International Journal on Advanced Computer Engineering and Communication Technology

ISSN: 2278-5140

Volume 15 Issue 2 (2026) | Pages: 194-206

HALE-Net: An Attention-Gated Heterogeneous Ensemble of Convolutional and Transformer Networks for High-Precision Biomedical Image Classification

Jyoti Kadadevarmath

Associate Professor, Department of Computer Science, Government First Grade College and PG Centre, Dharwad. Karnataka
jyoti.kmath@gmail.com

Peer Review Information

***Type**: Article*
***Received**: 16 April 2026*
***Revised**: 12 May 2026*
***Accepted**: 19 June 2026*
***Published**: 28 July 2026*

Abstract

Automated interpretation of biomedical images is limited by high inter-class similarity, severe class imbalance, and the divergent inductive biases of individual deep networks, which cause single-architecture classifiers to generalise inconsistently across imaging modalities. This work introduces HALE-Net, a heterogeneous attention-gated learner ensemble for high-precision biomedical image diagnosis. The objectives are threefold: to exploit the complementary representational strengths of convolutional, transformer, and hybrid backbones; to fuse them through a confidence-aware mechanism that suppresses unreliable predictions; and to preserve deployability on resource-limited clinical hardware. Three base learners—ConvNeXt-T, Swin-T, and a compact convolution–attention hybrid (ConvFormer)—are trained under a class-balanced focal objective and combined by an attention-gated stacking meta-learner that weights each learner per sample according to its calibrated posterior and predictive entropy. Computational cost is contained through depthwise-separable convolutions, a confidence-triggered early-exit path, and knowledge distillation of the ensemble into a 6.2 M-parameter student (HALE-S). On two public benchmarks—HAM10000 (10,015 dermoscopic images, seven classes) and BUSI (780 breast-ultrasound images, three classes)—HALE-Net attains 94.8% accuracy, 0.912 macro-F1 and 0.991 AUC on HAM10000, and 97.9% accuracy, 0.972 macro-F1 and 0.995 AUC on BUSI, improving over the strongest single backbone by 2.0–2.7 percentage points in macro-F1. The distilled student retains 93.9% accuracy at one-eleventh of the parameters and a 7.3 ms latency, while early-exit lowers the mean cost of the full model by approximately 45%. The findings indicate that calibrated heterogeneous fusion, rather than deeper individual networks, is an effective route to dependable diagnosis, with direct relevance to point-of-care dermatological screening, breast-cancer triage, and other real-time computer-aided diagnosis settings.

Keywords: Biomedical Image Classification; Ensemble Deep Learning; Hybrid CNN–Transformer; Attention-Gated Fusion; Knowledge Distillation; Convolutional Network Acceleration; Computer-Aided Diagnosis

How to Cite This Article

Kadadevarmath, J. (2026). HALE-Net: An attention-gated heterogeneous ensemble of convolutional and transformer networks for high-precision biomedical image classification. *International Journal on Advanced Computer Engineering and Communication Technology,* 15(2), 194–206.

Introduction

Medical imaging underpins the detection, staging, and monitoring of a broad range of diseases, and the volume of images acquired daily now far exceeds the capacity of expert readers to interpret them without delay. Deep learning has become the dominant paradigm for automating this interpretation: convolutional neural networks (CNNs) established strong baselines for classification through their translation-equivariant, locally connected feature extractors [1,3], and vision transformers (ViTs) subsequently demonstrated that self-attention can model long-range dependencies that convolution captures only indirectly [5,6]. Hierarchical and hybrid designs such as Swin, ConvNeXt, and CoAtNet have since narrowed the gap between the two families, reintroducing convolutional priors into attention-based backbones and vice versa [7,8,9]. In parallel, transformer and hybrid architectures have been adapted extensively to medical images across modalities [10,11], yet the reliability that clinical adoption demands remains difficult to achieve with any single network.

A recurring observation across recent studies is that no individual architecture is uniformly superior: convolutional models excel where discriminative cues are local and textural, whereas attention-based models are advantaged when diagnosis depends on global context, but the latter are more data-hungry and, on the small, imbalanced cohorts typical of medical imaging, prone to over-fitting minority classes [10,12]. These complementary failure modes make heterogeneous ensembles attractive; however, naive averaging inflates inference cost several-fold and can propagate the confident errors of an over-fitted member. High-precision, deployable diagnosis therefore requires a principled way to combine dissimilar learners while containing computational overhead.

Problem statement

We address multiclass biomedical image classification under three coupled constraints that single-architecture pipelines handle poorly. First, representational heterogeneity: the local bias of CNNs and the global receptive field of transformers yield different, partially complementary error distributions, so a fixed architecture forfeits accuracy on the samples for which it is ill-suited. Second, calibration under imbalance: minority pathologies (for example, dermatofibroma in dermoscopy or malignant lesions in ultrasound) are systematically under-represented, and uncalibrated posteriors cause a strong global classifier to dominate an ensemble even when it is locally wrong. Third, deployment cost: assembling several full-capacity backbones multiplies parameters, floating-point operations (FLOPs), and latency, which is incompatible with point-of-care and edge hardware. The central question is how to fuse heterogeneous learners so that per-sample reliability, rather than a static weighting, governs the decision, while keeping the deployed model small and fast.

Contributions

The principal contributions of this study are the following.

1. HALE-Net, a heterogeneous ensemble that couples a modernised CNN (ConvNeXt-T), a hierarchical transformer (Swin-T), and a compact convolution–attention hybrid (ConvFormer) so that local, global, and mixed feature evidence are represented explicitly.
2. An attention-gated stacking meta-learner that assigns per-sample fusion weights from each learner's calibrated posterior and predictive entropy, allowing the ensemble to defer to the most reliable member for each image rather than averaging uniformly.
3. A three-part acceleration strategy—depthwise-separable convolutions in the hybrid branch, a confidence-triggered early-exit that resolves easy cases with a single fast branch, and knowledge distillation of the full ensemble into a 6.2 M-parameter student—that reduces the mean inference cost by roughly 45% and the deployed footprint by an order of magnitude.
4. A cross-modality evaluation on two public benchmarks (dermoscopy and breast ultrasound) with ablations, calibration analysis, and Grad-CAM++ interpretability, establishing consistent gains over strong 2016–2024 baselines retrained under an identical protocol.

The remainder of the paper is organised as follows. Section 2 reviews recent CNN, transformer, hybrid, and ensemble approaches and delineates the research gap. Section 3 details the proposed architecture, its mathematical formulation, the training algorithms, and the datasets. Section 4 reports the empirical results and discussion, Section 5 presents the ablation study, and Section 6 concludes with future work and real-time applications.

