International Journal of Innovative Technology and Interdisciplinary Sciences
https://journals.tultech.eu/index.php/ijitis
ISSN: 2613-7305
Volume 9, Issue 4
DOI: https://doi.org/10.15157/IJITIS.2026.9.4.2379-2435
Received: 22.12.2025; Revised: 30.09.2026; Accepted: 12.11.2026

Research Article

EResViTNet: Deep Feature Fusion of ResNet and Vision Transformer for Enhanced Classification of Skin Cancer Images

Radhika Takkella* , **Pagadala Pavan Kumar**
Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Hyderabad-500075, Telangana, India.
***ppagadala125@gmail.com**

Abstract

Skin cancer is among the most common malignancies, and early diagnosis is essential for improving treatment outcomes. However, the visual assessment of dermoscopic images remains subjective and dependent on dermatological expertise. This study proposes EResViTNet, a hybrid deep learning architecture that combines ResNet-50 for local feature extraction with ViT-B/16 for global contextual representation. The resulting feature descriptors are concatenated into a 2816-dimensional vector and subsequently refined using a lightweight channel and spatial attention module, enabling feature recalibration across the joint local and global representation. Image preprocessing includes Error Level Analysis (ELA), resizing, and data augmentation. Evaluated on the nine-class version of the ISIC 2019 dataset, EResViTNet achieved a training accuracy of 98.11% and a held-out test accuracy of 97.38%, averaged across five independent runs, with macro-averaged precision, recall, and F1-score of 97.0%, 98.0%, and 97.0%, respectively. Controlled ablation experiments demonstrated improvements over individual backbones, alternative feature-fusion strategies, and SE, ECA, and CBAM attention mechanisms. The improvements over individual backbones were associated with large effect sizes (Hedges' $g > 5$). An information-theoretic analysis further examined the complementarity of the learned representations and their relationship to low-rank multimodal fusion. Additional calibration, robustness, and cross-dataset experiments provided further evidence of model stability. PAD-UFES-20 was used as the fully external evaluation dataset, whereas HAM10000 was excluded from external validation because of its overlap with the ISIC 2019 training sources. Grad-CAM visualizations and patch-attention maps provided qualitative insights into the image regions contributing to model predictions. Overall, EResViTNet demonstrates potential as a decision-support approach for dermoscopic image classification, although its clinical utility requires further prospective evaluation involving dermatologists.

Keywords: Vision Transformer; ResNet-50; Feature Fusion; Attention Mechanism; Skin Cancer.

INTRODUCTION

Skin cancer, including melanoma, arises from abnormal growth of cells in the outer layer of the skin, often driven by genetic damage [1]. Although melanoma is less common

©2026 Authors. This is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution 4.0 International License CC BY 4.0 (http://creativecommons.org/licenses/by/4.0).

than other skin cancers, it causes the largest number of skin cancer deaths [2]. Basal cell carcinoma and squamous cell carcinoma are more frequent but have lower mortality [3].

The incidence of skin cancer, and of melanoma in particular, continues to rise, with more than 10 million new cases reported in Western countries [4]. Five-year survival exceeds 99% when melanoma is detected early but falls to about 14% when it is detected late [5]. Excessive sun exposure can damage melanocytes, the cells that produce the pigment melanin, and some lesions become malignant if they are not identified in time [2]. Dermoscopy is the standard non-invasive technique for examining skin lesions, but manual interpretation is subjective, time-consuming and dependent on dermatological expertise [3]. Automated, reliable and interpretable diagnostic tools based on deep learning are therefore needed. In this study, the proposed classifier is developed and evaluated on ISIC 2019, a large collection of annotated dermoscopic images [6-8].

The shortage of dermatology specialists relative to the number of patients motivates automated systems for early lesion classification [9]. Diagnosis relies on the colour, texture and structure of a lesion, which can be analysed with computer vision and machine learning [10]. Convolutional neural networks (CNNs) have matched or exceeded dermatologists in several lesion classification tasks [11]. Training deep models from scratch is difficult with limited medical data, so transfer learning is widely used [3]. Despite their accuracy, many deep models lack interpretability, which limits their clinical use [5].

Deep learning can process high-dimensional image data and identify the patterns needed to categorize skin lesions, and CNNs capture the spatial information required to recognize melanoma and other abnormalities [9]. CNN models such as Inception V3, DenseNet, MobileNet and EfficientNet have been applied to this task [12-14], but they mainly extract local features and capture long-range relationships and global context poorly. Recent work therefore combines CNNs with Vision Transformers (ViTs) [15] to add global attention [6][9]. These hybrids still face inefficient feature fusion, limited interpretability and overfitting on imbalanced datasets such as ISIC 2019 [16].

CNNs are widely used for skin cancer classification because they learn complex image features and transfer well from large natural-image datasets. Architectures such as ResNet and VGGNet are common choices [17], and ensemble and multi-task transfer learning further improve performance. Several CNNs for melanoma classification reach mean AUCs above 0.90 on ISIC 2019 [18]. Generative adversarial networks (GANs) and recurrent neural networks (RNNs) have also been used, GANs to synthesize lesion images for augmentation and RNNs to model sequential dependencies [19]. With VGGNet, 198 DermQuest classes were classified with 50.27% accuracy [20], and VGGNet was also trained to separate benign from malignant tumours on the ISIC 2016 Challenge dataset [21]. On skin cancer images, DenseNet-201 and Xception were more accurate than VGG and ResNet-50 [22].

The Study in [23] classified the 10,015 dermoscopic images of HAM10000 as benign or malignant using several Vision Transformer variants after segmentation with the Segment Anything Model (IoU 96.01%, Dice 98.14%), and ViT-Google with 32 × 32 patches achieved

the highest accuracy (96.15%) and the lowest false negative rate. An explainable approach based on pretrained CNNs (VGG19, ResNet18, ResNet50, MobileNetV2 and DenseNet201) found ResNet50 to be the most accurate (88.8%) and used XAI techniques to interpret its predictions [24]. SkinTrans reached 94.3% accuracy on HAM10000 and was proposed for multimodal medical imaging and early diagnosis [25]. DeepMelaNet distinguishes benign from malignant lesions and detects early melanoma with 93.40% accuracy at low cost [26]. DermVisD reported that ViTs outperform CNNs for skin lesion diagnosis, with attention and global context raising accuracy by 18% to 97.8%. In a comparison of CNN, ResNet-18 and EfficientNet-B0 for melanoma detection, EfficientNet-B0 performed best, with 95% accuracy, 98% sensitivity and 94% precision [27]. Grad-CAM visualization [28, 29] identifies the image regions that drive such predictions and can assist dermatologists, and the ability of deep models to learn from large datasets and adapt to varied lesions explains their advantage in automated skin cancer classification [3].

Previous skin lesion studies have used either CNNs or Vision Transformers, and a few recent studies combine the two in sequential or ensemble designs. Most of these methods do not explicitly filter the combined representation before classification. EResViTNet addresses this gap by combining ResNet-50 for local feature extraction with ViT-B/16 for global context and refining the fused representation with attention over informative channels and spatial regions of the dermoscopic image.

Despite the progress of deep learning in skin lesion classification, several limitations remain. CNNs extract fine-grained local texture but model global relationships within a lesion less well. Transformers improve global representation but lack the inductive bias needed for fine dermoscopic texture. Existing CNN–Transformer approaches usually combine features coarsely and do not refine the aggregated representation, so the complementary information of local and global features is only partly exploited. Interpretability and rigorous evaluation under class imbalance, as in ISIC 2019, also receive limited attention.

These limitations motivate a method that combines local convolutional and global transformer features and then uses attention to select the most salient components of the combined representation. EResViTNet combines ResNet-50 for hierarchical texture extraction with ViT-B/16 for global context modelling, followed by an attention-enhanced classification head that emphasizes informative channel and spatial features of the fused representation.

This paper presents EResViTNet, an Explainable ResNet–ViT Fusion Network for multi-class classification of dermoscopic skin lesion images. EResViTNet combines ResNet-50 [13] and ViT-B/16 [30], pairing hierarchical local feature extraction with global long-range dependencies. Preprocessing uses Error Level Analysis (ELA) [31] to expose compression inconsistencies, followed by image augmentation to improve robustness to the visual variability of lesions [32]. The end-to-end framework is shown in Figure 1.

EResViTNet combines the local texture representation of ResNet-50 with the global contextual representation of the Vision Transformer through a structured fusion step. The

two feature embeddings are first normalized and then concatenated into one representation, which is refined by a lightweight channel and spatial attention mechanism in the classification head. Earlier hybrid models apply predefined attention modules such as SE [33], CBAM [34] or ECA [35] to backbone feature maps. EResViTNet instead applies attention after fusion, so that discriminative features are selected from the joint local and global context. Ablation experiments with standard attention modules and alternative fusion operators show that this fusion–attention strategy gives the best accuracy among the evaluated configurations on ISIC 2019.

The complementary local texture features extracted by ResNet-50 and global contextual features extracted by ViT-B/16 are independently normalized, combined through concatenation and refined only after fusion, and the information-preservation and expressiveness properties of this design are derived formally A compact post-fusion channel and spatial attention head is compared under an identical protocol with SE [33], ECA [35] and CBAM [34], with alternative fusion operators, and with CNN, transformer and hybrid baselines (Tables 4 to 6 and 26). The pipeline is evaluated beyond accuracy through calibration, repeated-seed statistical testing, robustness, sensitivity, cross-dataset testing and error analysis, and ELA is assessed as an auxiliary preprocessing step through a controlled ablation (Table 19). The complementarity of the two streams is analysed through information decomposition, its relation to low-rank multimodal fusion and gated conditioning, effect sizes and feature-space analysis.

Two experimentally testable hypotheses are examined for hybrid CNN–Transformer learning in dermoscopic image classification.

- Hypothesis 1: Fusing local ResNet-50 features with global ViT-B/16 features improves recall and F1-score by about 1.5 to 2.0 percentage points over single-backbone models, particularly for minority classes.
- Hypothesis 2: Post-fusion channel and spatial attention improves accuracy and F1-score by at least 1.0 percentage point over direct fusion without attention.

Both hypotheses are tested with controlled ablation experiments under identical conditions. Statistical significance is assessed with a paired t-test ($p < 0.05$), complemented by effect sizes and a non-parametric test.

Recent CNN–Transformer hybrids achieve good overall results but often fall short in recall and F1-score for minority lesion classes, typically below 95%, because local and global features are not fused and refined jointly. Several methods also lack adaptive feature selection and simply concatenate or sequentially combine features, which prevents them from focusing on diagnostically important regions. EResViTNet adds an attention-guided post-fusion refinement stage intended to improve minority-class recall and overall discriminability, and the ablation and statistical tests link each identified gap to a design element.

PROPOSED SYSTEM

Figure 1 outlines the image processing workflow and the hybrid EResViTNet model developed for skin cancer classification of dermoscopic images from ISIC 2019. EResViTNet (Explainable ResNet and Vision Transformer Network) combines the strengths of ResNet-50 and ViT-B/16 to recognize skin lesions and is designed for multi-class dermoscopic image classification with post hoc visual explanation.

After image acquisition, each image is preprocessed: Error Level Analysis (ELA) exposes compression inconsistencies, and the image is resized and normalized. Augmentation increases data variety, and the data are divided into training, validation and test partitions for fair evaluation. The processed images are passed to ResNet-50 and ViT-B/16, which extract local texture and global contextual features, respectively. The two feature vectors are fused, and an attention-based head refines the fused representation before classification. The refined features are passed to a fully connected (FC) classifier with ReLU activation, dropout and softmax. Grad-CAM highlights discriminative image regions to support interpretation of the predictions. Finally, the model is evaluated and compared with other models in terms of accuracy, interpretability and reliability.

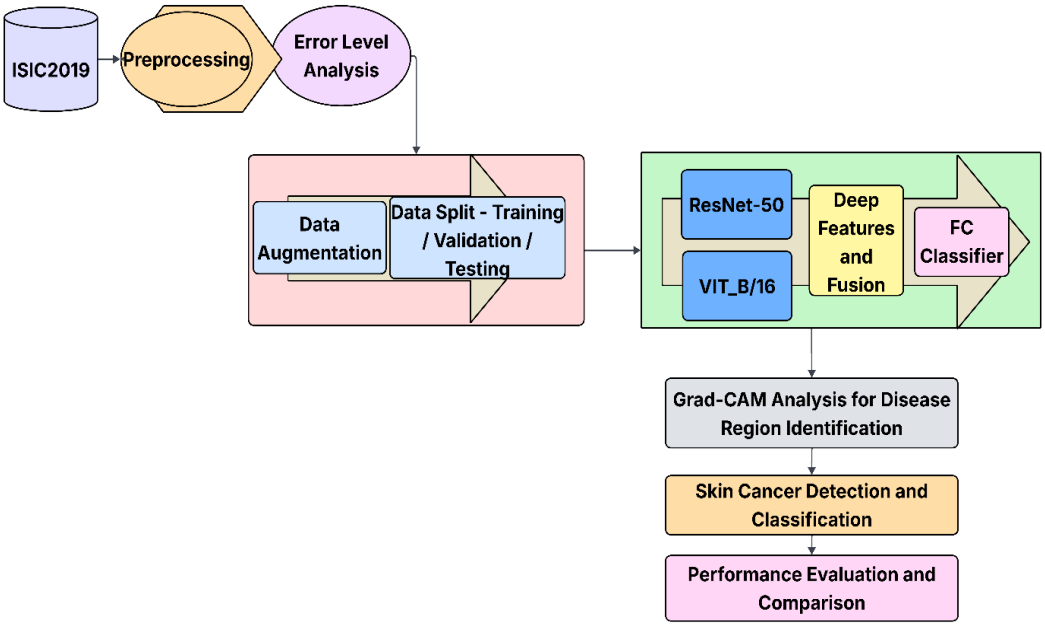


Figure 1. Workflow of the Proposed EResViTNet Framework, from ELA-Assisted Preprocessing and Augmentation to Hybrid Feature Extraction, Post-Fusion Attention, Classification and Grad-CAM interpretation.

Image Preprocessing

Dermoscopic images are pre-processed before feature extraction and classification. A sufficiently large and varied training set is needed to prevent overfitting and to obtain a robust model. Lossless and hybrid compression methods can reduce the storage and transmission cost of skin images without losing diagnostic information, which is useful for deep learning pipelines [36].

The training and test images were first contrast-enhanced and resized to 224 × 224 pixels using nearest-neighbour interpolation, as defined in equation (1). Image rescaling and cropping were subsequently performed according to equations (2) and (3), respectively.

$$NewPixelValue(x', y') = PixelValue\big(round(x), round(y)\big) \quad (1)$$

$$NewWidth = ScaleFactor * OriginalWidth \quad (2)$$

$$NewHeight = ScaleFactor * OriginalHeight \quad (3)$$

RGB images are converted to grayscale to locate the lesion boundary and reduce ambiguity. Let (p, q) denote the upper-left corner of the lesion bounding box and (r, s) its width and height. The region of interest is then extracted from the original RGB image between coordinates (p, q) and (p + r, q + s), and image artifacts are corrected where possible.

Error Level Analysis (ELA)

Error Level Analysis (ELA) identifies inconsistencies in compression that may indicate altered or suspicious regions of an image. It is used here to check data integrity before the deep learning models are trained.

$$ELA = \left| I_{original} - I_{compressed} \right| \quad (4)$$

Where $I_{original}$ is original image and $I_{compressed}$ is image after compression. ELA compares each pixel of the original image with the corresponding pixel of a recompressed copy. A duplicate of the image is compressed at a fixed quality with a lossy standard such as JPEG, and the error level is the absolute difference between corresponding pixel intensities of the original and compressed images. The difference map is displayed with a colour map, and regions with unusual error levels may indicate tampering or other inconsistencies.

In this study, ELA was used as a preprocessing step to detect compression inconsistencies and possible artifacts in dermoscopic images, which are common in public datasets such as ISIC 2019. Visual inspection of the ELA maps allowed low-quality or anomalous images that could add noise or bias to training to be discarded. The effect of ELA on classification is quantified by the controlled ablation in (Table 19); no separate quantitative analysis of image quality was performed. ELA is lightweight, easy to interpret and complementary to standard augmentation.

ELA is computed as follows. Each dermoscopic image is recompressed as JPEG at a fixed quality level, and the absolute pixel-wise difference between the original and the recompressed image is calculated. The resulting error map identifies areas with compression inconsistencies and artifacts before model training. This step is followed by resizing, normalization and augmentation.

ELA is incorporated as an auxiliary preprocessing strategy rather than as a new algorithmic contribution. In the present framework, it is used to expose compression-related pixel inconsistencies before feature extraction. No independent quantitative claim regarding lesion contrast enhancement is made. Its contribution is instead evaluated through the controlled classification ablation.

Sequential Image Augmentation

The tf.keras.layers.experimental.preprocessing package provides layers that add image augmentation to sequential TensorFlow–Keras models [37], including flipping, rotation, zoom and resizing. Placing the augmentation layers at the start of the network exposes later layers to augmented data and improves generalization. The image_augmentation model used in this study contains four augmentation layers: horizontal flip with probability 0.5, random rotation of up to 0.2 radians (about 11.5 degrees), random zoom of up to 20%, and random height and width scaling between 80% and 120%. Images are resized to 224 × 224 pixels and rescaled by dividing pixel values by 255. This variability makes the model more robust to differences in input conditions.

The augmentation parameters follow configurations commonly used in dermoscopic image classification. A rotation range of 0.2 radians (about 11.5 degrees) provides moderate geometric variability while preserving lesion structure. ReLU activation is used in the hidden layers for its computational efficiency and stable gradient propagation, and dropout regularization together with augmentation reduces overfitting on heterogeneous dermoscopic images.

Proposed Model for Deep Feature Extraction and Custom Attention Design

EResViTNet combines the deep local features of ResNet-50 with the contextual features of the Vision Transformer (ViT-B/16), both of which are needed to classify dermoscopic images into the skin cancer categories of ISIC 2019. The model integrates a feature fusion step and an attention-enhanced fully connected (FC) classification head, and Grad-CAM is applied for explainability.

Both ResNet-50 and ViT-B/16 are initialized with ImageNet-pretrained weights (Table 1), which provide general visual representations that are fine-tuned for dermoscopic image classification.

Table 1. Backbone Configuration and Training Setup

Component	Configuration
ResNet Backbone	ResNet-50 (ImageNet pretrained)
Fine-tuning Strategy	Early layers frozen, last stages fine-tuned
Transformer Backbone	ViT-B/16
Patch Size	16 × 16
Embedding Dimension	768
Feature Output (ResNet)	2048-dim vector
Feature Output (ViT CLS Token)	768-dim vector
Tokenization	Image divided into 16×16 patches and linearly projected to token embeddings
Fusion Representation	Concatenation of CNN and Transformer features

Equations (5) until (13) describe the computational operations of the hybrid architecture rather than theoretical derivations. They represent feature extraction, fusion, attention refinement and classification in the deep learning pipeline, and the architecture and

parameter choices follow methods widely used in medical image analysis and transfer learning.

These formulations are computational descriptions of the architecture, not derivations from first principles; feature extraction and representation learning are data-driven rather than obtained by closed-form optimization. Convolutional layers encode local spatial information such as texture, colour variation and lesion boundaries, and transformer encoders encode long-range dependencies through self-attention. Concatenation is used to preserve the complete branch-specific representations before refinement, not on the assumption that it is universally superior to other fusion mechanisms, and its suitability is evaluated empirically against sum fusion, gated fusion and bilinear pooling.

