representation layer. Every component-removal condition improved validation AUC relative to the full model, with the largest gain after removing the multi-scale Transformer. These results do not support the claim that every proposed component is beneficial; instead, they identify a concrete direction for simplifying the architecture. The explainability workflow generated Grad-CAM, LIME and SHAP outputs and recorded attention and occlusion artifacts. A limited quantitative audit recovered 28 top-20 energy-concentration values across four cases. Uncertainty analysis included an MC-dropout diagnostic, ensemble probability dispersion, temperature-scaled entropy and uncertainty quartiles. Calibration was evaluated with Brier score and ten-bin ECE. The original fairness objective specified Fitzpatrick I–VI skin tones; however, validated skin-tone labels were not established in the executed

metadata audit, so subgroup analysis was limited to the available age and sex variables. For the retrained ablation study, the executed notebook used a physical batch size of 2 with gradient accumulation of 8 (effective batch size 16), a learning rate of 1×10^{-5}, weight decay of 1×10^{-4}, and 10 training epochs.

3.7 Use of Artificial Intelligence

Generative artificial intelligence tools were used to assist with manuscript language editing, structural organization, formatting, and policy-compliance review. They were not used to generate or fabricate experimental data, and the reported numerical results were checked against the executed research notebook and source materials. The authors reviewed and verified the manuscript and remain fully responsible for its accuracy, integrity, interpretation and conclusions.

3.8 Metrics and Statistical Analysis

The five primary classification metrics were accuracy, precision, sensitivity, F1-score and ROC-AUC. Specificity and MCC were additionally reported where available, with Brier score and ECE used for probability quality. Segmentation was assessed with Dice, IoU, boundary accuracy and Hausdorff distance. Existing bootstrap artifacts used 2,000 iterations and a 95% confidence level for ROC-AUC/PR-AUC analyses across the 18 model–dataset cells. MelanoFusion was compared with each benchmark on frozen HAM10000 predictions using paired McNemar, Wilcoxon signed-rank and paired t-tests. Statistical significance was assessed at a two-sided $\alpha = 0.05$ threshold; exact p-values are reported where available. The subgroup analyses used Pearson's chi-square test for categorical associations, with Fisher's exact test additionally reported for the sex comparison.

4 Results

4.1 Lesion Segmentation

U-Net achieved Dice 0.878, IoU 0.807, boundary accuracy 0.585 and Hausdorff distance 20.02 px on HAM10000. On PH^2, Dice was 0.830, IoU 0.741, boundary accuracy 0.283 and Hausdorff distance 30.15 px. These results support the segmentation module as a measurable computational endpoint, while not implying that segmentation performance automatically improves classification.

4.2 Six-Model, Three-Dataset Classification Benchmark

Table 1
Classification performance across the six evaluated architectures and three datasets.

Model / dataset	Accuracy	Precision	Sensitivity	F1	ROC-AUC	Specificity
DenseNet121 / HAM	0.707	0.231	0.702	0.348	0.758	0.707
ResNet101 / HAM	0.718	0.223	0.619	0.328	0.734	0.730
EfficientNetV2-S / HAM	0.777	0.266	0.571	0.363	0.741	0.803
ConvNeXt-Tiny / HAM	0.383	0.138	0.869	0.239	0.737	0.322
Swin Transformer / HAM	0.301	0.131	0.935	0.229	0.770	0.222
MelanoFusion / HAM	0.819	0.250	0.315	0.279	0.687	0.882
DenseNet121 / ISIC	0.886	0.490	0.140	0.218	0.758	0.981
ResNet101 / ISIC	0.880	0.000	0.000	0.000	0.286	0.992
EfficientNetV2-S / ISIC	0.901	0.714	0.205	0.318	0.761	0.990
ConvNeXt-Tiny / ISIC	0.749	0.286	0.813	0.423	0.849	0.741
Swin Transformer / ISIC	0.701	0.260	0.889	0.402	0.872	0.677
MelanoFusion / ISIC	0.890	0.727	0.047	0.088	0.740	0.998
DenseNet121 / PH²	0.815	0.636	0.175	0.275	0.734	0.975
ResNet101 / PH²	0.800	0.000	0.000	0.000	0.405	1.000
EfficientNetV2-S / PH²	0.840	1.000	0.200	0.333	0.809	1.000
ConvNeXt-Tiny / PH²	0.840	0.567	0.850	0.680	0.930	0.838
Swin Transformer / PH²	0.860	0.636	0.700	0.667	0.900	0.900
MelanoFusion / PH²	0.800	0.000	0.000	0.000	0.844	1.000