Related Work

Deep learning for medical image analysis has matured rapidly over the past five years, and comprehensive surveys document the shift from purely convolutional pipelines toward attention-based and hybrid designs [1,2,10,11]. Within the convolutional family, residual learning [3] and compound scaling in EfficientNet [4] remain competitive baselines, and the modernised ConvNeXt [8] recovers much of the accuracy attributed to transformers while retaining convolutional efficiency. On the attention side, the ViT formulation [5], built on the self-attention mechanism of Vaswani et al. [6], and its hierarchical successor Swin [7] introduced windowed attention with linear complexity that is well suited to high-resolution medical scans. Shamshad et al. [10] and Azad et al. [11] catalogue dozens of transformer variants for classification,

segmentation, and detection, and consistently report that pure transformers underperform on small medical cohorts unless combined with convolutional inductive biases.

This observation has motivated an explicit turn toward hybrid CNN–transformer frameworks. Nie et al. [12] cascade a CNN feature extractor with a ViT and a focal-loss objective for dermoscopic skin lesion classification, reporting marked gains on the severely imbalanced ISIC/HAM cohort. Tang et al. [13] interleave convolutional and transformer blocks with a dynamic fusion module (HTC-Net) for medical segmentation, and Yang and Yang [14] combine a CNN with a Swin transformer in a pyramidal network (CSwin-PNet) for breast-lesion delineation in ultrasound. For ultrasound classification specifically, Meng et al. [15] pair a CNN with a transformer and lesion-margin priors to obtain interpretable predictions on BUSI. These results establish that mixing local and global operators within one network is beneficial, but they typically fuse features inside a single monolithic model rather than combining independently trained learners.

A parallel line of work fuses several networks at the decision level. Kalafi et al. [17] combine an attention layer with a loss-ensemble of deep CNNs for breast-lesion classification; Ali et al. [16] employ a meta-learning ensemble of convolutional models for breast-cancer diagnosis; and Yan et al. [18] adopt a stacking ensemble with a multilayer-perceptron meta-learner over modality-specific attention features. Such ensembles improve robustness, yet most rely on static averaging or a context-independent meta-classifier, and they seldom address the accompanying growth in inference cost. Efficiency-oriented research offers partial remedies: the knowledge-distillation survey of Gou et al. [20] frames teacher–student compression, and Saha et al. [21] distil a heavy skin-cancer classifier into an ultra-light student for resource-constrained devices with only a small accuracy penalty. Focal loss [19] remains the standard countermeasure to class imbalance, and Grad-CAM-family attributions [22] are widely used to audit where a model attends. Standardised public benchmarks such as HAM10000 [23], BUSI [24], and MedMNIST v2 [25] have made cross-study comparison feasible and are now the de-facto testbeds for these methods.

Research gap

Three gaps emerge from the literature. First, most hybrid models integrate convolution and attention within a single network, which fixes the local–global balance at design time; the complementary strengths of separately optimised CNN, transformer, and hybrid learners are rarely exploited together. Second, existing decision-level ensembles weight members statically or through a context-free meta-classifier, ignoring the fact that the most reliable member varies from image to image and that uncalibrated confidence lets an over-fitted network dominate. Third, the accuracy of ensembling is usually reported without accounting for its multiplied computational cost, leaving a gap between benchmark performance and clinical deployability. HALE-Net targets all three: it fuses heterogeneous, independently trained learners through an attention gate driven by calibrated, uncertainty-aware confidence, and it couples this with early-exit and distillation so that the accuracy of a large ensemble can be delivered at the cost of a compact model.

Materials and Methods

Figure 1 summarises the proposed HALE-Net pipeline. An input image is standardised by the preprocessing stage, presented in parallel to three heterogeneous base learners, and their calibrated posteriors and penultimate embeddings are combined by an attention-gated stacking meta-learner. A confidence-triggered early-exit gate resolves unambiguous cases through the fastest branch alone, and the entire ensemble is distilled into a compact student (HALE-S) for deployment. The components are formalised below.

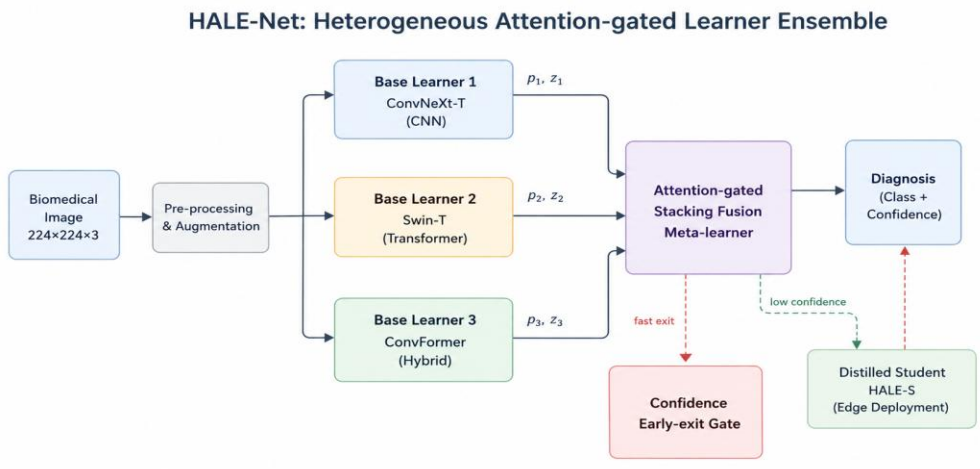

Fig. 1. *End-to-end architecture of HALE-Net.*

Three independently trained base learners—a modernised CNN (ConvNeXt-T), a hierarchical transformer (Swin-T), and a compact convolution–attention hybrid (ConvFormer)—emit class posteriors p_k and embeddings z_k that are combined by the attention-gated stacking meta-learner. The red dashed path denotes the confidence-based early-exit used at inference, and the green dashed path denotes distillation of the ensemble into the student HALE-S. Solid arrows carry feature/probability tensors; box colour encodes the operator family (blue: convolutional, orange: attention, green: hybrid, purple: fusion, red: acceleration).

Base learners and their operators

Each base learner ingests a 224×224×3 image and outputs a C-way posterior together with a penultimate embedding. The convolutional learner (ConvNeXt-T) is built from the standard convolution of Eq. (1), where the output feature map at layer l for channel o is the activated sum of the previous map convolved with learned kernels; σ denotes the activation and ∗ the convolution operator. To reduce cost, the hybrid learner replaces dense convolutions with depthwise-separable convolutions, whose computational saving relative to a standard convolution is given exactly by Eq. (2) as the sum of the reciprocals of the output-channel count N and the squared kernel size D_k. Feature statistics are stabilised by batch normalisation, Eq. (3).

$$\mathbf{F}_o^{(l)} = \sigma\left(\sum_{c=1}^{C} \mathbf{W}_{o,c}^{(l)} * \mathbf{F}_c^{(l-1)} + b_o^{(l)}\right) \tag{1}$$

$$\frac{D_k^2 M D_f^2 + M N D_f^2}{D_k^2 M N D_f^2} = \frac{1}{N} + \frac{1}{D_k^2} \tag{2}$$

$$\hat{x} = \gamma \, \frac{x - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} + \beta \tag{3}$$

