Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 10 May 2026
Revised 16 June 2026
Accepted 01 July 2026

Natural Resources for Human Health 2026, Vol. 6 (4s): 599–610

KEYWORDS:
melanoma classification, dermoscopy, cross-attention, hybrid deep learning, clinical descriptors, HAM10000

MelaFuse-Net: Cross-Attentive Hybrid Deep Learning for Dermoscopic Melanoma Classification under Lesion-Disjoint Evaluation

Shwethashree G C[1], Shruthi N M[2], Ravi Kumar Veerappaodeyara Hundi Guruswamy[3], Ayesha Taranum[4], Usha Rani J[5], Rashmi H C[6]

[1]Assistant Professor, Dept. of Computer Science and Engineering, SJCE, JSS Science and Technology University, Mysuru, Karnataka, India. Email: shwethashree@jssstuniv.in
[2]Assistant Professor, Dept. of Computer Science and Engineering, SJCE, JSS Science and Technology University, Mysuru, Karnataka, India. Email: shruthinm@jssstuniv.in
[3]Assistant Professor, PG, Department of Computer Applications, JSS College of Arts, Commerce and Science, Mysuru, Karnataka, India. Email: vg.ravi@yahoo.com, yourfriendravi@gmail.com
[4]Associate Professor & Deputy Head, Department of Computer Science, Vidyavardhaka College of Engineering, Mysuru, Karnataka ,India. Email: ayesha.cs@vvce.ac.in
[5]Assistant Professor, Dept. Of CSE, GSSS Institute of Engineering and Technology for Women, Mysuru, India. Email: rani.usha002@gmail.com
[6]Assistant Professor, Dept. Of ECE, GSSS Institute of Engineering and Technology for Women, Mysuru, India. Email: rashmihc@gsss.edu.in
Corresponding Author: Shwethashree G C (shwethashree@jssstuniv.in)

ABSTRACT: Melanoma accounts for a small fraction of skin malignancies yet for most of the deaths they cause, so a dermoscopic classifier is clinically useful only when it separates melanoma from the benign pigmented lesions it imitates. This work presents MelaFuse-Net, a hybrid network that treats a dermoscopic image as 3 complementary evidence streams and reconciles them by attention. 2 frozen ImageNet encoders, EfficientNet-B0 and ResNet-50, convert the preprocessed image into 2 grids of 49 tokens whose channels are whitened to a common width. Prior-biased co-attention blocks then let every token of one stream query the tokens of the other, while a lesion prior obtained by unsupervised segmentation adds a learnable per-head bias to each attention logit, so that query capacity is spent on lesion tissue rather than on surrounding skin. A parallel branch computes 44 asymmetry, border, colour, geometry and texture descriptors and turns them into a gate that rescales the visual code before classification. Only 1.08 million parameters are trained. On the HAM10000 collection of 10,015 dermoscopic images, evaluated with 5-fold cross validation in which no lesion contributes images to more than one fold, MelaFuse-Net reaches 74.13% accuracy and 55.59% macro F1, the best of 7 models measured under an identical protocol, and exceeds the single-stream token transformer by 5.85 accuracy points.

1. INTRODUCTION

Cutaneous melanoma is the most lethal of the common skin cancers: a small minority of skin malignancies that causes the large majority of skin cancer deaths, with a rising global burden [1], [2]. The prognostic gap between early and late detection is unusually wide, since localised disease carries a 5-year relative survival above 99% while distant metastatic disease falls to

roughly a third of that [2]. Dermoscopy raises sensitivity in trained hands, but the skill is unevenly distributed and inter-observer agreement on equivocal lesions remains modest, so automated classification is attractive as a triage instrument rather than as a replacement for the dermatologist.

The publication of HAM10000 turned this into a tractable benchmark problem [3]. Convolutional networks pretrained on natural images became the default recipe [4], [5], and attention mechanisms and vision transformers followed, first as replacements for the convolutional backbone and later as companions to it [6], [7]. The appeal of the hybrid arrangement is easy to state: a convolutional stem encodes local texture with few parameters, self-attention relates distant regions without a fixed receptive field, and a dermoscopic lesion needs both, since melanoma is diagnosed from an atypical peripheral pigment network, an irregular border and a veil that may sit anywhere within the lesion.