4.3 Dataset-Specific Behaviour

On HAM10000, MelanoFusion achieved the highest accuracy (0.819) and specificity (0.882), while Swin Transformer achieved the highest sensitivity (0.935) and ROC-AUC (0.770). On ISIC 2018, Swin Transformer achieved the strongest sensitivity (0.889) and ROC-AUC (0.872); MelanoFusion had high specificity (0.998) and precision (0.727) but sensitivity of only 0.047 at the evaluated threshold. On PH², ConvNeXt-Tiny achieved the strongest sensitivity (0.850), F1-score (0.680) and ROC-AUC (0.930); Swin Transformer achieved 0.900 ROC-AUC. MelanoFusion achieved 0.844 ROC-AUC but zero sensitivity and F1 at the evaluated threshold.

4.4 Statistical Comparison and Bootstrap Analysis

On frozen HAM10000 predictions, all five benchmark comparisons with MelanoFusion were statistically significant in the executed paired analyses at the unadjusted two-sided α = 0.05 threshold. Bootstrap ROC-AUC differences were − 0.054 versus EfficientNetV2-S (95% CI [-0.093, -0.014], p = 0.007), -0.050 versus ConvNeXt-Tiny ([-0.088, -0.011], p = 0.014), -0.082 versus Swin Transformer ([-0.124, -0.042], p < 0.001), -0.071 versus DenseNet121 ([-0.111, -0.032], p = 0.001) and − 0.047 versus ResNet101 ([-0.087, -0.004], p = 0.034).

4.5 Calibration and Uncertainty

Table 2
Calibration performance measured by expected calibration error and Brier score.

Model	HAM ECE	HAM Brier	ISIC ECE	ISIC Brier	PH² ECE	PH² Brier
DenseNet121	0.259	0.218	0.045	0.093	0.114	0.150
ResNet101	0.248	0.209	0.148	0.127	0.199	0.200
EfficientNetV2-S	0.185	0.170	0.071	0.089	0.150	0.145
ConvNeXt-Tiny	0.554	0.475	0.201	0.164	0.113	0.112
Swin Transformer	0.606	0.502	0.273	0.204	0.072	0.097
MelanoFusion	0.116	0.136	0.092	0.104	0.182	0.180

MelanoFusion had the lowest HAM10000 ECE (0.116) and Brier score (0.136) among the evaluated models, despite lower HAM10000 AUC. In the uncertainty analysis, the lowest-uncertainty quartile had a 5.82% error rate (22/378), whereas the highest-uncertainty quartile had a 39.68% error rate (150/378). The association between uncertainty quartile and classification error was evaluated using Pearson's chi-square test ($\chi^2 = 181.817763$, df = 3, $p = 3.57 \times 10^{-39}$). The MC-dropout diagnostic produced zero probability variation for the evaluated checkpoint and was not interpreted as informative stochastic uncertainty.

4.6 Fairness-Related Subgroup Analysis

Available subgroup analysis showed age-group sensitivities of 0.095 for age < 40 (n = 353), 0.275 for age 40–59 (n = 469) and 0.385 for age 60+ (n = 682), corresponding to false-negative rates of 0.905, 0.725 and 0.615. Female sensitivity was 0.273 (n = 665) and male sensitivity 0.336 (n = 839). The age-group association was evaluated using Pearson's chi-square test and was statistically significant ($\chi^2 = 7.2877286107$, df = 2, p = 0.0261511), whereas the sex association was not (Pearson's $\chi^2 = 0.4289584$, df = 1, p = 0.5125005; Fisher's exact p = 0.4804980). These results do not establish Fitzpatrick I–VI fairness because validated skin-tone labels were unavailable.

4.7 Explainability and Ablation

The executed workflow generated Grad-CAM, LIME, SHAP, attention and occlusion artifacts, with 28 top-20 energy-concentration values recovered across four cases. These outputs support computational inspection but do not establish clinician-validated explanation faithfulness.

