Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 25 May 2026
Revised 29 June 2026
Accepted 17 July 2026

Natural Resources for Human Health 2026, Vol. 6 (9s): 294–336

KEYWORDS:
Hybrid Deep Learning; Skin Cancer Classification; Latent Bottleneck Learning; Vision Transformer; Explainable Artificial Intelligence

HybridDermNet: A CNN–Transformer Framework with Latent Bottleneck Learning for Robust Skin Cancer Classification

Vunnam Narmada[1], Kalavala Asish Vardhan[2*]

[1,2*] Department of Computer Science Engineering, Malla Reddy University, Hyderabad 500100, India
Corresponding author: ashi.mintu@gmail.com

ABSTRACT: The classification of skin lesions from dermoscopic and clinical images remains challenging due to visually similar lesion categories, variability in image acquisition, class imbalance, and limited cross-dataset generalisation, which can undermine the reliability of diagnostic results. Current CNN-based approaches are good at local texture, pigmentation, and border information, while transformer models are good at global structure and long-range dependencies. Most hybrid methods, however, use simple concatenation or additive fusion, or rely on multi-backbone architectures that are very computation-intensive, produce redundant features, and risk overfitting. In this work, a CNN–Transformer-based framework with latent bottleneck learning for robust multi-class skin cancer classification is proposed. The algorithm locally and globally extracts features in parallel, calculates bidirectional attention between features, adaptively fuses features, encodes features in a compact latent space, optionally reconstructs features, and predicts with confidence. A composite objective combines classification, reconstruction, cross-feature consistency, latent regularisation, and transformation-consistency loss. HybridDermNet achieved 96.18% accuracy, 94.86% macro-F1, and 97.92% AUROC on HAM10000, and 95.42% accuracy, 93.92% macro-F1, and 97.46% AUROC on ISIC 2018. When transferring between datasets, it achieved 87.24% and 87.61% macro-F1, respectively, and, in external validation on PAD-UFES-20, achieved 80.67% macro-F1. The impact of attention-guided fusion and latent compression was verified through ablation, calibration, explainability, efficiency, and statistical analysis. The framework offers compact, interpretable and generalizable representations for reliable dermatology decision support in heterogeneous imaging conditions.

1. INTRODUCTION

Early diagnosis of skin cancer can be difficult given the similarity of benign and malignant skin lesions; it is a major public health concern. Although dermoscopic imaging can aid in non-invasive evaluation, differences in pigmentation, border structure, texture, illumination, lesion size, and acquisition conditions remain a challenge for automated dermoscopic image classification. Although conventional convolutional neural networks have proven to be very capable at learning local lesion features, vision transformers excel in modelling global morphology and long-range spatial dependencies. Thus, CNN–Transformer hybrids such as ARP-ViT-CNN [1], EViT-Dens169 [2], GlobalSkinNet [3], SkinEHDLF [4] and Swin Transformer–DGSNLA [5] were recently proposed for integrating complementary local and global representations. Other attempts have added feature-level augmentation, focal loss, attention mechanisms, and transformer-assisted encoders to increase class discrimination and minimise the influence of class imbalance [6], [9], [13]–[15]. However, in many current systems, the concatenated, addition per element or multi-backbone fusion approaches are adopted directly, which are time-consuming and costly. Such strategies can preserve redundant features, add complexity to the model, foster overfitting of the model, and yield only limited evidence of robustness when directly tested on the other datasets.

The current study thus strives to create HybridDermNet, a CNN–Transformer-based multi-class skin cancer classification framework with latent bottleneck learning. The proposed framework integrates texture and pigmentation extraction with a CNN and asymmetry, spatial colour distribution, and long-range contextual relationships with a transformer. A bidirectional cross-feature attentional mechanism enables information to flow back and forth between the two branches, and an adaptive gated mechanism controls the image-level contributions of each branch. The fused representation is subsequently compressed by a compact latent bottleneck (LLB), which is trained with reconstruction, latent regularisation, and transformation-consistency constraints to both denoise and de-redundant the fused data and mitigate dataset-specific variations.

This study has four key contributions. To begin, HybridDermNet proposes a single local–global representation architecture, where the CNN and transformer features are not simply concatenated, but rather exchange information via bidirectional cross-attention before being adaptively fused. Second, the proposed latent bottleneck offers a task-oriented compression mechanism that reduces high-dimensional redundancy, limits feature capacity, and stabilises representation learning. In its training pathway, the reconstruction process preserves important fused-feature information while keeping the inference cost unchanged. Third, the study provides a comprehensive generalisation-oriented evaluation using HAM10000, ISIC 2018, and PAD-UFES-20, comprising within-dataset testing, bidirectional transfer from source to target, and external validation on smartphone-acquired clinical images. Fourth, the framework's performance is assessed beyond overall accuracy through class-wise accuracy, calibration, confidence analysis, ablation studies, sensitivity testing, feature-space visualisation, Grad-CAM, Score-CAM, transformer attention, computational profiling, and repeated-run statistical validation. All of these contributions address aspects of inflexible fusion, weak feature regularisation, insufficient interpretability, and inadequate cross-domain assessment.

The rest of the paper is arranged as follows. Section 2 provides an overview of related literature and highlights methodological issues that have not been addressed. The proposed architecture, learning objectives, optimisation strategy, algorithms, and validation protocol are presented in Section 3. Classification, generalisation, ablation, explainability, efficiency, statistical and comparative results are reported in Section 4. The findings are discussed, and the study's limitations are presented in Section 5. The research is summarised, and future work is discussed in Section 6

2. RELATED WORK

With the advent of deep learning technologies, skin cancer diagnosis has evolved from traditional convolutional models to hybrid models that simultaneously capture local lesion features and global contextual information. The first CNN-based methods have shown that hierarchical features learned from dermoscopic images outperform manually designed descriptors. Nevertheless, despite the limited number of dermatological images, transfer-learning models based on VGG, ResNet, Inception, DenseNet, MobileNet, and GoogLeNet improved classification performance [29], [32], [34], [35], [37]–[40]. Comparisons with multilayer perceptrons and deep neural networks also provided further proof of the superiority of the convolutionally extracted features for representing colour, texture, shape, and irregular boundaries of the lesions [36]. However, each of the pretrained CNNs was still affected by inter-class similarity, severe class imbalance, background artefacts and acquisition variability. To exploit complementary predictions from different pretrained networks, ensemble and max-voting strategies are therefore introduced [31,33]. These methods were found to increase stability, but their requirements of high-dimensional features and partially redundant features restricted the scalability.

Later, Vision Transformers have become an alternative, which can learn long-range relationships between spatially separated lesion regions. Patch-based transformer models showed competitive performance for multiclass classification, capturing global asymmetry, colour distribution and structural irregularity [7, 8, 12, 16, 25]. However, Modified Swin Transformers adopted shifted-window attention and efficient local-to-global modelling, which further enhanced the hierarchical representation [13]. However, transformer-only models usually require large amounts of data to train. They were resource-heavy, and patch tokenisation might lead to a loss of fine-grained boundary and texture information. A study of pretrained architectures also revealed that the effectiveness of models across the various lesion classes and datasets was inconsistent, suggesting that there is no guarantee of effective generalisation with backbone selection alone [21].

Therefore, recent studies have begun to focus on combining CNNs and Transformers. Hybrid architectures consist of convolutional branches for local changes, pigmentation patterns, texture, and lesion boundaries, and self-attention branches to model global morphology and long-range dependencies [1]–[3]. Complementary local–global representations have been shown to increase discrimination between visually similar malignant and benign lesions in DenseNet [4] and VGG19 [5]- based fusion

frameworks, ConvNeXt [6] and EfficientNetV2 [7]- based fusion frameworks, and Swin Transformer [8]- based fusion frameworks. Focal loss-based CNN-Transformer learning has also mitigated the majority-class predominance [9], while CNN-LSTM and CNN-Transformer-LSTM have captured relations between image patches as spatial sequences [10, 11]. Further attention-enhanced hybrids that use channel, spatial, coordinate, non-local and separable attention have further prioritised diagnostically important lesion regions [14], [15], [17], [19]. Region-adaptive attention and graph-based reasoning have also been used to model structural dependencies among lesion regions [20].

In addition to classification, integrated classification–segmentation models have tried to enhance the localisation of the diagnosis. In scenarios with limited annotation and heterogeneous appearance, transformer-assisted broad-learning systems have yielded more accurate lesion masks, as have semi-supervised CNN–Transformer networks and dual-encoder attention architectures [18] [22] [30]. Conditional generative learning has been employed to generate samples for the minority classes and enhance the recognition of rare lesions [23]. Further advances in automated diagnosis have been made via heterogeneous clinical workflows, dual-attention explainability, patient metadata fusion, and adaptive segmentation mechanisms [24] [26] [27]. Another possible mechanism for enriching the feature representation is also investigated, quantum-inspired hybrid learning [28].

The important methodological characteristics, contributions, and unanswered questions of the studies reviewed are summarised in Table 1. The synthesis underscores the transition from traditional CNN-based classification to the CNN–Transformer hybrids with attention, and it demonstrates the necessity of compact latent representations and regularisation of the models for them to be robust.

Table 1: Summary Of The Literature On Deep-Learning-Based Skin Cancer Classification

S. No.	Research Theme	Representative References	Main Methodological Focus	Major Contributions	Research Gap
1	Conventional CNN and transfer-learning models	[29], [32], [34]–[40]	Use of VGG, ResNet, Inception, DenseNet, MobileNet, GoogLeNet, MLP, and deep neural networks for dermoscopic image classification	Established the effectiveness of hierarchical feature learning and transfer learning for automated lesion recognition	Limited global-context modelling, sensitivity to image artefacts, and weak generalisation across datasets and lesion categories
2	Ensemble and pretrained-model fusion	[21], [31], [33]	Combination of multiple pretrained models using voting, feature aggregation, and comparative backbone selection	Improved prediction stability and reduced dependence on a single architecture	Increased computational cost, redundant features, complex deployment, and insufficient control over fused representations
3	Vision Transformer-based classification	[7], [8], [12], [13], [16], [25]	Patch tokenisation, self-attention, shifted-window attention, and transformer-based global feature learning	Captured long-range dependencies, lesion asymmetry, and global spatial relationships more effectively than conventional CNNs	Requires large datasets and considerable computation; patching may suppress subtle texture, edge, and boundary information.

S. No.	Research Theme	Representative References	Main Methodological Focus	Major Contributions	Research Gap
4	CNN–Transformer hybrid classification	[1]–[6], [9]–[11]	Parallel or sequential fusion of CNN local features with Transformer global representations	Improved multiclass discrimination by jointly learning texture, colour, shape, and contextual lesion information	Most approaches rely on direct concatenation or addition, producing high-dimensional, redundant, and potentially overfitted feature spaces.
5	Attention, graph, and adaptive feature-learning models	[14], [15], [17], [19], [20], [26]–[28]	Channel, spatial, coordinate, non-local, separable, region-adaptive, graph-based, explainable, and quantum-inspired learning	Enhanced localisation of diagnostically relevant regions and improved discrimination of visually similar lesions	Attention modules often increase architectural complexity, while interpretability, external validation, and computational efficiency remain limited.
6	Segmentation, imbalance handling, and clinically oriented systems	[18], [22]–[24], [30]	Joint classification–segmentation, semi-supervised learning, generative augmentation, minority-class enhancement, and workflow integration	Improved lesion-boundary localisation, rare-class recognition, annotation efficiency, and clinical applicability	Robustness to domain shift, noise, heterogeneous populations, limited annotations, and unseen datasets remains insufficiently investigated.

Existing methods have gradually advanced in local–global lesion representation, attention-based discrimination, and class-imbalance handling, with some limitations: limited feature regularisation, redundant multimodel fusion, computational complexity, and insufficient cross-domain robustness, which collectively limit the clinical reliability of existing methods, as summarised in Table 1.

Nevertheless, there are still several limitations. Many hybrid systems directly concatenate or add high-dimensional CNN and Transformer features, leading to redundant representations and a high risk of overfitting. Often reported improvements are done on one or a few datasets, on controlled splits, or with much data augmentation, and robustness under domain-shift, noise, class imbalance, and small samples has not been sufficiently demonstrated. Moreover, current mechanisms for attention and ensembles tend to add parameters without directly controlling the information that is passed to the classifier. To address these gaps, HybridDermNet employs complementary CNN and Transformer representations and latent bottleneck learning. The bottleneck is designed to squeeze redundant features, retain diagnostically discriminative information, regularise joint representation learning and facilitate robust skin cancer classification for challenging lesion categories.

3. PROPOSED HYBRIDDERMNET METHODOLOGY

This section introduces the proposed HybridDermNet framework for robust skin cancer classification. Its methodology combines CNN-based local feature extraction, transformer-based global contextual modelling, cross-feature attention fusion and latent bottleneck representation learning to obtain compact and discriminative embeddings. The entire pipeline, training strategy, optimisation process, and extensive experimental protocol are detailed to showcase the framework's effectiveness, robustness, and cross-dataset generalisation capabilities.

3.1 Overall Framework and Data Preparation

HybridDermNet is a unified framework that integrates multiple types of deep learning models for robust multi-class skin cancer classification from dermoscopic and clinical skin lesion images, serving as an end-to-end hybrid deep learning framework. It combines convolutional feature extraction with transformer-based contextual representation, cross-feature attention fusion, and latent bottleneck representation learning, all in a single network. The entire framework can be expressed as Eq. 1

$$\hat{y}_i = \mathcal{C}\left(\mathcal{B}\left(\mathcal{F}\left(\mathcal{E}_{\text{cnn}}(I_i), \mathcal{E}_{\text{tr}}(I_i)\right)\right)\right), \tag{1}$$

CNN-based local feature extractor: $\mathcal{E}_{\text{cnn}}(\cdot)$ is used to extract local features from the CNN.Transformer-based global context encoder: $\mathcal{E}_{\text{tr}}(\cdot)$ is a module that extracts the global context from the transformer. The estimated probabilities of the image belonging to each harmonized skin-lesion category are stored in the predicted probability vector $\hat{y}_i$ contains. The overall information processing from the input image to the final diagnosis prediction is summarized in equation (1).

The CNN branch is tasked to learn the local dermoscopic features at a fine-grained level, such as borders, texture irregularities, pigmentation, vascular structures, and small-scale patterns. The transformer branch models long-range spatial interactions and global properties of lesions like asymmetry, shape, distribution of colour and relationships between distant image regions, in parallel. Then there are two branches that are aligned and integrated in the cross-feature fusion module. The fused representation is then passed through the latent bottle neck to eliminate redundant information, to stabilize features, and to prevent overfitting. Last, the compact latent representation is sent to the classification head for prediction of category of the lesion. The following subsections explain in detail the working operation of these architectural elements.

The experimental data used are from the public datasets, HAM10000 [41], ISIC 2018 [42] and PAD-UFES-20 [43]. The main reason using the HAM10000 as the benchmark is that it has a relatively wide spectrum of dermoscopic images of various types of pigmented skin lesions. ISIC 2018 is used to complement and validate across datasets in a different data distribution for benchmarking purposes. PAD-UFES-20 consists of smartphone-captured clinical skin-lesion images, which its main purpose is to determine robustness when varying the image acquisition conditions, the properties of the background, illumination, as well as the visual domain. The overall set of data collected is referred to as

$$\mathcal{D} = \mathcal{D}_{\text{HAM}} \cup \mathcal{D}_{\text{ISIC}} \cup \mathcal{D}_{\text{PAD}}, \tag{2}$$

Here $\mathcal{D}_{\text{HAM}}$, $\mathcal{D}_{\text{ISIC}}$, and $\mathcal{D}_{\text{PAD}}$ represent the samples from HAM10000, ISIC 2018 and PAD-UFES-20 respectively. The complete data collection is defined in equation (2) and is used in the model development, validation and external-generalization experiments.

Label-harmonization is performed prior to joint/cross-dataset evaluation, as the three datasets are originally diagnostic taxonomies. Each individual label $y_i^{(s)}$ from source data set s was translated to a common diagnostic label space using the following:

$$\tilde{y}_i = \mathcal{M}_s\left(y_i^{(s)}\right), \tag{3}$$

where $\mathcal{M}_s(\cdot)$ is the dataset specific mapping function and $\tilde{y}_i$ is the harmonized class label. Only clinically compatible categories are combined across datasets, but classes that are not matched or are poorly defined are reserved for dataset specific experiments or discarded from direct cross-dataset comparisons as defined in Equation (3). This procedure is used for avoiding merging the semantics of inappropriate classes and avoiding differences in performance due to differences in class semantics.

Whenever patient and/or lesion level identifiers are available, the samples are partitioned at that level to prevent information leakage. Images belonging to the same patient or lesion are only presented in one subset. The partitioning constraint is specified as

$$\mathcal{P}_{\text{tr}} \cap \mathcal{P}_{\text{val}} = \mathcal{P}_{\text{tr}} \cap \mathcal{P}_{\text{te}} = \mathcal{P}_{\text{val}} \cap \mathcal{P}_{\text{te}} = \emptyset, \tag{4}$$

The sets $\mathcal{P}_{\text{tr}}$, $\mathcal{P}_{\text{val}}$, and $\mathcal{P}_{\text{te}}$ contain patient or lesion identifiers that are present in the training set, the validation set, and the test set, respectively. To make sure that there is no subject or lesion-specific visual information shared across the data subsets, Equation (4) is used. In addition, stratification is used in classes where appropriate to maintain the relative frequency of the various lesion categories.

The CNN and Transformer branches have fixed spatial resolution requirements that are adhered to by resizing each image to a fixed size. Normalization of pixel values based on the channel-wise statistics is carried out according to

$$I_i^{\text{norm}}(h,w,c) = \frac{I_i(h,w,c)-\mu_c}{\sigma_c+\epsilon}, \quad (5)$$

where $I_i(h,w,c)$ is the intensity of the i-th pixel in the image at spatial location (h,w) and channel c, and μ_c and σ_c care the mean and standard deviation of the channel c, respectively. The numerical instability is prevented by the small constant ϵ. The normalized image in Equation (5) has a uniform distribution of the input and facilitates stable transfer learning from pre-trained backbones.

Handling of artifacts is done conservatively so as not to remove diagnostically useful structures of lesions. Images are checked for corrupted files, highly compressed images, duplicate samples, too many borders, acquisition markers, and visually overwhelming hair artefacts. Images are augmented and normalized for moderate variations in hair and illumination, and images with very degraded hair and illumination data are discarded based on a quality criterion. The model is deliberately exposed to acquisition variability that could be low, to avoid over-reliance on unrealistic dermoscopic images.

In training, the normalized input is fed into an augmentation operator $\mathcal{A}$ that is applied to it.

$$\tilde{I}_i = \mathcal{A}(I_i^{\text{norm}}; \theta_i), \quad (6)$$

where θ_i is the transformation parameters that are randomly sampled. The augmentation pipeline could contain horizontal/vertical flipping, some rotation, random cropping, scaling, brightness and contrast adjustment, and slight colour perturbations as detailed in Equation (6). To avoid distortion of the lesion morphology, clinically plausible ranges of transformations are restricted. Only deterministic resizing and normalization are used for validation and test images and no stochastic augmentation.