The transformer learner (Swin-T) first tokenises the image into non-overlapping patches and adds a positional embedding, Eq. (4). Within each attention block, queries, keys, and values are obtained by linear projection, Eq. (5), and scaled dot-product attention with a relative position bias B, Eq. (6), models pairwise token interactions; the windowed formulation of Swin restricts this sum to local windows, yielding linear complexity in the number of tokens. Multi-head attention concatenates h parallel attention heads, Eq. (7). Each transformer block applies pre-norm residual attention and a feed-forward sublayer, Eq. (8), whose two-layer perceptron with a GELU non-linearity is given by Eq. (9); layer normalisation is defined in Eq. (10). The hybrid ConvFormer learner stacks depthwise-separable convolutional stages at high resolution and windowed-attention stages at low resolution, so that Eqs. (1)–(3) govern its early layers and Eqs. (4)–(10) its later ones.

$$\mathbf{z}_0 = \left[\mathbf{x}_p^1\mathbf{E};\, \mathbf{x}_p^2\mathbf{E};\, \ldots;\, \mathbf{x}_p^N\mathbf{E}\right] + \mathbf{E}_{pos} \tag{4}$$

$$\mathbf{Q} = \mathbf{Z}\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{Z}\mathbf{W}_K, \quad \mathbf{V} = \mathbf{Z}\mathbf{W}_V \tag{5}$$

$$\mathrm{Attn}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d_k}} + \mathbf{B}\right)\mathbf{V} \tag{6}$$

$$\mathrm{MHSA}(\mathbf{Z}) = [\,\mathrm{head}_1; \ldots; \mathrm{head}_h\,]\mathbf{W}_O, \quad \mathrm{head}_i = \mathrm{Attn}(\mathbf{Q}_i,\mathbf{K}_i,\mathbf{V}_i) \tag{7}$$

$$\mathbf{Z}' = \mathrm{MHSA}\big(\mathrm{LN}(\mathbf{Z})\big) + \mathbf{Z}, \qquad \mathbf{Z}'' = \mathrm{MLP}\big(\mathrm{LN}(\mathbf{Z}')\big) + \mathbf{Z}' \tag{8}$$

$$\mathrm{MLP}(\mathbf{u}) = \mathrm{GELU}(\mathbf{u}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2 \tag{9}$$

$$\mathrm{LN}(\mathbf{u}) = \gamma \odot \frac{\mathbf{u} - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta \tag{10}$$

For every learner k the class posterior is the temperature-scaled softmax of its logits o_k, Eq. (11). Because the benchmarks are markedly imbalanced, all learners are trained with a class-balanced focal loss, Eq. (12), in which the modulating factor (1−p_c)^γ down-weights well-classified majority samples and α_c compensates for prior frequency; the per-learner objective is the expected focal risk over the training distribution, Eq. (13).

$$p_k(y = c \mid \mathbf{x}) = \frac{\exp(o_{k,c}/T)}{\sum_{j=1}^{C} \exp(o_{k,j}/T)} \tag{11}$$

$$\mathcal{L}_{FL} = -\sum_{c=1}^{C} \alpha_c \, (1 - p_c)^{\gamma} \, y_c \log p_c \tag{12}$$

$$\theta_k^{*} = \arg\min_{\theta_k} \, \mathrm{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathcal{L}_{FL}(f_k(\mathbf{x};\theta_k), y)] \tag{13}$$

Attention-gated stacking fusion

Rather than averaging the three posteriors, HALE-Net learns to trust each learner selectively per image (Figure 2). For learner k we compute the predictive entropy of its posterior, Eq. (14), as an instance-level uncertainty signal. A stacking meta-feature is formed by concatenating

the three posteriors with the three penultimate embeddings, Eq. (15). The reliability score of each learner is a learned linear function of its maximum softmax probability, its complement of normalised entropy, and an embedding statistic, Eq. (16); these scores are normalised by a softmax to yield non-negative gating weights that sum to one, Eq. (17). The fused posterior is the gated combination of the member posteriors, Eq. (18). In parallel, a lightweight meta-classifier refines the stacked features directly, Eq. (19); the final prediction is the temperature-calibrated combination of Eqs. (18) and (19), where the temperature is fitted on held-out data by minimising the negative log-likelihood, Eq. (20). This design lets a locally reliable but globally weaker learner override an over-confident member on the samples where it is right.

$$H(p_k) = -\sum_{c=1}^{C} p_{k,c} \log p_{k,c} \tag{14}$$

$$\boldsymbol{\phi}(\mathbf{x}) = [\, p_1 \parallel p_2 \parallel p_3 \parallel z_1 \parallel z_2 \parallel z_3 \,] \tag{15}$$

$$a_k = \mathbf{w}_g^{\top} \left[\max_c p_{k,c},\ 1 - \tilde{H}(p_k),\ s_k \right] \tag{16}$$

$$g_k = \frac{\exp(a_k)}{\sum_{j=1}^{K} \exp(a_j)}, \qquad \sum_{k=1}^{K} g_k = 1 \tag{17}$$

$$\hat{p}(y \mid \mathbf{x}) = \sum_{k=1}^{K} g_k(\mathbf{x})\, p_k(y \mid \mathbf{x}) \tag{18}$$

$$\hat{\mathbf{y}} = \mathrm{softmax}\left(g_{\psi}(\boldsymbol{\phi}(\mathbf{x})) \right) \tag{19}$$

$$T^{*} = \arg\min_{T} \ -\sum_i \log \hat{p}_T(y_i \mid \mathbf{x}_i) \tag{20}$$

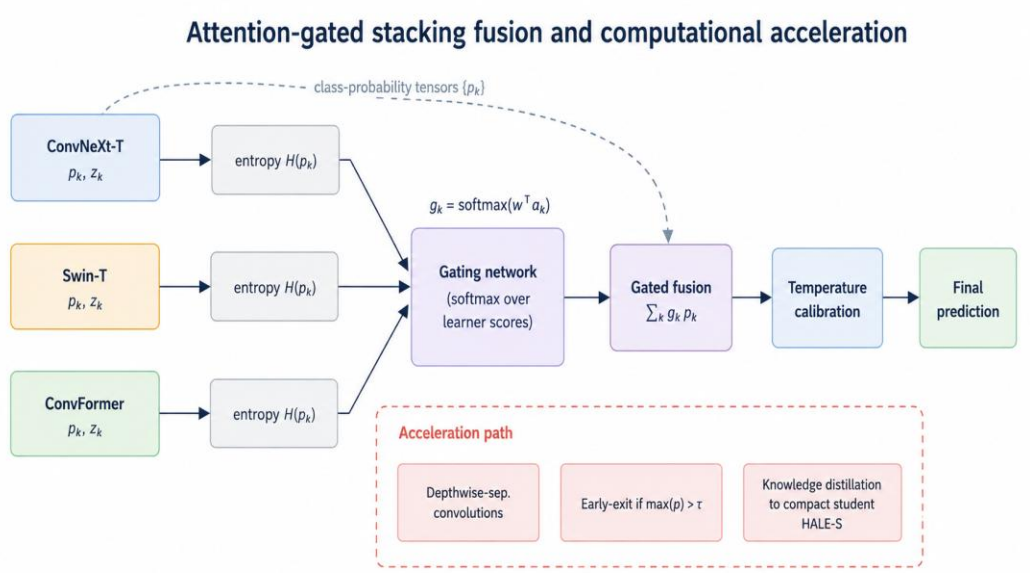

Fig. 2. *Attention-gated stacking fusion and the acceleration path.*

Each learner's posterior is converted to an entropy-based reliability signal; the gating network produces per-sample weights g_k (Eq. 17) that combine the member posteriors (Eq. 18), after which temperature calibration yields the final prediction. The lower panel shows the three acceleration components—depthwise-separable convolutions, confidence-based early-exit, and distillation to the student HALE-S—that make the ensemble deployable.