Table 3
Retrained MelanoFusion ablation study showing validation AUC and F1 changes after component removal.

MelanoFusion configuration	Validation AUC	Validation F1	ΔAUC vs full
Full model	0.681109	0.266332	0.000000
Without cross attention	0.716513	0.351648	+ 0.035405
Without lesion attention	0.713698	0.330508	+ 0.032590
Without feature pyramid	0.721588	0.340037	+ 0.040480
Without multi-scale transformer	0.763603	0.364621	+ 0.082494
Without uncertainty head	0.699659	0.276119	+ 0.018550
Without fairness-oriented representation layer	0.725741	0.341198	+ 0.044633

Every tested component-removal condition improved validation AUC, with the largest improvement after removing the multi-scale transformer (0.681109 to 0.763603). This finding indicates that the executed full configuration may contain redundancy, capacity mismatch or optimization burden; it does not establish that the removed components are universally detrimental.

4.8 Operating-Point Analysis

At a threshold of 0.15 selected to satisfy specificity $\geq$ 0.70, MelanoFusion sensitivity was 0.539, specificity 0.738, PPV 0.204, NPV 0.928 and F1 0.296. At the threshold maximizing validation F1 (0.25), sensitivity was 0.455, specificity 0.812 and F1 0.307. The evaluated Swin Transformer point at threshold 0.90 yielded sensitivity 0.581, specificity 0.789 and F1 0.355. These results demonstrate that threshold choice materially changes the interpretation of model behaviour.

4.9 Computational Complexity and Inference Time

The executed test-set timing logs were used to compare single-image inference cost across the six evaluated architectures on the same evaluation hardware. The mean measured inference time per image was 3.33 ms for DenseNet121, 3.62 ms for EfficientNetV2-S, 4.40 ms for Swin Transformer, 6.13 ms for ResNet101, and 8.38 ms for ConvNeXt-Tiny. MelanoFusion recorded the highest inference cost (12.28 ms/image), consistent with its larger parameter count (65,287,373 parameters) and its additional cross-attention, lesion-attention, feature-pyramid and Transformer-encoder stages relative to either of its two constituent backbones evaluated alone.

Inference time is reported here as a computational-efficiency measurement on the executed evaluation hardware and does not by itself indicate deployment readiness; a model with favourable discrimination,

calibration and inference cost may still require external validation and clinical evaluation, as noted throughout this study.

5 Discussion

The central finding is that MelanoFusion did not achieve the highest overall discrimination. Swin Transformer achieved the highest mean ROC-AUC across the three datasets (0.847), followed by ConvNeXt-Tiny (0.839), while MelanoFusion ranked fourth (0.757). This result is important because it shows that adding architectural components does not automatically improve melanoma classification under the evaluated training configuration.

The results also show that model quality is multi-dimensional. MelanoFusion achieved the highest HAM10000 accuracy and specificity and favourable HAM10000 calibration, whereas Swin Transformer achieved substantially higher sensitivity and ROC-AUC. In an imbalanced medical classification problem, this distinction is important because high accuracy can coexist with clinically important false negatives.

The operating-point analysis further shows that binary decisions cannot be interpreted from ROC-AUC alone. Previous studies have reported competitive or superior AI performance under specific dermoscopic evaluation settings [3], [4], [14]. However, dermatologist confidence and AI confidence may provide complementary information for diagnostic decision-making [26]. External testing further demonstrates the importance of distribution shift. MelanoFusion obtained AUCs of 0.740 on ISIC 2018 and 0.844 on PH², but its thresholded sensitivity was 0.047 and 0.000, respectively, at the evaluated operating points. These findings support the use of independent datasets when assessing robustness [1], [2], [26].

The ablation study provides the clearest direction for the next model iteration. Every component-removal condition improved validation AUC, and removing the multi-scale Transformer produced the largest gain. The current architecture may therefore contain unnecessary capacity or may be difficult to optimize under the available data and training regime. Future work should test a simpler MelanoFusion variant and replicate the ablation across multiple random seeds and independent datasets.

The uncertainty analysis provides a useful reliability signal: the error rate increased from 5.82% in the lowest uncertainty quartile to 39.68% in the highest. However, the MC-dropout diagnostic produced zero probability variation for the evaluated checkpoint, indicating that this particular stochastic uncertainty implementation was not informative for that checkpoint. A more robust uncertainty protocol should therefore be evaluated before deployment-oriented interpretation.