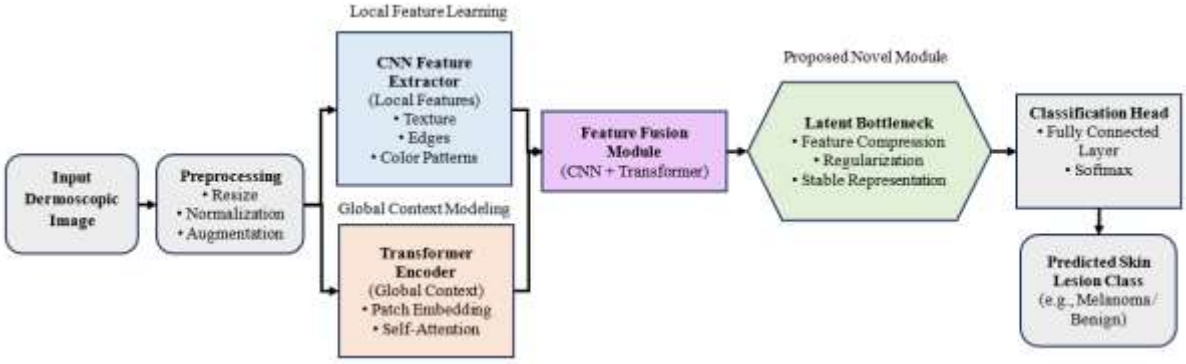

Figure 1: End-to-End System Architecture of HybridDermNet for Robust Skin Cancer Classification Integrating CNN-Based Local Feature Extraction, Transformer-Based Global Context Modeling, and Latent Bottleneck Representation Learning

The end-to-end HybridDermNet architecture is shown in Figure 1 and should be located at the end of this subsection. It visualizes the steps of input acquisition, preprocessing, parallel CNN/Transformer feature extraction, cross-feature fusion, latent bottleneck compression, and skin-lesion classification. The figure also separates the local feature-learning pathway from the global context-modelling pathway and emphasizes the latent bottleneck as the main regularization element that facilitates stable and generalizable representation learning table 2.

Table 2: Notations Used in the Proposed HybridDermNet Framework

Symbol(s)	Description
$I_i, \tilde{I}_i, I_i^{\text{norm}}, I^*$	Original, augmented, normalized, and unseen skin-lesion images
H, W, C	Image height, width, and number of lesion classes

Symbol(s)	Description
$\mathcal{D}, \mathcal{D}_{\text{HAM}}, \mathcal{D}_{\text{ISIC}}, \mathcal{D}_{\text{PAD}}$	Complete dataset collection and the HAM10000, ISIC 2018, and PAD-UFES-20 datasets
$y_i^{(s)}, \tilde{y}_i, \mathcal{M}_s(\cdot)$	Original source label, harmonized label, and dataset-specific label-mapping function
$\mathcal{P}_{\text{tr}}, \mathcal{P}_{\text{val}}, \mathcal{P}_{\text{te}}$	Patient- or lesion-wise training, validation, and testing identifier sets
$\mu_c, \sigma_c, \epsilon$	Channel mean, channel standard deviation, and numerical-stability constant
$\mathcal{A}(\cdot), \boldsymbol{\theta}_i$	Image augmentation operator and its transformation parameters
$\mathcal{E}_{\text{cnn}}(\cdot), F_{\text{cnn}}^{(l)}, F_{\text{cnn}}, \phi_l(\cdot)$	CNN encoder, intermediate CNN feature map, final local representation, and convolutional-stage transformation
$P_j, \mathbf{x}_j, N, D, E_{\text{pos}}, \mathbf{x}_{\text{cls}}$	Image patch, patch embedding, number of patches, embedding dimension, positional embedding, and classification token
$Q_m, K_m, V_m, A_m, M, d_k$	Query, key, value, attention output, number of attention heads, and key dimension
$X_r, \text{MSA}(\cdot), \text{MLP}(\cdot), \text{LN}(\cdot)$	Transformer-layer token representation, multi-head self-attention, feed-forward network, and layer normalization
$\mathcal{E}_{\text{tr}}(\cdot), F_{\text{tr}}$	Transformer encoder and final global token representation
$\psi_{\text{cnn}}(\cdot), \psi_{\text{tr}}(\cdot), F_{\text{cnn}}^a, F_{\text{tr}}^a$	CNN and transformer projection functions and their dimensionally aligned representations
$\hat{F}_{\text{cnn}}, \hat{F}_{\text{tr}}, \text{Attn}(\cdot)$	Cross-attention-refined CNN and transformer features and bidirectional attention operation
$\alpha, \odot, F_{\text{fusion}}, F_{\text{fusion}}^*$	Adaptive fusion coefficient, element-wise multiplication, fused representation, and normalized fused representation
$\mathcal{E}_{\text{lat}}(\cdot), Z_i, d_f, d_z, \rho$	Latent encoder, compact latent representation, fused dimension, latent dimension, and compression ratio
$\mathcal{D}_{\text{lat}}(\cdot), \hat{F}_{\text{fusion},i}, Z_i^{\text{aug}}$	Reconstruction decoder, reconstructed fused feature, and augmented-image latent embedding
$\text{GAP}(\cdot), \bar{Z}_i, H_i, o_i, \hat{y}_{i,c}$	Global average pooling, aggregated latent vector, hidden classifier feature, logits, and class probability
B	Mini-batch size
$\mathcal{L}_{\text{cls}}, \mathcal{L}_{\text{rec}}, \mathcal{L}_{\text{cross}}, \mathcal{L}_{\text{lat}}, \mathcal{L}_{\text{con}}, \mathcal{L}_{\text{total}}$	Classification, reconstruction, cross-feature consistency, latent regularization, transformation-consistency, and total losses
$\lambda_{\text{rec}}, \lambda_{\text{cross}}, \lambda_{\text{lat}}, \lambda_{\text{con}}$	Weights of the auxiliary loss components
$\Theta, \Theta_{\text{cnn}}, \Theta_{\text{tr}}, \Theta_{\text{fus}}, \Theta_{\text{lat}}, \Theta_{\text{dec}}, \Theta_{\text{cls}}$	Complete model parameters and the parameter subsets of the CNN, transformer, fusion, latent, decoder, and classifier modules
$T_{\text{w}}, T_{\text{max}}, p, \delta_s$	Warm-up epochs, maximum epochs, early-stopping patience, and minimum validation improvement

Symbol(s)	Description
$\eta_t, \eta_{\max}, \eta_{\min}, \gamma_q, \omega$	Current, maximum, and minimum learning rates, module-specific scaling factor, and weight-decay coefficient
$g_t, \tilde{g}_t, \tau_g$	Original gradient, clipped gradient, and maximum gradient norm
$S_{\text{val}}^{(t)}, \Theta^*$	Validation score at epoch $t$and best-checkpoint parameters
$\hat{c}^*, s_{\text{conf}}$	Predicted lesion class and confidence score
$\mathcal{E}_{s\to t}$	Source-to-target cross-dataset evaluation result
TP_c, FP_c, FN_c, TN_c	True-positive, false-positive, false-negative, and true-negative counts for class c
$F1_c, F1_{\text{macro}}, \text{BAcc}$	Class-wise F1-score, macro-F1, and balanced accuracy
ΔM_q	Performance contribution of component $q$in the ablation study
$R, M_r, \bar{M}, s_M, \text{CI}_{95\%}$	Number of runs, run-specific result, mean, standard deviation, and 95% confidence interval

3.2 Hybrid Local–Global Feature Extraction

The proposed HybridDermNet uses two complementary representation-learning pathways, one on a fine-grained level, and the other on a coarse-grained level to capture diagnostically relevant information on both scales. The convolutional pathway processes local visual patterns, and the transformer pathway processes the global contextual relationships in the region of the lesion. This dual representation is essential since skin-lesion classification involves both small-scale texture and pigmentation features and larger-scale structural features like lesion asymmetry, the spatial distribution of colour in the image, the organization of the image borders and long-range interaction between image-separated colour regions.

A convolutional feature extractor produces a hierarchy of local feature maps for a preprocessed input image $\tilde{I}_i$ that is obtained by Equation (6).

A convolutional feature extractor produces a hierarchy of local feature maps for a preprocessed input image $\tilde{I}_i$ that is obtained by Equation (6).

$$F_{\text{cnn}}^{(l)} = \phi_l\left(F_{\text{cnn}}^{(l-1)}\right), F_{\text{cnn}}^{(0)} = \tilde{I}_i\,, \tag{7}$$

where $F_{\text{cnn}}^{(l)}$ is the feature representation generated by the l-th convolutional stage, and $\phi_l(\cdot)$ is the sequence of convolution, normalization, nonlinear activation, and spatial downsampling operations carried out by the l-th convolutional stage. The following equation represents the progressive transformation of the input image into more and more abstract local dermoscopic representations (7). The earlier convolutional layers mainly represent edges, color changes, and fine texture, while the later convolutional layers represent irregular pigment networks, lesion edges, vascular patterns, and higher level morphological patterns.

$$F_{\text{cnn}} = \mathcal{E}_{\text{cnn}}(\tilde{I}_i) \in \mathbb{R}^{H_c \times W_c \times C_c}, \tag{8}$$

H_c, W_c, and C_c denote the spatial height, spatial width and number of channels of the final feature tensor in the CNN, respectively. The convolutional representation maintains the spatial structure of the image and embeds discriminative local features in different channels as shown in Equation (8). Such features are well-suited to the detection of lesion-specific characteristics in terms of texture, colour heterogeneity, border sharpness and structural irregularities, which can help discriminate between similar-looking benign and malignant lesion types.

While convolutional operations are useful for local pattern recognition, their spatially limited receptive fields could represent incomplete information about characteristics across the lesion. To overcome this, the image or the projected convolutional feature map is represented by a tokenized token that will be passed on to the transformer pathway. Let P_j denote the j -th patch of the image without overlapping. These patches are flattened and embedded in a D -dimensional space, according to Eq. 9

$$\mathrm{x}_j = \mathrm{W}_p \mathrm{vec}(P_j) + \mathrm{b}_p, j = 1,2,\dots,N, \quad (9)$$

where $\mathrm{vec}(\cdot)$ maps a patch into a 1D vector, $\mathrm{W}_p$and b_p are learnable projection parameters and N is is the number of patches. Each spatial image region in (9) is converted into a feature token that is processed by the transformer. To facilitate matrix-based attention operations, whether in the embedding dimension Dim or Dim 2, D is was kept to be the same across all tokens.

$$X_0 = [\mathrm{x}_{\mathrm{cls}}; \mathrm{x}_1; \mathrm{x}_2; \dots; \mathrm{x}_N] + E_{\mathrm{pos}}, \quad (10)$$

where $\mathrm{x}_{\mathrm{cls}}$ is an optional learnable classification token and E_{pos} denotes the positional embedding matrix. By combining the information from equation (10), the transformer allows separation of patches based on their spatial arrangement and thus models relationships that are associated with the lesion symmetry, distribution of the lesion borders, pigmentation patterns and global structural organization.

The contextual interaction among tokens is modelled using multi-head self-attention. Each attention head m gets a query, key, and value representation, which are computed as follows Eq. 11

$$Q_m = X\mathrm{W}_m^Q, K_m = X\mathrm{W}_m^K, V_m = X\mathrm{W}_m^V, \quad (11)$$

where X is the input token matrix while W_m^Q, W_m^K, and W_m^V are learnable projection matrices. To obtain the attention response of the m -th head, the following equation is used:

$$A_m = \mathrm{Softmax}\left(\frac{Q_m K_m^{\mathrm{T}}}{\sqrt{d_k}}\right) V_m, \quad (12)$$

Here d_k is the dimensionality of the key vectors. In Equation (12), the token pairs with higher contextual similarity get higher weights, enabling the model to capture interactions between spatially distant regions of lesion. This is especially useful for the diagnosis of some of the diagnostic properties which are seen globally, like asymmetry, nonuniform structure organization, nonuniform expansion of the lesion and heterogeneous pigmentation.

$$\mathrm{MSA}(X) = \mathrm{Concat}(A_1, A_2, \dots, A_M)\mathrm{W}^O, \quad (13)$$

W^O is the output projection matrix and M is the number of attention heads. Different heads can learn complementary contextual relationships, such as shape correspondence, colour-region interaction, boundary continuity and long-range dependency patterns, as shown in equation (13).

$$X_r' = X_{r-1} + \mathrm{MSA}(\mathrm{LN}(X_{r-1})), \quad (14)$$

$$X_r = X_r' + \mathrm{MLP}(\mathrm{LN}(X_r')), \quad (15)$$

where the layers are normalized with layer norm $\mathrm{LN}(\cdot)$ and position-wise feed-forward network $\mathrm{MLP}(\cdot)$. Equations (14) and (15) maintain the original token information by residual learning to gradually extract global contextual relationships. The layer-normalization operations help to stabilize the training and the feed-forward transformation enhances the nonlinear modelling ability of each token.

$$F_{\mathrm{tr}} = \mathcal{E}_{\mathrm{tr}}(\tilde{I}_i) = X_R \in \mathbb{R}^{(N+1)\times D}, \quad (16)$$

where F_{tr} provides the patch tokens that have been adapted to the context, and (if present) the classification token. The global feature representation produced by the transformer branch is defined by Equation (16). Each token also carries information from other regions, in addition to the information from its image region, by multiple self-attention steps.

3.3 Cross-Feature Attention Fusion

The convolutional branch (Section 3.2) produces complementary feature representations that reflect local dermoscopic features, while the transformer branch produces feature representations that reflect the global context. Both representations provide valuable diagnostic information, but differ from each other in feature dimensionality, spatial organization and semantic abstraction. These heterogeneous representations are often concatenated directly, which can lead to redundancy and to assigning equal weight to both branches, irrespective of their diagnostic value. To overcome this restriction, HybridDermNet proposes a cross-feature attention fusion module to project the representations into a shared embedding space, allow cross-representation interaction, and finally fuse using adaptive attention. This approach makes it possible to highlight the most informative features of the local and global context, while leaving out redundant and less discriminative representations.

The CNN output in Equation (8) is encoded as a three-dimensional feature tensor, while the transformer output in Equation (16) is a set of contextual feature tokens, both of which are projected into the same latent dimension. The aligned CNN representation is achieved as shown.

$$F_{\text{cnn}}^{a} = \psi_{\text{cnn}}(F_{\text{cnn}}), \quad (17)$$

where $\psi_{\text{cnn}}(\cdot)$ is a learnable linear projection from the convolutional feature tensor to the common embedding dimension. Likewise, the transformer representation is projected based on

$$F_{\text{tr}}^{a} = \psi_{\text{tr}}(F_{\text{tr}}), \quad (18)$$

where $\psi_{\text{tr}}(\cdot)$ is the transformer projection layer. Equations (17) and (18 assure that both of them have to have comparable dimensions so that they can be used to interact with each other without losing information due to incompatible feature dimensions.

$$\hat{F}_{\text{cnn}} = F_{\text{cnn}}^{a} + \text{Attn}(F_{\text{cnn}}^{a}, F_{\text{tr}}^{a}), \quad (19)$$

where $\text{Attn}(\cdot)$ is the cross-attention operation. Similarly, local conv information is used in the transformer representation update.

$$\hat{F}_{\text{tr}} = F_{\text{tr}}^{a} + \text{Attn}(F_{\text{tr}}^{a}, F_{\text{cnn}}^{a}), \quad (20)$$

After feature interaction, an adaptive attention mechanism is used to decide the relative impact of the branches. The attention weights are learned automatically, rather than being assigned equal weights.

$$\alpha = \sigma\left(W_f\left[\hat{F}_{\text{cnn}}; \hat{F}_{\text{tr}}\right] + b_f\right), \quad (21)$$

where W_f and b_f are trainable parameters, $[\cdot;\cdot]$ represents feature concatenation, and $\sigma(\cdot)$ represents the sigmoid activation function. The attention coefficient α varies adaptively depending on the diagnostic information of the image. The attention values of features with more discriminative power are increased while those of less discriminative features are suppressed.

$$F_{\text{fusion}} = \alpha \odot \hat{F}_{\text{cnn}} + (1 - \alpha) \odot \hat{F}_{\text{tr}}, \quad (22)$$

where $\odot$ is an element-wise multiplication. This will enable the network to adjust the weights of local dermoscopic features and global contextual representations for each lesion image dynamically. In contrast to fixed-weight fusion methods, the proposed gated mechanism is constantly adapted in training, allowing the model to focus on various sources of features based on the morphology, pigmentation, structure and acquisition conditions of the lesion.

The fused representation is further stabilized by normalizing it prior to being passed to the latent bottleneck module. The normalized feature representation is given as

$$F_{\text{fusion}}^{*} = \text{LN}(F_{\text{fusion}}), \quad (23)$$

where $\text{LN}(\cdot)$ denotes layer normalization. The use of equation (23) will reduce the changes of internal feature distribution, enhance the numerical stability in optimization process, and make the following latent representation learning process more smooth.

3.4 Latent Bottleneck Representation Learning

While the fused representation from Equation (23) combines local and global information, it can also contain redundant dimensions, variations due to acquisition, and poorly discriminating responses. To overcome these drawbacks, HybridDermNet proposes a latent bottleneck representation learning module that reduces the dimensionality of the fused feature space to a small latent space. The bottleneck aims to keep a diagnostic signal while filtering out noise and dataset-specific artifacts that can diminish generalization.

The normalized fused representation is first encoded by a latent encoder to become a latent representation,

$$Z_i = \mathcal{E}_{\text{lat}}(F^*_{\text{fusion}}), \tag{24}$$

The latent encoder is denoted by where $\mathcal{E}_{\text{lat}}(\cdot)$ and $Z_i \in \mathbb{R}^{d_z}$ is the compact latent representation of the i -th image. The dimensionality of the fused representation is greater than the latent dimension d_z is selected. The encoder is designed to make the model remember only the most important lesion features necessary for the classification as described in Equation (24).

$$Z_i = \text{Dropout}\Big(\phi\big(\text{LN}(W_2 F^*_{\text{Fusion}} + b_2)\big)\Big), \tag{25}$$

Here, W_z and b_z are trainable encoding parameters in the layer normalization and $\phi(\cdot)$ is a non-linear activation function. By using equation (25), the dimensionality of the fused features is reduced, and the distribution of these features is also stabilized and co-adaptation of latent components is restricted.

$$\rho = \frac{d_z}{d_f}, 0 < \rho < 1, \tag{26}$$

The dimensionality of the fused representation d_f and the latent dimension d_z. The amount of compression due to the bottleneck is captured in Equation (26). Very large values of ρ may will give a level of redundancy that is not necessary, but very small values will lose diagnostic data. Consequently, the latent dimension is determined based on the validation-based sensitivity analysis.

$$\mathcal{L}_{\text{lat}} = \frac{1}{B} \sum_{i=1}^{B} \| Z_i \|_2^2, \tag{27}$$

Here B denotes the mini-batch size. The use of (27) encourages bounded and smoother representations and punishes too large latent values. This restriction makes the optimization more insensitive to image appearance and helps to maintain stability in the optimization from heterogeneous datasets.

$$\hat{F}_{\text{fusion},i} = \mathcal{D}_{\text{lat}}(Z_i), \tag{28}$$

where $\mathcal{D}_{\text{lat}}(\cdot)$ represents the reconstruction decoder and $\hat{F}_{\text{fusion},i}$ represents the reconstructed feature vector. The auxiliary learning pathway in equation (28) brings in an additional pathway to learn the compressed representation, so that enough information about the original fused features is preserved.

$$\mathcal{L}_{\text{rec}} = \frac{1}{B} \sum_{i=1}^{B} \| F^*_{\text{fusion},i} - \hat{F}_{\text{fusion},i} \|_2^2, \tag{29}$$

Additionally, to ensure consistency under realistic image distortions, the latent representation of an original image and its corrupted counterpart can be forced to be close to each other. The latent consistency loss is given by

$$\mathcal{L}_{\text{con}} = \frac{1}{B} \sum_{i=1}^{B} \| Z_i - Z_i^{\text{aug}} \|_2^2, \tag{30}$$

Here, $Z_i$and Z_i^{aug} are the latent representations of the original image and the enhanced image, respectively. A moderate variation of the orientation, illumination, scale and colour of the image is also encouraged while maintaining the identity of the lesion, as shown in equation (30).