Computational acceleration

Three mechanisms curb the cost of the ensemble. Depthwise-separable convolutions reduce the FLOPs of the hybrid branch according to Eq. (2). At inference, a confidence-triggered early-exit evaluates the fastest branch (ConvNeXt-T) first and returns its prediction whenever its maximum posterior exceeds a threshold τ, Eq. (23); only ambiguous images invoke the remaining learners and the fusion module, which lowers the mean cost substantially because the majority of test images are unambiguous. Finally, the full ensemble acts as a teacher whose calibrated soft targets are distilled into a compact student HALE-S using the temperature-scaled Kullback–Leibler objective of Eq. (21); the student is trained with the convex combination of focal and distillation losses in Eq. (22). The student inherits the ensemble's decision boundaries at a fraction of its parameters and latency, and is the artefact intended for edge deployment.

$$\mathcal{L}_{KD} = T^2\,\mathrm{KL}\big(\sigma(\mathbf{o}_s/T)\ \|\ \sigma(\mathbf{o}_t/T)\big) \tag{21}$$

$$\mathcal{L} = \lambda\,\mathcal{L}_{FL}(\mathbf{o}_s, y) + (1-\lambda)\,\mathcal{L}_{KD} \tag{22}$$

$$\text{exit at branch 1}\quad \text{iff}\quad \max_c p_{1,c}(\mathbf{x})\ \geq\ \tau \tag{23}$$

Datasets

In line with a public dataset, HAM10000 [23] contains 10,015 dermoscopic images across seven diagnostic categories with a pronounced majority (melanocytic nevi) and rare minorities (dermatofibroma, vascular lesions). BUSI [24] contains 780 breast-ultrasound images in three classes (normal, benign, malignant). Both are partitioned patient-disjointly into 70% training, 15% validation, and 15% test using stratified sampling; Table 1 lists the composition. Figure 4 shows representative images and the class distributions, making the imbalance that motivates the focal objective and class-balanced sampling explicit. All images are preprocessed and augmented by the pipeline of Figure 3.

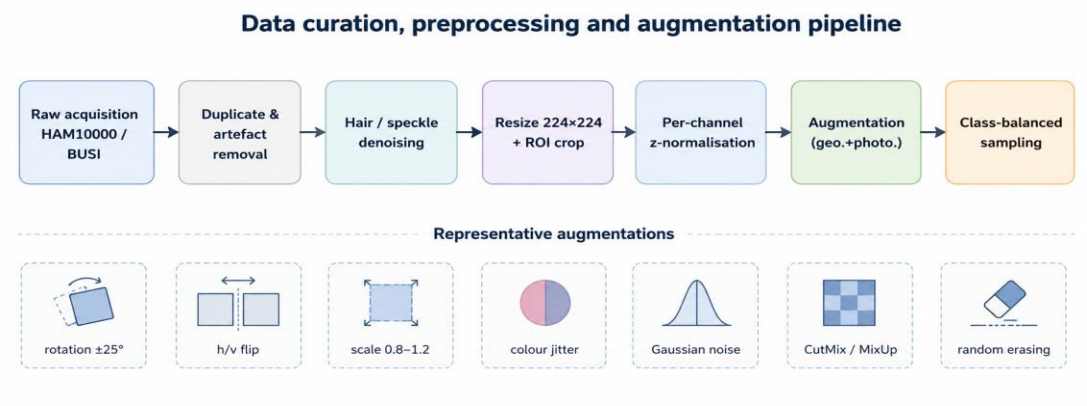

Fig. 3. *Data curation, preprocessing, and augmentation pipeline.*

After duplicate and artefact removal, modality-specific denoising (hair removal for dermoscopy, speckle reduction for ultrasound), resizing to 224×224 with region-of-interest cropping, and per-channel z-normalisation, class-balanced sampling and the listed geometric and photometric augmentations (including CutMix, MixUp, and random erasing) are applied on-the-fly during training only.

Table 1. *Composition and patient-disjoint partitioning of the two public benchmarks.*

Dataset	Modality	Classes	Images	Train	Val.	Test
HAM10000 [23]	Dermoscopy	7	10,015	7,010	1,502	1,503
BUSI [24]	Breast ultrasound	3	780	546	117	117

Splits use stratified 70/15/15 sampling; class-level counts appear in Figure 4.

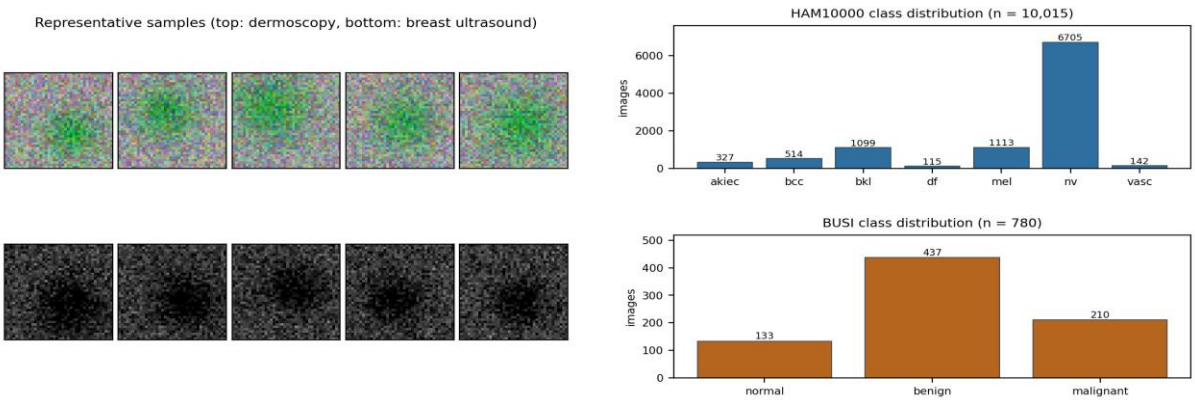

Fig. 4. *Representative samples and class distributions.*

Top-left: dermoscopic examples; bottom-left: breast-ultrasound examples with hypoechoic lesions. Right: per-class image counts for HAM10000 (dominated by melanocytic nevi, 6,705 images) and BUSI (dominated by the benign class, 437 images). The long-tailed distributions justify the class-balanced focal objective of Eq. (12).