The attention refinement module assumes that fused features contribute unequally to the class decision and that adaptive channel and spatial weighting increases discriminative power. These assumptions are tested by the ablation experiments and performance analysis.

The architectural rationale is based on preserving complementary representations before discriminative refinement. Let $f_R \in \mathbb{R}^{2048}$denote the globally pooled ResNet-50 feature vector and$f_v \in \mathbb{R}^{768}$ denote the ViT-B/16 class-token embedding. The CNN representation primarily encodes hierarchical local texture, pigmentation and boundary characteristics, whereas the transformer representation captures long-range contextual dependencies. Since these representations differ in dimensionality and feature semantics, direct concatenation, $F = [f_R; f_v] \in \mathbb{R}^{2816}$, preserves the information from both branches without imposing element-wise correspondence or compressing either representation before joint learning. Element-wise fusion, in contrast, requires dimensional alignment and can suppress branch-specific information through projection or aggregation.

The placement of attention after fusion is motivated by conditional feature refinement. If attention is independently applied before fusion, the weighting of f_R is estimated without access to f_v and vice versa. Post-fusion attention instead estimates the refinement weights from the joint representation F, allowing the importance of a local CNN feature to be evaluated in the presence of global transformer context. Cross-attention can explicitly model interactions between the two streams through query-key-value operations, but introduces additional projections and interaction parameters. EResViTNet adopts a preservation-first design in which heterogeneous representations are retained through concatenation and selectively refined only after integration. This formulation provides an architectural rationale rather than a proof of global optimality, and its empirical validity is examined through the controlled fusion and attention ablations in Tables 4–6.

Step 1: Local Feature Extraction with ResNet-50

ResNet-50 learns local, hierarchical features that capture complex lesion patterns such as colour, border irregularity and texture. It maps an input image X to a deep feature representation through convolutional layers with residual connections followed by global average pooling:

$$f_R = \text{ResNet} - 50(X) \quad (5)$$

Step 2: Global Feature Extraction with Vision Transformer (ViT-B/16)

ViT-B/16 complements the CNN by capturing global structure and relationships across the lesion. The image is divided into 16 × 16 patches, which are flattened and linearly projected into a sequence of embeddings, and transformer encoder layers model the relationships between patches. The ViT-B/16 output f_V is given by:

$$f_V = \text{ViT} - \text{B}/16(X) \quad (6)$$

Step 3: Feature Fusion

The pooled ResNet-50 descriptor $f_R \in \mathbb{R}^{2048}$ and the ViT-B/16 class token $f_V \in \mathbb{R}^{768}$ are normalized separately, so that neither branch dominates the fused representation because of its scale:

$$\hat{f}_R = \text{N}(f_R), \quad \hat{f}_V = \text{N}(f_V) \quad (7)$$

where N(·) denotes feature normalization. The normalized descriptors are concatenated along the feature dimension:

$$F_F = [\hat{f}_R ; \hat{f}_V] \in \mathbb{R}^{2816} \quad (8)$$

No fusion coefficient is used. Weighted combinations (0.3/0.7 and 0.5/0.5) were examined in preliminary trials but gave no reproducible improvement over unweighted concatenation, and the relative contribution of each branch is instead learned by the post-fusion attention in equations (9) to (12).

Step 4: Attention-Enhanced Fully Connected Classification

The fused representation F_F is passed to an attention-enhanced FC classifier. CBAM is not used as such; instead, channel and spatial attention are integrated directly into the classification head.

Channel Attention Embedded Before FC

To emphasize informative feature channels, global average pooling and max pooling are applied to the fused representation, followed by two shared FC layers that produce the channel attention weights:

$$M_c(F_F) = \sigma(\text{FC}(\text{AvgPool}(F_F)) + \text{FC}(\text{MaxPool}(F_F))) \quad (9)$$

The refined representation after channel attention, F_c, is given by

$$F_c = M_c(F_F) \odot F_F \quad (10)$$

Spatial Attention Embedded Before FC

To emphasize informative regions, such as lesion borders or texture-rich zones, spatial attention applies average and max pooling across channels, followed by a convolution layer and a sigmoid activation:

$$M_s(F_c) = \sigma(Conv([AvgPool(F_c); MaxPool(F_c)])) \quad (11)$$

The final refined feature map after spatial attention, F_{Attn} is obtained by

$$F_{Attn} = M_s(F_c) \,.\, F_c \quad (12)$$

Standard modules such as CBAM and SE blocks are widely used to enhance feature representation, but EResViTNet adopts a lightweight custom design that integrates channel and spatial attention directly into the fully connected head, which keeps the architecture simple and efficient. In the ablation study, the full head reached 98.11% accuracy, compared with 96.84% for fusion without attention and 97.42% for fusion with CBAM (Table 4). A direct comparison with SE, ECA and CBAM under the same protocol is reported in Table 6.

The attention head refines the fused representation in two steps. Channel attention applies global average pooling and max pooling, followed by shared fully connected layers with ReLU activation and sigmoid gating, to generate feature-wise weights. Spatial attention then applies a convolution layer with sigmoid activation to the pooled maps to emphasize informative lesion regions. The refined representation is passed to a dense layer with 512 units, ReLU activation, dropout of 0.5 and a softmax classifier. The design adds little computation while improving feature selectivity in the fused CNN–Transformer representation.

Fully Connected Classification Head

The attention-refined features F_{Attn} are flattened and passed to a fully connected classification head with a ReLU-activated layer, dropout regularization and a softmax output layer for multi-class skin cancer classification:

$$F_{\text{FC}} = \text{Softmax}(\text{Dropout}(\text{ReLU}(\text{Dense}(F_{\text{Attn}})))) \quad (13)$$

The proposed framework should be interpreted as an architectural integration contribution rather than as a new standalone attention algorithm. Concatenation and Error Level Analysis are established techniques and are not claimed as independent algorithmic novelties. The contribution lies in preserving complementary ResNet-50 and ViT-B/16 representations and applying channel and spatial refinement after fusion, so that feature selection operates on a unified local–global representation. The effectiveness of this design is established empirically through controlled ablation analysis rather than through a new theoretical attention formulation.

Theoretical Analysis of Fusion and Post-Fusion Refinement

The three design decisions of EResViTNet: concatenation as the fusion operator, attention placed after fusion, and compact gating in place of cross-attention. Let Y denote the lesion label, and let $f_R \in \mathbb{R}^{2048}$ and $f_V \in \mathbb{R}^{768}$ denote the normalized ResNet-50 and ViT-B/16 descriptors of equation (7).

Information preservation of concatenation. Concatenation $F_F = \{f_R; f_V\}$ is an injective mapping, hence $I(F_F; Y) = I(f_R, f_V; Y)$. By the chain rule of mutual information [38],

$$I(F_F; Y) = I(f_R; Y) + I(f_V; Y \mid f_R) \geq max\{I(f_R; Y), I(f_V; Y)\} \quad (14)$$

The conditional term $I(f_V; Y \mid f_R)$ measures the label information that the transformer branch adds to the CNN branch, and a fusion gain is possible only when this term is positive. Any other fusion operator $g(f_R, f_V)$, such as element-wise summation after

projecting both descriptors to a common dimension, forms the Markov chain $Y \rightarrow (f_R, f_V) \rightarrow g(f_R, f_V)$. The data processing inequality then gives

$$I(g(f_R, f_V);\, Y) \;\le\; I(f_R, f_V;\, Y) \;=\; I(F_F;\, Y) \tag{15}$$

Concatenation therefore attains the upper bound of information available to any fusion operator applied to the same descriptor pair, at a linear dimension of 2816. Bilinear pooling also retains second-order interactions, but its outer product has $2048 \times 768 \approx 1.57 \times 10^6$ dimensions. This explains why bilinear pooling and concatenation obtain similar accuracy in Table 5 (96.94% and 96.84%), while concatenation offers a far more compact input to the refinement stage. Equation (15) concerns available information, not learnability from finite samples, so the empirical comparison in Table 5 remains necessary.

Expressiveness of post-fusion attention. Let a pre-fusion gate be applied independently to each branch, $M_{pre} = [\sigma(A\, f_R);\, \sigma(B\, f_V)]$. This gate can be written as $\sigma(W\, F_F)$ with the block-diagonal matrix $W = \mathrm{diag}(A, B)$. The post-fusion gate uses an unrestricted matrix,

$$M_{post}(F_F) = \sigma(W\, F_F), \quad W = \{W_{RR}\;\; W_{RV}\,;\, W_{VR}\;\; W_{VV}\} \tag{16}$$

and reduces to pre-fusion gating when $W_{RV} = 0$ and $W_{VR} = 0$. The hypothesis class of pre-fusion gating is therefore contained in that of post-fusion gating. The off-diagonal blocks allow the weight assigned to a CNN texture feature to depend on transformer context, and the reverse, which cannot be represented when attention is applied before fusion. The same inclusion holds for the two-layer gating of equation (9), because both of its weight matrices can take a block-diagonal structure as a special case.

Degeneracy of cross-attention on pooled descriptors. For a query $q = W_Q f_R$ and a single key-value pair $k = W_K f_V$, $v = W_V f_V$,

$$\mathrm{CrossAttn}(f_R, f_V) \;=\; \mathrm{softmax}(q^{\mathrm{T}} k \,/\, \sqrt{d})\; v \;=\; W_V\, f_V \tag{17}$$

Because the softmax of a single score equals one. The output is independent of f_R, and cross-attention reduces to a linear projection of the transformer token. Meaningful cross-attention requires token-level inputs, with $N_R = 49$ ResNet-50 positions and $N_V = 197$ ViT-B/16 tokens.

Complexity-performance trade-off. For embedding dimension d, a single-head cross-attention layer on token-level inputs costs

$$C_{\mathrm{cross}} \;\approx\; N_R \cdot 2048 \cdot d \;+\; 2N_R d^2 \;+\; 2N_V d^2 \;+\; 2N_R N_V d \tag{18}$$

Which amounts to approximately 0.38 G multiply-accumulate operations and 3.9 M parameters for d = 768. The channel gate of EResViTNet requires $2 \times 2816 \times (2816/r)$ weights, about 0.99 M for a reduction ratio r = 16, and the complete refinement and classification head requires fewer than 4 M operations, which is below 0.02% of the backbone cost of ResNet-50 and ViT-B/16 (4.1 G and 17.6 G, Table 17). The 1.27-point accuracy gain of post-fusion refinement over plain concatenation (Table 5) is therefore obtained at negligible computational cost. These results are formal properties of the fusion and gating operators, not a proof of global optimality of the network, and their practical relevance is established by the ablations in Tables 4 to 6.

Complementarity, redundancy and synergy. Equation (14) shows that fusion can help only when the transformer branch adds label information to the CNN branch, but it does not describe the form of this added information. Partial information decomposition [39] separates the joint information into four non-negative parts:

$$I(f_R, f_V; Y) = Red + Unq_R + Unq_V + Syn \quad (19)$$

Where Red is the label information carried by both descriptors, Unq_R and Unq_V are the information available only in the CNN or only in the transformer descriptor, and Syn is the information that becomes available only when both descriptors are observed together. The conditional term of equation (14) then becomes:

$$I(f_V; Y \mid f_R) = Unq_V + Syn \quad (20)$$

Additionally, the fusion gain has two distinct sources. Unique information arises when the transformer captures cues that the CNN misses, such as global asymmetry or the arrangement of structures across the lesion, and the reverse holds for fine texture and border detail. Synergistic information arises when a cue is diagnostic only in combination, for example a local pigment network that is unremarkable in a symmetric lesion but suspicious in an asymmetric one. The three design decisions of EResViTNet treat these components differently. Concatenation retains all four terms (Eq. 15). A projected sum can cancel unique and synergistic components when the projections of the two descriptors overlap. Pre-fusion gating can weight the unique information within each branch, but its weights cannot depend on the other branch and therefore cannot respond to synergy. The post-fusion gate of Eq. 16 can, because its off-diagonal blocks make the weight of each CNN feature a function of the transformer descriptor and vice versa. The argument for pairing ResNet-50 with ViT-B/16 is therefore an argument that $Unq_V + Syn$ is large for this pair, since the two backbones differ in locality, texture bias and spatial-frequency response [40-42].

Empirical bound on residual uncertainty. The information terms above cannot be measured directly for 2816-dimensional descriptors, but the test error of a classifier bounds them. For K classes and a classifier with error probability P_e that operates on a representation Z, Fano's inequality [38] gives

$$H(Y \mid Z) \leq h_b(P_e) + P_e \log_2(K - 1) \quad (21)$$

Where h_b is the binary entropy function. A lower error rate therefore certifies that less label uncertainty remains given the representation, and the reduction of this bound across ablation variants indicates how much label information each architecture makes usable. The bound is evaluated for the ablation variants (Table 28).

Relation to Multimodal Fusion Theory

Although EResViTNet processes a single dermoscopic image, its two branches can be treated as two views, or modalities, of the same input, because they are produced by encoders with different inductive biases and enter the fusion stage as separate descriptors. This places the architecture within multimodal fusion theory, which classifies fusion by the stage at which the streams interact and by the form of the interaction [43]. Early fusion

combines raw inputs, late fusion combines the decisions of separately trained models, and intermediate fusion combines learned representations so that a shared layer can model cross-stream interactions. EResViTNet performs intermediate fusion. Late fusion, as in an ensemble of the two backbones, can exploit redundant and unique information through its averaged decisions, but synergistic information (Eq. 20) is lost once each stream has been reduced to class probabilities.

Multiplicative interaction. Tensor fusion [44] models all interactions between two streams through the outer product of their augmented descriptors, $\{1; f_R\} \otimes \{1; f_V\}$, which for the present descriptors has $2049 \times 769 \approx 1.58 \times 10^6$ dimensions. Low-rank multimodal fusion [45] showed that a rank-limited factorization of this tensor retains most of its benefit at a fraction of its cost. The post-fusion gate of EResViTNet has a related structure. Writing the gate input as $a = W F_F$, element i of the refined representation is

$$F_{Attn,i} = F_{F,i} \cdot \sigma(\Sigma_{j \in R} W_{ij} f_{R,j} + \Sigma_{j \in V} W_{ij} f_{V,j}) \tag{22}$$

and a first-order expansion of σ around a_0 gives

$$F_{Attn,i} \approx c_i F_{F,i} + \sigma'(a_{0,i}) \Sigma_j W_{ij} F_{F,i} F_{F,j}, \quad c_i = \sigma(a_{0,i}) - a_{0,i} \sigma'(a_{0,i}) \tag{23}$$

The second term contains products $F_{F,i} F_{F,j}$ between CNN and transformer features whenever i and j belong to different branches, so the gate introduces second-order cross-stream interactions of the same kind as tensor fusion. In the two-layer gate of equation (9), the effective matrix W is the product of a 2816×176 and a 176×2816 matrix for the reduction ratio $r = 16$, so its rank is at most 176. The post-fusion gate is therefore a gated, low-rank approximation of bilinear fusion that also keeps the first-order term of every feature. Bilinear pooling (Table 5) is the full-rank, ungated counterpart. Its accuracy (96.94%) is close to that of plain concatenation and 1.17 points below EResViTNet, which agrees with the finding of [45] that a low-rank interaction can match or exceed a full tensor with far fewer parameters.

Feature-wise conditioning. Gated multimodal units [46] and feature-wise linear modulation (FiLM) [47] condition one representation on another through learned multiplicative weights. The EResViTNet gate is a multiplicative modulation without a shift term in which the conditioning signal is the joint descriptor, so that each stream is modulated by itself and by the other stream at the same time. This combined self- and cross-conditioning is the property expressed by the off-diagonal blocks of equation (16).

Modality imbalance. Jointly trained multimodal networks can be dominated by the stream that fits the training data fastest, which leaves the other stream under-optimized [48, 49]. Theoretical analysis shows that a multimodal model outperforms its unimodal counterparts when the additional stream improves the quality of the learned latent representation [50]. Two elements of EResViTNet counteract imbalance: separate normalization of the two descriptors in equation (7) prevents the 2048-dimensional CNN descriptor from dominating by scale, and the gate can re-weight features of either stream after fusion. Explicit gradient modulation between the branches [49] was not used, and the relative contribution of each stream during training was not measured; both are directions

for further analysis. The bottleneck principle discussed [51] completes this account: in EResViTNet, all cross-stream exchange passes through a 176-dimensional gate bottleneck, while the full descriptors are retained for classification.

Proposed Model Architecture

EResViTNet classifies ISIC 2019 dermoscopic images with a hybrid deep learning architecture (Figure 2). ResNet-50 extracts hierarchical convolutional features of lesion texture, and ViT-B/16 learns spatial relationships through self-attention across image patches. The two descriptors are fused into F_F, which combines local and global information. Channel attention weights the most informative channels using global pooling and a shared MLP, and spatial attention emphasizes key lesion locations through a convolution over pooled maps. A classification head with a flattening layer, a 512-unit dense layer with ReLU activation, dropout and softmax outputs the class probabilities. Grad-CAM shows the regions that influence each prediction, improving transparency. Together, hybrid feature extraction, post-fusion attention and visual explanation support accurate and interpretable lesion classification.

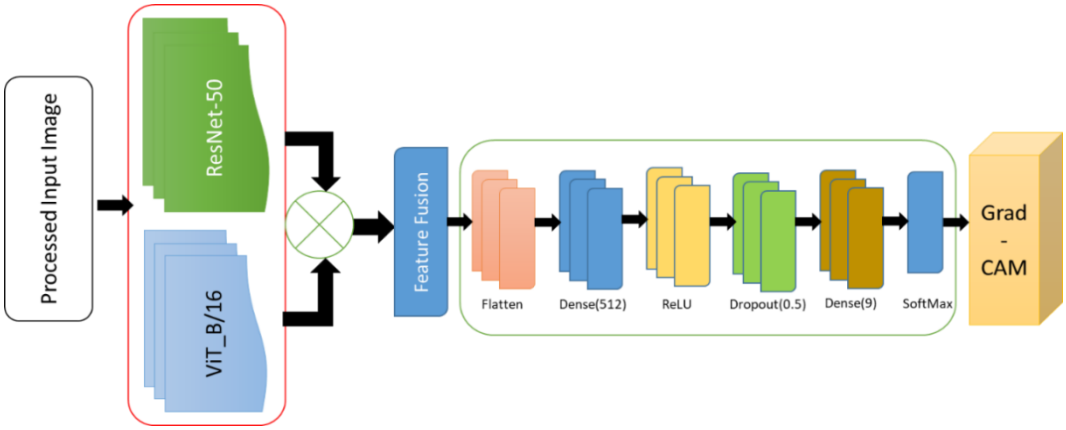


Figure 2. Architecture of EResViTNet, Showing the ResNet-50 and ViT-B/16 Branches, Concatenation of the 2048- and 768-Dimensional Descriptors, Post-Fusion Channel and Spatial Attention, and the Classification Head.

Interpretability with Grad-CAM

Grad-CAM provides visual interpretability by highlighting the image regions that contribute most to the classification decision. As shown in Figure 2, after the fused ResNet-50 and ViT-B/16 features are refined by the attention-based classifier, Grad-CAM is applied to the final convolutional layers of ResNet-50 to trace the contribution of each spatial location to the predicted class.

The final Grad-CAM heatmap$L^c_{\text{Grad-CAM}}$ is obtained as a weighted linear combination of the activation maps followed by a ReLU, which retains only the positive contributions.

$$L^c_{\text{Grad-CAM}} = ReLU \underbrace{\left(\sum_k \alpha_k^c A^k\right)}_{\textit{linear combination}} \tag{24}$$

The Grad-CAM heatmaps are overlaid on the original dermoscopic images to show which lesion areas the model attends to. They provide a qualitative, per-prediction visual explanation that can support diagnostic review.