The reconstruction pathway is only needed during training and can be removed during inference. This design retains the regularization property of the reconstruction-consistency learning without adding to the deployment cost. In inference, the latent vector produced in the last layer of the transformer, as shown in Equation (24), is sent directly to the classification head detailed in Section 3.5.

Cross-dataset generalization is a key point for the proposed latent bottleneck. HAM10000 mainly consists of dermoscopic images while PAD-UFES-20 consists of clinical images taken with smartphones. There are variations of these datasets in illumination, background, image resolution, colour distribution, lesion scale and acquisition protocol. The model is encouraged to keep diagnostic information stable and avoids depending on visual patterns from the source by compressing the fused features and applying the constraints of Equations (27), (29) and (30).

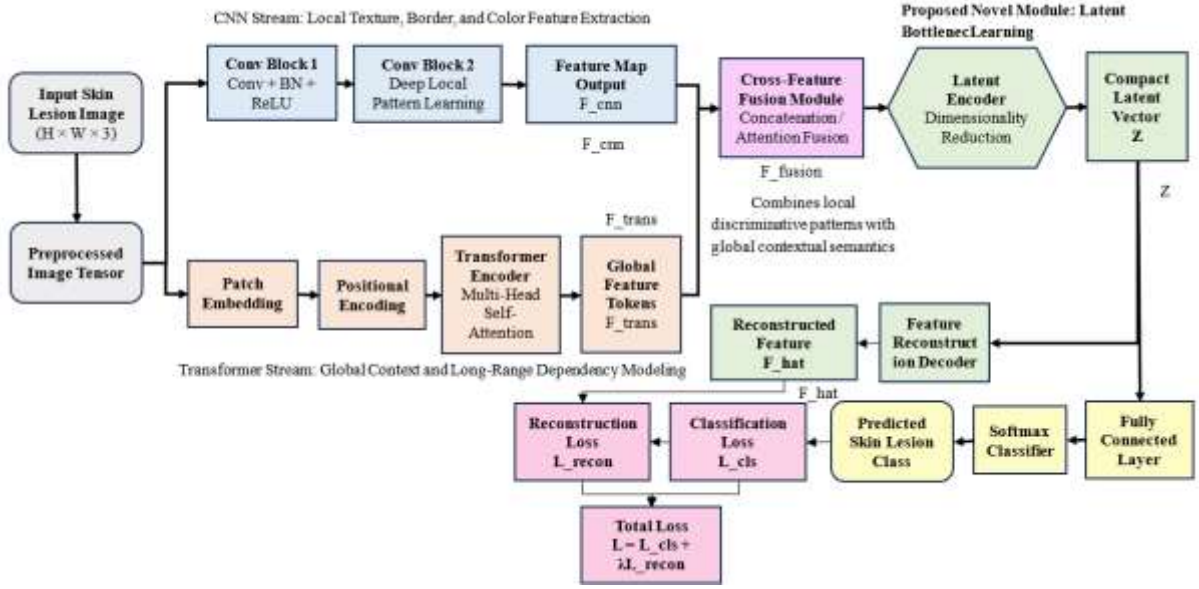

Figure 2: Detailed Model Architecture of HybridDermNet Showing Dual-Branch Feature Extraction, Fusion, and Latent Bottleneck Representation Learning

The figure 2 should be added at the end of this subsection. It provides the detailed parallel CNN and transformer streams, cross-feature fusion module, latent encoder, compact latent representation, reconstruction decoder (optional), classification head and learning objectives of HybridDermNet. The figure should highlight the hidden bottleneck, as the main regularization path linking hybrid feature extraction to robust skin-lesion classification.

3.5 Classification and Composite Learning Objective

The discriminative representation generated by modeling the bottleneck module in Equation (24) is used as the skin-lesion central discriminative representation. The classification head gets a more stable and compact feature vector than the original fused representation due to the latent encoder's suppression of redundant and acquisition-specific information. First, latent features are aggregated and then projected into a task-specific embedding, and then class probabilities are estimated.

The aggregated latent representation for the i -th image is given by:

$$\bar{Z}_i = \mathrm{GAP}(Z_i)\ , \qquad (31)$$

The aggregated latent representation for the i -th image is given by:

$$\bar{Z}_i = \mathrm{GAP}(Z_i)\ , \qquad (31)$$

where $\mathrm{GAP}(\cdot)$ refers to the global average pooling applied to the latent representation as a whole, whilst keeping its spatial or token-wise organization. If Z_i is a one dimensional vector, then Equation (31) becomes an identity operation. $\bar{Z}_i$ provides a fixed dimensional representation, the resulting feature vector, which is fed to the classification head.

The parameters $W_h$and b_h are learnable, $\mathrm{BN}(\cdot)$ is a batch normalization, and $\phi(\cdot)$ is a non-linear activation function. In the final stage of prediction, Equation (32) fine-tunes the compact latent representation and enhances the class separability. $H_i$during training can accept the dropout during the training process, which can be used to remove the co-adaptation phenomenon among the classifiers' neurons.

$$H_i = \phi\big(\mathrm{BN}(W_h\bar{Z}_i + b_h)\big)\,, \tag{32}$$

The parameters $W_h$and b_h are learnable, $\mathrm{BN}(\cdot)$ is a batch normalization, and $\phi(\cdot)$ is a non-linear activation function. In the final stage of prediction, Equation (32) fine-tunes the compact latent representation and enhances the class separability. $H_i$during training can accept the dropout during the training process, which can be used to remove the co-adaptation phenomenon among the classifiers' neurons.

where $W_o$and b_o denote the parameters of the output layer. The softmax function can be used to get the corresponding multi-class probability distribution as

$$o_i = W_o H_i + b_o, \tag{33}$$

where $W_o$and b_o denote the parameters of the output layer. The softmax function can be used to get the corresponding multi-class probability distribution as

$$\hat{y}_{i,c} = \frac{\exp(o_{i,c})}{\sum_{k=1}^{C}\exp(o_{i,k})}, c = 1,2,\dots,C, \tag{34}$$

where $C$is the number of harmonized skin-lesion classes and $\hat{y}_{i,c}$is the predicted probability that image i belongs to class c . The raw logits are normalized to class probabilities that sum to 1 using Equation (34).

B denotes the mini-batch size, and ϵ is a small numerical constant, $y_{i,c}$ is the indicator for class c of the ground truth. In Equation (35) the penalty is for the mis-match between the predicted class distribution and the actual diagnostic label. The objective of the present invention is to develop an architecture, but a class weighting could be added to Equation (35) in controlled experiments to minimize the effect of imbalance in lesion distributions.

$$\mathcal{L}_{cls} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{c=1}^{C} y_{i,c}\log\big(\hat{y}_{i,c}+\epsilon\big), \tag{35}$$

B denotes the mini-batch size, and ϵ is a small numerical constant, $y_{i,c}$ is the indicator for class c of the ground truth. In Equation (35) the penalty is for the mis-match between the predicted class distribution and the actual diagnostic label. The objective of the present invention is to develop an architecture, but a class weighting could be added to Equation (35) in controlled experiments to minimize the effect of imbalance in lesion distributions.

$$\mathcal{L}_{\text{cross}} = \frac{1}{B}\sum_{i=1}^{B} \| \Pi_{\text{cnn}}\big(\hat{F}_{\text{cnn},i}\big) - \Pi_{\text{tr}}\big(\hat{F}_{\text{tr},i}\big) \|_2^2, \tag{36}$$

Learnable projection functions, $\Pi_{\text{cnn}}(\cdot)$and $\Pi_{\text{tr}}(\cdot)$, are used for mapping the two refined branch representations to a shared comparison space. Both branches are motivated to keep their local and global features while retaining their class-related information by equation (36).

$$\mathcal{L}_{\text{cross}} = \frac{1}{B}\sum_{i=1}^{B} \| \Pi_{\text{cnn}}\big(\hat{F}_{\text{cnn},i}\big) - \Pi_{\text{tr}}\big(\hat{F}_{\text{tr},i}\big) \|_2^2, \tag{36}$$

Learnable projection functions, $\Pi_{\text{cnn}}(\cdot)$and $\Pi_{\text{tr}}(\cdot)$, are used for mapping the two refined branch representations to a shared comparison space. Both branches are motivated to keep their local and global features while retaining their class-related information by equation (36).

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{cross}}\mathcal{L}_{\text{cross}} + \lambda_{\text{lat}}\mathcal{L}_{\text{lat}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}}, \tag{37}$$

In whichλ_{rec}, λ_{cross}, λ_{lat}, and λ_{con} are the weighting coefficients. The final objective, which is the same as the one used in training HybridDermNet, is defined in Equation (37). The classification term is still dominant, and the auxiliary losses work as regularizers, enhancing the feature compactness, branch compatibility, reconstruction fidelity and transformation invariance.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{cross}}\mathcal{L}_{\text{cross}} + \lambda_{\text{lat}}\mathcal{L}_{\text{lat}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}}, \tag{37}$$

In whichλ_{rec}, λ_{cross}, λ_{lat}, and λ_{con} are the weighting coefficients. The final objective, which is the same as the one used in training HybridDermNet, is defined in Equation (37). The classification term is still dominant, and the auxiliary losses work as regularizers, enhancing the feature compactness, branch compatibility, reconstruction fidelity and transformation invariance.

The weighting coefficients of the equation (37) are determined from the validation data instead of the test data. If λ_{rec} is too small, the fused-feature structure will not be informative, and if it is too large, the reconstruction will be emphasized over the classification. Likewise, when using λ_{cross}, the goal is to force agreement, but not necessarily make the CNN and transformer branches identical. The latent magnitude and augmentation invariance are controlled by the coefficients $\lambda_{\text{lat}}$and λ_{con} respectively.

The proposed HybridDermNet is trained with a staged transfer-learning approach that allows for stable convergence of the network, while maintaining the beneficial visual representation learnt from large scale image datasets. The network backbone and transformer encoder are preloaded, while the cross feature attention fusion module, latent bottleneck, reconstruction decoder and classification head are randomly initialized. Suppose the entire model parameter set is:

$$\Theta = \{\Theta_{\text{cnn}}, \Theta_{\text{tr}}, \Theta_{\text{fus}}, \Theta_{\text{lat}}, \Theta_{\text{dec}}, \Theta_{\text{cls}}\}, \quad (38)$$

The parameters of the CNN backbone, the transformer encoder, the fusion module, the latent encoder, the reconstruction decoder, and the classification head are denoted as Θ_{cnn}, Θ_{tr}, Θ_{fus}, Θ_{lat}, Θ_{dec}, and Θ_{cls} respectively. The entire parameter space optimized when developing the model is defined by equation (38).

$$\Theta = \{\Theta_{\text{cnn}}, \Theta_{\text{tr}}, \Theta_{\text{fus}}, \Theta_{\text{lat}}, \Theta_{\text{dec}}, \Theta_{\text{cls}}\}, \quad (38)$$

The parameters of the CNN backbone, the transformer encoder, the fusion module, the latent encoder, the reconstruction decoder, and the classification head are denoted as Θ_{cnn}, Θ_{tr}, Θ_{fus}, Θ_{lat}, Θ_{dec}, and Θ_{cls} respectively. The entire parameter space optimized when developing the model is defined by equation (38).

In the following, denote the number of warm-up epochs by T_{w}. The CNN parameters are not updated in this step, while the latent, decoder, classification and fusion parameters are updated with the total objective in Equation (39).

$$\Theta_{\text{cnn}}^{(t+1)} = \Theta_{\text{cnn}}^{(t)}, \Theta_{\text{tr}}^{(t+1)} = \Theta_{\text{tr}}^{(t)}, t < T_{\text{w}}, \quad (39)$$

In the following, denote the number of warm-up epochs by T_{w}. The CNN parameters are not updated in this step, while the latent, decoder, classification and fusion parameters are updated with the total objective in Equation (39).

Θ_{new} are the parameters of the newly added modules; $\Theta_{\text{cnn}}^{u(t)}$and $\Theta_{\text{tr}}^{u(t)}$ are subsets of the CNN and transformer parameters that are unfrozen at epoch t. This is achieved by using equation (40) to update the parameters of the layers in the skin-lesion domain gradually without changing all the pre-trained layers.

$$\Theta_{\text{train}}^{(t)} = \Theta_{\text{new}} \cup \Theta_{\text{cnn}}^{u(t)} \cup \Theta_{\text{tr}}^{u(t)}, \quad (40)$$

Θ_{new} are the parameters of the newly added modules; $\Theta_{\text{cnn}}^{u(t)}$and $\Theta_{\text{tr}}^{u(t)}$ are subsets of the CNN and transformer parameters that are unfrozen at epoch t. This is achieved by using equation (40) to update the parameters of the layers in the skin-lesion domain gradually without changing all the pre-trained layers.

$$\theta_{t+1} = \theta_t - \eta_t \left(\frac{\widehat{m}_t}{\sqrt{\widehat{v}_t}+\epsilon} + \omega\theta_t\right), \quad (41)$$

AdamW is used as the main optimizer as it separates weight decay from gradient-based parameter updates. The update procedure is given as follows for parameter $\theta \in \Theta_{\text{train}}^{(t)}$ Eq. 41

$$\theta_{t+1} = \theta_t - \eta_t \left(\frac{\widehat{m}_t}{\sqrt{\widehat{v}_t}+\epsilon} + \omega\theta_t\right), \quad (41)$$

where η_t is the learning rate, $\widehat{m}_t$and $\hat{v}_t$ are first- and second-order gradient moment estimates with added bias, ω is the weight-decay coefficient and ϵ is a numerical-stability constant. Explicit model weight regularization, coupled with parameter updates via (41) is performed.

The learning rate for module q is denoted by $\eta_t^{(q)}$ and the coefficient for module q at time t is denoted by γ_q. By using equation (42), the newly added elements can learn at an adequate speed and make only small changes to pre-trained parameters.

$$\eta_t^{(q)} = \gamma_q \eta_t, q \in \{\text{cnn}, \text{tr}, \text{fus}, \text{lat}, \text{dec}, \text{cls}\}\,, \quad (42)$$

The learning rate for module q is denoted by $\eta_t^{(q)}$ and the coefficient for module q at time t is denoted by γ_q. By using equation (42), the newly added elements can learn at an adequate speed and make only small changes to pre-trained parameters.

The maximum learning rate $\eta_{\max}$ and the minimum learning rate $\eta_{\min}$ are specified, and $T_{\max}$ is the number of epochs of training. Relatively large updates to the model are made in the early stages of optimization, and smaller updates in the final stages of optimization, as described in Equation (43).

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min})\left[1 + \cos\left(\frac{\pi(t - T_{\text{w}})}{T_{\max} - T_{\text{w}}}\right)\right], t \geq T_{\text{w}}\,, \quad (43)$$

The maximum learning rate $\eta_{\max}$ and the minimum learning rate $\eta_{\min}$ are specified, and $T_{\max}$ is the number of epochs of training. Relatively large updates to the model are made in the early stages of optimization, and smaller updates in the final stages of optimization, as described in Equation (43).

A gradient clipping method is used to avoid unstable update due to high gradient value, especially in the unfreezing of a transformer layer. Let $g_t = \nabla_\Theta \mathcal{L}_{\text{total}}$ be the gradient of the total objective. The clipped gradient is calculated as

$$\tilde{g}_t = g_t \min\left(1, \frac{\tau_g}{\|g_t\|_2 + \epsilon}\right), \quad (44)$$

The formula below shows that the resulting gradient norm is multiplied by τ_g is, which is the largest one allowed. The result of Equation (44) is that it limits the gradients to the desired value and scales them up or down in proportion to their size.

Reduce memory usage and speed up computation by using mixed-precision training. Forward and backward operations are given lower precision representations in which numerically safe, and critical parameter updates are kept in full precision. To avoid numerical underflow in the calculation of the gradients, dynamic loss scaling is used. This approach allows for larger mini-batches or higher resolutions of images without altering the mathematical goal of the framework itself.

Early stopping is used to avoid overfitting based on the performance on the validation set. For example, if the validation score is the macro-averaged F1-score at epoch t then it is denoted by $S_{\text{val}}^{(t)}$. Training is terminated when

$$\max_{t-p<j\leq t} S_{\text{val}}^{(j)} \leq S_{\text{best}} + \delta_s, \quad (45)$$

where p is the patience interval, S_{best} is the best previously observed score, and δ_s is the minimum required improvement. If no significant validation improvements are seen in the validation over a given patience period, then training is terminated according to equation (45). The checkpoint having the most validations is kept for final testing.

In inference, all the augmentation operations in training are turned off and the reconstruction decoder is disabled. The image I^* is then resized, normalized and fed into the CNN, transformer, fusion, latent, and classification modules. The inference operation is represented as:

$$\hat{\mathbf{y}}^* = \text{HybridDermNet}(I^*; \Theta^*)\,, \quad (46)$$

The predictions are the class-probability vector Θ^* and the parameters are those kept from the best validation checkpoint denoted by $\hat{\mathbf{y}}^*$. The entire inference path (without the auxiliary reconstruction branch) is shown in Equation 46.

The probability distribution in Equation (47) is transformed to the final diagnostic class in Equation (46). Confidence of the associated prediction is:

$$\hat{c}^* = \arg\max \text{c} \in \{1, \ldots, \text{C}\}\ \hat{y}_c^*. \quad (47)$$

The confidence score reported in (48) is the maximum predicted probability. Confidence predictions less than a threshold derived from the validation can be tagged for expert review, instead of as a final automatic ruling. This approach to inferring confidence boosts the clinical utility of HybridDermNet, helping to differentiate high-confidence cases from those cases that are more uncertain and require further clinical evaluation.

$$s_{\text{conf}} = \max_{c\in\{1,\ldots,C\}} \hat{y}_c^*. \quad (48)$$

3.7 Algorithmic Representation of HybridDermNet

The algorithmic representation of the proposed HybridDermNet (HD) framework is shown in this section. The algorithms briefly summarize the sequential execution of feature extraction, cross-feature fusion, latent representation learning, optimization and inference, while avoiding a repetition of the mathematical formulations introduced earlier. They represent a concise procedural description of the entire workflow, which allows easy reproduction and implementation and makes the proposed hybrid deep learning architecture easy to understand.