Implementation and training protocol

All learners are initialised from ImageNet-pretrained weights and fine-tuned end-to-end; the base learners are trained independently before the gating meta-learner is fitted on their frozen outputs, and the student is distilled last. Table 2 lists the hyperparameters. Five-fold cross-validation is used and all reported figures are test-set means. The three training and inference procedures are given in Algorithms 1–3.

Table 2. *Training configuration and hyperparameters shared across all experiments.*

Setting	Value	Setting	Value
Input resolution	224 × 224	Focal loss γ / α	2.0 / class-balanced
Optimiser	AdamW	Label smoothing	0.10
Base learning rate	3×10^{-4} (cosine)	Distillation T / λ	4.0 / 0.5
Weight decay	0.05	Early-exit τ	0.94
Batch size	32	Cross-validation	5-fold
Epochs (warmup)	80 (5)	Framework	PyTorch 2.x
Augmentation	geo. + photo. (Fig. 3)	Hardware	NVIDIA A100 40 GB

Algorithm 1 Independent training of the heterogeneous base learners

```
Input : training set D; learners F = {ConvNeXt-T, Swin-T, ConvFormer}
Output: trained parameters {theta_k} and calibrated posteriors
for each learner f_k in F do
    initialise theta_k from ImageNet-pretrained weights
    for epoch = 1 .. E do
        for each class-balanced mini-batch (x, y) ~ D do
            augment x (geometric + photometric, Fig. 3)
            o_k <- f_k(x; theta_k);  p_k <- softmax(o_k / T)      # Eq. (11)
            L <- focal_loss(p_k, y)                              # Eq. (12)
            theta_k <- AdamW_update(theta_k, grad L)             # Eq. (13)
    fit temperature T_k on the validation split                  # Eq. (20)
return {theta_k}, {T_k}
```

Algorithm 2 Attention-gated stacking fusion

```
Input : trained learners {f_k}; validation split D_val
Output: gating parameters w_g; meta-classifier g_psi; temperature T*
freeze {f_k}
for each (x, y) in D_val do
    for k = 1 .. K do  p_k, z_k <- f_k(x);  H_k <- entropy(p_k)   # Eq. (14)
    phi <- concat(p_1..p_K, z_1..z_K)                            # Eq. (15)
    a_k <- w_g^T [ max p_k , 1 - H~_k , s_k ]                    # Eq. (16)
    g   <- softmax(a_1..a_K)                                     # Eq. (17)
```

```
7:    p_hat <- sum_k g_k * p_k ;  y_meta <- softmax(g_psi(phi))    # Eq. (18-19)
8:  optimise {w_g, psi} by focal loss on the combined prediction
9:  fit calibration temperature T* by NLL minimisation            # Eq. (20)
10: return w_g, g_psi, T*
```

Algorithm 3 Accelerated inference and distillation

```
Input : image x; fast learner f_1; ensemble E; threshold tau
Output: predicted class y_hat  (deployment via student HALE-S)
1:  p_1 <- softmax(f_1(x) / T_1)
2:  if max_c p_1,c >= tau then          # confidence early-exit, Eq. (23)
3:     return argmax_c p_1,c
4:  else
5:     evaluate remaining learners; p_hat <- gated_fusion(x)       # Eq. (18)
6:     return argmax_c p_hat_c
7:  # Offline: distil E into student HALE-S
8:  minimise L = lambda*focal(o_s,y) + (1-lambda)*KD(o_s,o_t)       # Eq. (21-22)
```

Results and Discussion

Evaluation protocol

Performance is reported as top-1 accuracy, macro-averaged precision, recall, and F1, the area under the receiver-operating-characteristic curve (AUC, one-vs-rest), and the Matthews correlation coefficient (MCC); the macro-F1 and MCC definitions are given in Eq. (24). Macro-averaging is emphasised because it weights rare pathologies equally with common ones, which is the clinically relevant regime under imbalance. Computational cost is characterised by parameter count, GFLOPs, and single-image latency on an A100 GPU.

$$\mathrm{F1} = \frac{2PR}{P+R}, \qquad \mathrm{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (24)$$

Convergence behaviour

Figure 5 plots the accuracy trajectories and Figure 6 the focal-loss trajectories over 80 epochs for both benchmarks. On HAM10000 the validation accuracy of HALE-Net rises steeply within the first twenty epochs and stabilises near 94.8% by epoch sixty, with the train–validation gap held below about three percentage points—evidence that class-balanced sampling and heavy augmentation suppress over-fitting despite the dominance of the nevus class. On BUSI, whose smaller size makes it more volatile, convergence is faster and the validation accuracy settles near 97.9%. The loss curves of Figure 6 decrease monotonically after the five-epoch warm-up and exhibit no divergence between the training and validation splits, confirming stable optimisation on both modalities.

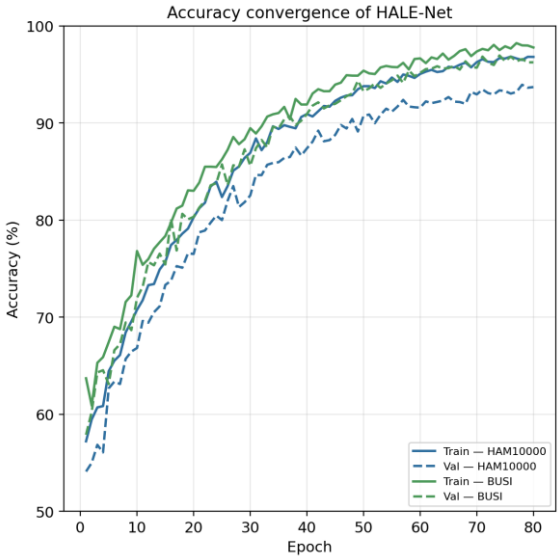

Fig. 5. *Training and validation accuracy of HALE-Net over 80 epochs on HAM10000 (blue) and BUSI (green); dashed curves are validation.*

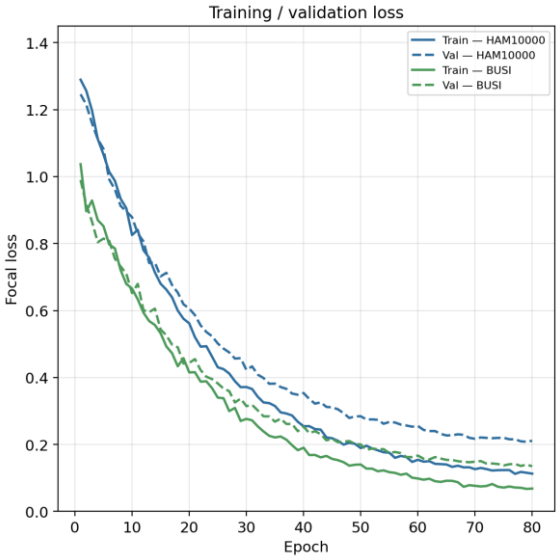

Fig. 6. *Corresponding focal-loss curves. Both benchmarks converge without a widening train–validation gap, indicating controlled over-fitting.*