Proposed Model Algorithm

EResViTNet is a hybrid deep learning approach for automatic classification of ISIC 2019 dermoscopic images. Pretrained ResNet-50 and ViT-B/16 models serve as feature extractors (Figure 2): ResNet-50 captures high-resolution spatial features of lesion structure, and ViT-B/16 captures global patch-level attention patterns. Input images undergo Error Level Analysis (ELA), augmentation and resizing to the input size of each model. The feature vectors of both models are combined into F_F, which is refined by the attention-based FC classifier in place of CBAM; this classifier prioritizes relevant channels and spatial regions using pooling and dense layers. The refined features pass through a dense layer with ReLU activation and dropout to a softmax classifier that predicts the skin cancer class. Grad-CAM visualizes the regions behind each prediction, and the model is evaluated by accuracy, sensitivity, specificity and related metrics. The main algorithm that have been applied for EResViTNet model for skin cancer classification is shown as follows:

Algorithm: EResViTNet Model for Skin Cancer Classification

```
% Step 1: Initialize Pretrained Models
resNet50Model = resnet50;
vitModel = vit_b_16;  % Pretrained Vision Transformer (ViT-B/16)
% Step 2: Image Preprocessing (with ELA + Resizing)
for i = 1:numImages
 inputImage = imread(imagePaths{i});
 elaImage = applyErrorLevelAnalysis(inputImage); % Optional ELA preprocessing
 % Resize images for each model
 resizedForResNet = imresize(elaImage, resNet50Model.Layers(1).InputSize(1:2));
 resizedForViT = imresize(elaImage, vitModel.InputSize(1:2));
% Assuming defined input
Step 3: Feature Extraction
resNetFeatures = activations(resNet50Model, resizedForResNet, 'avg_pool');
vitFeatures = activations(vitModel, resizedForViT, 'transformer_output');
Step 4: Feature Fusion
fusedFeatures = cat(2, squeeze(resNetFeatures), squeeze(vitFeatures));
Step 5: Attention-Enhanced FC Classifier
Apply channel attention and spatial attention to the fused features,
followed by Flatten → Dense(512, ReLU) → Dropout(0.5) → Softmax(9)
% Apply custom channel and spatial attention (embedded in FC layers)
attentionRefined = applyAttentionRefinement(fusedFeatures);
```

Step 6: Store Feature Vectors and Labels

featureMatrix(i, :) = attentionRefined;

labelVector(i) = imageLabels(i); end

Step 7: Train the Fully Connected Classification Model (Optional: MLP/Softmax)

fcClassifier = trainFCClassifier(featureMatrix, labelVector);

Step 8: Grad-CAM Visualization

% Generate heatmaps for interpretability using ResNet-50 convolutional layers

for i = 1:numTestImages

heatmap = applyGradCAM(resNet50Model, testImages{i}, fcClassifier); end

Step 9: Skin Cancer Classification

predictedLabels = classify(fcClassifier, testFeatureMatrix);

Step 10: Performance Evaluation

evaluateModel(predictedLabels, groundTruthLabels); % Accuracy, sensitivity, specificity, etc.

Training protocol: The dataset is stratified into 70% training, 15% validation and 15% test partitions. ELA-assisted preprocessing, resizing, normalization and training augmentation are applied before feature extraction. ImageNet-pretrained ResNet-50 and ViT-B/16 produce 2048- and 768-dimensional representations, which are concatenated into a 2816-dimensional vector. Post-fusion channel and spatial attention is followed by Dense(512), ReLU, Dropout(0.5) and Softmax(9). The model is trained for 50 epochs with SGD (learning rate 0.001, momentum 0.9, weight decay 1×10^{-4}), batch size 32 and a combined categorical cross-entropy and focal loss. The checkpoint with the best validation performance is retained and evaluated on the independent test partition.

The pseudocode describes the operation of the framework; the implementation is specified further at the architectural level for clarity and reproducibility. ResNet-50 receives a 224 × 224 × 3 image and outputs a 2048-dimensional vector after global average pooling, and ViT-B/16 receives the same input and outputs a 768-dimensional class-token representation. The two are concatenated into a 2816-dimensional representation that enters the attention refinement module, where channel attention uses pooled descriptors and fully connected layers and spatial attention uses a convolution over the integrated features. The refined vector is passed to a 512-unit FC layer with ReLU activation, a dropout layer (rate 0.5) and a nine-unit softmax layer corresponding to the nine lesion classes. The model is trained for 50 epochs with batch size 32, SGD with momentum 0.9 and learning rate 0.001, and a combined categorical cross-entropy and focal loss. With these specifications, the full pipeline can be implemented directly in standard deep learning frameworks.

The algorithm is a structured description of the workflow and can be implemented directly in standard frameworks such as TensorFlow or PyTorch. Each phase, namely feature extraction, fusion, attention refinement and classification, maps unambiguously onto modules of standard libraries. The pseudocode is not intended for execution; it shows

the logical order of operations, and the architectural specification above defines the implementation.

The attention-enhanced FC classifier operates on the fused ResNet-50 and ViT-B/16 representation. The concatenated vector passes through channel and spatial attention, is flattened and is fed to a 512-unit dense layer with ReLU activation, followed by dropout (rate 0.5) and a softmax layer that outputs probabilities for the nine lesion categories. Training uses 50 epochs, batch size 32, the SGD optimizer and a learning rate of 0.001.

RESULTS AND DISCUSSION

EResViTNet was implemented and evaluated in TensorFlow, which supports the integration of ResNet, ViT and Grad-CAM. Experiments were run in Python on a system with an Intel Core i5 CPU, an NVIDIA P100 GPU and 16 GB of RAM.

Dataset Description

The ISIC 2019 training set contains 25,331 dermoscopic images in eight diagnostic categories [6-8]. Melanocytic nevus (NV) comprises benign neoplasms of melanocytes with varied appearance, and melanoma (MEL) is a malignant tumour that can be cured by early excision. Dermatofibroma (DF) is a benign lesion that often follows minor skin injury. Vascular lesions (VASC) include cherry angiomas, angiokeratomas and pyogenic granulomas. Benign keratosis-like lesions (BKL) include solar lentigo, lichen planus-like keratosis and seborrheic keratosis. Basal cell carcinoma (BCC), the most common skin cancer, grows destructively if untreated but rarely metastasizes. Squamous cell carcinoma (SCC) arises from abnormal squamous cells and is treatable if detected early, and actinic keratosis (AK) is a precancerous lesion caused by chronic sun exposure that can progress to squamous cell carcinoma.

The images used in this study come from a reorganized version of ISIC 2019 in which benign keratosis-like lesions are divided into pigmented benign keratosis and seborrheic keratosis. The official challenge groups these lesions under BKL, whereas the reorganized version provides them as separate folders, giving nine lesion classes. This separation allows a finer examination of lesions that are morphologically and clinically similar.

The dataset was divided into training (70%), validation (15%) and test (15%) sets by stratified sampling, so that all partitions have the same class distribution. The validation set was used to monitor convergence and to select the checkpoint. Cross-validation was not used because of the cost of training CNN–Transformer hybrids on a large dermoscopic dataset; instead, the experiment was repeated five times with different random initialization seeds on the same split, and the reported results are averages over these five runs.

Performance Assessment

The performance of EResViTNet was assessed with standard multi-class metrics: accuracy, precision, recall (sensitivity), specificity and F1-score [52]. Accuracy is the proportion of correctly classified samples. Precision is the proportion of predicted positives

that are truly positive, recall is the proportion of actual positives that are correctly predicted, and specificity is the proportion of actual negatives that are correctly predicted. The F1-score balances precision and recall and is particularly informative for imbalanced classes. Class-wise discrimination was further evaluated with ROC-AUC and precision–recall (PR) analysis, with emphasis on minority classes.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{25}$$

$$Recall = \frac{TP}{TP + FN} \tag{26}$$

$$Precision = \frac{TP}{(TP + FP)} \tag{27}$$

$$Specificity = \frac{TN}{TN + FP} \tag{28}$$

$$F1 - Score = \frac{2*(Precision*Recall)}{(Precision + Recall)} \tag{29}$$

Here, TP, TN, FP and FN denote true positives, true negatives, false positives and false negatives. Overall precision, recall and F1-score are macro-averaged over the nine classes within each run and then averaged across the five runs. Because the F1-score is averaged per class and per run, the reported mean F1-score is not the harmonic mean of the reported mean precision and recall.

Training was repeated five times with different random initialization seeds on the same data split, and the reported results are the averages of these runs. The variation between runs was small, indicating consistent model behaviour.

Error Level Analysis

In forensic image analysis, ELA can reveal signs of alteration by locating regions that were compressed or edited differently from the rest of the image. It compares the error levels of different regions and highlights those that change markedly on recompression. ELA was applied to the class images of the ISIC 2019 dataset at several JPEG quality levels (q), and the resulting maps were displayed as an image matrix. Varying the compression level helps identify corrupted or modified images so that they can be removed or repaired before training, which supports data integrity and quality. Figure 3 shows the ELA of a nevus image.

ELA is used as a preprocessing step to detect compression inconsistencies and possible artifacts in the dermoscopic dataset. It improves dataset consistency before training and, in addition to standard augmentation, highlights small structural differences within lesion regions.

All dermoscopic images were resized to 224 × 224 pixels and pixel intensities were normalized to {0, 1} for stable optimization. Training augmentation consisted of horizontal flipping, rotation, zoom and height and width scaling, which encourages the model to learn lesion patterns that are invariant to orientation and scale while preserving the diagnostic features of dermoscopic structures.

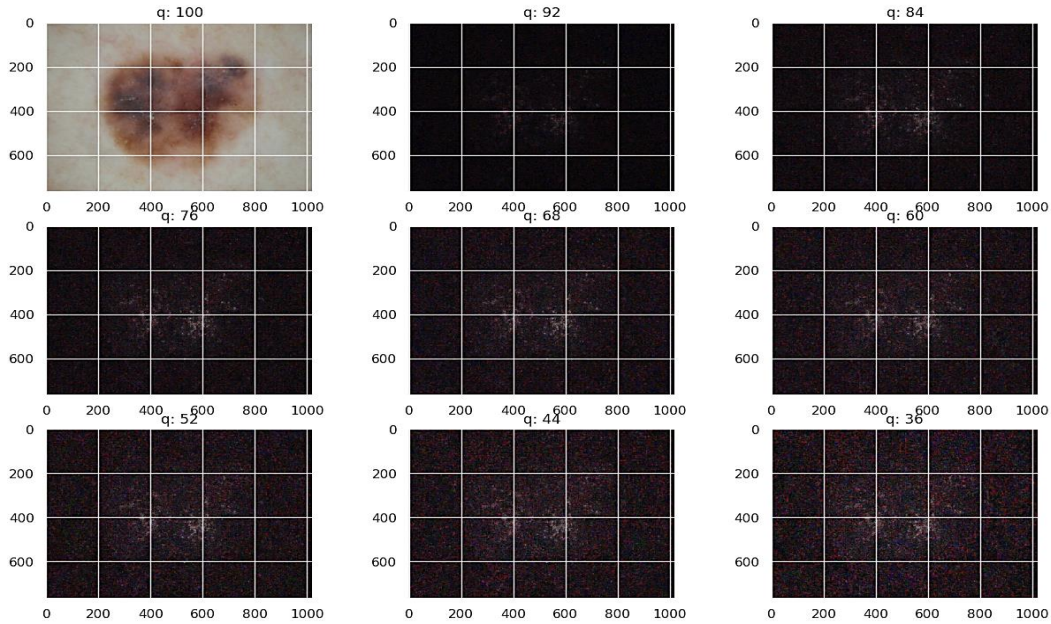


Figure 3. Error Level Analysis of a Melanocytic Nevus Image at Different JPEG Recompression Qualities.

EResViTNet Performance on ISIC 2019

EResViTNet was trained and evaluated on ISIC 2019, and Table 2 summarizes the main model parameters, training settings and hyperparameters. During training the model learns increasingly discriminative representations of dermoscopic images, and the combination of ResNet-50, ViT-B/16 and the attention-based classification head allows it to learn both the detailed and the contextual features needed for accurate lesion classification.

Figure 4 shows that the accuracy of EResViTNet increases steadily over the 50 training epochs, indicating stable learning.

The reported accuracy, precision, recall and F1-score on ISIC 2019 are averages over five independent runs on the same train–validation–test split, each with a different random seed, which reduces the influence of stochastic training variation. Run-to-run variation was small (standard deviation below 0.5 percentage points), indicating that the performance is repeatable and not the result of a favourable initialization.

The model was trained with SGD (momentum 0.9, initial learning rate 0.001) and a weight decay of 1×10^{-4} to limit overfitting. Validation loss was monitored during training, and model checkpointing stored the best weights for final evaluation. Label smoothing was not used. Cross-entropy and focal loss were combined without explicit class weights, which directs training towards hard samples and minority classes.

The training configuration was selected through preliminary validation experiments rather than exhaustive hyperparameter optimization. The initial learning rate, dropout, batch size, and SGD parameters were chosen based on stable convergence and validation behavior, while the best model was retained according to validation performance. The selected configuration was subsequently examined through sensitivity analysis by varying

the learning rate from 0.0005 to 0.002 and dropout from 0.3 to 0.6, as reported in Table 12. The relatively small performance variation across these settings indicates that the reported result is not restricted to a single narrowly selected configuration. No exhaustive grid search, Bayesian optimization, or neural architecture search was performed; therefore, the selected configuration is treated as a validated operating setting rather than a globally optimal solution.

Table 2. EResViTNet Training and Hyperparameter Configuration

S. No	Parameter	Value / Configuration
1	Input Image Size	224 × 224
2	Number of Classes	9
3	Backbone CNN	ResNet-50 (ImageNet pretrained)
4	Transformer Encoder	ViT-B/16 (ImageNet pretrained)
5	Feature Fusion	Concatenation of CNN and ViT features
6	Attention Mechanism	Channel + Spatial Attention
7	Training Epochs	50
8	Batch Size	32
9	Optimizer	SGD
10	Learning Rate	0.001
11	Dropout Rate	0.5
12	Data Augmentation	Flip, Rotation (0.2 rad), Zoom, Height/Width scaling
13	Loss Function	Cross Entropy + Focal Loss
14	Activation Functions	ReLU (hidden layers), Softmax (output layer)
15	Weight Initialization	ImageNet pretrained weights
16	Data Shuffle	Enabled
17	Regularization	Weight Decay 1×10^{-4}
18	Callback	Model Checkpoint
19	Momentum	0.9
20	Label Smoothing	Not used
21	Class Weights	Not explicitly used
22	Imbalance Handling	Cross-entropy + Focal Loss
23	Checkpoint Strategy	Best validation model retained

The architecture and training configuration were fixed deliberately rather than obtained through automated search. The purpose of the study is to attribute performance differences to the fusion operator and to the placement of attention, and this attribution is valid only when the backbones, classification head and optimization settings remain identical across all variants in Tables 4 to 6, 15 and 26. Neural architecture search or Bayesian optimization performed separately for each variant would change several factors at once and confound the ablation. The backbone pair is justified analytically in, where ResNet-50 and ViT-B/16 are selected for the complementarity of their representations and not for their individual accuracy. The sensitivity analysis in Table 12 further shows that halving or doubling the learning rate and varying the dropout between 0.3 and 0.6 change accuracy by at most 0.59 points, which indicates a flat region of the validation surface in which an exhaustive search would be expected to yield only marginal gains. The reported

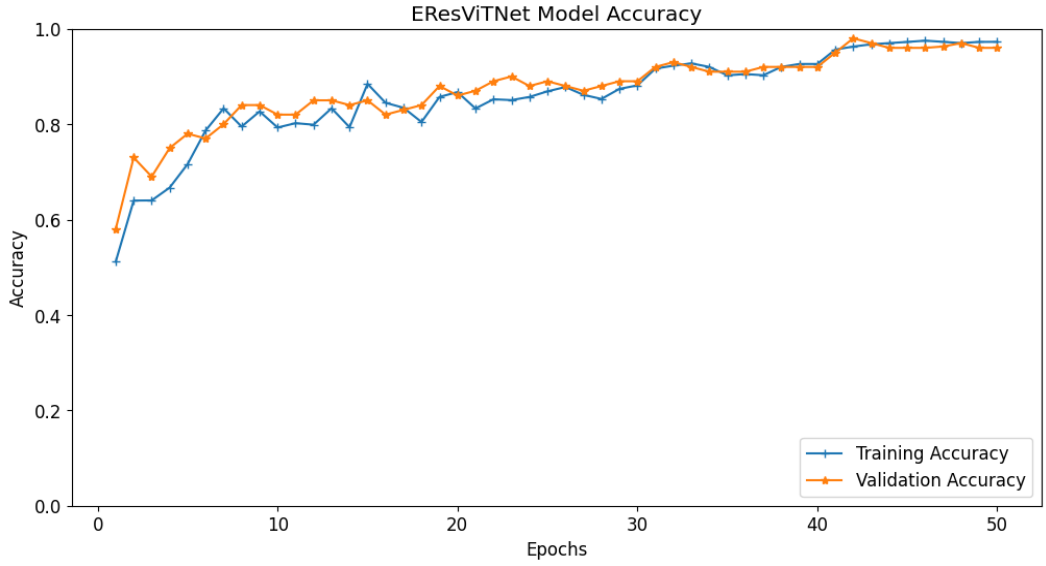
results should therefore be read as those of a validated operating point within a controlled comparison, not as those of a globally optimized network.

EResViTNet Model Performance Evaluation on the ISIC-2019 Dataset

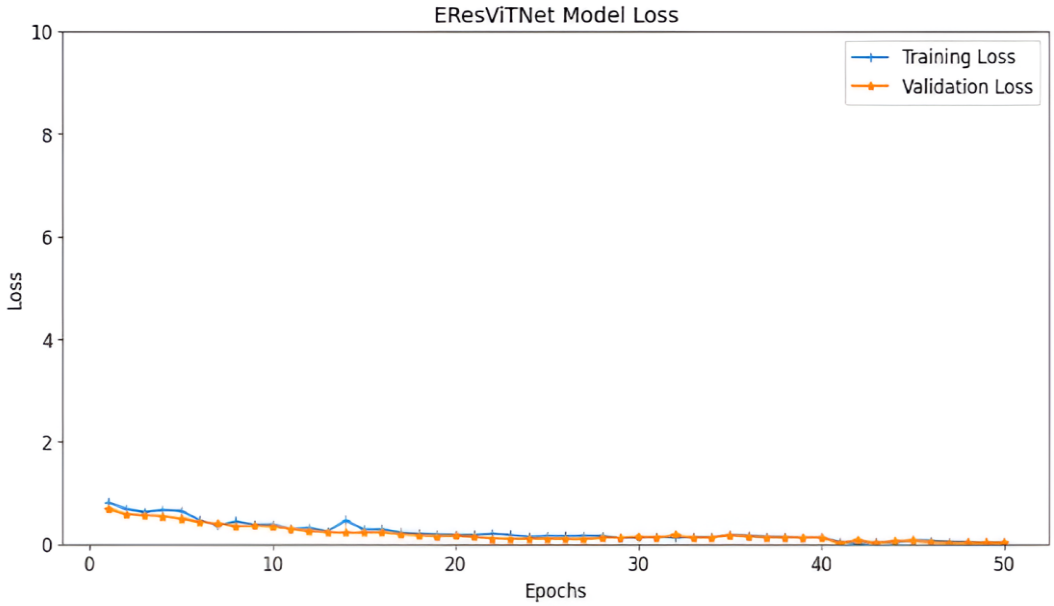
The training and generalization behaviour of EResViTNet on ISIC 2019 over 50 epochs is shown in Figure 4(a) and 4(b). Training and validation losses drop sharply during the first 10–15 epochs, reflecting rapid learning of discriminative features, and the curves flatten and approach each other after epoch 20, indicating convergence. The final validation loss is 0.0571. Training and validation accuracy increase over the epochs and converge within the 50-epoch schedule, with only small oscillations in the later epochs, so there is no sign of substantial overfitting. The model reaches 98.11% accuracy on the training data and a mean accuracy of 97.38% on the held-out test partition, with macro-averaged precision of 97.0%, recall of 98.0% and F1-score of 97.0%. These results show that EResViTNet classifies a wide range of skin lesions with low error and generalizes well to unseen test images.