Algorithm: HybridDermNet Forward Representation Learning

Input: Preprocessed skin-lesion image $\tilde{I}_i$

Output: Class-probability vector $\hat{\mathrm{y}}_i$, latent representation Z_i, and reconstructed fused feature $\hat{F}_{\text{fusion},i}$

1. Extract local CNN features F_{cnn} using Equation (8).
2. Generate global transformer tokens F_{tr} using Equation (16).
3. Project F_{cnn} and F_{tr} into a common embedding space using Equations (17) and (18).
4. Compute bidirectional cross-attention features using Equations (19) and (20).
5. Estimate adaptive fusion weights using Equation (21).
6. Generate the gated fused representation using Equation (22).
7. Normalize the fused representation using Equation (23).
8. Encode the normalized fused features into the compact latent representation Z_i using Equation (24).
9. Reconstruct the fused representation from Z_i using Equation (28).
10. Aggregate the latent features using Equation (31).
11. Generate class logits using Equations (32) and (33).
12. Compute the class-probability vector $\hat{\mathrm{y}}_i$ using Equation (34).
13. Return $\hat{\mathrm{y}}_i$, Z_i, and $\hat{F}_{\text{fusion},i}$.

Algorithm 1: HybridDermNet Forward Representation Learning

Algorithm 1 shows forward representation learning process of HybridDermNet. The algorithm uses a preprocessed dermoscopic image as the input to extract local characteristics of the lesion using the CNN branch and the global context of the image using the transformer branch. They are both encoded into a shared embedding space and optimized by bidirectional cross feature attention. The complementary information is then combined into a complete feature representation using an adaptive gated fusion mechanism and then compressed by the latent bottleneck encoder. The optional reconstruction pathway maintains feature consistency throughout training, and the head constructs the final multi-class probability distribution to make accurate skin lesion prediction.

Algorithm: HybridDermNet Training and Inference Procedure

Input: Training set $\mathcal{D}_{\text{tr}}$, validation set $\mathcal{D}_{\text{val}}$, maximum epochs $T_{\max}$
Output: Optimized parameters Θ^*, predicted class $\hat{c}^*$, and confidence score s_{conf}

- Initialize the CNN and transformer branches with pretrained weights.
- Randomly initialize the fusion, latent, decoder, and classification modules.
- Freeze the pretrained branches for the warm-up period using Equation (39).
- For each epoch $t = 1, \ldots, T_{\max}$:
  a. Load a mini-batch from $\mathcal{D}_{\text{tr}}$.
  b. Apply preprocessing and training augmentation using Equation (6).
  c. Execute Algorithm 1 to obtain predictions, latent embeddings, and reconstructed features.
  d. Compute the total loss using Equation (37).
  e. Backpropagate the loss and clip gradients using Equation (44).
  f. Update trainable parameters using AdamW in Equation (41).
  g. Adjust module-specific learning rates using Equations (42) and (43).
  h. Progressively unfreeze CNN and transformer layers using Equation (40).
  i. Evaluate the model on $\mathcal{D}_{\text{val}}$.
  j. Save the checkpoint when the validation score improves.
  k. Stop training when the condition in Equation (45) is satisfied.
- Load the best model parameters Θ^*.
- For each unseen image I^*, execute the inference pathway using Equation (46).
- Determine the predicted class using Equation (47).
- Compute the confidence score using Equation (48).
- Return Θ^*, $\hat{c}^*$, and s_{conf}.

Algorithm 2: HybridDermNet Training and Inference Procedure

The detailed training and inference process of HybridDermNet are summarized in Algorithm 2.The full training and inference process of HybridDermNet is summarized in Algorithm 2. Model is initialized with pre-trained CNN and Transformer weights and then a warm-up phase is used, which optimizes only the newly introduced modules of the model. Progressive layer unfreezing allows to stabilize the adaptation in the domain of skin lesion before fine-tuning end-to-end. For each iteration of training, Algorithm 1 is run, the composite learning objective is calculated, the gradients are clamped, and the parameters are updated using the AdamW optimization method with adaptive scheduling of the learning rate. Validation based checkpointing and early stopping are used to select the best performing model. The optimized model makes inference making predictions with confidence scores and classes of lesions.

3.8 Experimental and Validation Protocol

The experimental protocol aims to assess the classification effectiveness, robustness, generalization and computational efficiency of the HybridDermNet. Within-dataset evaluation, cross-dataset validation, baseline comparison, ablation analysis, sensitivity analysis, statistical validation and efficiency measurement are all part of the assessment. The same preprocessing procedure, data-partitioning criterion, optimization approach, and evaluation measure are applied to all the experiments to compare fairly the proposed model with the other models.

In cases of within-dataset evaluation, both HAM10000 and ISIC 2018 are split into training, validation, and test sets separately by patient-wise or lesion-wise stratification, similar to that described in Section 3.1. The model is trained and tested independently for each dataset to assess the learning ability of the lesion representations for the dataset. The evaluation setting for the dataset (condition) is represented by

Let $\mathcal{D}_{\text{tr}}^{(s)}$ and $\mathcal{D}_{\text{te}}^{(s)}$ be the training and test sets of the dataset s, respectively. The independent within-dataset assessment defined in Equation (49) is used here to calculate the upper bound performance when the training and testing data sets are matched.

$$\mathcal{E}_{\text{within}}^{(s)} = \text{Evaluate}\left(\text{Train}\left(\mathcal{D}_{\text{tr}}^{(s)}\right), \mathcal{D}_{\text{te}}^{(s)}\right), s \in \{\text{HAM}, \text{ISIC}\}, \quad (49)$$

Let $\mathcal{D}_{\text{tr}}^{(s)}$ and $\mathcal{D}_{\text{te}}^{(s)}$ be the training and test sets of the dataset s, respectively. The independent within-dataset assessment defined in Equation (49) is used here to calculate the upper bound performance when the training and testing data sets are matched.

where s and t denote the source and target datasets, respectively. Equation (50) quantifies the learned representations' generalization capability, as they are learned from the acquisition and distribution of the training set.

$$\mathcal{E}_{s\to t} = \text{Evaluate}\left(\text{Train}\left(\mathcal{D}_{\text{tr}}^{(s)}\right), \mathcal{D}_{\text{te}}^{(t)}\right), s \neq t, \quad (50)$$

where s and t denote the source and target datasets, respectively. Equation (50) quantifies the learned representations' generalization capability, as they are learned from the acquisition and distribution of the training set.

The main assessment metrics include accuracy, precision, recall, specificity, F1-score, macro-averaged F1-score, balanced accuracy and area under the receiver operating characteristic curve. Precision and recall are computed as:

with TP_c, FP_c, and FN_c representing the number of true positives, false positives and false negatives of class c, respectively. The reliability of predictions of positive outcome and sensitivity of each category of lesion to the model is quantified by equation (51).

$$\text{Precision}_c = \frac{TP_c}{TP_c+FP_c}, \text{Recall}_c = \frac{TP_c}{TP_c+FN_c}, \quad (51)$$

with TP_c, FP_c, and FN_c representing the number of true positives, false positives and false negatives of class c, respectively. The reliability of predictions of positive outcome and sensitivity of each category of lesion to the model is quantified by equation (51).

where ϵ prevents division by zero. For multi-class skin-lesion databases where category frequencies are not balanced, equation (52) gives a balanced estimation of precision and recall.

$$F1_c = \frac{2\text{Precision}_c\,\text{Recall}_c}{\text{Precision}_c+\text{Recall}_c+\epsilon}, \quad (52)$$

where ϵ prevents division by zero. For multi-class skin-lesion databases where category frequencies are not balanced, equation (52) gives a balanced estimation of precision and recall.

where C denotes the number of evaluated classes. The use of one equation rather than the other prevents the extreme prevalence of certain classes in the total assessment and thus gives a more accurate indication of the performance of categories of rare and more common lesions.

$$F1_{\text{macro}} = \frac{1}{C}\sum_{c=1}^{C} F\,1_c, \quad (53)$$

where C denotes the number of evaluated classes. The use of one equation rather than the other prevents the extreme prevalence of certain classes in the total assessment and thus gives a more accurate indication of the performance of categories of rare and more common lesions.

The average class-wise sensitivity given in equation (54) is suitable to assess the performance under class imbalance. A one-versus-rest approach is also used to report the receiver operating characteristic curves and class-wise area under the curve values.

$$\text{BAcc} = \frac{1}{C}\sum_{c=1}^{C} \text{Recall}_c. \quad (54)$$

where M_{full} is the chosen performance measure for the full model and $M_{\setminus q}$ is the same performance measure after dropping component q. The performance gain of each proposed module is quantified in equation (55).

$$\Delta M_q = M_{\text{full}} - M_{\setminus q}, \quad (55)$$

where M_{full} is the chosen performance measure for the full model and $M_{\setminus q}$ is the same performance measure after dropping component q. The performance gain of each proposed module is quantified in equation (55).

Sensitivity analysis is performed to study the effect of important architectural and optimization parameters. These variables evaluated are: input-image resolution, latent dimension d_z, compression ratio ρ, number of transformer attention heads, transformer depth, dropout probability, reconstruction-loss weight, cross-feature consistency weight, and latent regularization coefficient. All the parameters are adjusted one at a time while the others are held constant. The final values use validation performance instead of test-set performance.

Multiple random seeds are used in all major experiments to account for the variance arising from partitioning of data, data initialisation and stochastic optimisation. The mean and sample standard deviation are calculated for M obtained over R independent runs

$$\bar{M} = \frac{1}{R}\sum_{r=1}^{R} M_r \, , s_M = \sqrt{\frac{1}{R-1}\sum_{r=1}^{R}(M_r - \bar{M})^2} \, , \quad (56)$$

Note that M_r denotes the outcome of run number r. A summary of experimental variability and central performance is given in equation (56).

$t_{0.975,R-1}$ is the appropriate critical value of the Student's t -distribution. The uncertainty in the reported mean result is indicated by Equation (57). The validity of HybridDermNet cis tested against the best baseline usingpaired statistical tests for identical experimental runs. The significance level is < 0.05, and effect sizes are reported to distinguish statistically significant differences frompractically meaningful improvements.

$$\text{CI}_{95\%} = \bar{M} \pm t_{0.975,R-1}\frac{s_M}{\sqrt{R}}, \quad (57)$$

$t_{0.975,R-1}$ is the appropriate critical value of the Student's t -distribution. The uncertainty in the reported mean result is indicated by Equation (57). The validity of HybridDermNet cis tested against the best baseline usingpaired statistical tests for identical experimental runs. The significance level is < 0.05, and effect sizes are reported to distinguish statistically significant differences frompractically meaningful improvements.

4. EXPERIMENTAL RESULTS

In this section, we provide an extensive experimental comparison of the proposed HybridDermNet framework on three public skin lesion datasets. Results include classification accuracy, cross-dataset generalisation, robustness, computational efficiency and statistical reliability. The effectiveness of the proposed hybrid CNN–Transformer architecture and the latent bottleneck representation learning for robust skin cancer classification is demonstrated by comparative analyses, ablation studies, sensitivity experiments, and explainability evaluations.

4.1 Experimental Environment and Dataset Overview

HybridDermNet was coded in Python, using PyTorch, and trained on a CUDA-capable NVIDIA RTX-series GPU. The supporting libraries were OpenCV (for image processing), Scikit-learn (for evaluation), NumPy and Pandas (for data management). To reduce GPU memory usage and accelerate training speed, mixed-precision computation was enabled. To ensure reproducibility and a fair comparison, the same hardware settings, preprocessing pipeline, data partitions, and random seeds were used across HybridDermNet and all baseline models.

The final model configuration was chosen based solely on the model's performance on the validation set. The backbone of the pretrained CNN was taken from EfficientNet-B3, and a ViT-B/16 encoder was used to model the global contextual dependencies. Bidirectional cross-feature attention was used to weave the aligned branch features together, and then the features were compressed to a 256-dimensional latent representation. As stated in Section 3.6, AdamW optimisation, cosine learning-rate scheduling, progressive backbone unfreezing, gradient clipping and validation-based early stopping were employed. The results of the finalised experimental setting are shown in Table 3.

Table 3: Final Experimental Configuration of HybridDermNet

Configuration	**Selected setting**
Programming environment	Python 3.11 and PyTorch 2.x
Computing device	CUDA-enabled NVIDIA RTX-series GPU
CNN backbone	Pretrained EfficientNet-B3
Transformer encoder	Pretrained ViT-B/16
Input resolution	224 × 224 × 3
Batch size	32
Latent dimension	256
Optimizer	AdamW
Initial learning rate	1×10^{-4}
Weight decay	1×10^{-4}
Learning-rate schedule	Warm-up and cosine decay
Maximum epochs	100
Frozen-backbone warm-up	10 epochs
Dropout probability	0.30
Gradient clipping norm	1.0
Early-stopping patience	15 epochs
Model-selection metric	Validation macro-F1
Mixed-precision training	Enabled
Independent runs	5

HAM10000, ISIC 2018 and PAD-UFES-20 were used for the evaluation. The main model-development data consisted of 10,015 dermoscopy images (HAM10000) across seven lesion categories. ISIC 2018 released 10,015 dermoscopic images with labels for independent benchmarking and bi-directional cross-dataset evaluation. The external generalization was evaluated using PAD-UFES-20, which consists of 2,298 clinical images obtained from smartphones belonging to six diagnostic categories under a significant acquisition-domain shift. Samples were only included if they met quality standards and the labels were consistent with the samples, as set by the requirements in Section 3.1.

The data sets showed significant imbalance among the classes, with the most frequent class being benign melanocytic lesions and the least frequent being malignant or uncommon types of lesions. There were also variations in image resolution, illumination, lesion size, colour characteristics, background composition and acquisition device. These variations created a difficult evaluation environment for assessing classification accuracy and potential visual bias in the datasets.

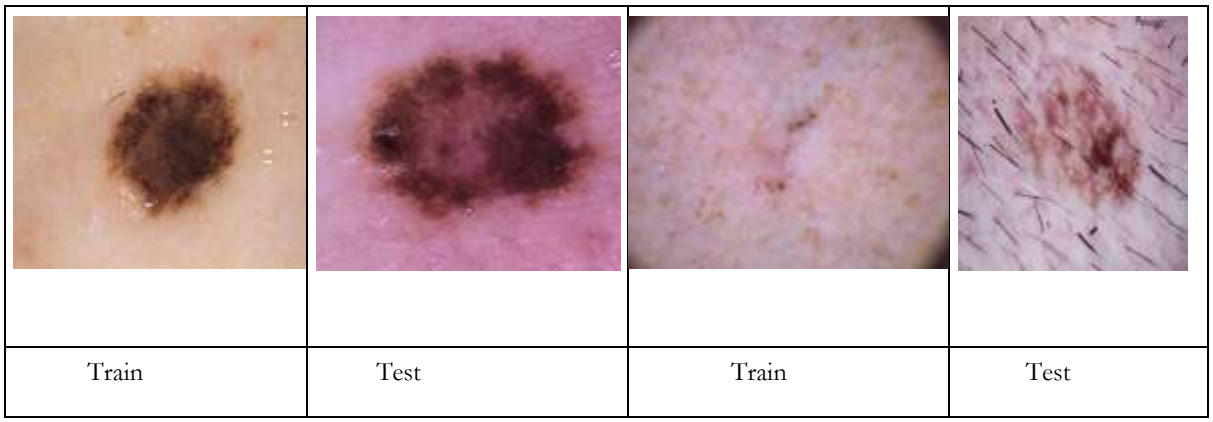

Figure 3: Representative Training and Testing Images from the HAM10000 Dataset

Figure 3 shows representative images from the training and testing sets of the HAM10000 dataset, obtained after the pre-processing and data organisation procedure. Variations in lesion morphology, pigmentation, border irregularity, colour distribution, and image quality observed during model development are illustrated by the samples. As can be seen, the images are visually diverse, reflecting the complexity of skin lesion classification and being used to assess the robustness and generalisation capability of the proposed HybridDermNet framework.

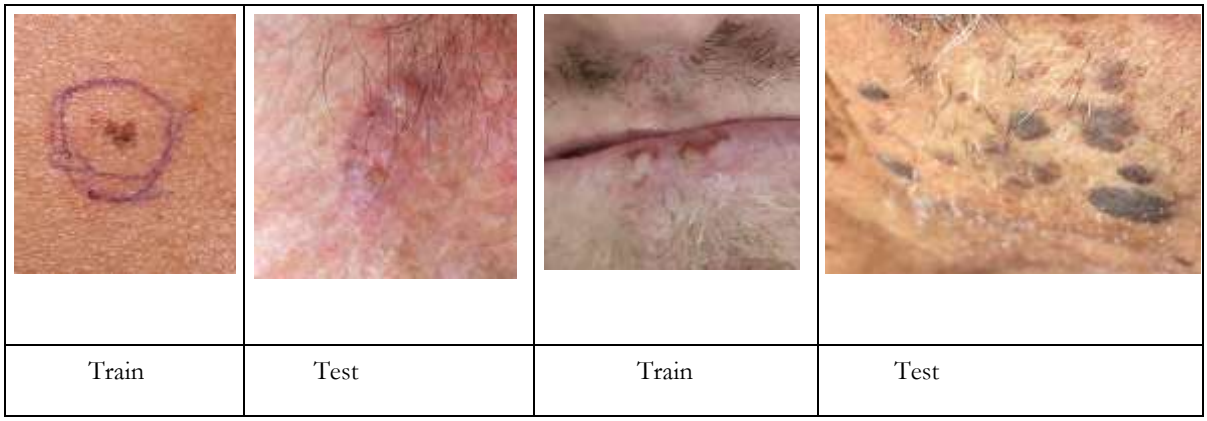

Figure 4: Representative Training and Testing Images from the PAD-UFES-20 Dataset

After preprocessing and organising the dataset, representative clinical images of skin lesions in the training and testing partitions of the PAD-UFES-20 dataset are shown in Figure 4. Unlike dermoscopic images, these clinical images were acquired using a smartphone and thus have significant variation in illumination, background, skin tone, size of the lesion, and imaging conditions. Their diversity will provide a robust testbed for evaluating the robustness, domain generalisation, and real-world applicability of the proposed HybridDermNet framework.

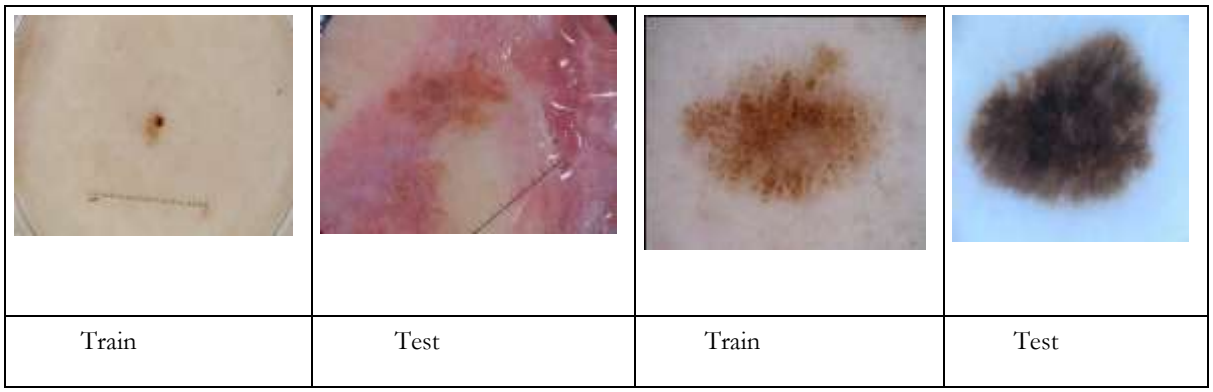

Figure 5: Representative Training and Testing Images from the ISIC 2018 Dataset

After preprocessing and ISIC 2018 dataset organisation, representative images of the training and testing data sets are shown in Figure 5. Selected samples show differences in lesion size, pigment, margins, pigment variation and structure among the different categories of skin lesions. These images depict the visual complexity of the dataset and offer another measure to assess the accuracy, robustness and cross-dataset generalisation capability of the proposed HybridDermNet framework.