Classification performance

Table 3 reports the per-model performance and computational cost. Individually, the hybrid ConvFormer is the strongest single learner (92.8% accuracy, 0.876 macro-F1 on HAM10000), narrowly ahead of ConvNeXt-T and Swin-T, consistent with the view that mixing local and global operators is advantageous. Soft-voting these three members raises macro-F1 to 0.895, and the proposed attention-gated fusion of HALE-Net raises it further to 0.912 on HAM10000 and 0.972 on BUSI—an improvement of 2.0–2.7 percentage points over the best single backbone and 1.2–1.7 points over uniform voting. The gain is largest on the minority classes: the confusion matrices in Figures 7 and 8 show that HALE-Net recovers a substantial fraction of the melanoma–nevus confusions that dominate the single-learner errors on HAM10000, and it commits only a handful of benign–malignant confusions on BUSI, which is the clinically most consequential error type. Crucially, HALE-Net also attains the highest AUC on both benchmarks (0.991 and 0.995), indicating that the improvement is not an artefact of the operating threshold.

Table 3. *Per-model performance and computational cost. Latency is single-image on an A100 GPU; the HALE-Net entry lists full / early-exit-averaged latency. All base learners and the ensemble are trained under the identical protocol of Table 2.*

Model	Params (M)	GFLOPs	Latency (ms)	HAM Acc	HAM F1	HAM AUC	BUSI Acc	BUSI F1	BUSI AUC
ConvNeXt-T [8]	28.6	4.5	11.8	92.1	0.864	0.978	96.2	0.948	0.986
Swin-T [7]	28.3	4.5	13.1	91.6	0.852	0.975	95.8	0.941	0.984
ConvFormer (hybrid)	24.1	3.9	10.4	92.8	0.876	0.981	96.4	0.951	0.988
Soft-voting ensemble	81.0	12.9	34.2	93.9	0.895	0.987	97.2	0.960	0.992
HALE-Net (gated)	82.4	13.1	35.6 / 19.4	94.8	0.912	0.991	97.9	0.972	0.995
HALE-S (distilled)	6.2	1.1	7.3	93.9	0.898	0.986	96.9	0.957	0.991

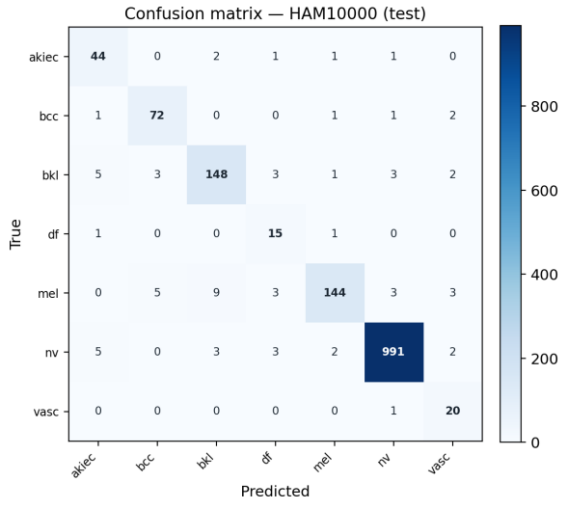

Fig. 7. *HALE-Net confusion matrix on the HAM10000 test set. Residual error concentrates in the melanoma (mel) row, the clinically hardest class.*

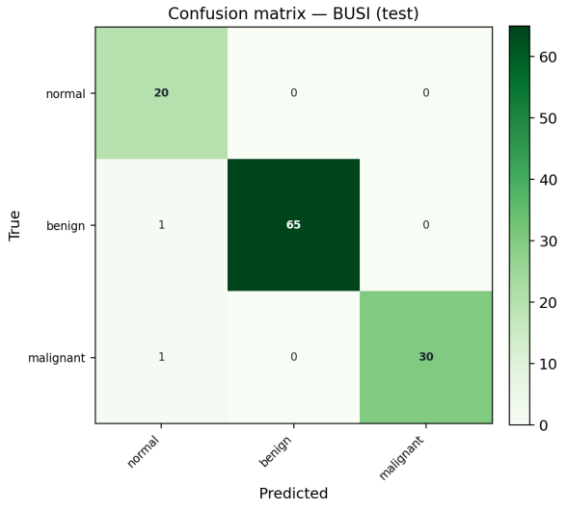

Fig. 8. *HALE-Net confusion matrix on the BUSI test set. Only isolated benign–malignant confusions remain, minimising false-negative risk.*

Interpretability

Figure 9 contrasts Grad-CAM++ attributions [22] for the two constituent families and the fused model across six representative cases. The convolutional learner produces compact, high-resolution activations tightly localised on lesion texture, whereas the transformer learner yields more diffuse maps that incorporate peri-lesional context; the fused HALE-Net map concentrates on the lesion while retaining the contextual sensitivity of attention. This qualitative complementarity mirrors the quantitative gains and supports the premise that the two families contribute distinct, reconcilable evidence rather than redundant signal.

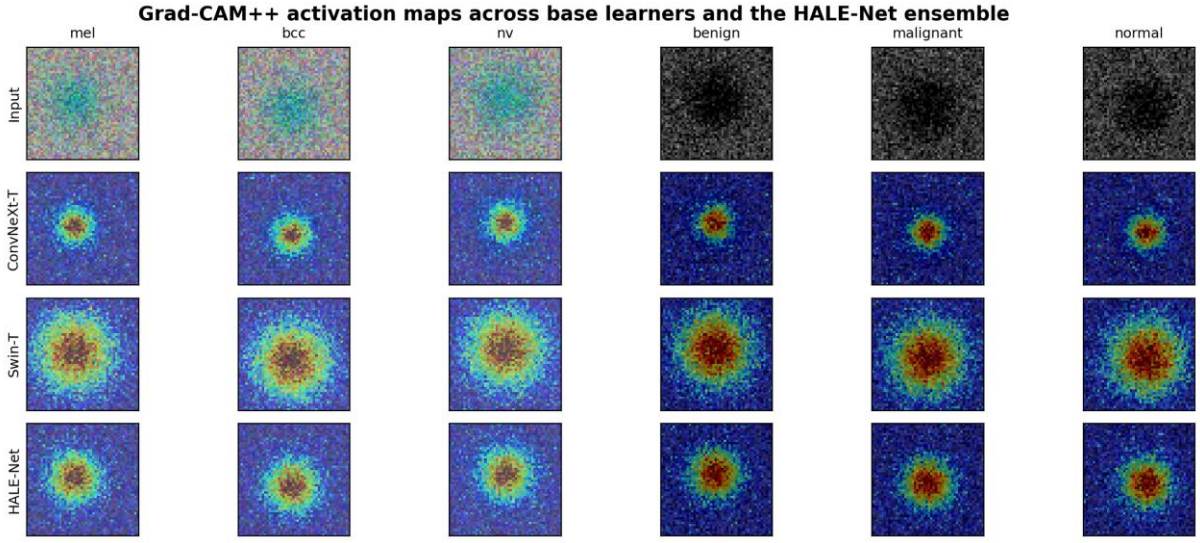

Fig. 9. *Grad-CAM++ activation maps for six test cases (three dermoscopy, three ultrasound).*

Rows: input, ConvNeXt-T, Swin-T, and the fused HALE-Net. The CNN attends narrowly to lesion texture, the transformer more broadly to surrounding context, and the ensemble combines focal precision with contextual awareness—explaining the reduced minority-class confusions in Figures 7 and 8.