The fairness findings should also be interpreted conservatively. The significant age-group difference in false-negative rate demonstrates the value of subgroup auditing, but the study does not establish fairness across Fitzpatrick I–VI skin tones because validated skin-tone labels were unavailable. The fairness-oriented representation component should therefore be regarded as an architectural design that requires dedicated subgroup data and metrics, rather than as evidence that fairness has been

established. Similarly, the explainability analysis provides computational interpretability evidence rather than proof of clinical trust. Grad-CAM, LIME, SHAP, attention and occlusion outputs can indicate whether predictions appear to rely on lesion regions, but visual plausibility alone does not establish explanation faithfulness. Future evaluation should quantify faithfulness and stability and include dermatologist assessment [26], [19], [20], [25].

Overall, the study supports a methodological conclusion: stronger discrimination, better calibration, useful uncertainty behaviour and architectural simplicity do not necessarily occur in the same model. The most defensible contribution is therefore a reproducible, multi-dimensional evaluation protocol together with an empirically supported direction for simplifying the next MelanoFusion iteration.

6 Limitations

- The study uses retrospective public datasets and does not establish prospective clinical effectiveness.
- PH² is comparatively small and its estimates should be interpreted cautiously.
- The primary 0.50 operating threshold is an evaluated threshold, not a clinically validated deployment threshold.
- Retrained ablation does not establish multi-seed reproducibility for every condition.
- Validated Fitzpatrick I–VI skin-tone labels were not established in the executed metadata audit.
- Explainability evidence is limited to generated artifacts and a small quantitative audit.
- MC-dropout probability variation was zero for the evaluated checkpoint.
- No dermatologist reader study, patient-outcome evaluation or prospective deployment study was performed.
- External computational evaluation should not be interpreted as regulatory approval or clinical validation.

7 Conclusion

This study presents a controlled computational evaluation of six melanoma classifiers across HAM10000, ISIC 2018 and PH². The workflow integrates preprocessing, lesion segmentation, benchmark classification, the proposed MelanoFusion hybrid model, explainability, calibration, uncertainty analysis, subgroup analysis and retrained ablation.

Swin Transformer achieved the highest three-dataset mean ROC-AUC, while ConvNeXt-Tiny performed particularly strongly on PH². MelanoFusion did not demonstrate overall discrimination superiority, but it showed favourable HAM10000 accuracy, specificity and calibration.

The uncertainty analysis showed a strong association between uncertainty and observed error, while subgroup analysis identified a statistically significant age-related difference in false-negative rate. Most

importantly for model development, every tested MelanoFusion component-removal condition improved validation AUC, with the largest improvement after removing the multi-scale Transformer. This indicates that the executed full architecture was not optimal under the tested training configuration and provides a concrete direction for simplification. The contribution is best characterized as an evidence-based, multi-dimensional evaluation rather than a claim of state-of-the-art performance.

Future work should test a simplified MelanoFusion variant with multi-seed replication, stronger explanation-faithfulness evaluation, validated skin-tone subgroup analysis and prospective human–AI assessment before clinical or regulatory claims are considered.

Declarations

Competing Interests

The authors declare no competing interests.

Ethics approval and consent to participate

This study analysed previously published, publicly available dermoscopic image datasets and did not involve new participant recruitment, direct participant interaction, or collection of new human data. No personally identifying participant information was collected by the authors for this study.

Consent for publication

Not applicable to this study because no new participant recruitment or collection of human data was performed.

Funding

No specific external funding is declared for this study.

Author Contribution

V. Purushothama Raju, Subhash R., T. Bharath Kumar, Smita Nirkhi, and Manish Motghare contributed to conceptualization, methodology, software, validation, formal analysis, investigation, visualization, and writing—original draft, writing—review and editing.

Acknowledgements

This work was conducted at Suvidha Foundation under the Internship Mentorship Program.