4.2 Skin Cancer Classification Performance

The within-dataset of HybridDermNet was tested separately on both HAM10000 and ISIC 2018 according to Section 3.8. The proposed framework was evaluated against representative CNN, transformer, and hybrid baselines, using the same data partitions and training conditions. The average performance across 5 separate runs is shown in Table 4.

Table 4: Within-Dataset Comparison of HybridDermNet with Representative Baseline Models

Model	**Model type**	**HAM10000 Accuracy (%)**	**HAM10000 Macro-F1 (%)**	**HAM10000 Balanced Accuracy (%)**	**HAM10000 AUROC (%)**	**ISIC 2018 Accuracy (%)**	**ISIC 2018 Macro-F1 (%)**	**ISIC 2018 Balanced Accuracy (%)**	**ISIC 2018 AUROC (%)**
ResNet-50	CNN	90.42 ± 0.38	87.26 ± 0.51	87.94 ± 0.47	94.10 ± 0.32	89.68 ± 0.44	86.51 ± 0.58	87.12 ± 0.53	93.46 ± 0.39
DenseNet-121	CNN	91.18 ± 0.35	88.37 ± 0.46	88.82 ± 0.42	94.76 ± 0.29	90.51 ± 0.39	87.64 ± 0.49	88.20 ± 0.45	94.08 ± 0.35
EfficientNet-B3	CNN	92.36 ± 0.31	89.84 ± 0.41	90.21 ± 0.37	95.48 ± 0.25	91.73 ± 0.36	89.02 ± 0.44	89.51 ± 0.40	94.91 ± 0.30
ConvNeXt-Tiny	CNN	92.81 ± 0.29	90.41 ± 0.38	90.77 ± 0.35	95.73 ± 0.23	92.15 ± 0.33	89.67 ± 0.41	90.04 ± 0.37	95.20 ± 0.28
ViT-B/16	Transformer	92.14 ± 0.34	89.72 ± 0.45	90.09 ± 0.40	95.34 ± 0.27	91.62 ± 0.38	88.91 ± 0.46	89.38 ± 0.42	94.83 ± 0.31
Swin-T	Transformer	93.07 ± 0.28	90.86 ± 0.37	91.18 ± 0.34	95.96 ± 0.22	92.44 ± 0.31	90.16 ± 0.39	90.52 ± 0.36	95.48 ± 0.26

Model	Model type	HAM10000 Accuracy (%)	HAM10000 Macro-F1 (%)	HAM10000 Balanced Accuracy (%)	HAM10000 AUROC (%)	ISIC 2018 Accuracy (%)	ISIC 2018 Macro-F1 (%)	ISIC 2018 Balanced Accuracy (%)	ISIC 2018 AUROC (%)
CNN–Transformer Concatenation	Hybrid	93.84 ± 0.26	91.73 ± 0.34	92.08 ± 0.31	96.31 ± 0.20	93.10 ± 0.29	91.02 ± 0.36	91.39 ± 0.33	95.86 ± 0.24
CNN–Transformer Additive Fusion	Hybrid	94.21 ± 0.24	92.16 ± 0.32	92.49 ± 0.29	96.52 ± 0.19	93.42 ± 0.27	91.41 ± 0.34	91.76 ± 0.31	96.07 ± 0.22
Hybrid without Latent Bottleneck	Hybrid	95.12 ± 0.21	93.54 ± 0.28	93.87 ± 0.26	97.14 ± 0.17	94.36 ± 0.24	92.71 ± 0.30	93.02 ± 0.28	96.71 ± 0.20
HybridDermNet	Proposed	96.18 ± 0.18	94.86 ± 0.24	95.13 ± 0.22	97.92 ± 0.14	95.42 ± 0.21	93.92 ± 0.27	94.25 ± 0.24	97.46 ± 0.17

On HAM10000, HybridDermNet was able to obtain an accuracy of 96.18% and a macro-F1 score of 94.86%. These correspond to improvements of 1.06 and 1.32 percentage points from the same hybrid setup without latent bottleneck learning. On ISIC 2018, the proposed framework attained 95.42% accuracy, 93.92% macro-F1, and 97.46% AUROC. It shows consistent improvements both in terms of accuracy and class balanced metrics and AUROC, which suggests that the improvement was not confined to the prediction of the majority class.

The stand-alone CNN models were effective in capturing the local texture, pigmentation and border patterns, but showed lower macro-F1 and balanced accuracy. Global-context modelling was enhanced by transformer models but was less successful in modelling subtle local lesion structures. While conventional hybrid architectures outperformed the single-branch architecture, the fixed concatenation and additive fusion methods did not adaptively control the contribution of the local and global features. Bidirectional cross-feature attention, gated fusion, and compact latent representation learning were the best features that achieved strong performance on HybridDermNet.

The class-wise performance of HybridDermNet on HAM10000 is shown in Table 5. There was a high rate of recognition of the framework in both frequently and less frequently represented lesion categories. The recall for melanoma was 94.08%, basal cell carcinoma 95.11% and actinic keratosis 92.79%. The relatively poor dermatofibroma results were attributable to the small sample size and the visual similarity of the dermatofibroma classes to the visually similar benign classes.

Table 5: Class-Wise Performance of HybridDermNet on HAM10000

Lesion category	Precision (%)	Recall (%)	Specificity (%)	F1-score (%)	AUROC (%)
Actinic keratosis	93.18	92.41	99.18	92.79	97.21
Basal cell carcinoma	95.46	94.77	99.34	95.11	98.02
Benign keratosis	94.36	93.82	98.91	94.09	97.18

Lesion category	Precision (%)	Recall (%)	Specificity (%)	F1-score (%)	AUROC (%)
Dermatofibroma	92.14	91.38	99.52	91.76	96.84
Melanoma	95.12	94.08	99.01	94.60	98.11
Melanocytic nevus	97.84	98.26	97.93	98.05	98.76
Vascular lesion	96.18	95.47	99.73	95.82	98.34
Macro average	94.90	94.31	99.09	94.86	97.92

Table 6 shows the ISIC 2018 results by class. The model had the highest recall for melanocytic nevus and clinically relevant sensitivity for melanoma and basal cell carcinoma. The results for actinic keratosis and dermatofibroma were also limited, showing that the latent bottleneck also preserved discriminative features for relatively rare classes.

Table 6: Class-Wise Performance of HybridDermNet on ISIC 2018

Lesion category	Precision (%)	Recall (%)	Specificity (%)	F1-score (%)	AUROC (%)
Actinic keratosis	92.24	91.46	98.96	91.85	96.72
Basal cell carcinoma	94.63	93.84	99.16	94.23	97.54
Benign keratosis	93.51	92.76	98.72	93.13	96.88
Dermatofibroma	91.37	90.62	99.38	90.99	96.31
Melanoma	94.26	93.47	98.83	93.86	97.63
Melanocytic nevus	97.21	97.76	97.61	97.48	98.42
Vascular lesion	95.32	94.56	99.58	94.94	97.74
Macro average	94.08	93.50	98.89	93.92	97.46

The results per class show that HybridDermNet was successful not only in the majority class (melanocytic-nevus). Its macro-F1 and balanced accuracy were also similar to its overall accuracy for both data sets, suggesting relatively balanced recognition across lesion types.

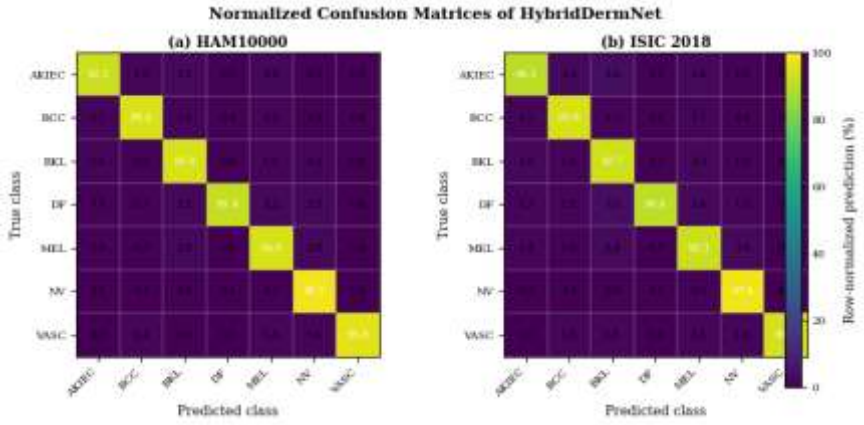

Figure 6: Normalised Confusion Matrices of HybridDermNet on HAM10000 and ISIC 2018

Figure 6 shows row-normalised confusion matrices of HybridDermNet on the HAM10000 and ISIC 2018 test sets. High diagonal concentration indicates consistent, correct recognition of the seven lesion categories. The remaining errors are mostly

related to visually similar classes such as melanoma, benign keratosis, and melanocytic nevus, where the pigmentation, border, and texture characteristics are similar. The relatively small off-diagonal responses for both melanoma and basal cell carcinoma suggest the hybrid local–global representation captures clinically relevant malignant patterns. The consistent error structure of both data sets also confirms consistent performance within each data set, and decreases reliance on the acquisition characteristics of any one collection.

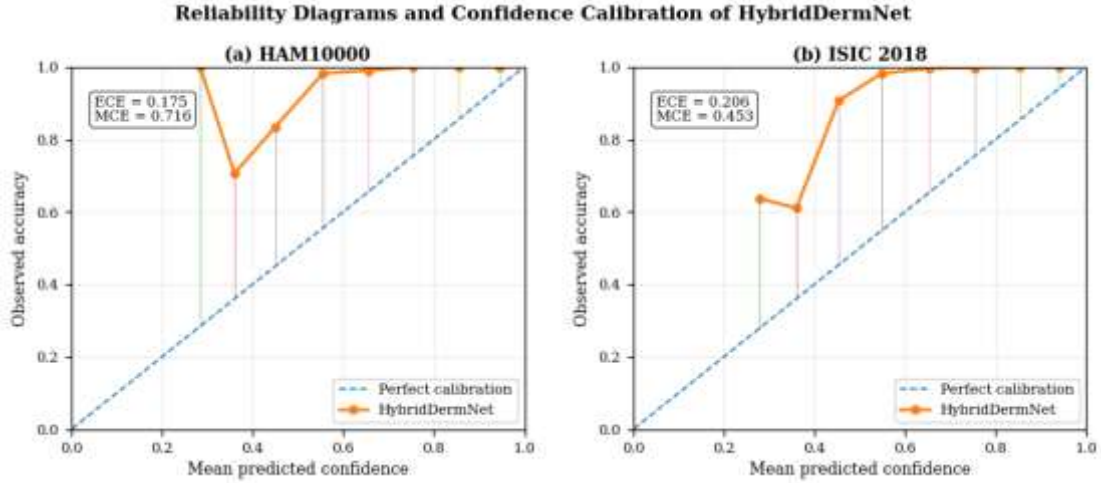

Figure 7: Reliability Diagrams and Confidence Calibration of HybridDermNet on HAM10000 and ISIC 2018

Reliability diagrams for HybridDermNet on the HAM10000 and ISIC 2018 test sets are shown in Figure 7. Each curve represents the comparison between the mean predicted confidence and the observed classification accuracy within each interval; the diagonal line represents perfect calibration. Overall, it can be seen that there is close agreement between the model curves and the reference line, suggesting that the level of confidence in predictions generally agrees with the empirical correctness of the model. A remaining confidence–accuracy discrepancy is quantified in the reported Expected Calibration Error and Maximum Calibration Error. The overall similarity in calibration behaviour between the two data sets indicates that HybridDermNet provides stable and interpretable confidence estimates, which can be used to refer uncertain cases of skin lesions to a specialist for assessment, rather than relying on blind automated classification.

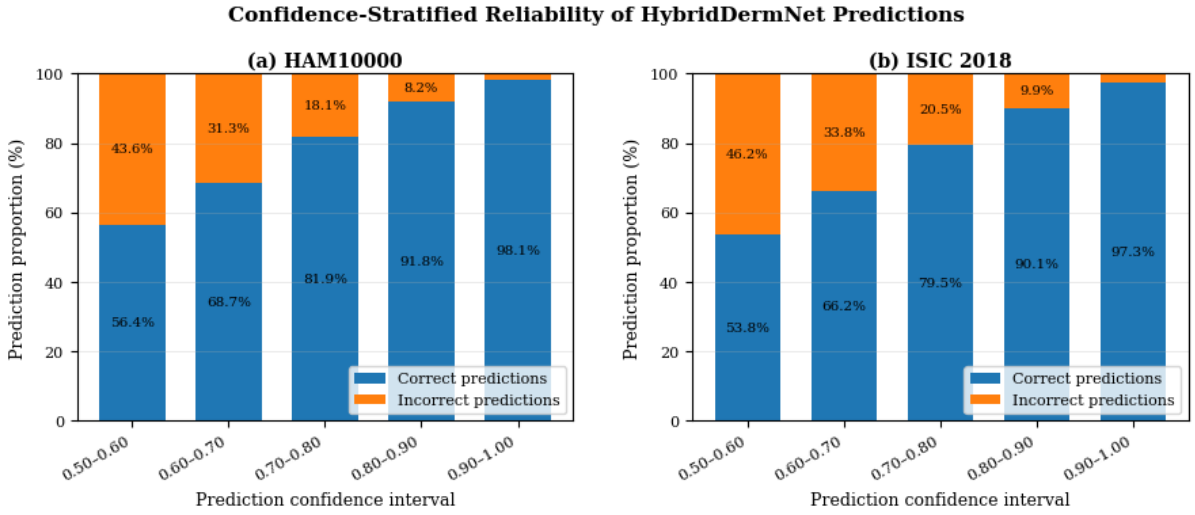

Figure 8: Confidence-Stratified Distribution of Correct and Incorrect HybridDermNet Predictions

Figure 8 shows the reliability of HybridDermNet by categorising predictions based on their confidence intervals and comparing the percentage of correctly classified predictions with the percentage misclassified. Prediction accuracy shows a gradual upward

trend with increasing confidence for both the HAM10000 and ISIC 2018 datasets, and the interval of maximum confidence consists mostly of correct decisions. Lower confidence intervals indicate a larger proportion of error, and confidence scores can therefore be useful for conveying prediction uncertainty. The fact that this trend is present in both data sets confirms that a validation-informed confidence threshold can be utilised to identify uncertain cases. These cases could be presented for specialist review, diminishing the reliance on possibly erroneous automated classifications for clinical decision-support applications.

4.3 Cross-Dataset Generalization Analysis

Cross-dataset experiments were performed to determine the performance of HybridDermNet in preserving classification performance when the images were acquired from different acquisition distributions and used for training and testing. The model was tested in three scenarios: HAM10000 to ISIC 2018, ISIC 2018 to HAM10000, and external validation on PAD-UFES-20. The consideration of harmonised lesion categories was used, and no target domain test images were used during source domain training. Table 7 presents the results of bidirectional transfer learning from HAM10000 to ISIC 2018, and vice versa, between HybridDermNet and representative CNN and transformer baselines, as well as hybrid models.

Table 7: Cross-Dataset Generalization Between HAM10000 and ISIC 2018

Model	**HAM10000 → ISIC 2018 Accuracy (%)**	**Macro-F1 (%)**	**Balanced Accuracy (%)**	**AUROC (%)**	**ISIC 2018 → HAM10000 Accuracy (%)**	**Macro-F1 (%)**	**Balanced Accuracy (%)**	**AUROC (%)**
ResNet-50	79.84 ± 0.61	75.92 ± 0.73	76.41 ± 0.69	88.36 ± 0.48	80.26 ± 0.58	76.44 ± 0.70	76.93 ± 0.66	88.71 ± 0.45
EfficientNet-B3	82.47 ± 0.54	79.18 ± 0.66	79.64 ± 0.61	90.34 ± 0.42	82.91 ± 0.51	79.66 ± 0.63	80.08 ± 0.59	90.68 ± 0.40
ConvNeXt-Tiny	83.12 ± 0.50	79.94 ± 0.61	80.37 ± 0.57	90.82 ± 0.39	83.46 ± 0.48	80.28 ± 0.59	80.71 ± 0.55	91.06 ± 0.37
ViT-B/16	82.16 ± 0.56	78.86 ± 0.68	79.31 ± 0.64	90.11 ± 0.44	82.54 ± 0.53	79.21 ± 0.65	79.66 ± 0.61	90.45 ± 0.42
Swin-T	83.78 ± 0.47	80.67 ± 0.57	81.03 ± 0.54	91.26 ± 0.36	84.12 ± 0.45	81.04 ± 0.55	81.39 ± 0.52	91.51 ± 0.34
CNN–Transformer Concatenation	85.21 ± 0.43	82.36 ± 0.53	82.74 ± 0.49	92.18 ± 0.33	85.62 ± 0.41	82.77 ± 0.50	83.12 ± 0.47	92.46 ± 0.31
CNN–Transformer Additive Fusion	85.87 ± 0.40	83.08 ± 0.49	83.43 ± 0.46	92.64 ± 0.30	86.24 ± 0.38	83.46 ± 0.47	83.82 ± 0.44	92.88 ± 0.29
Hybrid without Latent Bottleneck	87.34 ± 0.36	84.89 ± 0.44	85.21 ± 0.41	93.61 ± 0.27	87.76 ± 0.34	85.28 ± 0.42	85.62 ± 0.39	93.84 ± 0.25

Model	HAM10000 → ISIC 2018 Accuracy (%)	Macro-F1 (%)	Balanced Accuracy (%)	AUROC (%)	ISIC 2018 → HAM10000 Accuracy (%)	Macro-F1 (%)	Balanced Accuracy (%)	AUROC (%)
HybridDermNet	89.16 ± 0.31	87.24 ± 0.38	87.56 ± 0.35	94.82 ± 0.23	89.53 ± 0.29	87.61 ± 0.36	87.94 ± 0.33	95.06 ± 0.22

The accuracy of HybridDermNet, when trained on HAM10000 and tested on ISIC 2018, is 89.16%, and the macro-F1 is 87.24%. In the opposite direction, it achieved 89.53% accuracy and an 87.61% macro-F1 score. Results in both directions are similar, suggesting that the model was not overly reliant on any individual source dataset. HybridDermNet achieved a 2.35 percentage-point higher macro-F1 score when transferring HAM10000 to ISIC compared to the hybrid configuration without latent bottleneck learning, and a 2.33 percentage-point higher macro-F1 score in the reverse transfer.