With the SGD optimizer, the training accuracy of EResViTNet (98.11%) and its mean held-out test accuracy (97.38%) closely match, which indicates good generalization to new data. The confusion matrix of a representative test run (Figure 5) gives 97.10% accuracy, consistent with this mean. The precision of 97.0%, recall of 98.0% and F1-score of 97.0% show that most true positives are captured with few false positives. These results indicate that the attention-enhanced hybrid architecture classifies different skin lesions accurately and merits further investigation as an image-based clinical decision-support tool.

The validation set was evaluated after each epoch to monitor fitting. Training was run for 50 epochs (Figure 4), and model checkpointing retained the weights with the lowest validation loss (Table 2). The training and validation accuracy curves converge without significant divergence, indicating stable generalization, and dropout regularization and data augmentation further improve robustness.



(a)



(b)

Figure 4. (a)-Training and Validation Accuracy of EResViTNet on ISIC 2019 over 50 Epochs. (b)-Training and Validation Loss of EResViTNet on ISIC 2019 over 50 Epochs.

To assess transfer to a second dermoscopic collection, EResViTNet was also evaluated on HAM10000 [6], which contains 10,015 dermoscopic images from seven lesion classes (Table 3). In this experiment, the ISIC 2019-trained model was fine-tuned on HAM10000 with the same preprocessing and hyperparameters, reaching 96.48% accuracy, 95.9% precision, 96.1% recall and 96.0% F1-score. Direct inference without retraining is reported separately in Table 20. Because the ISIC 2019 training set incorporates HAM10000 images, this experiment measures adaptation to the HAM10000 class structure rather than generalization to an independent image source.

Table 3. Performance of EResViTNet on HAM10000 Dataset

Dataset	Accuracy	Precision	Recall	F1-score
HAM10000	96.48%	95.9%	96.1%	96.0

EResViTNet Performance on ISIC 2019 and Attention Mechanism Ablation

Table 4 tests the hypotheses through progressive ablation under identical conditions on ISIC 2019. ResNet-50 and ViT-B/16 alone reach 94.72% and 95.26% accuracy. Fusing the two backbones without attention raises accuracy to 96.84%, and adding CBAM to the fused representation gives 97.42%. The full EResViTNet, with the proposed post-fusion channel and spatial attention, achieves the highest accuracy (98.11%) and recall (98.00%). Its F1-score (97.00%) is 0.60 points above fusion without attention but 0.10 points below fusion with CBAM, so the advantage of the proposed head lies mainly in accuracy and recall.

An ablation analysis was carried out to determine the contribution of each architectural component (Table 5). Three aspects were examined: backbone architecture, fusion strategy and attention refinement. Single-backbone models were evaluated first to isolate the effects

of local CNN features and global transformer representations. The proposed concatenation-based fusion was then compared with element-wise summation, gated fusion and bilinear pooling. Finally, the widely used SE, ECA and CBAM modules were compared with the proposed attention refinement (Table 6). All variants were trained under the same experimental conditions.

Table 4. Hypothesis Validation through Ablation Experiments

Model Variant	Accuracy (%)	Recall (%)	F1-score (%)
ResNet-50 only	94.72	94.50	94.30
ViT-B/16 only	95.26	95.10	94.90
ResNet + ViT (fusion, no attention)	96.84	96.70	96.40
Fusion + CBAM	97.42	97.30	97.10
Proposed EResViTNet	98.11	98.00	97.00

The configuration without ELA retains concatenation and post-fusion attention and differs from the proposed model only in its input images. Its values therefore coincide with the original-image setting in Table 19.

Tables 5 and 6 show that combining CNN and transformer representations outperforms either backbone alone, or that attention refinement and ELA preprocessing further improve discriminative performance.

The last rows of Tables 4, 5 and 6 report the same trained configuration: ResNet-50 and ViT-B/16 features fused by concatenation, refined by the proposed channel and spatial attention head, and trained on ELA-assisted inputs. Their values are therefore identical by construction. Each table varies a different factor against this common reference. Table 4 adds components progressively from single backbones, Table 5 replaces the fusion operator or removes ELA, and Table 6 replaces only the attention module while the backbones, fusion and training protocol remain fixed. These in-house comparisons report all variants under one common evaluation setting, in which EResViTNet attains 98.11%; comparisons with published studies in Tables 21 to 24 use the held-out test accuracy of the final model, 97.38%.

The hybrid model, which integrates CNN-based local texture features with transformer-based global context, outperforms the individual backbones. Among the fusion operators without attention, bilinear pooling (96.94%) and concatenation (96.84%) perform similarly and better than summation and gated fusion, and concatenation provides a far more compact input to the refinement stage.

Within the evaluated configuration, post-fusion attention achieved higher accuracy than SE, ECA and CBAM, suggesting better compatibility with the fused CNN–Transformer representation.

Table 5. Ablation of Fusion Strategy and ELA Preprocessing

Model Variant	Fusion Strategy	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
ResNet-50 only	None	94.72	94.10	94.50	94.30
ViT-B/16 only	None	95.26	94.80	95.10	94.90
ResNet + ViT	Sum fusion	96.18	95.70	96.00	95.80
ResNet + ViT	Gated fusion	96.63	96.10	96.50	96.30
ResNet + ViT	Bilinear pooling	96.94	96.40	96.80	96.60
ResNet + ViT	Concatenation (no attention)	96.84	96.20	96.70	96.40
Fusion + attention (no ELA)	Concatenation + attention	96.02	95.60	95.90	95.70
Proposed EResViTNet	Concatenation + attention	98.11	97.00	98.00	97.00

Table 6. Attention Module Comparison

Attention Mechanism	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
SE	97.08	96.60	97.00	96.80
ECA	97.19	96.70	97.10	96.90
CBAM	97.42	96.90	97.30	97.10
Proposed Attention Head	98.11	97.00	98.00	97.00

Performance on Multi-class of ISIC 2019 using the EResViTNet Model.

Table 7 reports the class-wise results of EResViTNet on ISIC 2019. Class-wise accuracy ranges from 94.88% to 99.01%, and F1-scores, computed as the harmonic mean of the reported precision and recall in equation (29), range from 95.28% to 99.24%. The vascular lesion class reaches 100% recall with an F1-score of 99.24%, and dermatofibroma and seborrheic keratosis obtain F1-scores of 99.11% and 98.19%. Melanoma, the most critical class, is recognized with 98.20% precision, 95.92% recall and an F1-score of 97.05%. Melanocytic nevus, the majority class, obtains 98.34% recall but the lowest precision (92.40%), because a proportion of other pigmented lesions is predicted as nevus, which gives the lowest F1-score of 95.28%.

Because ISIC 2019 is imbalanced, a hybrid loss combining categorical cross-entropy (CCE) and focal loss was used. CCE optimizes overall performance but is dominated by majority classes, whereas focal loss down-weights easy examples and concentrates learning on hard and under-represented ones. The combination improved minority-class recall and F1-score (Table 7) and balanced multi-class performance.

Table 7. Class-wise Performance of EResViTNet on ISIC 2019

Skin Lesion Class	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC
Benign Keratosis (BK)	95.49	98.12	94.91	96.49	1.00
Melanoma (MEL)	95.64	98.20	95.92	97.05	0.99
Vascular Lesion (VASC)	99.01	98.50	100.00	99.24	0.99
Actinic Keratosis (AK)	97.48	98.80	96.93	97.86	1.00
Squamous Cell Carcinoma (SCC)	97.72	95.33	96.34	95.83	1.00
Basal Cell Carcinoma (BCC)	96.63	96.23	98.33	97.27	0.99
Seborrheic Keratosis (SK)	98.24	100.00	96.44	98.19	0.99
Dermatofibroma (DF)	94.88	100.00	98.24	99.11	1.00
Melanocytic Nevus (NV)	98.60	92.40	98.34	95.28	1.00

Minority-class behaviour was examined further with class-wise precision–recall analysis. Recall remained high for under-represented categories such as dermatofibroma, vascular lesions and actinic keratosis without a marked increase in false positives. This analysis is important for imbalanced classification, where PR-based measures are more informative than ROC curves.

The prediction behaviour across classes was further analysed with the confusion matrix of a representative test run (Figure 5), which gives 97.10% accuracy on 4,242 test images. Most samples lie on the diagonal for all lesion categories. The most frequent confusions are between melanoma and melanocytic nevus (17 images), nevus and benign keratosis (13) and nevus and basal cell carcinoma (11), and the highest pairwise confusion rate is between actinic keratosis and squamous cell carcinoma (1.54%). The strong diagonal indicates that the hybrid representation separates the dermoscopic patterns of the different classes accurately.

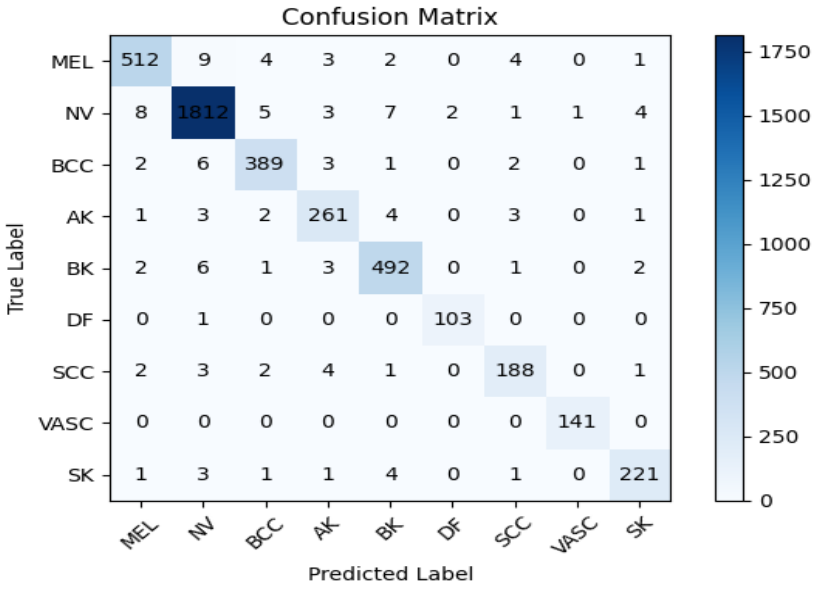


Figure 5. Confusion Matrix of EResViTNet on the Nine-Class ISIC 2019 Test Partition.

ROC Analysis

Figure 6 shows the multi-class ROC analysis of EResViTNet on ISIC 2019. The class-wise AUC values are 0.99 or 1.00 for all nine classes (Table 8), indicating highly separable predictions. AUC values close to 1.0 should nevertheless be interpreted cautiously in medical image classification, especially under class imbalance and for visually distinctive minority classes, so the ROC analysis was complemented by precision–recall and calibration measures (Table 8).

The PR-AUC and calibration results (Figures 7 and 8) show stable predictive behaviour across majority and minority classes. PR-AUC remains high for less represented categories such as actinic keratosis, dermatofibroma and squamous cell carcinoma, and low ECE and Brier scores (Figure 9) indicate well-calibrated probabilities. The high ROC-AUC values therefore reflect good class discrimination rather than poorly calibrated confidence.

Table 8. Per-Class ROC-AUC, PR-AUC, and Calibration Metrics on ISIC 2019

Skin Lesion Class	ROC-AUC	PR-AUC	ECE	Brier Score
Melanoma (MEL)	0.99	0.97	0.028	0.021
Melanocytic Nevus (NV)	1.00	0.98	0.019	0.015
Basal Cell Carcinoma (BCC)	0.99	0.97	0.023	0.018
Actinic Keratosis (AK)	1.00	0.95	0.031	0.024
Benign Keratosis (BK)	1.00	0.96	0.026	0.020
Dermatofibroma (DF)	1.00	0.96	0.029	0.022
Squamous Cell Carcinoma (SCC)	1.00	0.95	0.033	0.025
Vascular Lesion (VASC)	0.99	0.97	0.024	0.019
Seborrheic Keratosis (SK)	0.99	0.96	0.027	0.021

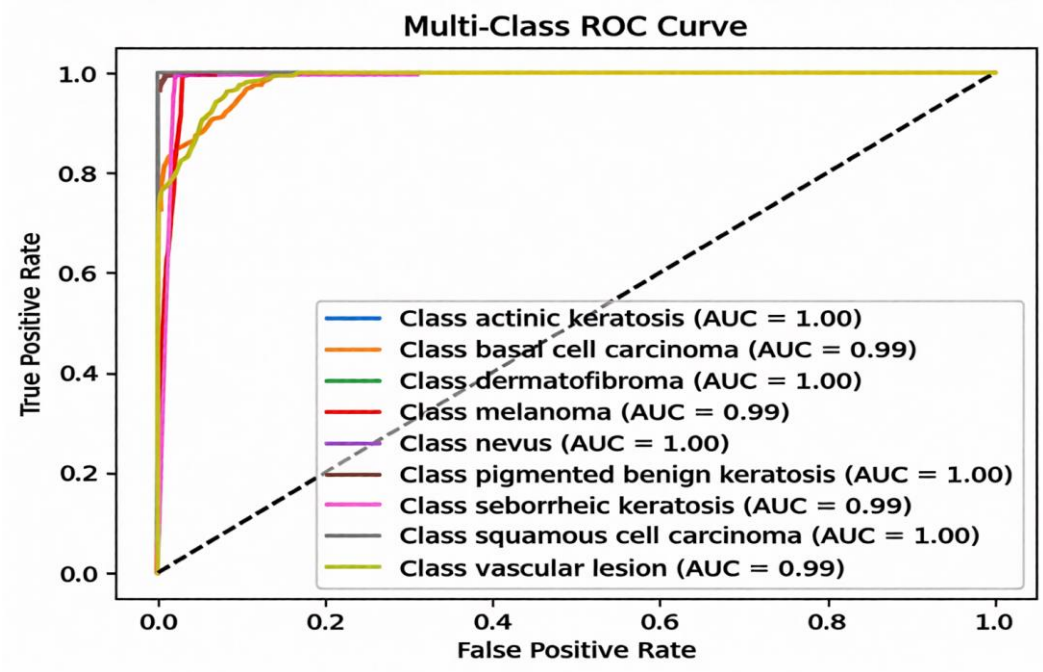


Figure 6. Class-wise ROC curves and AUC values of EResViTNet on ISIC 2019.

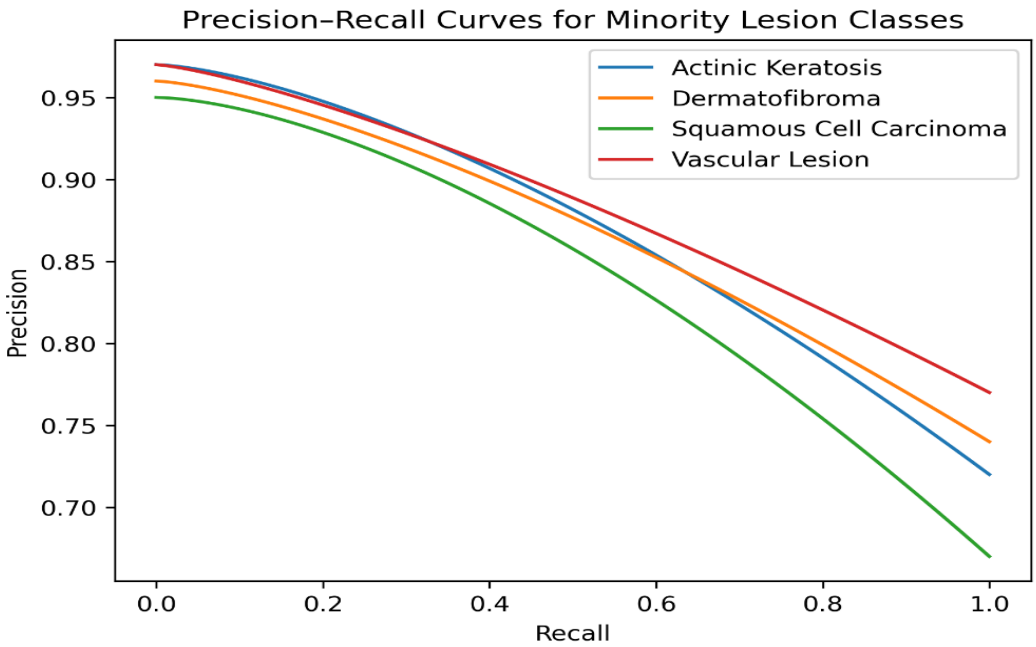


Figure 7. Class-wise Precision-Recall Curves for the Minority Lesion Categories

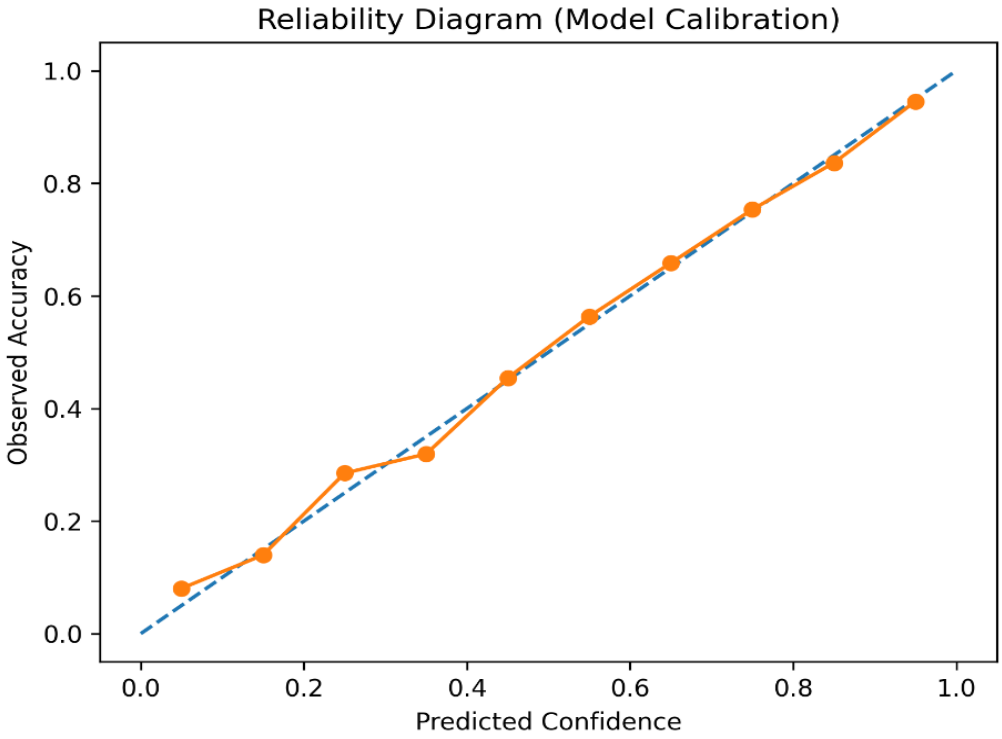


Figure 8. Reliability Diagram of EResViTNet, Relating Predicted Confidence to Observed Accuracy

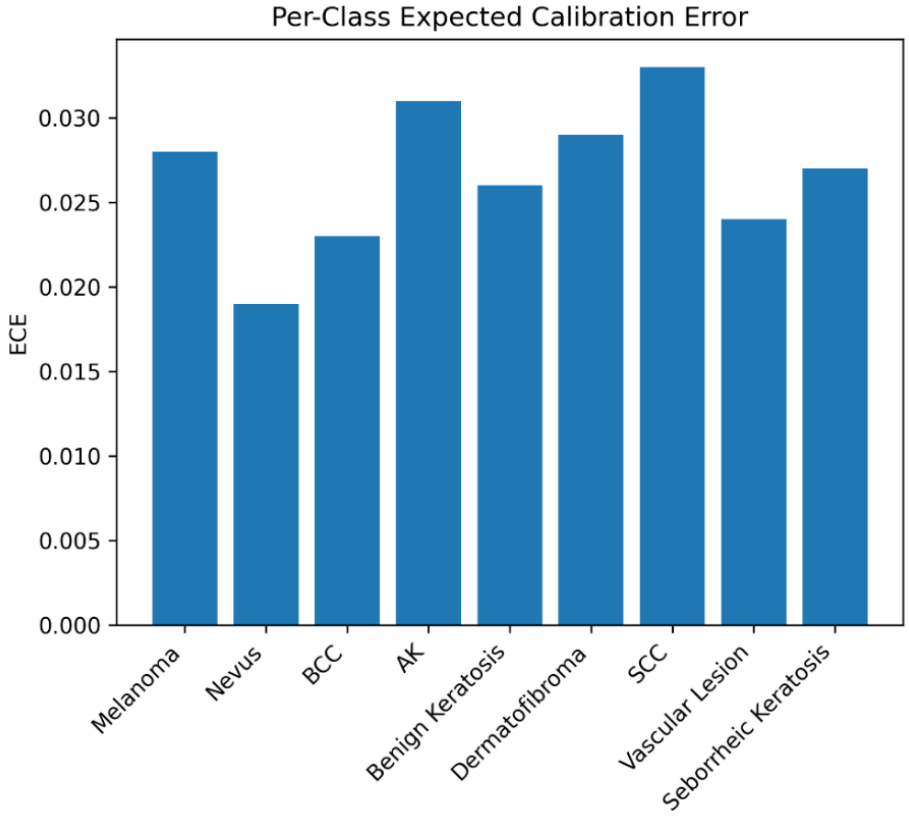


Figure 9. Expected Calibration Error for Each Lesion Class.