Data Availability

This study used three publicly available dermoscopic image datasets and publicly available lesion-segmentation annotations, each retained in its original third-party repository rather than redistributed by the authors. HAM10000 was used for model development, validation, held-out testing and lesion-segmentation training; the HAM10000 lesion-segmentation masks provided reference lesion masks; ISIC 2018 was used as an external classification benchmark; and PH² was used for external classification and lesion segmentation. All four datasets are publicly accessible and subject to their respective repositories' terms of use; no new data were generated during this study.Data Source:HAM10000: https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000HAM10000 lesion segmentations:https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DBW86TISIC 2018: https://challenge.isic-archive.com/data/PH²: https://www.kaggle.com/datasets/spacesurfer/ph2-dataset

Materials Availability

Not applicable.

Code Availability

The executed research notebook containing the training, evaluation, ablation, calibration, uncertainty, subgroup-analysis and figure/table-generation workflows is provided in the GitHub repository [https://github.com/Subhash-R11/MedFairXNet].

References

1. Ye, Z. et al. Deep learning algorithms for melanoma detection using dermoscopic images: a systematic review and meta-analysis. *Artif. Intell. Med.* **155**, 102934. https://doi.org/10.1016/j.artmed.2024.102934 (2024).
2. Ayas, S. Multiclass skin lesion classification in dermoscopic images using Swin Transformer model. *Neural Comput. Appl.* **35**, 6713–6722. https://doi.org/10.1007/s00521-022-08053-z (2023).
3. Salma, W. & Eltrass, A. S. Automated deep learning approach for classification of malignant melanoma and benign skin lesions. *Multimed Tools Appl.* **81**, 32643–32660. https://doi.org/10.1007/s11042-022-13081-x (2022).
4. Jaber, N. F. & Akbas, A. Melanoma skin cancer detection based on deep learning methods and binary Harris Hawk optimization. *Multimed Tools Appl.* **84**, 25709–25722. https://doi.org/10.1007/s11042-024-19864-8 (2025).

5. Vieira, J., Mendonça, F. & Morgado-Dias, F. Deep learning approaches for skin lesion detection. *Electronics* **14**, 2785. https://doi.org/10.3390/electronics14142785 (2025).
6. He, X. et al. Fully transformer network for skin lesion analysis. *Med. Image Anal.* **77**, 102357. https://doi.org/10.1016/j.media.2022.102357 (2022).
7. Nie, Y., Sommella, P., Carratù, M., O'Nils, M. & Lundgren, J. A deep CNN Transformer hybrid model for skin lesion classification of dermoscopic images using focal loss. *Diagnostics* **13**, 72. https://doi.org/10.3390/diagnostics13010072 (2023).
8. Tang, K. et al. SkinSwinViT: a lightweight transformer-based method for multiclass skin lesion classification with enhanced generalization capabilities. *Appl. Sci.* **14**, 4005. https://doi.org/10.3390/app14104005 (2024).
9. Wei, M., Wu, Q., Ji, H., Wang, J. & Lyu, T. A skin disease classification model based on DenseNet and ConvNeXt fusion. *Electronics* 12, 438. (2023). https://doi.org/10.3390/electronics12020438
10. Lee, H., Cho, S., Song, J., Kim, H. & Shin, Y. An enhanced approach using AGS network for skin cancer classification. *Sensors* **25**, 394. https://doi.org/10.3390/s25020394 (2025).
11. Jalili, A., Sajedi, H., Tabrizchi, H. & Mosavi, A. Skin cancer classification using DenseNet. IEEE 22nd Jubilee International Symposium on Intelligent Systems and Informatics (SISY), 333–340. (2024). https://doi.org/10.1109/SISY62279.2024.10737516 (2024).
12. Gallazzi, M. et al. A large dataset to enhance skin cancer classification with transformer-based deep neural networks. *IEEE Access.* **12**, 109544–109559. https://doi.org/10.1109/ACCESS.2024.3439365 (2024).
13. Chiu, T. M., Chi, I. C., Li, Y. C. & Tseng, M. H. Deep ensemble learning for multiclass skin lesion classification. *Bioengineering* **12**, 934. https://doi.org/10.3390/bioengineering12090934 (2025).
14. Selvaraj, K. M., Gnanagurusubbiah, S., Roy, R. R., Peter, J., Balu, S. & J. H. & Enhancing skin lesion classification with advanced deep learning ensemble models: a path towards accurate medical diagnostics. *Curr. Probl. Cancer*. **49**, 101077. https://doi.org/10.1016/j.currproblcancer.2024.101077 (2024).
15. Shamsi, A. et al. A novel uncertainty-aware deep learning technique with an application on skin cancer diagnosis. *Neural Comput. Appl.* **35**, 22179–22188. https://doi.org/10.1007/s00521-023-08930-1 (2023).
16. Fayyad, J., Alijani, S. & Najjaran, H. Empirical validation of Conformal Prediction for trustworthy skin lesions classification. *Comput. Methods Programs Biomed.* **253**, 108231. https://doi.org/10.1016/j.cmpb.2024.108231 (2024).
17. Bhattacharyya, S., Pal, U. & Chakraborti, T. Conformal uncertainty quantification to evaluate predictive fairness of foundation AI model for skin lesion classes across patient demographics. *Health Inf. Sci. Syst.* **14**, 18. https://doi.org/10.1007/s13755-025-00412-z (2026).
18. Tjiu, J. W. & Lu, C. F. Equity and generalizability of artificial intelligence for skin-lesion diagnosis using clinical, dermoscopic, and smartphone images: a systematic review and meta-analysis. *Medicina* **61**, 2186. https://doi.org/10.3390/medicina61122186 (2025).