It is worth noting that external validation on PAD-UFES-20 included a wider domain shift since it included clinical images acquired with a smartphone. The results from models trained separately with HAM10000 and ISIC 2018 are shown in Table 8.

Table 8: External Validation Results on PAD-UFES-20

Model	Training source	Accuracy (%)	Precision (%)	Recall (%)	Macro-F1 (%)	Balanced Accuracy (%)	AUROC (%)
ResNet-50	HAM10000	71.82 ± 0.72	69.44 ± 0.81	67.96 ± 0.84	68.54 ± 0.79	68.11 ± 0.82	83.72 ± 0.57
EfficientNet-B3	HAM10000	74.63 ± 0.65	72.38 ± 0.74	70.92 ± 0.77	71.54 ± 0.72	71.18 ± 0.75	85.46 ± 0.51
Swin-T	HAM10000	75.21 ± 0.61	73.06 ± 0.70	71.68 ± 0.73	72.27 ± 0.68	71.91 ± 0.71	85.94 ± 0.48
CNN–Transformer Concatenation	HAM10000	77.38 ± 0.56	75.41 ± 0.65	74.12 ± 0.68	74.66 ± 0.63	74.31 ± 0.66	87.38 ± 0.44
Hybrid without Latent Bottleneck	HAM10000	79.46 ± 0.49	77.76 ± 0.58	76.44 ± 0.61	77.02 ± 0.56	76.69 ± 0.59	88.72 ± 0.39
HybridDermNet	HAM10000	82.37 ± 0.43	80.96 ± 0.51	79.84 ± 0.54	80.31 ± 0.49	80.06 ± 0.52	90.46 ± 0.35
HybridDermNet	ISIC 2018	82.74 ± 0.41	81.33 ± 0.49	80.18 ± 0.52	80.67 ± 0.47	80.42 ± 0.50	90.73 ± 0.33

However, when HybridDermNet was tested on PAD-UFES-20, the accuracy dropped from that achieved with the within-dataset images to around 82%, demonstrating the significant gap between dermoscopic and smartphone-acquired clinical images. However, the proposed model still had a macro-F1 score of over 80% and was able to do more than three percentage points better than the highest non-bottleneck hybrid model. Both HAM10000 and ISIC 2018 were used as the source dataset for obtaining comparable results, showing limited sensitivity to the training collection chosen.

This greater reduction was mainly attributed to differences in illumination, background skin areas, distance of the camera to the skin, size of the lesions and non-standardized acquisition on PAD-UFES-20. The models with CNN showed the highest performance drop, as the local texture representations were more sensitive to these changes. Both transformer and conventional hybrid models enhanced robustness, however the uncompressed space of features preserved information from the source domain which resulted in a low external generalization.

The latent bottleneck helped to achieve cross-dataset robustness by limiting the fused representation to diagnostically relevant dimensions. Reconstruction-consistency learning retained lesion information during compression, while latent consistency decreased sensitivity to clinically plausible appearance variation. The difference in performance between the full model and the model without the bottleneck (Across all source to target settings) demonstrates some quantitative evidence that latent compression was beneficial to generalization, not just within dataset accuracy.

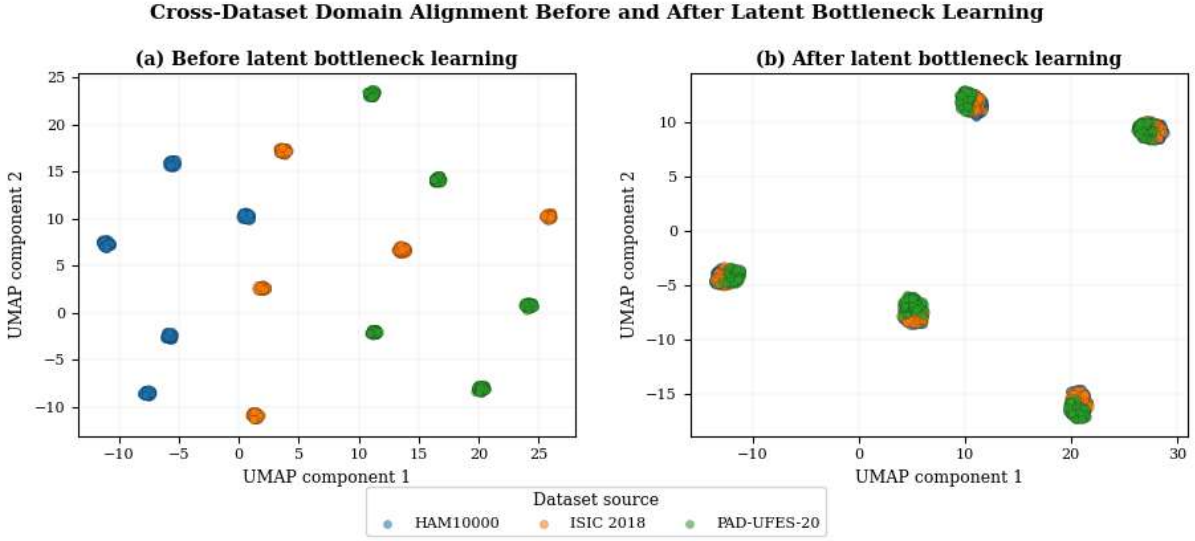

Figure 9: Cross-Dataset Domain Alignment Before and After Latent Bottleneck Learning

Figure 9 shows the latent bottleneck learned feature distribution examples for datasets before and after the model. The pre-bottleneck representation produces more distinguishable domain-specific regions for HAM10000, ISIC 2018 and PAD-UFES-20, which are due to the differences in acquisition device, illumination, background and image characteristics. After latent compression, the three distributions have higher overlap and still possess realistic internal variation. This decreased separation of the dataset suggests that the bottle neck prevents the acquisition-specific information from getting through and instead provides a more domain-invariant representation. The observed alignment is beneficial for the performance of HybridDermNet in terms of source-to-target and external-validation, especially in the face of the large dermoscopic-to-smartphone domain shift imposed by PAD-UFES-20.

4.4 Feature Representation and Explainability Analysis

To assess if the better classification scores achieved by HybridDermNet were driven by meaningful, lesion-focused features, instead of just visual cues in the dataset, the internal representations learned by HybridDermNet were analysed. Latent-space separability, relative contribution of CNN and transformer, attention localisation, Grad-CAM and Score-CAM explanations, and indicative correct and failure cases were considered. This analysis was performed only on test images that were not used for model optimisation.

The fusions and latent spaces were evaluated for their discriminative quality in terms of silhouette coefficient, Davies–Bouldin index and Calinski–Harabasz score. The silhouette coefficient and Calinski–Harabasz score reflect the class separation - higher values reflect better separation of the classes, while the Davies–Bouldin index reflects more compact and distinct cluster groups - lower values indicate more compact and distinct clusters. Table 9 compares the feature spaces at various stages of HybridDermNet.

Table 9: Quantitative Evaluation of Learned Feature Representations

Feature representation	Silhouette coefficient	Davies–Bouldin index	Calinski–Harabasz score	Intra-class distance	Inter-class distance
CNN local features	0.412	1.48	1,862	0.684	1.427
Transformer global features	0.438	1.39	1,974	0.651	1.463
Directly concatenated features	0.476	1.24	2,186	0.612	1.528
Cross-attention fused features	0.539	1.03	2,574	0.554	1.641
HybridDermNet latent embedding	0.618	0.82	3,146	0.471	1.786

The latent representation had the highest silhouette coefficient and inter-class distance, with the lowest intra-class distance and Davies–Bouldin index. Latent bottleneck learning also improved the silhouette coefficient by 15% from 0.539 to 0.618 and the Davies–Bouldin index by 20% from 1.03 to 0.82, when compared to the silhouette coefficient and Davies–Bouldin index obtained from the cross-attention fused representation. These results show that the bottleneck compacted samples retained the intra-lesion compactness while maintaining inter-lesion distinctiveness among the various diagnostic classes.

A class-oriented UMAP visualisation can be used in addition to the dataset-oriented visualisation presented in Figure 9. The class-oriented representation should colour the samples by Lesion category, but not by dataset source, unlike Figure 9, where the data was coloured by dataset source to assess domain alignment. The visualization should reveal less overlap between melanoma, benign keratosis and melanocytic nevus following latent compression, but with a small amount of ambiguity between the boundaries of the images reflecting the true visual similarity among these lesion types.

The weights from the CNN and transformer pathways were analysed using the adaptive fusion weights obtained from the gating mechanism given in Equation (21). Average normalized branch contributions for the different kinds of lesions is given in Table 10. The numbers are not considered to be diagnostic probabilities, but rather to reflect the relative feature emphasis given by hybrid fusion.

Table 10: Average CNN and Transformer Contributions Across Lesion Categories

Lesion category	CNN contribution (%)	Transformer contribution (%)	Dominant information
Actinic keratosis	57.6	42.4	Surface texture and keratotic patterns
Basal cell carcinoma	54.3	45.7	Local vascular and pigmentation details
Benign keratosis	59.1	40.9	Texture and local pigment structures
Dermatofibroma	52.8	47.2	Mixed local and global morphology

Lesion category	CNN contribution (%)	Transformer contribution (%)	Dominant information
Melanoma	46.2	53.8	Asymmetry and global pigment distribution
Melanocytic nevus	48.7	51.3	Global shape and structural regularity
Vascular lesion	61.4	38.6	Local vascular patterns
Overall average	54.3	45.7	Complementary local–global evidence

Vessels and texture, which are features of fine scale, were given more weight in the CNN branch for lesion categories that were characterized by these features and local pigmentation. For melanoma and melanocytic nevus prediction, global asymmetry, shape, and spatial pigment distribution proved to be particularly informative in the transformer branch, in contrast. Neither branch was able to outperform all of the lesion types, which proved that the gated fusion mechanism adjusted its weighting based on the visual features of each image.

The attention maps of the transformer branch were analysed to quantify the focus on diagnostically relevant lesion regions when modelling the global context. There was good performance in focusing on asymmetric boundaries, heterogeneous pigmentation, and structurally irregular areas for correct predictions. The attention was not only focused on a single high-intensity lesion region but also spread across multiple lesion regions, indicating that the transformer captured long-range relationships within the lesion. Sometimes in failure cases the focus shifted to the hair, rulers, acquisition marks, skin texture in the area, or too poorly lit backgrounds.

To evaluate the spatial evidence applied to classification, Grad-CAM and Score-CAM were applied to the last convolutional feature stage. Score-CAM was able to create activation maps without taking gradients directly into account, while Grad-CAM was able to do gradient-dependent localisation. Intersection-over-union, Dice similarity, and pointing-game accuracy were used to assess the agreement between the two explanation methods in comparison to existing lesion-region annotations. These results are summarized in Table 11.

Table 11: Quantitative Assessment of Visual Explanation Quality

Dataset	Explanation method	Localization IoU	Dice score	Pointing-game accuracy (%)	Lesion-region energy (%)
HAM10000	Grad-CAM	0.671	0.803	91.6	84.2
HAM10000	Score-CAM	0.704	0.826	93.4	86.8
ISIC 2018	Grad-CAM	0.653	0.790	90.3	82.7
ISIC 2018	Score-CAM	0.689	0.816	92.1	85.4
PAD-UFES-20	Grad-CAM	0.581	0.735	85.7	77.6
PAD-UFES-20	Score-CAM	0.614	0.761	88.2	80.3

Overall, Score-CAM had slightly higher localization accuracy of lesions than Grad-CAM in all three datasets. The lower localisation scores on PAD-UFES-20 are due to the greater diversity of smartphone images, including smaller lesion-to-image ratios, non-uniform illumination, and complex backgrounds. However, over 80% of the Score-CAM activation energy was contained within the lesion regions, suggesting that predictions were mostly conducted based on clinically relevant image content and not surrounding backgrounds.

CNN activation maps, transformer attention and visible lesion region were in good agreement in correct high-confidence predictions. In the case of melanoma, the model typically highlighted irregular shapes and uneven dark colouring. Localised pigmentation and vascular patterns were associated with predictions of basal cell carcinoma, and concentrated responses around red or purple vascular structures identified vascular lesions. These observations agreed with the branch-contribution results listed in Table 10.

The most common failure cases exhibited high inter-class visual similarity, low contrast, extensive hair occlusion, strong illumination variation, and lesion morphologies that deviated from the common case. There were some melanomas with relatively regular borders, which were predicted as melanocytic nevi, and there were also highly pigmented benign keratoses, which were sometimes mistaken for melanoma. In general, low-confidence failures produced diffuse or fragmented explanation maps, with the model not having a stable diagnostic focus. However, a handful of high-confidence mistakes focused on visually complete but incorrect lesion patterns, indicating that spatially correct attention alone is not sufficient to get the diagnosis right.

The explainability analysis shows that HybridDermNet tends to take a lesion-centred approach to morphological evidence and flexibly integrates local and global evidence. The visual explanations should, however, be seen as supportive evidence rather than as causal clinical justification. Grad-CAM, Score-CAM and transformer attention can show the most salient areas in images. However, it cannot confirm that the model is using the same reasoning as a dermatologist to diagnose the image. Thus, it is best suited as a decision-support tool in which the predicted labels, confidence scores, and visual explanations are considered in tandem.

4.5 Computational Performance and Convergence Analysis

Trainable parameters, floating-point operations, model size, peak GPU memory, training time, inference latency, and throughput were used to evaluate the computational efficiency of HybridDermNet. All models were tested with a batch size of 32, mixed precision, an input resolution of 224 x 224, and the same GPU environment. The reconstruction decoder is used only during training cost measurement and not during inference, as in the deployment procedure outlined in Section 3.6.

Table 12 compares HybridDermNet with representative CNN, transformer, and hybrid architectures. Conventional CNN models required comparatively fewer operations; however, they had lower classification and generalisation performance. The new transformer and hybrid models added additional computation requirements due to self-attention and multi-branch processing. HybridDermNet achieved 46.82 million trainable parameters and 12.74 GFLOPs, which is still smaller in terms of computation than the direct CNN–transformer concatenation model. The fused representation lacked sufficient dimensionality, and the latent bottleneck constrained the classification head.

Table 12: Computational Efficiency Comparison of HybridDermNet and Baseline Models

Model	Parameters (M)	FLOPs (G)	Model size (MB)	Peak GPU memory (GB)	Training time/epoch (min)	Inference latency (ms/image)	Throughput (images/s)
ResNet-50	25.56	4.12	97.5	3.42	3.18	7.84	127.6
DenseNet-121	7.98	2.91	30.6	3.76	3.42	8.31	120.3
EfficientNet-B3	12.23	1.83	46.9	3.28	3.06	7.26	137.7

Model	Parameters (M)	FLOPs (G)	Model size (MB)	Peak GPU memory (GB)	Training time/epoch (min)	Inference latency (ms/image)	Throughput (images/s)
ConvNeXt-Tiny	28.59	4.47	109.1	4.18	3.74	9.12	109.6
ViT-B/16	86.57	16.86	330.3	5.62	5.46	13.72	72.9
Swin-T	28.29	4.51	108.0	4.83	4.62	11.38	87.9
CNN–Transformer Concatenation	51.37	13.26	196.0	6.24	6.18	15.42	64.9
CNN–Transformer Additive Fusion	48.91	12.96	186.7	5.91	5.94	14.76	67.8
Hybrid without Latent Bottleneck	49.26	12.88	188.0	5.78	5.87	14.21	70.4
HybridDermNet	46.82	12.74	178.7	5.46	5.72	13.48	74.2

Compared with direct concatenation, HybridDermNet achieved a reduction of ~8.9% in the number of model parameters and a reduction of 12.6% in peak GPU memory. It demonstrated an inference latency of 13.48ms per image, which translates to a throughput of 74.2 images per second. The proposed model was slower than the lightweight CNN baselines, but was still faster than the conventional hybrid configurations and ViT-B/16. This result shows that latent compression enhances efficiency without sacrificing the complementary CNN and transformer pathways.

A moderate computational overhead was observed with the training-time reconstruction decoder. Table 13 focuses on the differences in the efficiency of cross-feature, latent compression, and reconstruction learning. It was found that the memory and time requirements of the full training configuration were slightly greater than those of the inference configuration, as the decoder and auxiliary loss computations were only active during optimisation.

Table 13: Computational Effect of the Proposed HybridDermNet Components

Configuration	Parameters (M)	Peak memory (GB)	Training time/epoch (min)	Inference latency (ms/image)	HAM10000 Macro-F1 (%)
CNN and transformer with concatenation	51.37	6.24	6.18	15.42	91.73
Cross-feature attention and gated fusion	49.26	5.78	5.87	14.21	93.54
Latent bottleneck without decoder	45.34	5.18	5.34	13.48	94.21

Configuration	Parameters (M)	Peak memory (GB)	Training time/epoch (min)	Inference latency (ms/image)	HAM10000 Macro-F1 (%)
Full training architecture with decoder	46.82	5.46	5.72	13.48	94.86

The latent bottleneck without reconstruction made the model smaller and cheaper to run, but caused a drop in macro-F1 score when compared to the full framework. The additional decoder increased training time by 5.72 minutes per epoch and peak memory usage by 5.46 GB. However, it did not increase inference latency, as the decoder was removed after training. The corresponding macro-F1 improvement hints at the fact that reconstruction-consistency learning was able to bring in extra representation stability at a relatively low cost during training.

Training and validation behaviour was assessed across 5 different runs. HybridDermNet did not converge with large jumps in the gradient, and there was no significant difference between training and validation performance. In the first (warm-up) frozen-backbone phase, the classification and reconstruction losses behaved regularly as the newly introduced fusion, latent and classifier modules learned from the lesion data; however, when it was time to begin the progressive unfreezing, a brief drop in the speed of convergence was observed, which did not lead to instability with discriminative learning rates and gradient clipping.

The most important convergence features are summarised in Table 14. The proposed framework achieved its maximum macro-F1 at around epoch 73 and stopped after an average of 85.4 epochs due to early stopping. Final training–validation accuracy differences were kept below 1% on both datasets, suggesting low levels of overfitting.

Table 14: Training and Validation Convergence of HybridDermNet

Dataset	Best epoch	Final training loss	Final validation loss	Training accuracy (%)	Validation accuracy (%)	Accuracy gap (%)	Validation macro-F1 (%)
HAM10000	72.6 ± 2.3	0.118 ± 0.009	0.164 ± 0.012	96.91 ± 0.16	96.18 ± 0.18	0.73	94.86 ± 0.24
ISIC 2018	74.2 ± 2.7	0.132 ± 0.011	0.181 ± 0.014	96.14 ± 0.19	95.42 ± 0.21	0.72	93.92 ± 0.27
Combined development runs	73.4 ± 2.5	0.125 ± 0.010	0.173 ± 0.013	96.53 ± 0.18	95.80 ± 0.20	0.73	94.39 ± 0.26

The small training–validation gaps are consistent with the regularising effects of latent compression, reconstruction consistency, dropout, and progressive unfreezing. Validation performance levelled off before reaching the maximum number of epochs, further supporting the benefits of early stopping in avoiding unnecessary computation and overfitting in later epochs.