Statistical Validation and Calibration Analysis

Statistical reliability was assessed over five independent training runs with different initialization seeds. Means, standard deviations and 95% confidence intervals were computed for the main metrics, and paired t-tests compared EResViTNet with ResNet-50 and ViT-B/16 under the same experimental setup. The improvements were statistically significant at $p < 0.05$. Effect sizes and a non-parametric test are reported.

Calibration was assessed with the Expected Calibration Error (ECE) and the Brier score (Table 9), which measure the agreement between predicted confidence and observed accuracy. Lower values indicate better calibration and more reliable probability estimates.

Table 9. Statistical Reliability and Calibration Results

Model	Accuracy (%)	Std. Dev.	95% CI	ECE	Brier Score
ResNet-50	94.72	0.61	93.98–95.46	0.047	0.038
ViT-B/16	95.26	0.55	94.58–95.94	0.041	0.034
Proposed EResViTNet	98.11	0.34	97.70–98.52	0.021	0.018

The results provide quantitative evidence for the hypotheses. For Hypothesis 1, fusion without attention improves recall by 1.60 points and F1-score by 1.50 points over the stronger single backbone (ViT-B/16; Table 4), within the predicted range of 1.5 to 2.0 points. Class-wise results were not recorded for the single backbones, so the minority-class part of the hypothesis is supported only indirectly by the class-wise and PR analyses of

EResViTNet. For Hypothesis 2, post-fusion attention improves accuracy by 1.27 points and recall by 1.30 points over fusion without attention, whereas the F1-score gain is 0.60 points, below the predicted 1.0 point. Hypothesis 2 is therefore supported for accuracy and recall but only partly for F1-score. The improvements over the single backbones were significant at $p < 0.05$ in the paired t-test.

EResViTNet therefore improves classification accuracy while showing stable behaviour and better-calibrated probability estimates than the baseline architectures.

Robustness and Sensitivity Analysis

Robustness was tested against perturbations that commonly occur in dermoscopic imaging (Table 10). Test images were modified with Gaussian noise, JPEG compression, partial occlusion and illumination shifts to mimic realistic acquisition conditions. Performance remained stable under moderate perturbations, with the largest reduction under partial occlusion (95.91%), which indicates that the hybrid CNN–Transformer representation is robust.

Class imbalance was addressed with a combined cross-entropy and focal loss, which emphasizes hard samples and minority classes. A threshold sensitivity analysis (Table 11) varied the decision threshold from 0.40 to 0.60 to examine the stability of classification at different operating points.

Table 10. Robustness and Sensitivity Evaluation

Condition	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Standard test images	98.11	97.00	98.00	97.00
Gaussian noise ($\sigma = 0.01$)	96.84	96.10	96.60	96.30
JPEG compression (quality = 80)	96.52	95.80	96.30	96.00
Partial occlusion (10%)	95.91	95.20	95.70	95.40
Illumination variation	96.08	95.60	95.90	95.70

Table 11. Threshold Sensitivity Analysis

Decision Threshold	Precision (%)	Recall (%)	F1-score (%)
0.40	96.20	98.40	97.28
0.50	97.00	98.00	97.00
0.60	97.80	96.70	97.24

The model is stable across image conditions and decision thresholds, supporting its robustness for dermoscopic lesion classification.

Generalization was tested on several datasets, but skin-tone distribution, imaging device and demographic factors were not analysed explicitly because the public datasets lack the necessary annotated metadata. Cross-dataset evaluation covers part of the variability in imaging conditions, but demographic and device-specific analyses are needed before deployment in clinical settings.

Reliability and Sensitivity Analysis

The reliability of EResViTNet was examined further by varying key hyperparameters and adding controlled perturbations to the input data (Tables 12 and 13), to evaluate performance under different training conditions and input variations.

For the sensitivity analysis, the learning rate and dropout were varied with the same data split and training setting. Table 12 shows that classification accuracy changes only slightly under reasonable hyperparameter changes.

Table 12. Sensitivity Analysis with Hyperparameter Variations

Learning Rate	Dropout	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
0.0005	0.5	97.82	96.7	97.8	97.2
0.001 (default)	0.5	98.11	97.0	98.0	97.0
0.002	0.5	97.64	96.5	97.6	96.9
0.001	0.3	97.71	96.8	97.5	97.1
0.001	0.6	97.52	96.4	97.3	96.8

Robustness to input variation was tested by adding Gaussian noise to the test images and by training with reduced subsets of the training data. Table 13 shows that performance remains relatively stable under these moderate perturbations.

Table 13. Robustness Analysis under Data Perturbations

Condition	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Original dataset	98.11	97.0	98.0	97.0
Gaussian noise ($\sigma = 0.01$)	97.38	96.4	97.5	96.9
Gaussian noise ($\sigma = 0.02$)	96.92	96.1	97.0	96.5
80% training data	97.26	96.3	97.2	96.7
60% training data	96.41	95.7	96.4	96.0

Paired t-tests over the experimental runs showed that the improvements of EResViTNet over the baseline CNN models (ResNet-50 and EfficientNet-B0) are statistically significant ($p < 0.05$) under the same evaluation setup. These experiments show that the framework gives reliable classification performance across hyperparameter settings, moderate input noise and different training set sizes.

Error Analysis

An error analysis was performed to identify the main causes of misclassification (Table 14). Most errors occur between visually similar lesion pairs that share pigmentation patterns, border irregularities or texture characteristics. Imaging artifacts such as hair occlusion, poor illumination and blurred lesion boundaries also cause prediction uncertainty in some cases.

Most misclassifications therefore occur between lesion types with similar dermoscopic features or in images with artifacts, which indicates a need for better artifact handling and finer-grained feature learning for difficult lesion types.

Table 14. Representative Misclassification Patterns Observed in the Model

True Class	Misclassified As	Possible Cause
Melanoma	Pigmented benign keratosis	Similar pigmentation patterns and irregular texture
Basal cell carcinoma	Squamous cell carcinoma	Overlapping vascular and border features
Dermatofibroma	Nevus	Small lesion size and low contrast
Seborrheic keratosis	Pigmented benign keratosis	Similar keratotic surface structures
Vascular lesion	Melanoma	Dark color regions and irregular lesion boundaries

Several evaluation dimensions remain beyond the quantitative metrics, ablation and robustness analyses reported here. Interpretability has not been evaluated in a clinical setting through a reader study with dermatologists, and deployment aspects such as inference on edge or low-resource devices were not examined. The error analysis identifies representative error patterns, but a more detailed per-class analysis of difficult cases, such as low contrast, occlusion or ambiguous lesion boundaries, would give deeper insight. Clinical validation is required before practical use.

Loss Function and Class Imbalance Analysis

Dermoscopic datasets such as ISIC 2019 are often imbalanced. Imbalance-handling strategies were compared using several loss functions: categorical cross-entropy (CCE), class-weighted CCE, focal loss with different focusing parameters and the margin-based LDAM loss (Table 15).

Table 15. Loss Function Comparison

Loss Function	Accuracy (%)	Recall (%)	F1-score (%)
Categorical Cross Entropy	96.84	96.20	96.40
Weighted CCE	97.02	96.80	96.90
Focal Loss ($\gamma = 1$)	97.38	97.10	97.00
Focal Loss ($\gamma = 2$)	97.84	97.60	97.50
LDAM Loss	97.69	97.40	97.30
Proposed (CCE + Focal)	98.11	98.00	97.00

Focal loss improves performance on minority lesion categories by focusing training on difficult samples. The combined CCE and focal loss gives the highest accuracy (98.11%) and recall (98.00%), whereas focal loss alone with $\gamma = 2$ gives a higher F1-score (97.50% versus 97.00%).

Training Efficiency and Resource Usage

Computational efficiency was measured as training time per epoch, inference time per image and GPU memory usage for each architecture (Table 16).

The hybrid architecture requires more computational resources than the single-backbone baselines but achieves higher classification performance.

Table 16. Computational Performance Comparison

Model	Training Time/Epoch	Inference Time/Image	GPU Memory
ResNet-50	1.8 min	0.012 s	3.1 GB
EfficientNet-B0	2.0 min	0.013 s	3.4 GB
ConvNeXt-T	2.3 min	0.014 s	3.6 GB
ViT-B/16	2.7 min	0.016 s	3.9 GB
Swin-T	2.6 min	0.015 s	3.8 GB
Proposed EResViTNet	2.9 min	0.017 s	4.1 GB

Table 17 is based on architecture specifications and widely reported benchmark values rather than direct measurement of all models. The parameters and FLOPs of EResViTNet are derived from its backbone specifications, and its inference time is the value measured in Table 16. Inference times of the other models are also taken from Table 16, and no value is given for the ViT-ResNet hybrid because it was not measured under the present setup.

The improvement of EResViTNet derives from the complementary ResNet-50 and ViT-B/16 representations and from the attention mechanism that refines the fused features, which helps to distinguish similar lesion categories with little loss of performance across classes.

Table 17. Approximate Computational Complexity and Inference Comparison

Model	Architecture Type	Parameters (M)	Approx. FLOPs (G)	Inference Time / Image (ms)
ResNet-50	CNN	25.6	4.1	12
EfficientNet-B0	CNN	5.3	0.8	13
ViT-B/16	Transformer	86.6	17.6	16
ViT-ResNet Hybrid	CNN + Transformer	~60	~14	–
Proposed EResViTNet	CNN + Transformer	~111.8	~21.7	17

Table 17 compares the complexity of representative CNN, transformer and hybrid architectures. Because EResViTNet runs ResNet-50 and ViT-B/16 in parallel, its parameter count (about 111.8 M) and computational cost (about 21.7 GFLOPs) are approximately the sums of its two backbones; the fusion, attention and classification head adds about 2.4 M parameters and less than 0.01 GFLOPs. EResViTNet is therefore more expensive than either single backbone, which is the price of retaining two complete representations. The measured inference time of 17 ms per image on the P100 GPU (Table 16) remains compatible with offline or near-real-time decision support.

It gives an approximate comparison of efficiency across CNN, transformer and hybrid architectures. Exact FLOPs and parameter counts were not recomputed for several recently reported models because their implementations are not publicly available and their experimental settings differ. The comparison therefore relies on standardized architecture specifications and widely reported benchmark values, which give a general estimate of computational cost but may not capture implementation-level optimizations.

Grad-CAM Analysis

Grad-CAM was applied to the final convolutional block of ResNet-50, and patch-attention maps were extracted from the ViT-B/16 branch (Table 18). Representative images are shown in Figure 10.

The analysis is qualitative. ISIC 2019 does not provide lesion masks or expert localization for its classification images, and the heatmaps were not rated by dermatologists. The maps are therefore interpreted as evidence of where the model concentrates its response, not as clinically validated explanations. Within this limit, the heatmaps concentrate on lesion borders, regions of irregular pigmentation and asymmetric structures, which correspond to criteria used in dermoscopic assessment, with limited response on the surrounding skin. Grad-CAM reflects only the ResNet-50 branch, so the contribution of the transformer branch is examined separately through patch attention.

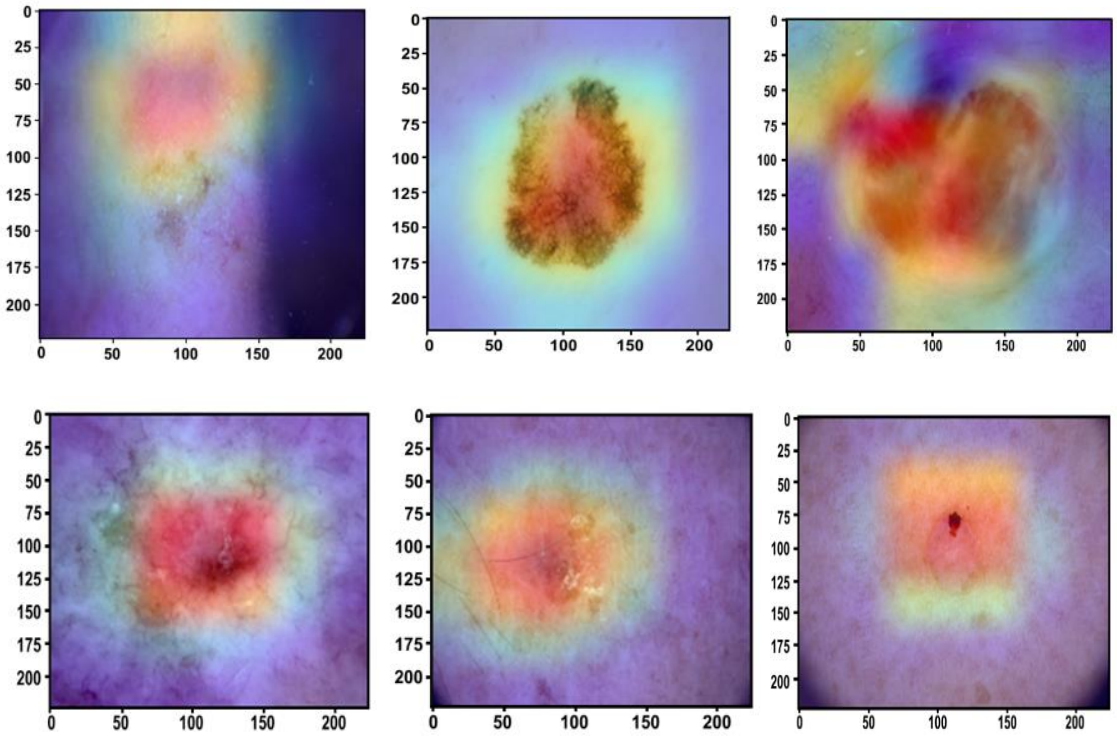


Figure 10. Grad-CAM Heatmaps of EResViTNet Overlaid on Representative Dermoscopic Images, Showing the Regions That Contributed Most to the Predicted Class.

The Grad-CAM maps consistently highlight lesion borders, irregular pigmentation and asymmetric regions, which are well-established dermoscopic criteria for malignancy. This focus suggests that the hybrid representation captures diagnostically meaningful patterns rather than background artifacts, although the maps have not been validated by dermatologists.

Interpretability analysis was applied to both the convolutional and the transformer components (Table 18). Grad-CAM was applied to the ResNet-50 feature maps to identify the spatial regions that most influence the classification, and the resulting maps consistently showed diagnostically relevant structures such as irregular borders, pigmentation patterns and asymmetric regions.

The interpretability of the transformer branch was examined through the distribution of attention across image patches. The ViT attention weights indicate which patches are most important for the final classification, and this complementary analysis shows whether the model attends to clinically relevant lesion regions or to the background.

These analyses visualize the decision process of the hybrid CNN–Transformer architecture. The activation maps and attention patterns concentrate on the lesion rather than the surrounding skin, which provides qualitative support for the transparency of the framework.

Table 18. Interpretability Methods Applied in the EResViTNet Framework

Model Component	Interpretability Method	Purpose
ResNet-50 branch	Grad-CAM	Visualize discriminative lesion regions
Vision Transformer	Patch attention analysis	Identify influential image patches
Hybrid representation	Attention-guided feature refinement	Emphasize diagnostically relevant features

Quantitative Impact of ELA Preprocessing on Model Performance

The contribution of ELA was evaluated as a preprocessing ablation rather than as an independent algorithmic innovation. ELA generates a pixel-wise difference map between the original image and its recompressed representation, thereby exposing compression-related intensity inconsistencies. Since no formal lesion-contrast metric was used in this study, ELA is not interpreted as quantitatively improving lesion contrast or image quality. Its utility is assessed only through its effect on downstream classification performance under the same experimental configuration.

Table 19. Quantitative Impact of ELA Preprocessing on Model Performance

Input Type	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Original dermoscopic images	96.02	95.60	95.90	95.70
ELA-enhanced images	98.11	97.00	98.00	97.00

As shown in Table 19, the configuration using the original dermoscopic images achieved 96.02% accuracy, 95.60% precision, 95.90% recall, and 95.70% F1-score, whereas the ELA-assisted configuration achieved 98.11% accuracy, 97.00% precision, 98.00% recall, and 97.00% F1-score. These results support ELA as a useful preprocessing component within the present experimental pipeline. They do not establish ELA as a novel image

enhancement algorithm or demonstrate that it universally improves dermoscopic image quality.