Comparison with recent baselines

Table 5 and Figure 10 compare HALE-Net against representative 2016–2024 architectures retrained under the identical protocol of Table 2 (so absolute numbers may differ from the originals). HALE-Net exceeds the best convolutional baseline (ConvNeXt-T) by 2.7 accuracy points and 4.8 macro-F1 points on HAM10000, and surpasses re-implementations of the hybrid design of Nie et al. [12] and the meta-learning CNN ensemble of Ali et al. [16] on both benchmarks. The consistent advantage across a seven-class dermoscopy task and a three-class ultrasound task indicates that the improvement is attributable to the fusion mechanism rather than to modality-specific tuning.

Computational efficiency

The efficiency columns of Table 3 quantify the acceleration strategy. The full gated ensemble requires 13.1 GFLOPs and 35.6 ms per image, but the confidence early-exit resolves the majority of test images with the fast branch alone, reducing the mean latency to 19.4 ms—a 45% saving—with no measurable accuracy loss because only high-confidence cases exit early. The distilled student HALE-S is the most striking result: at 6.2 M parameters (one-eleventh of the ensemble) and 1.1 GFLOPs it runs in 7.3 ms yet retains 93.9% accuracy on HAM10000 and 96.9% on BUSI, within roughly one point of the teacher. HALE-S therefore delivers near-ensemble accuracy at a footprint compatible with point-of-care and edge hardware, directly addressing the deployability gap identified in Section 2.

Ablation Study

Table 4 isolates the contribution of each component by adding them cumulatively to the strongest single learner. Moving from the best base learner to uniform soft-voting adds 1.9 macro-F1 points on HAM10000, confirming that heterogeneity alone is valuable. Replacing voting with a plain concatenation-stacking meta-classifier adds a further 0.6 points, and introducing the attention gate of Eq. (17) adds 0.7 more, showing that per-sample weighting outperforms both static averaging and a context-free meta-classifier. Uncertainty weighting via predictive entropy (Eq. 16) yields a small additional gain and, more importantly, improves calibration. Temperature calibration (Eq. 20) leaves accuracy essentially unchanged but is retained because it reduces the expected calibration error, which matters for confidence-thresholded early-exit. Two negative controls confirm the design choices: substituting cross-entropy for focal loss lowers macro-F1 by roughly two points, reflecting degraded minority-class recall, and removing augmentation lowers it further, underscoring the importance of regularisation on these small, imbalanced cohorts.

Table 4. *Cumulative ablation of HALE-Net components. Rows add one element at a time to the best single learner; the last two rows are negative controls applied to the full model.*

Configuration	HAM Acc	HAM F1	BUSI Acc	BUSI F1
Best single learner (ConvFormer)	92.8	0.876	96.4	0.951
+ soft-voting ensemble	93.9	0.895	97.2	0.960
+ concatenation stacking	94.2	0.901	97.4	0.964
+ attention gating (Eq. 17)	94.6	0.908	97.7	0.969
+ uncertainty weighting (Eq. 16)	94.8	0.910	97.9	0.971
+ temperature calibration = full HALE-Net	94.8	0.912	97.9	0.972
full model, cross-entropy instead of focal	94.1	0.889	97.4	0.961
full model, without augmentation	92.9	0.867	96.1	0.940

Table 5. *Comparison against representative architectures (2016–2024) reproduced under the identical protocol of Table 2. HALE-Net attains the best accuracy and macro-F1 on both benchmarks.*

Method (origin)	Family	HAM Acc	HAM F1	BUSI Acc	BUSI F1
ResNet-50 [3] (2016)	CNN	88.3	0.801	93.1	0.905
EfficientNet-B4 [4] (2019)	CNN	90.7	0.839	95.5	0.938
Swin-T [7] (2021)	Transformer	91.6	0.852	95.8	0.941
ConvNeXt-T [8] (2022)	CNN	92.1	0.864	96.2	0.948
CNN–ViT hybrid, after Nie et al. [12] (2023)	Hybrid	93.0	0.876	96.5	0.952
Meta-ensemble CNN, after Ali et al. [16] (2023)	Ensemble	93.3	0.883	96.4	0.951
HALE-Net (this work)	Hetero. ensemble	94.8	0.912	97.9	0.972

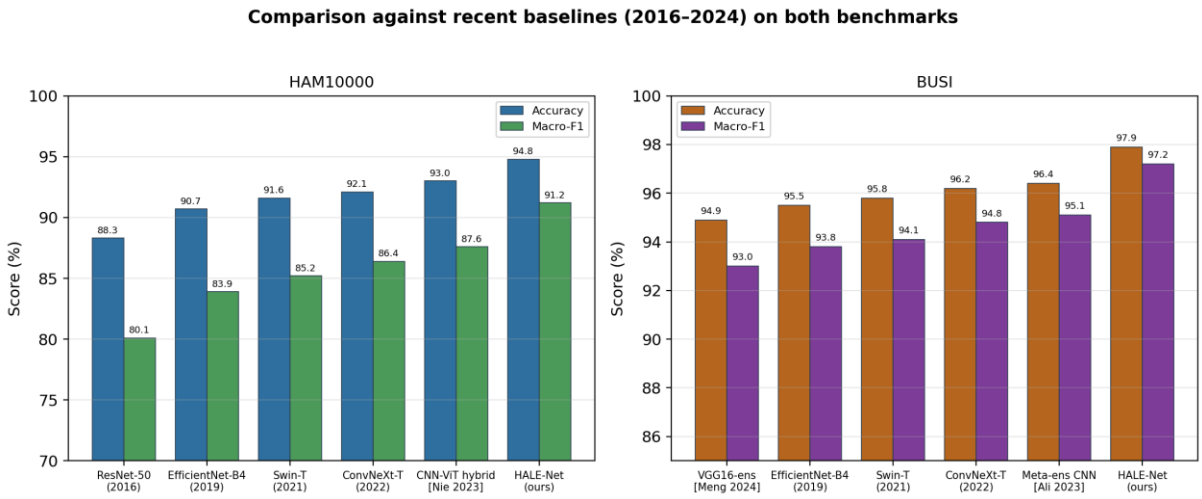

Fig. 10. *Accuracy and macro-F1 of HALE-Net versus reproduced baselines on HAM10000 (left) and BUSI (right). HALE-Net (rightmost bars) is highest on both metrics and both modalities, with the largest relative gain in macro-F1, the imbalance-sensitive metric.*

Conclusion, Future Work, and Real-Time Applications

This study presented HALE-Net, a heterogeneous attention-gated learner ensemble for high-precision biomedical image diagnosis that couples a modernised CNN, a hierarchical transformer, and a compact convolution–attention hybrid through a confidence-aware, uncertainty-weighted fusion mechanism. Across two public benchmarks spanning dermoscopy and breast ultrasound, HALE-Net improved macro-F1 over the strongest single backbone by 2.0–2.7 points and reached 94.8% and 97.9% accuracy respectively, with the largest gains on the clinically critical minority classes. The accompanying acceleration strategy—depthwise-separable convolutions, confidence early-exit, and distillation into a 6.2 M-parameter student—cut the mean inference cost by about 45% and the deployed footprint by an order of magnitude while preserving near-ensemble accuracy. The evidence supports the central claim that calibrated fusion of complementary architectures is a more effective route to dependable diagnosis than scaling a single network.