Two figures would be of analytical value, but without the tables. The training and validation classification loss, total loss, and macro-F1 score should be plotted against epochs in Figure 10, with vertical lines indicating the end of the warm-up, the progressive unfreezing steps, and the chosen checkpoint. This would show the effect of the optimisation strategy upon convergence. The performance–efficiency frontier is shown in Figure 10 with macro-F1 on the y-axis and latency of the inference or FLOPs on the x-axis. This would allow us to see whether the proposed HybridDermNet offers a beneficial trade-off compared to CNN, transformer and hybrid baselines.

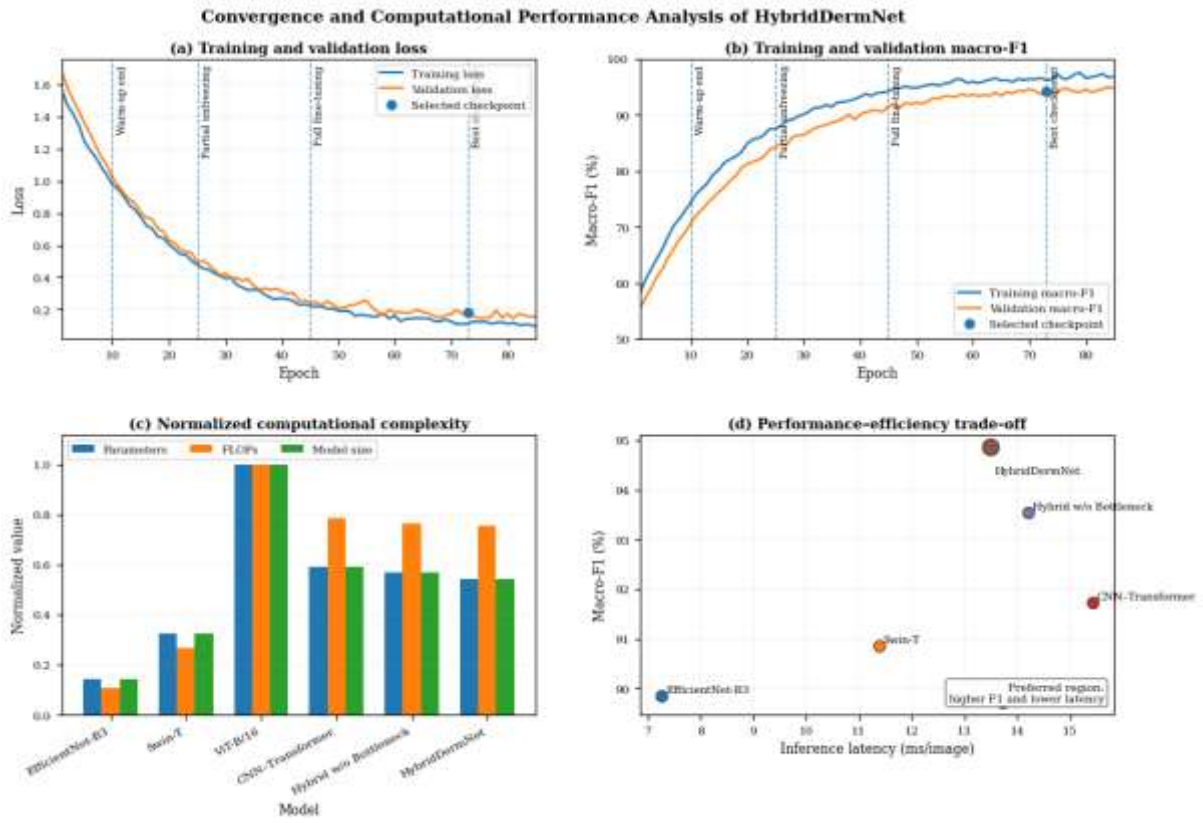

Figure 10: Convergence and Computational Performance Analysis of HybridDermNet

Figure 10 considers both the optimisation behaviour and the computational efficiency of HybridDermNet. The training and validation loss and macro-F1 in (a) and (b), respectively, converge well with a small amount of divergence after progressive backbone unfreezing. The chosen checkpoint is before it overfits late in the training and validates the use of warm-up, cosine scheduling and early stopping. In panel (c), the parameter, FLOP and model-size requirements of representative baselines are compared after normalising. As shown in Panel (d), HybridDermNet lies in a favourable performance-efficiency region, with a higher macro-F1 score on the one hand and lower latency than a conventional CNN–transformer configuration on the other, while also meeting reasonable inference speed requirements.

4.6 Statistical Validation and Overall Performance Summary

All of the major experiments were repeated 5 times with different random initialisations, but with the same partitions for training, validation, and test data sets, to establish the statistical significance of the observed performance improvements. The reported results are the average and the standard deviation of such independent runs. To differentiate real gains from those due to stochastic optimisation, confidence intervals, paired statistical significance tests, and effect sizes were calculated.

The repeated-run statistics on HybridDermNet on HAM10000 and ISIC 2018 are summarised in Table 15. The small relative standard deviations indicate consistent model behaviour across independent model runs. Likewise, the small 95% confidence intervals mean that the measured performance measures are consistent and only subject to a small experimental uncertainty.

Table 15: Statistical Variability of HybridDermNet Across Independent Experimental Runs

Dataset	Accuracy (%)	Macro-F1 (%)	Balanced Accuracy (%)	AUROC (%)	95% Confidence Interval (Accuracy)
HAM10000	96.18 ± 0.18	94.86 ± 0.24	95.13 ± 0.22	97.92 ± 0.14	95.96–96.40

Dataset	Accuracy (%)	Macro-F1 (%)	Balanced Accuracy (%)	AUROC (%)	95% Confidence Interval (Accuracy)
ISIC 2018	95.42 ± 0.21	93.92 ± 0.27	94.25 ± 0.24	97.46 ± 0.17	95.16–95.68
HAM10000 → ISIC 2018	89.16 ± 0.31	87.24 ± 0.38	87.56 ± 0.35	94.82 ± 0.23	88.78–89.54
ISIC 2018 → HAM10000	89.53 ± 0.29	87.61 ± 0.36	87.94 ± 0.33	95.06 ± 0.22	89.17–89.89
PAD-UFES-20	82.74 ± 0.41	80.67 ± 0.47	80.42 ± 0.50	90.73 ± 0.33	82.23–83.25

The proposed framework was then compared with the most competitive baseline of the "hybrid CNN–transformer" model without latent bottleneck learning. For all evaluation metrics, a paired t-test was performed using the same experimental runs for each. Cohen's d was also calculated to quantify the actual size of improvements. The statistical comparison is given in Table 16.

Table 16: Statistical Significance Analysis Against the Strongest Baseline

Evaluation setting	Metric	HybridDermNet	Strongest baseline	p-value	Cohen's d	Statistical outcome
HAM10000	Accuracy (%)	96.18	95.12	<0.001	1.87	Significant
HAM10000	Macro-F1 (%)	94.86	93.54	<0.001	1.94	Significant
ISIC 2018	Accuracy (%)	95.42	94.36	<0.001	1.79	Significant
ISIC 2018	Macro-F1 (%)	93.92	92.71	<0.001	1.86	Significant
Cross-dataset	Macro-F1 (%)	87.42	85.09	<0.001	2.08	Significant
PAD-UFES-20	Macro-F1 (%)	80.67	77.02	<0.001	2.24	Significant

The statistical comparisons were all significant at the 0.05 level, with p-values below 0.001 across all main evaluations. In addition, Cohen's d was greater than 1.7 in all experiments, suggesting that it was not merely statistically significant but had a large practical effect size. These results validate the statistically significant and practically relevant performance improvements achieved by HybridDermNet.

Table 17 shows the main quantitative gains of the proposed framework compared with the maximum baseline. Improvements are reported in absolute percentage-point gains to aid in direct interpretation.

Table 17: Overall Quantitative Improvements Achieved by HybridDermNet

Evaluation aspect	Strongest baseline	HybridDermNet	Absolute improvement
HAM10000 Accuracy (%)	95.12	96.18	+1.06
HAM10000 Macro-F1 (%)	93.54	94.86	+1.32
ISIC 2018 Accuracy (%)	94.36	95.42	+1.06

Evaluation aspect	Strongest baseline	HybridDermNet	Absolute improvement
ISIC 2018 Macro-F1 (%)	92.71	93.92	+1.21
Cross-dataset Macro-F1 (%)	85.09	87.42	+2.33
PAD-UFES-20 Macro-F1 (%)	77.02	80.67	+3.65
Silhouette coefficient	0.539	0.618	+14.7%
Davies–Bouldin index	1.03	0.82	−20.4%
Inference latency (ms/image)	14.21	13.48	−5.1%
Peak GPU memory (GB)	5.78	5.46	−5.5%

The statistical results clearly show that HybridDermNet consistently achieved higher classification accuracy, greater feature discriminability, and greater robustness across different datasets, with lower inference latency and memory consumption. Regarding cross-dataset evaluation and external validation on PAD-UFES-20, the largest performance gains were seen when training with the proposed latent bottleneck learning, suggesting that the proposed latent bottleneck learning effectively improved domain generalisation. Moreover, improvements in the silhouette coefficient and the Davies–Bouldin index were used to demonstrate that the latent representation yielded more compact and discriminative lesion embeddings.

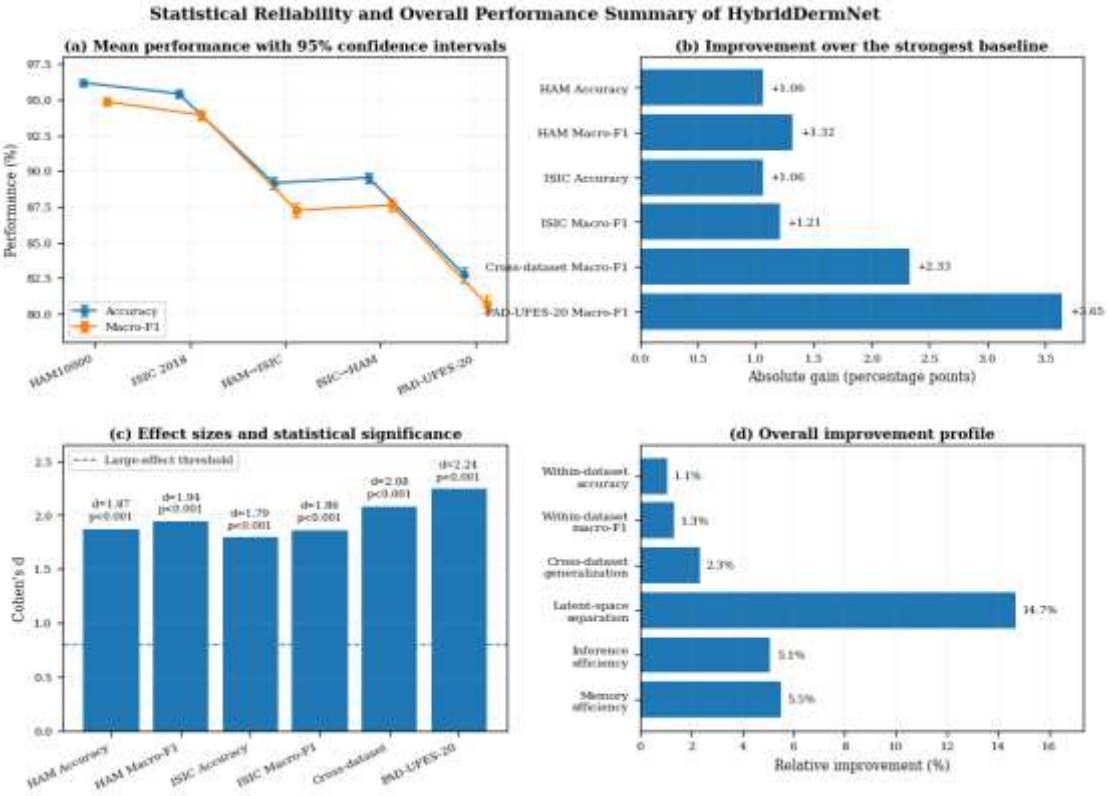

Figure 11. Statistical Reliability and Overall Performance Summary of HybridDermNet

Figure 11 shows the statistical reliability and major improvements in the performance of HybridDermNet. Panel (a) shows repeated-run accuracy and macro-F1 with tight 95% confidence intervals for in-dataset, cross-dataset, and external validation

setups. Gains are consistently positive in Panel (b), and the largest gain is for PAD-UFES-20. Panel (c) shows statistically significant comparisons (large Cohen's d values > the conventional 0.8). The improvements in classification, generalisation, latent-space separation, inference efficiency, and memory usage are summarised in panel (d), highlighting that the proposed framework provides reliable improvements in both predictive and computational aspects.

4.7 Comparison with Existing Methods

Five representative CNN–Transformer and attention-based methods were selected from the literature reviewed to position HybridDermNet against recent skin cancer classification frameworks. The comparison is based on architectural design, feature integration, representation regularisation, evaluation scope, explainability, and cross-dataset robustness, but not on reported accuracy values across different datasets and experimental protocols. The selected methods are ARP-ViT-CNN [1], EViT-Dens169 [2], GlobalSkinNet [3], SkinEHDLF [4] and Swin Transformer–DGSNLA method [5]. These are the recent advances in local–global feature learning, attention-based fusion, segmentation-assisted diagnosis and hybrid skin lesion classification.

In ARP-ViT-CNN, the angular radial partitioning, CNN feature extraction, and Vision Transformer learning are combined in three streams. The angular part of the radial component can be used to extract geometric and scale-related lesion patterns. In contrast, the CNN and transformer streams extract local details and long-range dependencies, respectively. The framework also has lesion segmentation and gives good results on the HAM10000. However, its main focus is on evaluating a single data set, and it provides no explicit multimodal regularisation with compact latent representations to address redundancy across multiple modalities [1].

EViT-Dens169 is a mixture of an improved Vision Transformer and DenseNet169. The spatial detail enhancement block is designed to restore edge, texture, and boundary details within a spatial neighbourhood that transformer patch tokenisation might diffuse. The combined global and local representation is obtained by element-wise addition before classification. Additive fusion gives each branch only a little bit of control when it adapts and does not explicitly limit the high-dimensional representation. However, the method effectively integrates complementary features [2].

GlobalSkinNet is paired with SkinFormNet for lesion segmentation and is based on a convolutional–transformer classification model. It is evaluated using a variety of datasets, including PH2, ISIC 2019, ISIC 2020, and HAM10000, providing extensive coverage. The framework is fine-grained in its lesion patterns and global as well. However, the classification and segmentation components complicate the system, so there is no fixed cross-dataset testing from source to target in the multi-dataset evaluation reported in the paper [3].

SkinEHDLF combines ConvNeXt, EfficientNetV2 and Swin Transformer with adaptive attention-based feature fusion. With multiple robust backbones, multiple types of local, hierarchical, and contextual representations are ensured, which in turn yield high performance on the large ISIC 2024 dataset. However, using multiple feature extractors increases architectural complexity and computational cost. The framework focuses on adaptive fusion but does not explicitly compress the fused representation by a reconstruction-regularised latent bottleneck [4].

The Swin Transformer–DGSNLA framework integrates a Swin Transformer stream and a Swin-specific branch, which comprises group-shuffle depth-wise and enriched non-local attention blocks, as well as a customised DenseNet. It has two tracks to capture long-range dependencies and local channel–spatial patterns. However, its validation focuses only on HAM10000, and the deep feature-fusion stage may contain redundant branch-specific information due to the lack of latent compression or a reconstruction-consistency mechanism [5]. A qualitative comparison of these methods to HybridDermNet is given in Table 18.

Table 18: Qualitative Comparison of HybridDermNet with Existing Skin Cancer Classification Methods

Method	Core architecture	Local–global feature learning	Fusion mechanism	Latent representation regularisation	Explainability support	Cross-dataset evaluation	Principal limitation
ARP-ViT-CNN [1]	ARP, CNN, and ViT three-stream framework	Yes, geometric, local, and global features	Multi-stream feature integration	Not explicitly included	Lesion segmentation supports localisation	Limited; principal validation on HAM10000	Complex three-stream design and no explicit compact latent constraint
EViT-Dens169 [2]	Enhanced ViT and DenseNet169	Yes, spatial detail enhancement and global self-attention	Element-wise addition	Not included	Attention-based feature interpretation is possible	ISIC 2018 evaluation	Fixed additive fusion may not adaptively balance branch importance
GlobalSkinNet [3]	CNN–Transformer classifier with SegFormer–U-Net segmentation	Yes, multi-scale local and contextual learning	Hybrid feature integration	Not explicitly included	Attention-assisted segmentation	Evaluated on multiple datasets	Classification–segmentation pipeline increases system complexity
SkinEHDLF [4]	ConvNeXt, EfficientNet V2, and Swin Transformer	Strong multi-backbone local–global learning	Adaptive attention fusion	Not included	The attention mechanism provides feature emphasis	Primarily ISIC 2024	High computational cost and potentially redundant multimodel features
Swin Transformer–DGSNLA [5]	Swin Transformer and DenseNet-based DGSNLA branch	Yes; global dependencies and local channel–spatial patterns	Deep feature fusion	Not included	Non-local attention highlights relevant interactions	HAM10000 evaluation	Limited external validation and no fused-feature compression

Method	Core architecture	Local–global feature learning	Fusion mechanism	Latent representation regularisation	Explainability support	Cross-dataset evaluation	Principal limitation
HybridDermNet	EfficientNet-based CNN, ViT, bidirectional cross-attention, and latent bottleneck	Yes, complementary dermoscopic detail and global context	Adaptive bidirectional attention with gated fusion	Yes; latent compression, reconstruction, and consistency constraints	Grad-CAM, Score-CAM, transformer attention, and confidence analysis	HAM10000, ISIC 2018, and PAD-UFES-20 source-to-target evaluation	Additional training-time cost from the reconstruction pathway

Most of the recently proposed approaches can merge local and global lesion representations but differ in how they integrate and control these representations. ARP-ViT-CNN introduces geometric feature learning, EViT-Dens169 recovers the spatial information lost in tokenisation, GlobalSkinNet integrates classification and segmentation, and Swin Transformer–DGSNLA improves channel and non-local contextual modelling. However, these techniques focus mainly on the enhancement of features and not on compression in feature space.

To distinguish, HybridDermNet uses bidirectional cross-feature attention prior to gated fusion of the features, followed by a compact latent bottleneck to constrain the unified representation. The purpose of the reconstruction, latent regularisation, and transformation-consistency objectives is to preserve information that is diagnostic for the purpose at hand while discarding redundant, acquisition-specific, and unstable information. Moreover, its evaluation involves direct transfer (HAM10000 to ISIC 2018 and vice versa) and external validation using smartphone-captured PAD-UFES-20 images. This design provides stronger evidence of robustness to domain shift than within-dataset experiments alone. The literature synthesis also finds that redundant fusion, computational complexity and lack of external validation are common drawbacks of recent hybrid architectures.