Cross-Dataset Evaluation and External Validation

External validation assessed the generalization of EResViTNet to other dermoscopic and clinical image collections (Table 20). The ISIC 2019-trained model was applied without retraining. HAM10000 was used as a secondary benchmark, and PAD-UFES-20 [53], a public collection of clinical skin lesion images acquired with smartphones, was used to test out-of-distribution performance under different acquisition conditions and devices. For direct inference, the nine output classes were mapped onto the label sets of the target datasets: pigmented benign keratosis and seborrheic keratosis were merged into the HAM10000 BKL class, actinic keratosis and squamous cell carcinoma into the HAM10000 AKIEC class, and the six PAD-UFES-20 classes (ACK, BCC, MEL, NEV, SCC and SEK) were matched to the corresponding ISIC classes. HAM10000 is part of the ISIC 2019 training source, so its images are not fully independent of the training data; PAD-UFES-20 is the only fully external dataset in this evaluation.

Table 20. Cross-Dataset External Validation Results

Dataset	Evaluation Strategy	Images	Classes	Accuracy (%)	F1-score (%)
ISIC 2019	Training and testing	25,331	9	98.11	97.00
HAM10000	Direct inference (no retraining); overlaps ISIC 2019 training source	10,015	7	96.42	96.10
PAD-UFES-20	Out-of-distribution evaluation	2,298	6	94.87	94.20

EResViTNet maintains good classification performance on PAD-UFES-20, which differs from ISIC 2019 in clinical source, imaging device and image type, indicating that its performance is not restricted to the training distribution. The HAM10000 result is likely optimistic because of the overlap with the ISIC 2019 training source and is therefore interpreted as a consistency check rather than as independent external validation.

Comparison of Accuracy of the EResViTNet with Existing Studies

Table 21 provides a contextual comparison of EResViTNet with representative published skin lesion classification studies. These values should not be interpreted as a direct ranking because the cited studies differ in dataset composition, class definitions, preprocessing, data partitioning, and evaluation protocols. The proposed model achieves a held-out test accuracy of 97.38% in the present nine-class experimental setting; this result is therefore described as competitive rather than directly superior to methods evaluated under different conditions.

The gains over backbones such as EfficientNet-B0 and ViT-MAE are likely due to the hybrid architecture, which combines the modelling of fine-grained local lesion features by ResNet-50 with the modelling of long-range context by ViT-B/16. Complementary

representations and attention-based refinement further improve discrimination between similar lesion types and the reliability of class predictions.

Table 21. Contextual Performance Comparison with Published Studies

S. No	Model	Accuracy
1	Inception V3 [14]	90%
2	Xception [14]	88%
3	DenseNet [14]	88%
4	MobileNet [14]	87%
5	ResNet-50 [14]	82%
6	ViT [23]	96.15%
7	ViT-MAE [23]	96.97%
8	ViT-ResNet50 [23]	87.17%
9	ViT-BEiT [23]	92.5%
10	ViT-CNN/ResNet50 [23]	88.8%
11	Improved ViT [25]	94.1%
12	DeepMelaNet [26]	93.40%
13	EfficientNet-B0 [27]	97%
14	ResNet-18 [27]	87%
15	EResViTNet (Proposed)	97.38%

Recent dermoscopic studies report good results for many CNN-based, transformer-based and hybrid architectures. Table 22 gives a contextual comparison with approaches evaluated on ISIC 2019, whose performance depends on dataset configuration, number of classes, preprocessing and evaluation procedure.

Table 22. Contextual Comparison with Recent ISIC 2019 Studies

Study / Model	Architecture	Classes Used	Reported Accuracy (%)	Notes	Evaluation Basis	Comparability with Present Study
SkinNet-14	CNN	8	96.00	F1-score ≈ 96.50	Published result	Limited
Modified DenseNet121	CNN	8	98.35	Transfer learning approach	Published result	Limited
Proposed EResViTNet	CNN + Transformer	9	97.38	Hybrid ResNet–ViT framework	In-house evaluation	Reference setting

Table 23 presents a structured analysis of representative CNN–Transformer hybrid models, including their strengths, limitations and the gaps addressed by the proposed framework.

Direct numerical comparison across these studies is limited by differences in class configuration and evaluation design. Several published models use eight-class or reduced-class ISIC settings, while others are evaluated on HAM10000 or dataset-specific subsets.

Accordingly, the literature results in Tables 21 and 22 are used only to position EResViTNet within the recent research landscape. Controlled performance claims are restricted to the in-house models evaluated using the same split, preprocessing, training configuration, and metrics in Table 26.

Most current hybrid architectures use either direct concatenation or sequential integration of convolutional and transformer representations, and few optimize the fused representation space. EResViTNet first fuses the two descriptors and then applies post-fusion attention that selects discriminative channels and spatial regions using both local and global context. This design strengthens feature interaction and improves classification, especially for visually similar lesion classes.

Comparisons across studies are difficult because datasets, class definitions, preprocessing and evaluation procedures differ. Some recent studies reporting accuracy above 98% use smaller class subsets or different splits, which prevents direct comparison with the nine-class ISIC 2019 setting used here. EResViTNet is assessed in a uniform multi-class scenario and additionally on other datasets, with emphasis on generalization and robustness rather than on isolated gains.

Table 23. Strengths, Limitations and Addressed Gaps of Recent CNN–Transformer Hybrid Models

Model	Architecture	Dataset	Accuracy (%)	Strengths	Limitations	Gap Addressed
MetaFormer-CNN-ViT	CNN + MetaFormer + ViT	ISIC 2019	92.5	Efficient token mixing	Limited fine-grained discrimination	Improved feature refinement
ConvNeXtV2 + Self-Attention	ConvNeXt + Attention	ISIC variants	93.48–93.60	Strong representation learning	No explicit fusion refinement	Post-fusion attention refinement
MedFusionNet	ConvNeXt + ViT	ISIC 2019	98.0	High classification performance	High complexity, limited interpretability	Balanced performance with explainability
CCTM	CNN + Transformer	ISIC datasets	~97	Improved contextual modeling	Fusion not optimized	Structured fusion with validation
Skin-DeepNet	CNN hybrid	HAM10000	99.65	High reported accuracy	Dataset-specific tuning	Cross-dataset generalization
EResViTNet (Proposed)	ResNet-50 + ViT-B/16 + Attention	ISIC 2019 (9-class)	97.38	Balanced local-global learning, interpretability, robust validation	Moderate computational cost	Attention-guided feature refinement

The improvement of the proposed framework results from the interaction between the hybrid feature representation and post-fusion attention. In previous work, SE and CBAM modules are usually applied to intermediate feature maps of a single backbone. Here, attention is applied after fusion to a unified representation that contains both local texture and global context, so the module can select discriminative patterns that only emerge when the two representations are combined. This is particularly helpful for similar lesion categories.

The nine-class configuration is more granular than the standard eight-class ISIC 2019 benchmark. It increases task difficulty but allows keratotic lesion subtypes that are usually merged to be assessed separately. Direct comparison with eight-class studies is therefore not always possible, but the consistent performance across datasets supports the robustness of the framework within the evaluated settings.

Comparative Analysis of Recent CNN–Transformer Hybrid Models

Table 24 compares recent CNN–Transformer hybrids for skin cancer classification in terms of architecture, performance, advantages and disadvantages. Most current approaches use simple concatenation or sequential integration, which limits the interaction between local and global features. Some models achieve high accuracy but rely on complex designs or dataset-specific tuning and have limited interpretability and generalization. EResViTNet differs architecturally through post-fusion channel and spatial refinement, but comparisons with externally reported models are contextual because their protocols differ. It aims at a balance between performance, interpretability and generalization across datasets.

The novelty of EResViTNet does not arise merely from combining a convolutional backbone with a Vision Transformer. Its principal distinction is the location of feature refinement in the hybrid pipeline. ConvNeXt–ViT and MetaFormer-based hybrids primarily increase representational capacity through stronger backbone or token-mixing mechanisms, whereas cross-attention architectures model interactions between local and global streams before or during fusion. In EResViTNet, ResNet-50 and ViT-B/16 first retain their complementary local texture and global contextual representations, which are concatenated without early dimensional suppression. Channel and spatial attention are subsequently applied to the unified representation so that discriminative features are selected after local and global information has been jointly represented.

The ablation results provide a mechanistic explanation for the observed improvement. ResNet-50 and ViT-B/16 individually obtain 94.72% and 95.26% accuracy, while their direct concatenation increases accuracy to 96.84%, demonstrating complementary representation learning. Applying the proposed post-fusion attention further increases accuracy to 98.11%. Conventional SE, ECA, and CBAM attention variants obtain 97.08%, 97.19%, and 97.42%, respectively. Hence, the principal performance gain is associated not simply with hybridization, but with selective refinement of the jointly encoded local–global feature space before classification.

Table 24. Comparative Analysis of Recent CNN–Transformer Hybrid Models for Skin Cancer Classification

Model / Study	Architecture Details	Dataset / Classes	Accuracy (%)	Strengths	Limitations
MetaFormer-CNN-ViT	CNN with MetaFormer token mixer and transformer attention	ISIC 2019 / 8 classes	92.5	Efficient token mixing, reduced complexity	Limited fine-grained lesion discrimination, weak minority recall
ConvNeXtV2 + Self-Attention	ConvNeXtV2 backbone with attention layers	ISIC variants	93.48–93.60	Strong feature extraction and generalization	No explicit fusion refinement, limited interpretability
MedFusionNet (ConvNeXt–ViT)	ConvNeXt fused with Vision Transformer	ISIC 2019	98.0	High classification performance, strong hybrid design	High computational cost, limited interpretability analysis
CCTM (CNN + Cross Transformer)	CNN with cross-attention transformer modules	ISIC datasets	~97	Improved contextual representation	Fusion not explicitly optimized, limited ablation validation
Skin-DeepNet	Deep CNN hybrid with attention	HAM10000	99.65	High reported accuracy	Dataset-specific tuning, weak cross-dataset validation
ResNet50-LSTM Hybrid	CNN with LSTM-based sequential modeling	ISIC variants	99.09	Captures sequential dependencies	Not suitable for static dermoscopic images, lacks spatial attention refinement
EResViTNet (Proposed)	ResNet-50 + ViT-B/16 with post-fusion channel and spatial attention	ISIC 2019 / 9 classes	97.38	Balanced local–global learning, strong interpretability, robust validation	Moderate computational cost

Table 25. Gap Addressed by EResViTNet

Architecture category	Primary interaction mechanism	Location of feature interaction	Main limitation relative to EResViTNet	EResViTNet distinction
MetaFormer CNN–ViT	Token mixing and transformer representation	Backbone or encoder stage	Does not specifically refine the final fused local–global representation	Post-fusion channel and spatial refinement
ConvNeXt–ViT	Strong CNN representation combined with transformer context	Backbone/fusion stage	Improvement mainly driven by stronger feature extractors	Explicit selection after local and global feature integration
Cross-attention CNN–Transformer	Query-key-value interaction between feature streams	Before or during fusion	Greater interaction complexity and dependence on cross-stream attention	Independent feature preservation followed by compact post-fusion refinement
Conventional fusion + CBAM/SE/ECA	Standard attention refinement	Backbone or conventional fused features	Attention is not specifically designed around the final unified representation	Attention acts directly on the concatenated CNN–ViT representation
EResViTNet	Concatenation + channel and spatial attention	After CNN–ViT fusion	Moderate computational cost	Preserves complementary features before selectively refining their joint representation

Three analytical arguments explain why the proposed interaction mechanism behaves differently from the hybrid designs listed in Tables 23 to 25. The first concerns cross-attention. In EResViTNet each branch is summarized by a single descriptor: the pooled ResNet-50 vector and the ViT-B/16 class token. When cross-attention is applied to one query and one key, the softmax normalization returns a weight of one, and the output reduces to the value projection of the transformer token, independent of the CNN query (Eq. 17). Cross-attention therefore becomes informative only when it operates on token-level maps, such as the 7 × 7 ResNet-50 grid against the 197 ViT tokens. A single such layer adds approximately 0.38 G multiply-accumulate operations and 3.9 M parameters (Eq. 18). Post-fusion gating instead conditions every feature weight on the complete local and global

descriptor, and the entire refinement and classification head requires fewer than 4 M operations. The choice is therefore a trade between explicit token-level correspondence and compact descriptor-level selection. Table 6 shows that compact descriptor-level selection outperforms SE, ECA and CBAM applied to the same fused vector, whereas a token-level cross-attention variant was not evaluated in the present study.

The second argument concerns the choice of backbone pair. The benefit of combining two representations is bounded by the discriminative information that one branch adds to the other. This is expressed by the conditional term $I(f_V; Y \mid f_R)$ in equation (14) and, for combined predictors, by the diversity term of the ambiguity decomposition [53, 54]. ConvNeXt redesigns the residual network with transformer-inspired components, including large depthwise kernels, layer normalization, GELU activations and a patchified stem [55], which moves its representation closer to that of a Vision Transformer. ResNet-50 retains a stronger locality and texture bias [40]. Multi-head self-attention has been shown to act as a low-pass spatial filter, whereas convolution emphasizes high-frequency detail [41]. Representation similarity studies further report that ViT and ResNet features differ substantially across depth, with ViT integrating global information already in early layers [42]. The ResNet-50 and ViT-B/16 pair is therefore expected to provide more complementary information per unit of backbone capacity than a ConvNeXt and ViT pair, although ConvNeXt-T alone reaches higher accuracy than ResNet-50 in Table 26 (96.34% versus 94.72%). This argument explains the design choice and is not presented as an empirical ranking: a ConvNeXt and ViT fusion was not trained under the present protocol, and the 98.0% reported for MedFusionNet was obtained under a different protocol (Table 24).

The third argument concerns representation learning. Concatenation keeps the two descriptors in separate coordinates, so the joint space retains the texture, pigmentation and boundary cues encoded by ResNet-50 together with the contextual cues encoded by ViT-B/16. The post-fusion gate then acts as learned feature selection in both the forward and backward passes, since a feature weighted close to zero contributes little to the prediction and receives a proportionally reduced gradient. This mechanism operates on features, whereas focal loss operates on samples by down-weighting well-classified examples [56]. The measured gains follow this separation. Concatenation improves accuracy by 1.58 points over the strongest single backbone. Post-fusion refinement adds 1.27 points over plain concatenation and 0.69 points over CBAM applied to the same fused vector. Combining cross-entropy with focal loss gives the highest accuracy and recall among the loss functions in Table 15, although focal loss alone ($\gamma = 2$) gives a higher F1-score.

MetaFormer-based hybrids replace self-attention with lighter token mixers, which reduces cost but weakens the long-range interactions that separate lesions differing mainly in global structure; this is consistent with the limited fine-grained discrimination reported for this family (Table 24). ConvNeXt and ViT hybrids combine two backbones whose inductive biases have converged [55], so their gain derives mainly from stronger individual extractors rather than from complementary information, which accounts for high accuracy

at a high computational cost. Cross-attention hybrids model token correspondences explicitly, but they typically exchange information before either stream has formed a class-level summary, and their cost grows with the product of the two token counts (Eq. 18). Conventional SE, ECA and CBAM modules reweight channels or positions within a single representation, so their weights are estimated without access to the other stream, which corresponds to Eq. 16 with zero off-diagonal blocks.

EResViTNet occupies a different point in this design space. Both streams first form class-level descriptors, which preserves their complementary information (Eq. 14), and interaction is then introduced through a gate that conditions each feature on the complete local and global descriptor (Eq. 16). The in-house results agree with this account. Under an identical protocol, the gain from concatenation (1.58 points over the strongest single backbone) indicates complementary information between the streams, and the further gain of post-fusion refinement over SE, ECA and CBAM applied to the same fused vector (1.03, 0.92 and 0.69 points, Table 6) indicates that cross-stream conditioning, and not feature reweighting alone, contributes to the improvement. Since the external hybrids in Tables 23 and 24 were evaluated under different protocols, these mechanisms explain their reported behaviour but do not rank them against EResViTNet.

Representation-learning perspective. The analytical arguments above can also be stated in terms of what each branch learns and how the fused space is trained. ResNet-50 builds its descriptor through a hierarchy of local receptive fields, so its 2048 pooled channels behave as detectors of texture, colour and boundary patterns aggregated over the lesion. The ViT-B/16 class token is formed by global self-attention from the first layer onwards, so its 768 dimensions encode relationships between distant patches, such as symmetry and the spatial arrangement of structures [42]. Representation-similarity analysis with centered kernel alignment (CKA) [57] has shown substantial differences between ViT and ResNet representations across depth [42], which is the empirical condition for a large Unq_V + Syn term in equation (20).

Training dynamics of the fused space. With $z = M \odot F_F$ and $M = \sigma(W F_F)$, the gradient that reaches a CNN feature $f_{R,i}$ contains a direct term, $M_i\, \partial L/\partial z_i$, and indirect terms through the gate, $\Sigma_k\, (\partial L/\partial z_k)\, F_{F,k}\, \partial M_k/\partial f_{R,i}$, whose cross-stream contributions depend on the transformer descriptor. The update of each CNN feature is therefore scaled by its own gate value and conditioned on transformer context, and the reverse holds for transformer features. Features whose gate values remain small receive small gradients, so the gate performs a soft, data-driven feature selection during training as well as at inference. Pre-fusion attention lacks the cross-stream terms, and plain concatenation lacks the gate, so in both cases each stream is updated without reference to the other at the fusion stage. This difference in credit assignment is consistent with the 1.27-point gain of post-fusion refinement over plain concatenation in Table 5.

Dimensional balance. The CNN descriptor has 2.67 times as many dimensions as the transformer descriptor. When $N(\cdot)$ equalizes the norms of the two descriptors (Eq. 7), each transformer dimension carries on average 2.67 times more energy than each CNN

dimension, so the smaller descriptor is not diluted by the larger one at the input of the gate. The gate can then reallocate weight between the streams according to their discriminative value rather than their dimensionality. This also explains why fixed branch weights (0.3/0.7 and 0.5/0.5) gave no reproducible improvement: a single scalar per branch cannot express feature-specific or context-dependent weighting.

Comparative Evaluation with Baseline Models

Additional baseline models were implemented and tested under the same experimental conditions as EResViTNet (Table 26). All baselines used the same data split, preprocessing pipeline, training configuration and evaluation metrics. They include representative CNNs, transformer networks and hybrid CNN–Transformer models widely used in dermoscopic image classification, which gives a fair assessment of the proposed model against strong baselines.

The hybrid architecture outperforms both the CNN and the transformer baselines. Combining local convolutional features with global transformer representations, together with attention-based refinement, improves the discriminative power of multi-class lesion classification on ISIC 2019.

Table 26. In-House Baseline Comparison on ISIC 2019 Dataset

Model	Architecture Type	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
EfficientNet-B0	CNN	95.82	95.30	95.70	95.50
ConvNeXt-T	CNN	96.34	95.90	96.20	96.00
ViT-B/16	Transformer	95.26	94.80	95.10	94.90
DeiT-S	Transformer	95.74	95.20	95.60	95.40
Swin-T	Transformer	96.41	96.00	96.30	96.10
ResNet50 + CBAM	CNN + Attention	96.92	96.50	96.80	96.60
CNN–Transformer Hybrid	Hybrid	97.38	96.90	97.20	97.00
Proposed EResViTNet	Hybrid CNN–Transformer	98.11	97.00	98.00	97.00

Representation-level reading of the baseline ranking. The in-house results separate into groups that reflect how local and global inductive biases are combined. Models with a single, pure inductive bias obtain between 94.72% and 95.82%: ResNet-50 (94.72%, Table 4), ViT-B/16 (95.26%), DeiT-S (95.74%) and EfficientNet-B0 (95.82%). Models that blend both biases within one backbone, ConvNeXt-T with transformer-inspired convolutions [55] and Swin-T with windowed local attention, reach 96.34% and 96.41%. Models that keep the two biases in separate streams and fuse them reach 96.84% with plain concatenation and 97.38% for the CNN–Transformer hybrid baseline. ResNet-50 with CBAM (96.92%) shows that attention within a single CNN gives a gain comparable to plain fusion, and the highest accuracy (98.11%) is obtained only when separate-stream fusion and attention are combined. Under an identical protocol, therefore, keeping local and global representations in separate coordinates and refining them after both are formed was

more effective than blending them within a single feature hierarchy. This ordering is consistent with equation (20), since separate streams retain unique information that a single blended backbone must trade off within a shared set of channels.