19. Metta, C. et al. Improving trust and confidence in medical skin lesion diagnosis through explainable deep learning. *Int. J. Data Sci. Anal.* **20**, 183–195. https://doi.org/10.1007/s41060-023-00401-z (2025).
20. Whewell, J., Peters, R. & Kulon, J. Explainable artificial intelligence for skin lesion classification: a comprehensive review of methods and challenges. *Technologies* **14**, 391. https://doi.org/10.3390/technologies14070391 (2026).
21. Kartal, M. S. & Polat, Ö. Segmentation of skin lesions using U-Net with EfficientNetB7 backbone. 2022 Innovations in Intelligent Systems and Applications Conference (ASYU). (2022). https://doi.org/10.1109/ASYU56188.2022.9925369
22. Karimi, A., Faez, K. & Nazari, S. DEU-Net: dual-encoder U-Net for automated skin lesion segmentation. *IEEE Access.* **11**, 134804–134821. https://doi.org/10.1109/ACCESS.2023.3337528 (2023).
23. Li, X. et al. USL-Net: uncertainty self-learning network for unsupervised skin lesion segmentation. *Biomed. Signal. Process. Control*. **89**, 105769. https://doi.org/10.1016/j.bspc.2023.105769 (2024).
24. Khan, S., Khan, A. & Teng, Y. DFF-UNet: a lightweight deep feature fusion U-Net model for skin lesion segmentation. *IEEE Trans. Instrum. Meas.* **74**, 5030214. https://doi.org/10.1109/TIM.2025.3565715 (2025).
25. Eita, M. A. & Rizk, H. A trustworthy and explainable deep learning framework for skin lesion detection in smart dermatology. *Eng. Appl. Artif. Intell.* **159**, 111594. https://doi.org/10.1016/j.engappai.2025.111594 (2025).
26. Kommoss, K. S. et al. Observational study investigating the level of support from a convolutional neural network in face and scalp lesions deemed diagnostically unclear by dermatologists. *Eur. J. Cancer*. **185**, 53–60. https://doi.org/10.1016/j.ejca.2023.02.025 (2023).
27. Khanna, S., Tyagi, A., Sharma, S. & Bharti, A. K. Pixels to prognosis: unveiling skin lesion patterns through Swin Transformer. *Procedia Comput. Sci.* **245**, 193–201. https://doi.org/10.1016/j.procs.2024.10.243 (2024).
28. Thwin, S. M. & Park, H. S. Skin lesion classification using a deep ensemble model. *Appl. Sci.* **14**, 5599. https://doi.org/10.3390/app14135599 (2024).
29. Gouabou, A. C. F. et al. Computer aided diagnosis of melanoma using deep neural networks and game theory: application on dermoscopic images of skin lesions. *Int. J. Mol. Sci.* **23**, 13838. https://doi.org/10.3390/ijms232213838 (2022).
30. Alfi, I. A., Rahman, M. M., Shorfuzzaman, M. & Nazir, A. A non-invasive interpretable diagnosis of melanoma skin cancer using deep learning and ensemble stacking of machine learning models. *Diagnostics* **12**, 726. https://doi.org/10.3390/diagnostics12030726 (2022).