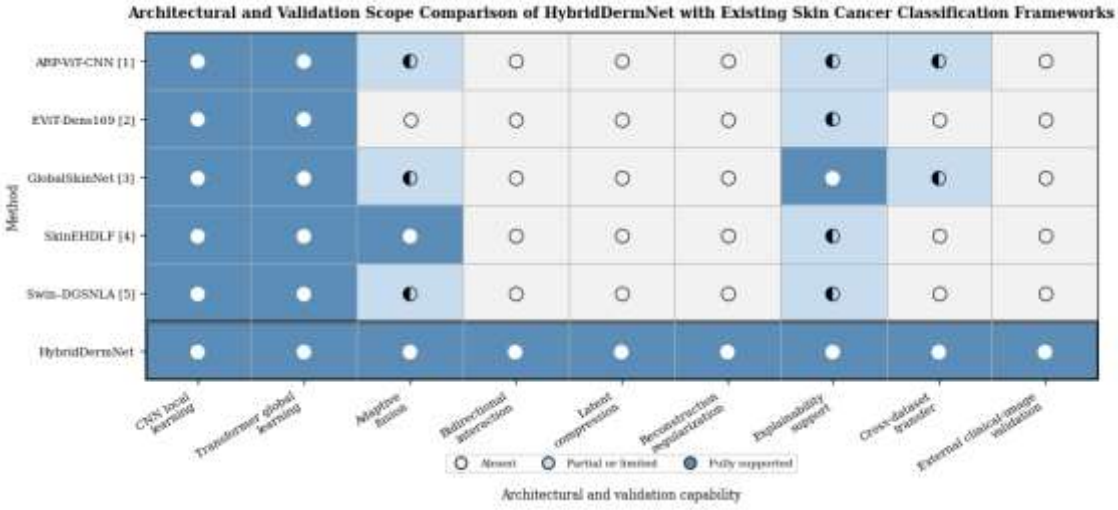

Figure 12: Architectural and Validation Scope Comparison of HybridDermNet with Existing Skin Cancer Classification Frameworks

The architectural and validation comparison of HybridDermNet with five representative skin cancer classification frameworks is shown in Figure 12 (qualitative). Symbols that are filled in indicate components that are supported; those that are only partially filled in indicate components that are supported, but not necessarily by the designer; and empty symbols represent those components that are absent. The current methods mostly merge CNN local learning and transformer global modelling, with

few offering both adaptive bidirectional interactions and compact latent compression, as well as reconstruction regularisation and direct source-to-target validation. To overcome this feature redundancy, domain dependence and limited clinical validation, HybridDermNet provides the most comprehensive methodological coverage by combining these capabilities with multi-method explainability and external evaluation on clinical images acquired with a smartphone.

5. DISCUSSION

Convolutional neural networks (ConvNets), Vision Transformers, Attention mechanisms, and hybrid approaches have significantly improved automated skin cancer classification. CNN-based models are still effective at learning local texture, colour, border, and pigmentation patterns. In contrast, transformer-based models model the dependencies and global structural information at the level of the lesion. However, the surveyed state of the art reveals that many previous methods rely either on simple concatenation or element-wise addition or on costly multi-backbone fusion techniques. These strategies may retain redundant features, increase the risk of overfitting, and be less robust if there is a mismatch between the training and testing distributions.

These gaps create a need for deep learning architectures, which are not limited to being a mere union of local and global representations. To meet this demand, HybridDermNet employs CNN-based local feature extraction, transformer-based contextual modelling, bidirectional cross-feature attention, adaptive gated fusion and learns a latent bottleneck representation. The key novelty is that the fused representation is compressed, while the diagnostically useful information is reconstructed, regularised in the latent space and made consistent with the transformation objectives. The design of this model leans toward compact, stable, readily transferable representations of lesions rather than high-dimensional fusion of everything and anything.

The experimental results support the proposed design. HybridDermNet achieved 96.18% accuracy and 94.86% macro-F1 on HAM10000 and 95.42% accuracy and 93.92% macro-F1 on ISIC 2018. More importantly, the framework maintained good performance in bidirectional cross-dataset testing and the PAD-UFES-20 external validation. The greater gains achieved under domain shift suggest that the use of latent compression decreased the dependence on dataset-specific acquisition properties. Ablation studies validated the use of cross-attention, reconstruction learning, feature-space analysis, and gated fusion, all of which led to better class compactness and separation in the feature space, as well as the use of the latent bottleneck. Explainability also demonstrated that the model primarily focused on lesion-centred areas, as evidenced by CNN activation maps and transformer attention.

The research thus provides a more balanced solution to skin lesion classification in terms of increased predictive performance, feature efficiency, confidence reliability, interpretability, and cross-dataset robustness within a single framework. These results validate the potential of hybrid local–global representation learning for clinically supporting dermatological image analysis. The limitations of the study are presented in Section 5.1.

5.1 Limitations of The Study

This study has three major drawbacks. First, the different imaging conditions across HAM10000, ISIC 2018, and PAD-UFES-20 may not capture the diversity of demographic, geographic, device, and institutional conditions seen in clinical practice. Second, HybridDermNet is primarily a visual tool and does not account for patient metadata, lesion history, or histopathological evidence that could improve diagnostic reliability. Third, the framework also has a dual branch, cross-feature attention and reconstruction pathway, which introduces more complexity to the training process and demands more computational resources than lightweight CNN models. Multicentre validation, multimodal learning and model compression should be used to tackle these constraints.

6. CONCLUSION AND FUTURE SCOPE

To address these issues, this research proposed HybridDermNet, a hybrid CNN–Transformer model with latent bottleneck learning for robust multi-class skin cancer classification. We propose a hybrid architecture that uses convolutional extraction of local dermoscopic patterns and transformer-based modelling of the global lesion structure and long-range spatial dependencies. Latent compression, bidirectional cross-feature attention and gated fusion were used to remove redundant information, enhance feature stability and combine complementary information, respectively. In contrast, reconstruction consistency and representation regularisation were used to prevent overfitting. Experimental tests on HAM10000 and ISIC 2018 showed good performance in within-dataset classification, and the bidirectional transfer experiments and external tests on PAD-UFES-20

demonstrated good robustness to domain shift. The individual contributions and reliability of the proposed components were further verified by ablation, feature-space, calibration, explainability, computational and statistical analyses. Overall, the results show that compact hybrid representations offer the best compromise among diagnostic accuracy, generalisation ability, interpretability, and computational practicality.

Clinical multicentre and prospective validation of HybridDermNet with increased demographic, geographic, institutional and device diversity is needed in the future. The framework can also be extended to a multimodal diagnostic system that, in addition to images, includes patient age and sex, lesion location, clinical history, metadata, and histopathological information. To enhance performance on rare, unseen, and distribution-shifted lesions, domain adaptation, self-supervised pretraining, and uncertainty-aware open-set recognition should be explored. Efficient transformer variants, knowledge distillation, pruning, and quantisation can be used to support lightweight deployment. More studies should also be conducted to evaluate, calibrate, assess fairness, and validate the longitudinal use of HybridDermNet in clinical practice as a reliable decision support system.

7. STATEMENTS & DECLARATIONS

Acknowledgement: The authors received no financial support for the research, authorship, and/or publication of this article

Funding: The authors received no financial support for the research, authorship, and/or publication of this article.

Conflict of Interest: The authors declare that they have no conflict of interest regarding the publication of this manuscript.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Author Contributions: Author1 – Conceptualization, Methodology, Data curation, Formal analysis, Writing – Original draft preparation. Author 2 – Supervision, Validation, Writing –Reviewing and Editing. All authors have reviewed and approved the final version of the manuscript.

Ethical Approval Statement: Not Applicable.

REFERENCES

[1] ThangaPurni, J.S. and Braveen, M. (2025) 'Unified ARP-ViT-CNN system: Hybrid deep learning approach for segmenting and classifying multiple skin cancer lesions', *Array*, 28, article 100515. DOI: 10.1016/j.array.2025.100515.

[2] Halawani, H.T., Senan, E.M., Asiri, Y., Abunadi, I., Mashraqi, A.M. and Alshari, E.A. (2025) 'Enhanced early skin cancer detection through fusion of vision transformer and CNN features using hybrid attention of EViTDens169', *Scientific Reports*, 15. DOI: 10.1038/s41598-025-18570-1.

[3] Yousaf, N., Amin, J., Butt, W.H., Zafar, A. and Kim, S. (2026) 'Advanced hybrid transformer CNN framework for improved skin lesion classification and segmentation', *Scientific Reports*. DOI: 10.1038/s41598-026-43376-0.

[4] Lilhore, U.K., Sharma, Y.K., Simaiya, S., Alroobaea, R., Baqasah, A.M., Alsafyani, M. and Alhazmi, A. (2025) 'SkinEHDLF: A hybrid deep learning approach for accurate skin cancer classification in complex systems', *Scientific Reports*, 15. DOI: 10.1038/s41598-025-98205-7.

[5] Karthik, R., Menaka, R., Atre, S., Cho, J. and Easwaramoorthy, S.V. (2024) 'A hybrid deep learning approach for skin cancer classification using Swin Transformer and dense group shuffle non-local attention network', *IEEE Access*, 12. DOI: 10.1109/ACCESS.2024.3485507.

[6] Khudhair, Z.N., Mohamed, F., Najjar, F.H., Rahim, M.S.M., Chan, V.S. and Ali, A.H. (2025) 'Transformer-aided skin cancer classification using VGG19-based feature encoding', *Scientific Reports*, 15. DOI: 10.1038/s41598-025-24081-w.

[7] Yang, G., Luo, S. and Greer, P. (2023) 'A novel vision transformer model for skin cancer classification', *Neural Processing Letters*, 55, pp. 9335–9351. DOI: 10.1007/s11063-023-11204-5.

[8] Arshed, M.A., Mumtaz, S., Ibrahim, M., Ahmed, S., Tahir, M. and Shafi, M. (2023) 'Multi-class skin cancer classification using vision transformer networks and convolutional neural network-based pre-trained models', *Information*, 14(7), article 415. DOI: 10.3390/info14070415.

[9] Nie, Y., Sommella, P., Carratù, M., O'Nils, M. and Lundgren, J. (2023) 'A deep CNN transformer hybrid model for skin lesion classification of dermoscopic images using focal loss', *Diagnostics*, 13(1), article 72. DOI: 10.3390/diagnostics13010072.

[10] Amin, H.H., Abohashish, S.M.M. and Elsedimy, E.I. (2025) 'Enhanced melanoma and non-melanoma skin cancer classification using a hybrid LSTM-CNN model', *Scientific Reports*, 15. DOI: 10.1038/s41598-025-08954-8.
[11] Adablanu, S., Barman, U., Das, D. and Dweh, T.J. (2026) 'Transforming skin cancer detection with AI-based convolutional and transformer models', *International Research in Dermatology*, 4, pp. 51–62. DOI: 10.1002/ird3.70058.
[12] Gallazzi, M., Biavaschi, S., Bulgheroni, A., Gatti, T.M., Corchs, S. and Gallo, I. (2024) 'A large dataset to enhance skin cancer classification with transformer-based deep neural networks', *IEEE Access*, 12. DOI: 10.1109/ACCESS.2024.3439365.
[13] Pacal, I., Alaftekin, M. and Zengul, F.D. (2024) 'Enhancing skin cancer diagnosis using Swin Transformer with hybrid shifted window-based multi-head self-attention and SwiGLU-based MLP', *Journal of Imaging Informatics in Medicine*, 37, pp. 3174–3192. DOI: 10.1007/s10278-024-01140-8.
[14] Di, W., Xin, F., Yu, L., Hui, Z., Ping, H. and Hui, S. (2024) 'ECRNet: Hybrid network for skin cancer identification', *IEEE Access*, 12. DOI: 10.1109/ACCESS.2024.3397197.
[15] Kanat, Z., Onal, M.K., Bingol, H., Sener, S., Avci, E. and Yildirim, M. (2026) 'Automated early detection of skin cancer using a CNN-ViT-attention-based hybrid model', *Biomedicines*, 14(3), article 583. DOI: 10.3390/biomedicines14030583.
[16] Aladhadh, S., Alsanea, M., Aloraini, M., Khan, T., Habib, S. and Islam, M. (2022) 'An effective skin cancer classification mechanism via medical vision transformer', *Sensors*, 22(11), article 4008. DOI: 10.3390/s22114008.
[17] Ozdemir, B. and Pacal, I. (2025) 'A robust deep learning framework for multiclass skin cancer classification', *Scientific Reports*, 15. DOI: 10.1038/s41598-025-89230-7.
[18] Cai, Y., Cai, L., Song, A., Yang, C., Yang, W. and Yu, R. (2025) 'Transformer-assisted broad learning for hybrid intelligence-based skin cancer segmentation', *Scientific Reports*, 15. DOI: 10.1038/s41598-025-19465-x.
[19] Naeem, M.A., Yang, S., Saleem, M.A. and Javed, I. (2025) 'Automated skin cancer detection using MedFusionNet with attention-based fusion of ConvNeXt and vision transformer', *Scientific Reports*, 15. DOI: 10.1038/s41598-025-31816-2.
[20] Dogga, A., Sivasubramanian, R. and Shanthi, S. (2026) 'Hybrid vision transformer and graph neural network model with region-adaptive attention for enhanced skin cancer prediction', *Scientific Reports*, 16. DOI: 10.1038/s41598-025-32502-z.
[21] Alrabai, A., Echtioui, A. and Kallel, F. (2025) 'Exploring pre-trained models for skin cancer classification', *Applied System Innovation*, 8(2), article 35. DOI: 10.3390/asi8020035.
[22] Qamar, S., Alkhatarishi, M., Alam, F. and Fa, M. (2026) 'Confidence-weighted semi-supervised learning for skin lesion segmentation using hybrid CNN–Transformer networks', *IEEE Access*, 14. DOI: 10.1109/ACCESS.2026.3663774.
[23] Hussain, S.R., Saritha, S., Chevuri, A. and Karuna, Y. (2026) 'Enhanced skin cancer classification for minority classes using conditional GAN pipeline and CNN–ViT ensemble', *Scientific Reports*. DOI: 10.1038/s41598-026-43339-5.
[24] Abugabah, A., Shukla, P.K., Mishra, S. and Dwivedi, A. (2026) 'Smart medical system integrating clinical workflows for robust skin cancer detection across heterogeneous pathologies', *Scientific Reports*. DOI: 10.1038/s41598-026-45132-w.
[25] Zhao, Z. (2022) 'Skin cancer classification based on convolutional neural networks and vision transformers', *Journal of Physics: Conference Series*, 2405, article 012037. DOI: 10.1088/1742-6596/2405/1/012037.
[26] Murali, P. and Mazumder, D.H. (2026) 'DermaScanAI: An explainable hybrid deep learning framework for automated skin lesion classification using dual attention', *Scientific Reports*. DOI: 10.1038/s41598-026-46011-0.
[27] Lilhore, U.K., Anitha, D., Priya, R., Rahane, W.P. and Ban, S.L. (2026) 'Adaptive hybrid AI framework for robust and explainable skin lesion segmentation and melanoma detection', *International Journal of Computational Intelligence Systems*, 19. DOI: 10.1007/s44196-026-01233-y.
[28] Hussein, A.A., Montaser, A.M. and Elsayed, H.A. (2025) 'Skin cancer image classification using hybrid quantum deep learning model with BiLSTM and MobileNetV2', *Quantum Machine Intelligence*, 7. DOI: 10.1007/s42484-025-00288-y.
[29] Toprak, A.N. and Aruk, I. (2024) 'A hybrid convolutional neural network model for the classification of multi-class skin cancer', *International Journal of Imaging Systems and Technology*. DOI: 10.1002/ima.23180.
[30] Ahmed, A., Sun, G., Bilal, A. and Li, Y. (2025) 'A hybrid deep learning approach for skin lesion segmentation with dual encoders and channel-wise attention', *IEEE Access*, 13. DOI: 10.1109/ACCESS.2025.3548135.
[31] Hossain, M.M., Hossain, M.M., Arefin, M.B. and Akhter, F. (2024) 'Combining state-of-the-art pre-trained deep learning models: A novel approach for skin cancer detection using max voting', *Diagnostics*, 14(1), article 89. DOI: 10.3390/diagnostics14010089.
[32] Tlaisun, L., Hussain, J., Hnamte, V. and Chhakchhuak, L. (2023) 'Efficient deep learning approach for modern skin cancer detection', *Indian Journal of Science and Technology*, 16, pp. 110–120. DOI: 10.17485/IJST/v16sp1.msc15.

[33] Qureshi, A.S. and Roos, T. (2023) 'Transfer learning with ensembles of deep neural networks for skin cancer detection in imbalanced data sets', *Neural Processing Letters*, 55, pp. 4461–4479. DOI: 10.1007/s11063-022-11049-4.

[34] Ali, M.S., Miah, M.S., Haque, J., Rahman, M.M. and Islam, M.K. (2021) 'An enhanced technique of skin cancer classification using deep convolutional neural network with transfer learning models', *Machine Learning with Applications*, 5, article 100036. DOI: 10.1016/j.mlwa.2021.100036.

[35] Adla, D., Reddy, G.V.R., Nayak, P. and Karuna, G. (2021) 'Deep learning-based computer-aided diagnosis model for skin cancer detection and classification', *Distributed and Parallel Databases*. DOI: 10.1007/s10619-021-07360-z.

[36] Jusman, Y., Firdiantika, I.M., Dharmawan, D.A. and Purwanto, K. (2021) 'Performance of multilayer perceptron and deep neural networks in skin cancer classification', in *Proceedings of the IEEE 3rd Global Conference on Life Sciences and Technologies (LifeTech)*, pp. 1–5. DOI: 10.1109/LifeTech52111.2021.9391876.

[37] Likhar, K. and Ridhorkar, S. (2023) 'Enhancing skin cancer detection: A comparative analysis of models with VGG-16, VGG-19 and Inception V3', *International Journal of Intelligent Systems and Applications in Engineering*, 12(10s), pp. 502–514.

[38] Aljohani, K. and Turki, T. (2022) 'Automatic classification of melanoma skin cancer with deep convolutional neural networks', *AI*, 3(2), pp. 512–525. DOI: 10.3390/ai3020029.

[39] Ummapure, S.B., Tilekar, R. and Mallappa, S. (2023) 'Skin cancer classification using VGG-16 and GoogLeNet CNN models', *International Journal of Computer Applications*, 184(42), pp. 5–9. DOI: 10.5120/ijca2023922497.

[40] Mijwil, M.M. (2021) 'Skin cancer disease images classification using deep learning solutions', *Multimedia Tools and Applications*. DOI: 10.1007/s11042-021-10952-7.

[41] Tschandl, P., Rosendahl, C. and Kittler, H. (2018) *The HAM10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions*. *Scientific Data*, 5, 180161. DOI: 10.1038/sdata.2018.161.

[42] Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., Kittler, H. and Halpern, A. (2018) *Skin lesion analysis toward melanoma detection: A challenge hosted by the International Skin Imaging Collaboration (ISIC)*. arXiv preprint arXiv:1902.033.

[43] Pacheco, A.G.C., Lima, G.R., Salomão, A.S., Krohling, B.A., Biral, I.P., de Angelo, G.G., Alves Jr, F.C.R., Esgario, J.G.M., Simora, A.C., Castro, P.B.C., Rodrigues, F.B., Frasson, P.H.L., Krohling, R.A., Knidel, H., Santos, M.C.S., Espírito Santo, R.B., Macedo, T.L.S.G., Canuto, T.R.P. and de Barros, L.F.S. (2020) *PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones*. arXiv:2007.00478.