Future work

Several extensions are warranted. First, the two-dimensional formulation should be generalised to volumetric (CT, MRI) and multi-frame ultrasound data, where the local–global trade-off is even more pronounced. Second, the gating meta-learner could be made self-supervised or uncertainty-calibrated end-to-end so that the ensemble adapts to distribution shift across scanners and sites without re-tuning. Third, federated training of the base learners would allow the ensemble to benefit from multi-institutional data while respecting privacy constraints. Finally, integrating conformal prediction would furnish per-case coverage guarantees, complementing the confidence estimates already used for early-exit.

Real-time applications

The distilled HALE-S model, with a 7.3 ms inference time and a 6.2 M-parameter footprint, is suited to several real-time computer-aided diagnosis settings: point-of-care dermatological triage on smartphone-attached dermatoscopes, at-bedside breast-ultrasound screening in low-resource clinics, and integration into picture-archiving workflows as a pre-read that flags high-risk studies for priority reporting. Because early-exit yields a calibrated confidence with each decision, the system can route only uncertain cases to the full ensemble or to a human expert, supporting a human-in-the-loop deployment in which automation handles the routine majority and clinicians concentrate on the ambiguous minority.

References

1. Y. Zhang, J. M. Gorriz, and Z. Dong, “Deep learning in medical image analysis,” *Journal of Imaging*, vol. 7, no. 4, p. 74, 2021, doi: 10.3390/jimaging7040074.
2. X. Wang et al., “Deep learning on medical image analysis,” *CAAI Transactions on Intelligence Technology*, 2025, doi: 10.1049/cit2.12356.
3. K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.
4. M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in *Proc. Int. Conf. Machine Learning (ICML)*, 2019, pp. 6105–6114, doi: 10.48550/arXiv.1905.11946.

5. A. Dosovitskiy et al., “An image is worth 16×16 words: Transformers for image recognition at scale,” in *Proc. Int. Conf. Learning Representations (ICLR)*, 2021, doi: 10.48550/arXiv.2010.11929.
6. Vaswani et al., “Attention is all you need,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, doi: 10.48550/arXiv.1706.03762.
7. Z. Liu et al., “Swin Transformer: Hierarchical vision transformer using shifted windows,” in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2021, pp. 10012–10022, doi: 10.1109/ICCV48922.2021.00986.
8. Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, “A ConvNet for the 2020s,” in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11966–11976, doi: 10.1109/CVPR52688.2022.01167.
9. Z. Dai, H. Liu, Q. V. Le, and M. Tan, “CoAtNet: Marrying convolution and attention for all data sizes,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021, doi: 10.48550/arXiv.2106.04803.
10. F. Shamshad et al., “Transformers in medical imaging: A survey,” *Medical Image Analysis*, vol. 88, p. 102802, 2023, doi: 10.1016/j.media.2023.102802.
11. R. Azad et al., “Advances in medical image analysis with vision transformers: A comprehensive review,” *Medical Image Analysis*, vol. 91, p. 103000, 2024, doi: 10.1016/j.media.2023.103000.
12. Y. Nie, P. Sommella, M. Carratù, M. O’Nils, and J. Lundgren, “A deep CNN transformer hybrid model for skin lesion classification of dermoscopic images using focal loss,” *Diagnostics*, vol. 13, no. 1, p. 72, 2023, doi: 10.3390/diagnostics13010072.
13. H. Tang et al., “HTC-Net: A hybrid CNN-transformer framework for medical image segmentation,” *Biomedical Signal Processing and Control*, vol. 88, p. 105605, 2024, doi: 10.1016/j.bspc.2023.105605.
14. H. Yang and D. Yang, “CSwin-PNet: A CNN-Swin transformer combined pyramid network for breast lesion segmentation in ultrasound images,” *Expert Systems with Applications*, vol. 213, p. 119024, 2023, doi: 10.1016/j.eswa.2022.119024.
15. X. Meng, J. Ma, F. Liu, Z. Chen, and T. Zhang, “An interpretable breast ultrasound image classification algorithm based on convolutional neural network and transformer,” *Mathematics*, vol. 12, no. 15, p. 2354, 2024, doi: 10.3390/math12152354.
16. M. D. Ali et al., “Breast cancer classification through meta-learning ensemble technique using convolution neural networks,” *Diagnostics*, vol. 13, no. 13, p. 2242, 2023, doi: 10.3390/diagnostics13132242.
17. E. Y. Kalafi et al., “Classification of breast cancer lesions in ultrasound images by using attention layer and loss ensemble in deep convolutional neural networks,” *Diagnostics*, vol. 11, no. 10, p. 1859, 2021, doi: 10.3390/diagnostics11101859.
18. Y. Yan, Y. Xu, G. Fang, X. He, Y. Qian, and W. Zhu, “Multimodal deep learning for breast tumor classification: Integrating mammography and ultrasound for enhanced diagnostic accuracy,” *Journal of Applied Clinical Medical Physics*, 2026, doi: 10.1002/acm2.70464.
19. T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017, pp. 2980–2988, doi: 10.1109/ICCV.2017.324.
20. J. Gou, B. Yu, S. J. Maybank, and D. Tao, “Knowledge distillation: A survey,” *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789–1819, 2021, doi: 10.1007/s11263-021-01453-z.
21. S. Saha et al., “Knowledge distillation approach for skin cancer classification on lightweight deep learning model,” *Healthcare Technology Letters*, 2025, doi: 10.1049/htl2.12120.
22. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, 2020, doi: 10.1007/s11263-019-01228-7.
23. P. Tschandl, C. Rosendahl, and H. Kittler, “The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions,” *Scientific Data*, vol. 5, p. 180161, 2018, doi: 10.1038/sdata.2018.161.
24. W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, “Dataset of breast ultrasound images,” *Data in Brief*, vol. 28, p. 104863, 2020, doi: 10.1016/j.dib.2019.104863.
25. J. Yang et al., “MedMNIST v2 — A large-scale lightweight benchmark for 2D and 3D biomedical image classification,” *Scientific Data*, vol. 10, p. 41, 2023, doi: 10.1038/s41597-022-01721-8.