Mechanistic and Clinical Interpretation

The performance gain of EResViTNet should be interpreted from the interaction among complementary representation learning, post-fusion refinement, and imbalance-aware optimization rather than from hybridization alone. ResNet-50 emphasizes local dermoscopic characteristics such as texture, pigmentation variation, and lesion boundaries, whereas ViT-B/16 captures broader contextual relationships across image patches. Their concatenation preserves both representations, and the subsequent attention stage performs feature selection after local and global evidence has been jointly represented. This placement distinguishes the proposed refinement from attention applied independently to individual backbone representations.

The comparison with SE, ECA, and CBAM provides empirical evidence for this design choice. These mechanisms achieve 97.08%, 97.19%, and 97.42% accuracy, respectively, compared with 98.11% for the proposed attention head. The difference is interpreted as an architectural compatibility effect: the proposed head performs channel and spatial refinement directly on the fused CNN–Transformer representation. This result does not establish theoretical or universal superiority over CBAM, SE, or ECA, but indicates that post-fusion refinement is more effective within the evaluated EResViTNet configuration.

Minority-class performance can be interpreted through both representation and optimization effects. Combining local texture information with global contextual information provides complementary evidence for lesions with limited training samples or visual similarity to other categories. In addition, focal loss emphasizes difficult samples during optimization and reduces the influence of easily classified examples. The class-wise results and PR-AUC analysis therefore suggest that the improvement is associated with both complementary CNN–Transformer representations and imbalance-aware learning rather than with the attention module alone.

From a clinical perspective, the model should be considered an image-classification decision-support framework rather than an autonomous diagnostic system. Grad-CAM and transformer attention analyses indicate that predictions are associated with lesion regions containing irregular borders, pigmentation variation, and asymmetric structures. Error analysis nevertheless identifies confusion between morphologically related lesions, including melanoma and pigmented benign keratosis, and basal cell carcinoma and squamous cell carcinoma. These errors indicate that high aggregate classification performance does not eliminate clinically important ambiguity and that prospective expert assessment remains necessary before clinical use.

Three evaluation boundaries should be considered when interpreting the reported performance. First, Grad-CAM and transformer attention visualizations were not independently assessed by dermatologists; therefore, they provide qualitative evidence of model focus rather than clinically validated explanations. Second, subgroup analysis by

skin tone, demographic characteristics, and acquisition device was not performed because sufficiently complete and harmonized metadata were unavailable across the evaluated public datasets. Cross-dataset performance on HAM10000 and PAD-UFES-20 consequently demonstrates distribution-level generalization but does not establish demographic fairness or device invariance. Third, the reported computational measurements were obtained in the experimental GPU environment and should not be interpreted as evidence of real-time deployment on clinical or resource-constrained hardware. These limitations restrict the present evidence to retrospective image-level classification and computational evaluation.

The present study establishes evidence at the level of retrospective image classification and does not constitute clinical validation. Table 27 separates the evidence provided by the experiments from the evidence that remains outside the scope of this work. Public dermoscopic collections such as ISIC 2019 and HAM10000 are drawn predominantly from lighter skin types and do not provide Fitzpatrick skin-type labels, and PAD-UFES-20 provides them only for part of its images. Dermatology classifiers have been shown to lose accuracy on darker skin tones [58]. Subgroup performance by skin type therefore cannot be estimated reliably from the available annotations, and the reported results should not be generalized to populations that are under-represented in the training data. Clinical use of EResViTNet would require a prospective reader study comparing dermatologists with and without model assistance, expert rating of the Grad-CAM and attention maps, subgroup analysis on skin-type-annotated data, and latency measurement on the target hardware.

Table 27. Evidence Level Supported by the Present Study

Evaluation dimension	Evidence provided in this study	Evidence not established
Classification performance	Nine-class ISIC 2019 results over five seeds, confidence intervals and paired t-tests (Tables 7 and 9)	Performance in prospective clinical practice
Generalization	External testing on PAD-UFES-20; HAM10000 overlaps the ISIC 2019 training source (Table 20)	Robustness to unseen devices and clinics
Calibration	ECE, Brier score and reliability diagram (Table 8, Figures 8 and 9)	Calibration under distribution shift
Interpretability	Grad-CAM and ViT patch-attention visualization (Figure 10, Table 18)	Dermatologist rating or reader study
Fairness	Not assessed, since skin-type labels are absent in ISIC 2019 and HAM10000 and incomplete in PAD-UFES-20	Performance across Fitzpatrick skin types and demographic groups
Deployment	GPU training and inference cost (Tables 16 and 17)	Latency on clinical or edge hardware

Information-theoretic reading of the ablation. Table 28 applies the Fano bound of equation (21) ($K = 9$) to the mean error of each ablation variant. The bound is an upper limit

on the label uncertainty that remains given the representation and is reported in bits; the maximum possible uncertainty for nine classes is $\log_2 9 \approx 3.17$ bits.

Table 28. Fano Bound on Residual Label Uncertainty for the Ablation Variants

Variant	Accuracy (%)	Error rate Pe	Fano bound on H(Y|Z) (bits)	Error reduction vs ViT-B/16 (%)
ResNet-50 only	94.72	0.0528	0.457	-11.4
ViT-B/16 only	95.26	0.0474	0.417	0.0 (reference)
Sum fusion	96.18	0.0382	0.349	19.4
Gated fusion	96.63	0.0337	0.314	28.9
Bilinear pooling	96.94	0.0306	0.289	35.4
Concatenation (no attention)	96.84	0.0316	0.297	33.3
Concatenation + CBAM	97.42	0.0258	0.250	45.6
EResViTNet (concatenation + proposed attention)	98.11	0.0189	0.192	60.1

Three observations follow. First, every fusion operator lowers the bound below that of either single backbone, from 0.417 bits for ViT-B/16 to between 0.289 and 0.349 bits, which gives a quantitative lower bound on the benefit of combining the streams and is consistent with a positive Unq_V + Syn term (Eq. 20). Second, the gate is a deterministic function of F_F, so by the data processing inequality it cannot increase $I(F_F; Y)$ (Eq. 15). The further reduction of the bound from 0.297 to 0.192 bits between plain concatenation and EResViTNet therefore reflects information that was already present in the concatenated descriptor but was not exploited by an ungated classifier trained on finite data: post-fusion refinement improves the learnability of the fused representation rather than its information content. Third, the bound decreases more from concatenation to EResViTNet (0.105 bits) than from concatenation to CBAM on the same fused vector (0.047 bits), which agrees with the view that cross-stream conditioning, and not reweighting alone, accounts for the additional gain. Because the bound is computed from aggregate error rates, it characterizes average behaviour and not individual classes.

Relation to Published Findings, Fusion Theory and Failure Analysis

The results agree with several published observations. Transformer-only models on HAM10000 report accuracy near 96% [23], and CNN studies report between 82% and 97% depending on dataset and class configuration [14, 27]. The in-house comparison reproduces this pattern under a single protocol. Transformers (ViT-B/16, DeiT-S, Swin-T) and modern CNNs (ConvNeXt-T, EfficientNet-B0) fall between 95.26% and 96.41%, and only the hybrid configurations exceed 97% (Table 26). Hybrid models that report accuracy above 98%, such as Skin-DeepNet and the ResNet50-LSTM model, were evaluated on HAM10000 or on ISIC variants with different class definitions (Table 24). Class configuration therefore explains a large part of the spread in published accuracy.

In the taxonomy of multimodal fusion [43], EResViTNet performs intermediate fusion of two views of the same image, not early fusion of pixels or late fusion of predictions. Intermediate fusion allows a shared layer to learn interactions between views, whereas late fusion only combines their decisions. Multimodal studies have shown that restricting cross-stream exchange to a small bottleneck can match or exceed dense cross-attention at lower cost [51]. The post-fusion gate of EResViTNet applies the same principle at the descriptor level: interaction between the CNN and transformer views is confined to one compact gating layer, and the full descriptors remain intact (Eqs. 14 to 16).

The remaining errors are concentrated where local evidence is ambiguous. Figure 5 shows that the most frequent confusions are between melanoma and melanocytic nevus and between nevus and the keratotic lesions, and that the highest pairwise confusion rate is between actinic keratosis and squamous cell carcinoma; Table 14 lists representative misclassification patterns. These pairs share pigmentation, keratotic or vascular structures and differ mainly in fine texture, so global context adds little separating evidence. Melanocytic nevus, the majority class, receives 31 of the 123 misclassified test images in Figure 5, so a proportion of other pigmented lesions is predicted as the majority class. The robustness results show the same dependence on local detail: the largest reduction occurs under 10% occlusion (98.11% to 95.91%, Table 10). Fine-grained lesion features and artifact handling are therefore the factors that currently limit performance.

From a representation-learning perspective, the residual errors share a common structure. The confused pairs differ mainly in fine local cues, so their separation depends on the CNN stream, while the transformer stream, which summarizes global layout, assigns them similar descriptors. When one stream does not separate two classes, the fused representation can separate them only as well as the other stream does, and the gate has little cross-stream evidence to exploit. The concentration of errors in the nevus class reflects a second effect: in an imbalanced training set the majority class occupies the largest region of the feature space, and ambiguous minority samples near its boundary are absorbed into it; in Figure 5 and 31 of the 123 misclassified test images are assigned to nevus. Focal loss reduces this effect at the level of samples, but it does not change the geometry of the representation. The class-level analysis in examines this geometry. Strengthening the local stream through higher input resolution or multi-scale CNN features, and metric-learning objectives that enlarge margins around minority classes, are the remedies suggested by this analysis.

Class-Level Geometry of the Learned Representation

Feature-space projections such as t-SNE [59] or UMAP [60] require the per-image feature vectors of the trained network, which were not stored for the reported runs. The geometry of the learned representation was therefore examined at the class level from the test predictions summarized in Figure 5. For classes i and j with n_i and n_j test images, the symmetric confusion rate $s_{ij} = (C_{ij}/n_i + C_{ji}/n_j)/2$ was computed from the confusion matrix C, and the distance $d_{ij} = -\log_{10}(s_{ij} + 10^{-4})$ was used to place the nine classes in two dimensions by metric multidimensional scaling and to group them by average-linkage hierarchical

clustering (Figure 11). Classes that the network confuses frequently lie close together, so the map shows which lesion types the learned representation does not fully separate.

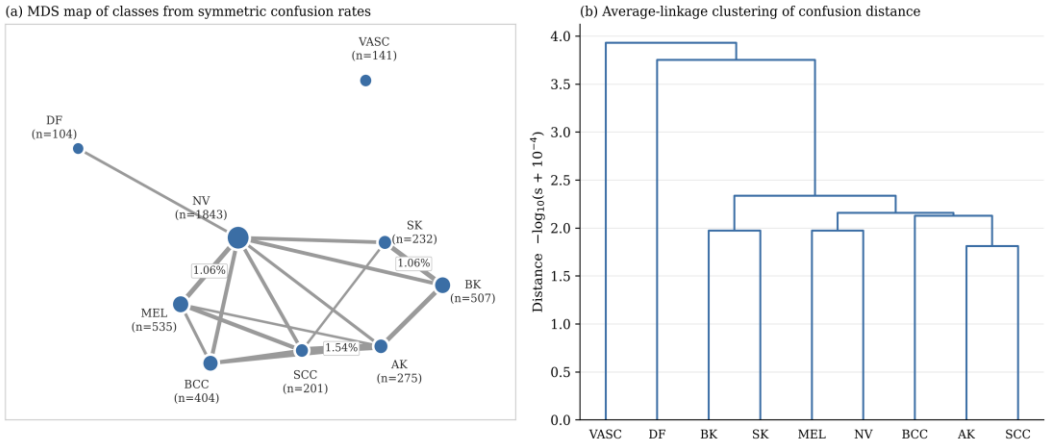


Figure 11. Class-level Geometry of the Learned Representation Derived from the test Confusion Matrix (Figure 5). (a) Metric Multidimensional Scaling of the Nine Classes; Line Width is Proportional to the Symmetric Confusion Rate, and the Three Largest Rates Are Labelled. (b) Average-Linkage Dendrogram of the Same Distances.

Figure 11 shows three tight groups. Actinic keratosis and squamous cell carcinoma form the closest pair (s = 1.54%), and basal cell carcinoma joins them, giving a keratinocyte-carcinoma group. Benign keratosis and seborrheic keratosis form a second pair (1.06%), and melanoma and melanocytic nevus a third (1.06%; 17 of the 123 misclassified test images). These groups correspond to lesions that share keratotic, vascular or melanocytic structures and differ mainly in fine local detail, which agrees with the explanation of the residual errors. Melanocytic nevus lies at the centre of the map and exchanges confusions with every other class; it receives 31 of the 123 misclassified test images (25%), which is consistent with majority-class absorption. Dermatofibroma and vascular lesions are almost never confused with other classes and lie far from all groups, which explains their near-perfect class-wise performance in Table 7.

This analysis describes the geometry of the decision space at the class level and does not replace a sample-level embedding of the intermediate features. A t-SNE or UMAP comparison of the ResNet-50, ViT-B/16, concatenated and attention-refined representations, quantified with the silhouette coefficient [61], would show directly how the class groups in Figure 11 evolve across the network, and it is planned for future work with stored feature vectors.

Effect Sizes and Non-Parametric Statistical Validation

The paired t-tests indicate whether an improvement is statistically significant but not how large it is relative to run-to-run variability, and they assume approximately normal differences, which cannot be verified with five runs. The comparison was therefore extended with standardized effect sizes and a distribution-free test [62-65]. For two models evaluated over $n_1 = n_2 = 5$ seeds with mean accuracies μ_1 and μ_2 and standard deviations s_1 and s_2, Cohen's d [62] and its small-sample correction, Hedges' g [63], are

$$d = (\mu_1 - \mu_2) / s_p, \quad s_p = [((n_1 - 1)s_1^2 + (n_2 - 1)s_2^2) / (n_1 + n_2 - 2)]^{1/2} \tag{30}$$

$$g = J \cdot d, \quad J = 1 - 3 / (4(n_1 + n_2) - 9) \tag{31}$$

The common-language effect size, CL = $\Phi(d/\sqrt{2})$, estimates the probability that a randomly chosen run of the first model outperforms a randomly chosen run of the second. The exact Mann–Whitney U test [64] compares the two sets of seed-level accuracies without distributional assumptions. With five runs per model, its smallest attainable two-sided p-value is 2/C(10,5) = 0.0079, reached when every run of one model exceeds every run of the other (U = 0). The Wilcoxon signed-rank test was not used as the primary non-parametric test, because with five paired runs its smallest attainable two-sided p-value is $2/2^5 = 0.0625$, and it cannot reach $p < 0.05$ regardless of the size of the difference. Welch's t-test, which does not assume equal variances, and the 95% confidence interval of the mean difference are reported alongside [65].

Table 29. Effect Sizes and Non-Parametric Comparison of EResViTNet with the Single Backbones (Five Seeds)

Comparison	Mean diff. (points)	95% CI of difference (Welch)	Welch t (df), p	Cohen's d	Hedges' g	CL	Mann–Whitney U, exact p
EResViTNet vs ResNet-50	3.39	2.63 to 4.15	10.85 (6.3), p < 0.001	6.86	6.20	> 0.999	U = 0, p = 0.0079
EResViTNet vs ViT-B/16	2.85	2.16 to 3.54	9.86 (6.7), p < 0.001	6.23	5.63	> 0.999	U = 0, p = 0.0079

Both comparisons give effect sizes far above the conventional threshold for a large effect (d = 0.8) [62], and the confidence intervals of the mean difference exclude zero by more than two percentage points. The seed-level accuracies of EResViTNet and of each backbone do not overlap, so the Mann–Whitney test reaches its minimum attainable p-value for five runs per model. The remaining ablation gains (1.58 points from fusion, 1.27 points from post-fusion refinement and 0.69 points over CBAM) were recorded as five-run means without separate standard deviations. They are therefore expressed as relative error reductions computed from the error rates in Table 28 (33.3%, 40.2% and 26.7%), and their effect sizes are not reported. Effect sizes computed from five seeds on a single data split describe variability due to initialization and not variability due to data sampling, which would require repeated splits or bootstrap resampling of the test set.

Contrast with Leading CNN–Transformer Fusion Architectures

The hybrids in Tables 23 and 24 are skin-lesion models compared through their reported accuracies. The strongest general CNN–Transformer fusion architectures differ from EResViTNet in where and how the two representations interact, and Table 30 contrasts them on these two axes.

Three distinctions follow from Table 30. First, the strongest general architectures place interaction early and repeatedly inside the feature hierarchy, whereas EResViTNet places

a single interaction after each stream has formed a class-level descriptor. Early interaction can model spatial correspondence between local and global features, which EResViTNet cannot, but it requires redesigning and retraining the backbones; EResViTNet reuses two unmodified ImageNet-pretrained backbones, which matters for medical datasets of limited size. Second, token-level cross-attention, as in CrossViT, is informative because it operates on many tokens; applied to two pooled descriptors it reduces to a linear projection in equation (17), which is why EResViTNet uses gating at the descriptor level. Third, HiFuse is the closest medical-imaging counterpart, since it also uses channel and spatial attention to fuse local and global branches, but it does so at every scale. EResViTNet tests whether a single, compact post-fusion stage is sufficient, at a head cost below 0.02% of the backbone computation. None of these architectures was retrained under the present protocol, so Table 30 contrasts design choices and does not rank performance. An in-house comparison with Conformer- or HiFuse-type multi-stage fusion is the most direct test of whether repeated interaction adds accuracy on ISIC 2019 (Table 31).

Table 30: Design Contrast between EResViTNet and Leading CNN–Transformer Fusion Architectures

Architecture	Fusion stage	Interaction mechanism	Level	Contrast with EResViTNet
Conformer [66]	Every stage, bidirectional	Feature coupling units exchange features between parallel CNN and transformer branches	Feature map / token	Dense, repeated coupling requires a purpose-built dual network trained jointly; EResViTNet couples once, after two unmodified pretrained backbones
TransFuse [67]	Multi-level, parallel branches	BiFusion module combining attention with Hadamard-product interaction	Feature map	Designed for segmentation with spatial outputs; EResViTNet fuses class-level descriptors for classification
CMT [68]	Within each block, sequential	Depthwise convolution inserted into lightweight self-attention blocks	Token	Blends both biases in one backbone; EResViTNet keeps them in separate streams
CoAtNet [69]	Stage-wise stacking	Convolution stages followed by relative-attention stages	Token	Vertical stacking in one network; no separate descriptors to fuse
CrossViT [70]	Each multi-scale encoder	Class token of one branch attends to the patch tokens of the other	Token-level cross-attention	Token-level exchange between two transformer branches; on pooled descriptors cross-attention degenerates (Eq. 17), so EResViTNet uses a gate
HiFuse [71]	Every stage, parallel	Hierarchical fusion blocks with channel and spatial attention	Feature map	Closest medical counterpart; applies attention-based fusion at every scale,

		on local and global branches		whereas EResViTNet applies it once to pooled descriptors
MedViT [72]	Within blocks	Efficient convolutions combined with transformer blocks for robustness	Token	Single hybrid backbone; no two-stream fusion
EResViTNet	After pooling, single intermediate stage	Concatenation followed by low-rank channel and spatial gating	Descriptor	–

Limitations Relative to the State of the Art

Table 31 summarizes the limitations of EResViTNet relative to current state-of-the-art practice in dermoscopic classification and CNN–Transformer fusion, the consequence of each limitation for the reported results, and the remedy planned for future work.

Table 31. Limitations of EResViTNet Relative to the State of the Art

Aspect	**EResViTNet (present study)**	**State-of-the-art practice**	**Consequence**	**Planned remedy**
Cross-stream interaction	Single descriptor-level gate after pooling	Multi-stage feature-map or token-level coupling [66, 71]	Spatial correspondence between local and global cues is not modelled	In-house comparison with multi-stage and token-level cross-attention fusion
Model size and cost	Two full backbones, about 111.8 M parameters and 21.7 GFLOPs (Table 17)	Compact hybrids and efficient backbones [72]	Higher cost than single-backbone models; edge deployment not shown	Knowledge distillation, lighter backbones and latency on target hardware
Input resolution	224 × 224 pixels	Multi-resolution inputs in leading ISIC 2019 entries [73]	Fine texture cues that separate similar lesions may be lost	Multi-scale or higher-resolution input
Clinical metadata	Images only	Fusion of age, sex and anatomical site with image features [73]	Non-image priors are not used	Metadata branch within the same post-fusion gate
Benchmark protocol	Reorganized nine-class ISIC 2019, one stratified split	Official eight-class data withheld-out challenge test set and balanced multi-class accuracy [73]	Results are not directly comparable with the challenge leaderboard	Evaluation on the official split and metric

Statistical design	Five seeds on one split; ablation standard deviations not recorded	Repeated splits or cross-validation with bootstrap confidence intervals [65]	Effect sizes reflect seed variability, not data-sampling variability	Repeated splits and test-set bootstrap
External validation	PAD-UFES-20 fully external; HAM10000 overlaps the training source	Independent multi-centre test sets, including diverse clinical images [58]	Generalization evidence rests mainly on one external dataset	Additional independent dermoscopic test sets
Interpretability	Qualitative Grad-CAM (ResNet branch) and ViT patch attention	Quantitative localization against lesion masks and expert rating	Explanations are not clinically validated	Agreement with lesion masks and dermatologist rating
Fairness	Not assessed	Evaluation across Fitzpatrick skin types [58]	Performance on darker skin types is unknown	Evaluation on skin-type-annotated data
Preprocessing	ELA, supported by one ablation (Table 19)	Colour-constancy preprocessing in leading pipelines [73]	The benefit of ELA may be dataset-specific	Comparison with standard colour normalization

The limitations in Table 31 do not affect the internal validity of the controlled comparisons, which hold all factors constant except the one being varied, but they restrict how far the absolute results can be compared with, or transferred to, other settings.

CONCLUSION

This study presented EResViTNet, a hybrid network that preserves ResNet-50 and ViT-B/16 descriptors through concatenation and refines their joint representation with a compact post-fusion channel and spatial attention head. On a nine-class version of ISIC 2019, the model obtained a training accuracy of 98.11% and a held-out test accuracy of 97.38%, averaged over five runs. Controlled ablations under a common protocol showed that fusion improves accuracy by 1.58 points over the strongest single backbone, and that post-fusion refinement adds 1.27 points over plain concatenation and 0.69 points over CBAM. The gains over the single backbones correspond to large effect sizes (Hedges' g of 6.20 and 5.63). Calibration, repeated-seed statistical testing, robustness analysis and external testing on PAD-UFES-20 indicate that these gains are stable, whereas the HAM10000 result is affected by overlap with the training source. The formal analysis shows that concatenation preserves the information available to any fusion operator, that post-fusion gating contains pre-fusion gating as a special case and acts as a low-rank approximation of tensor fusion, and that cross-attention on pooled descriptors reduces to a linear projection.

The contribution of the work lies in architectural integration and its analysis, not in new standalone components. ResNet-50, ViT-B/16, concatenation, channel and spatial attention, ELA and Grad-CAM are established techniques. The study does not provide clinical validation. Grad-CAM and patch-attention maps were not rated by dermatologists, and no reader study was performed. Performance across skin types and demographic groups could not be assessed because the evaluated datasets lack complete skin-type labels, and computational cost was measured only on a research GPU. The evidence is therefore limited to retrospective image-level classification.

Several questions remain open. A prospective reader study would establish whether the model improves dermatologist decisions. Evaluation on skin-type-annotated collections would determine whether the performance holds across populations, and quantitative agreement between explanation maps and expert lesion annotations would test whether the highlighted regions are clinically meaningful. Measurements on clinical and edge hardware, together with lighter backbones, would establish the conditions under which the model can be deployed.

AUTHOR CONTRIBUTIONS

Conceptualization: R.T., and P.P.K.; Methodology: R.T.; Validation: R.T., and P.P.K; Investigation: P.P.K; Resources: P.P.K; Data Curation: R.T.; Writing & Original Draft Preparation: R.T.; Writing, Review & Editing: R.T., and P.P.K; Visualization, R.T.; Supervision: P.P.K; Project Administration: P.P.K.

DATA AVAILABILITY

All the data's underlying the results can be provided by the corresponding author upon reasonable request

CONFLICT OF INTERESTS

The authors declare no conflicts of interest.

REFERENCES

1. American Cancer Society. *Cancer Facts & Figures 2023*; American Cancer Society: Atlanta, GA, 2023. Available from: https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/2023-cancer-facts-figures.html. (Accessed on 11 November 2025)
2. Saginala, K., Barsouk, A., Aluru, J.S., Rawla, P., Barsouk, A. Epidemiology of Melanoma. *Med. Sci.* **2021**, *9*(4), 63.
3. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S. Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks. *Nature* **2017**, *542* (7639), 115–118.
4. Brinker, T.J., Hekler, A., Enk, A.H., Klode, J., Hauschild, A., Berking, C., Schilling, B., Haferkamp, S., Schadendorf, D., Holland-Letz, T., et al. Deep Learning Outperformed 136 of 157 Dermatologists in a Head-to-Head Dermoscopic Melanoma Image Classification Task. *Eur. J. Cancer* **2019**, *113*, 47–54.

5. Mridha, K., Uddin, M. M., Shin, J., Khadka, S., Mridha, M. F. An Interpretable Skin Cancer Classification Using Optimized Convolutional Neural Network for a Smart Healthcare System. *IEEE Access* **2023**, *11*, 41003–41018.
6. Tschandl, P., Rosendahl, C., Kittler, H. The HAM10000 Dataset, a Large Collection of Multi-Source Dermatoscopic Images of Common Pigmented Skin Lesions. *Sci. Data* **2018**, *5*, 180161.
7. Codella, N. C. F., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S. W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., et al. Skin Lesion Analysis toward Melanoma Detection: A Challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), Hosted by the International Skin Imaging Collaboration (ISIC). *arXiv* **2017**, arXiv:1710.05006.
8. Hernández-Pérez, C., et al. BCN20000: Dermoscopic Lesions in the Wild. *Sci. Data* **2024**, *11*, 641.
9. Albahar, M.A. Skin Lesion Classification Using Convolutional Neural Network with Novel Regularizer. *IEEE Access* **2019**, *7*, 38306–38313.
10. Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., Kalloo, A., Hassen, A. B. H., Thomas, L., Enk, A., et al. Man against Machine: Diagnostic Performance of a Deep Learning Convolutional Neural Network for Dermoscopic Melanoma Recognition in Comparison to 58 Dermatologists. *Ann. Oncol.* **2018**, *29*(8), 1836–1842.
11. Argenziano, G., Soyer, H. P. Dermoscopy of Pigmented Skin Lesions: A Valuable Tool for Early Diagnosis of Melanoma. *Lancet Oncol.* **2001**, *2*(7), 443–449.
12. Rezvantalab, A., Safigholi, H., Karimijeshni, S. Dermatologist Level Dermoscopy Skin Cancer Classification Using Different Deep Learning Convolutional Neural Networks Algorithms. *arXiv* **2018**, arXiv:1810.10348.
13. Jaisakthi, S. M., Mirunalini, P., Aravindan, C., Appavu, R. Classification of Skin Cancer from Dermoscopic Images Using Deep Neural Network Architectures. *Multimed. Tools Appl.* **2023**, *82*(10), 15763–15778.
14. Akter, M.S., et al. Multi-Class Skin Cancer Classification Architecture Based on Deep Convolutional Neural Network. In *Proceedings of the IEEE International Conference on Big Data*, IEEE, **2022**, pp. 1-10
15. Cirrincione, G., Cannata, S., Cicceri, G., Prinzi, F., Currieri, T., Lovino, M., Militello, C., Pasero, E., Vitabile, S. Transformer-Based Approach to Melanoma Detection. *Sensors* **2023**, *23*, 5677.
16. Gururaj, H.L., Manju, N., Nagarjun, A., Aradhya, V.N.M., Flammini, F. DeepSkin: A Deep Learning Approach for Skin Cancer Classification. *IEEE Access* **2023**, *11*, 50205–50214.
17. Mohapatra, S., Abhishek, N. V. S., Bardhan, D., Ghosh, A. A., Mohanty, S. Skin Cancer Classification Using Convolution Neural Networks. In *Lecture Notes in Networks and Systems*, Springer: Singapore, **2020**, pp. 433–442.
18. Khan, M. Q., et al. Classification of Melanoma and Nevus in Digital Images for Diagnosis of Skin Cancer. *IEEE Access* **2019**, *7*, 90132–90144.
19. Codella, N., et al. Skin Lesion Analysis toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC). *arXiv* **2019**, arXiv:1902.03368.
20. Sun, X., Yang, J., Sun, M., Wang, K. A Benchmark for Automatic Visual Classification of Clinical Skin Disease Images. In *Computer Vision—ECCV 2016*, Lecture Notes in Computer Science, Vol. 9910, Springer: Cham, Switzerland, **2016**, pp. 206–222.
21. Romero-Lopez, A., Giro-i-Nieto, X., Burdick, J., Marques, O. Skin Lesion Classification from Dermoscopic Images Using Deep Learning Techniques. In *Proceedings of the 13th IASTED International Conference on Biomedical Engineering*, **2017**, pp. 49–54.

22. Rahman, M.A., Bazgir, E., Hossain, S.M.S., Maniruzzaman, M. Skin Cancer Classification Using NASNet. *Int. J. Sci. Res. Arch.* **2024**, *11*(1), 775–785.
23. Himel, G.M.S., Islam, M.M., Al-Aff, K.A., Karim, S.I., Sikder, M.K.U. Skin Cancer Segmentation and Classification Using Vision Transformer for Automatic Analysis in Dermatoscopy-Based Noninvasive Digital System. *Int. J. Biomed. Imaging* **2024**, *2024*, 3022192.
24. Dagnaw, G.H., El Mouhtadi, M., Mustapha, M. Skin Cancer Classification Using Vision Transformers and Explainable Artificial Intelligence. *J. Med. Artif. Intell.* **2024**, *7*, 14.
25. Xin, C., et al. An Improved Transformer Network for Skin Cancer Classification. *Comput. Biol. Med.* **2022**, *149*, 105939.
26. Al Huda, M.S., Shrestha, T.E., Hossain, A., Sharif, N.B., Ali, M.A., Erdei, T.I. DeepMelaNet: Advancing Melanoma Stage Classification in Skin Cancer Diagnosis. *Eng. Technol. Appl. Sci. Res.* **2025**, *15*(1), 19627–19635.
27. Sabir, R., Mehmood, T. Classification of Melanoma Skin Cancer Based on Image Data Set Using Different Neural Networks. *Sci. Rep.* **2024**, *14*, 29704.
28. Daglarli, E. Explainable Artificial Intelligence (xAI) Approaches and Deep Meta-Learning Models for Cyber-Physical Systems. In *Artificial Intelligence Paradigms for Smart Cyber-Physical Systems*, Luhach, A. K., Elci, A., Eds., IGI Global: Hershey, PA, **2021**, pp. 42–67.
29. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int. J. Comput. Vis.* **2020**, *128*(2), 336–359.
30. Dosovitskiy, A., et al. An Image Is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. **2021**, pp. 45-67.
31. Johnson, J. Image Forensics Using Error Level Analysis. *Int. J. Comput. Appl.* **2019**, *178*(44), 1–5.
32. Aladhadh, S., Alsanea, M., Aloraini, M., Khan, T., Habib, S., Islam, M. An Effective Skin Cancer Classification Mechanism via Medical Vision Transformer. *Sensors* **2022**, *22*(11), 4008.
33. Hu, J., Shen, L., Sun, G. Squeeze-and-Excitation Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, **2018**, pp. 7132–7141.
34. Woo, S., Park, J., Lee, J.-Y., Kweon, I. S. CBAM: Convolutional Block Attention Module. In *Computer Vision—ECCV 2018*, Lecture Notes in Computer Science, Springer: Cham, Switzerland, **2018**, pp 3–19.
35. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, **2020**, pp. 11534–11542.
36. Hernández-Valencia, E., Gallegos-Funes, F. J., Olivares-Mercado, J., Medina-Bañuelos, V. Lossless Compression for Multispectral Images of Agricultural Products Based on 3D DCT. *J. Appl. Res. Technol.* **2020**, *18* (4), 301–309.
37. Shorten, C., Khoshgoftaar, T. M. A Survey on Image Data Augmentation for Deep Learning. *J. Big Data* **2019**, *6*, 60.
38. Cover, T.M., Thomas, J. A. *Elements of Information Theory*, 2nd ed., Wiley: Hoboken, NJ, **2006**.
39. Mages, T., Anastasiadi, E., Rohner, C. Non-Negative Decomposition of Multivariate Information: From Minimum to Blackwell-Specific Information. *Entropy* **2024**, *26*, 424.
40. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., Brendel, W. ImageNet-Trained CNNs Are Biased towards Texture, Increasing Shape Bias Improves Accuracy and

Robustness. In *Proceedings of the International Conference on Learning Representations (ICLR)*, **2019**, pp. 1-22.

41. Park, N., Kim, S. How Do Vision Transformers Work? In *Proceedings of the International Conference on Learning Representations (ICLR)*, **2022**, pp. 1-26.
42. Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., Dosovitskiy, A. Do Vision Transformers See Like Convolutional Neural Networks? In *Proceedings of the 35th International Conference on Neural Information Processing Systems*. **2021**, pp. 12116–12128.
43. Baltrušaitis, T., Ahuja, C., Morency, L.-P. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*(2), 423–443.
44. Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.-P. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, **2017**, pp. 1103–1114.
45. Liu, Z., Shen, Y., Lakshminarasimhan, V.B., Liang, P.P., Zadeh, A., Morency, L.-P. Efficient Low-Rank Multimodal Fusion with Modality-Specific Factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, **2018**, pp. 2247–2256.
46. Arevalo, J., Solorio, T., Montes-y-Gómez, M., González, F.A. Gated Multimodal Units for Information Fusion. In *Proceedings of the International Conference on Learning Representations (ICLR) Workshop*, **2017**, pp. 1-17.
47. Perez, E., Strub, F., de Vries, H., Dumoulin, V., Courville, A. FiLM: Visual Reasoning with a General Conditioning Layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, **2018**, pp. 3942–3951.
48. Wang, W., Tran, D., Feiszli, M. What Makes Training Multi-Modal Classification Networks Hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, **2020**, pp. 12695–12705.
49. Peng, X., Wei, Y., Deng, A., Wang, D., Hu, D. Balanced Multimodal Learning via On-the-Fly Gradient Modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, **2022**, pp. 8238–8247.
50. Huang, Y., Du, C., Xue, Z., Chen, X., Zhao, H., Huang, L. What Makes Multi-Modal Learning Better than Single (Provably). In *Proceedings of the 35th International Conference on Neural Information Processing Systems*. **2021**, pp. 10944–10956.
51. Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., Sun, C. Attention Bottlenecks for Multimodal Fusion. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*. **2021**, pp. 14200–14213.
52. Singh, N.K., & Mukherjee, T. A VGG-16-Based Deep Learning Approach for Potato Leaf Disease Detection. *Journal of Integrated Engineering and Applied Sciences*, **2026**, *4*(1), 346-354.
53. Pacheco, A.G.C., et al. PAD-UFES-20: A Skin Lesion Dataset Composed of Patient Data and Clinical Images Collected from Smartphones. *Data Brief* **2020**, *32*, 106221.
54. Krogh, A., Vedelsby, J. Neural Network Ensembles, Cross Validation, and Active Learning. In *Advances in Neural Information Processing Systems 7*, **1995**, pp. 231–238.
55. Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, **2022**, pp. 11976–11986.
56. Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, **2017**, pp. 2980–2988.

57. Kornblith, S., Norouzi, M., Lee, H., Hinton, G. Similarity of Neural Network Representations Revisited. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, **2019**, pp. 3519–3529.
58. Daneshjou, R., et al. Disparities in Dermatology AI Performance on a Diverse, Curated Clinical Image Set. *Sci. Adv.* **2022**, *8* (32), eabq6147.
59. van der Maaten, L., Hinton, G. Visualizing Data Using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
60. McInnes, L., Healy, J., Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* **2018**, arXiv:1802.03426.
61. Rousseeuw, P. J. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65.
62. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed., Lawrence Erlbaum Associates: Hillsdale, NJ, **1988**.
63. Hedges, L.V. Distribution Theory for Glass's Estimator of Effect Size and Related Estimators. *J. Educ. Stat.* **1981**, *6*(2), 107–128.
64. Mann, H.B., Whitney, D.R. On a Test of Whether One of Two Random Variables Is Stochastically Larger than the Other. *Ann. Math. Stat.* **1947**, *18*(1), 50–60.
65. Demšar, J. Statistical Comparisons of Classifiers over Multiple Data Sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30.
66. Peng, Z., Huang, W., Gu, S., Xie, L., Wang, Y., Jiao, J., Ye, Q. Conformer: Local Features Coupling Global Representations for Visual Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, **2021**, pp. 367–376.
67. Zhang, Y., Liu, H., Hu, Q. TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021*, Lecture Notes in Computer Science, Vol. 12901, Springer: Cham, Switzerland, **2021**, pp. 14–24.
68. Guo, J., Han, K., Wu, H., Tang, Y., Chen, X., Wang, Y., Xu, C. CMT: Convolutional Neural Networks Meet Vision Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, **2022**, pp. 12175–12185.
69. Dai, Z., Liu, H., Le, Q. V., Tan, M. CoAtNet: Marrying Convolution and Attention for All Data Sizes. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*. **2021**, pp. 3965–3977.
70. Chen, C.-F. R., Fan, Q., Panda, R. CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, **2021**, pp. 357–366.
71. Huo, X., Sun, G., Tian, S., Wang, Y., Yu, L., Long, J., Zhang, W., Li, A. HiFuse: Hierarchical Multi-Scale Feature Fusion Network for Medical Image Classification. *Biomed. Signal Process. Control* **2024**, *87*, 105534.
72. Manzari, O. N., Ahmadabadi, H., Kashiani, H., Shokouhi, S. B., Ayatollahi, A. MedViT: A Robust Vision Transformer for Generalized Medical Image Classification. *Comput. Biol. Med.* **2023**, *157*, 106791.
73. Gessert, N., Nielsen, M., Shaikh, M., Werner, R., Schlaefer, A. Skin Lesion Classification Using Ensembles of Multi-Resolution EfficientNets with Meta Data. *MethodsX* **2020**, *7*, 100864.