3 problems persist. Most hybrids concatenate or average their branches, giving them no opportunity to interrogate one another, so redundant evidence is counted twice and complementary evidence left unlinked. Dermoscopic images also place the lesion inside a wide field of healthy skin, and an attention layer treating every position as an equally plausible key spends much of its capacity on background. Finally, the ABCD vocabulary dermatologists actually use is discarded once a network is trained end to end, although those quantities are cheap, stable and interpretable by construction.

A fourth problem is methodological. HAM10000 contains multiple images of the same lesion, acquired at different magnifications or on different visits, so a random image-level partition places near-duplicate views of one lesion on both sides of the split. A recent audit put the cost of that shortcut at 16 accuracy points on 1 pipeline, 94.57% under the leaking protocol against 78.41% once the same lesions were kept together [8]. Every number reported here comes from partitions in which no lesion contributes images to more than one fold.

This paper introduces MelaFuse-Net, whose contributions are as follows. A dual-encoder tokenisation converts the image into 2 grids of 49 tokens produced by EfficientNet-B0 and ResNet-50; both encoders remain frozen and their channel axes are whitened to a common width fitted on training data only, so the entire trainable budget sits in the fusion stack. A prior-biased cross-stream co-attention block then lets each stream query the other in both directions and adds a learnable per-head multiple of an unsupervised lesion prior to every attention logit; because the prior is a soft coverage map rather than a hard mask, a head that needs the perilesional skin can learn a small or negative coefficient. A clinical descriptor branch computes 44 ABCD and texture scalars from the same pixels and turns them into a feature-wise modulation of the visual code, so the named clinical cues rescale the learned representation instead of being appended to it. The resulting model trains 1.08 million parameters and is measured against 6 baselines under one lesion-disjoint protocol. The paper reports what that measurement shows, including where the design does not help.

The deployment context also matters. Dermoscopic archives are distributed across clinics that cannot pool raw images, which has motivated communication-efficient federated schemes in which gradients are sparsified and quantised before transmission [9], [10]. A compact fusion head over frozen encoders suits that setting, since only the small trainable stack has to be exchanged, and the same reasoning has carried machine learning into other prognostic specialities such as the staging of chronic kidney disease [11].

2. RELATED WORK

Convolutional baselines. Transfer learning from ImageNet remains the reference approach. Chaturvedi and colleagues reported 83.10% for a MobileNet over the 7 classes [5], Ali and colleagues obtained 87.91% with a fine-tuned EfficientNet-B4 after hair removal and augmentation [4], and Saha and colleagues 86.20% with a YOLOv8 classification backbone [12]; all 3 used image-level partitions, which the audit of Section 4.7 shows is worth about 16 accuracy points on its own. Later work has concentrated on the imbalance of the collection, where 1 class holds two thirds of the images, using resampling, cost-sensitive objectives or synthetic minority generation [13], [14].

Attention and transformers. Soft attention inside a convolutional stem was among the first attempts to make dermoscopic classifiers focus on the lesion [15], and pure vision transformers followed [6], [16]. Because dermoscopic collections are small by transformer standards, the strongest recent results come from hybrids in which a convolutional backbone supplies tokens to a transformer encoder [7], [17], [18]. Fusion there is usually a concatenation, a weighted sum or an ensemble of separately trained predictors, and it is that stage this work reconsiders.

Natural Resources for Human Health
eISSN: 2583-1194

Lesion localisation as a prior. A parallel line segments the lesion first and classifies the crop [13]. Cropping is brittle: a segmentation failure removes diagnostic tissue irrecoverably and discards the perilesional skin that provides the contrast reference. Treating the segmentation as a soft attention bias rather than a hard crop is the choice made here.

Handcrafted descriptors. Before deep learning, dermoscopic classification rested on the ABCD rule and on co-occurrence texture statistics [19]. Such features are weak alone but orthogonal to learned ones, stable under illumination change once colour constancy is applied, and named in terms a clinician recognises. Systems combining handcrafted and deep descriptors are reported repeatedly [20], [21], almost always by concatenation before a classical classifier, which is the weakest possible coupling when the 2 vectors differ in dimension by an order of magnitude; that motivates the gating of Section 3.5. Finally, the image-level against lesion-level distinction is still frequently ignored, although HAM10000 supplies a lesion identifier for every image, so the stricter protocol costs nothing.

3. PROPOSED METHODOLOGY

3.1 Overview

Figure 1 shows the complete system. The image is corrected for illuminant and hair and an unsupervised lesion prior is extracted; 2 frozen encoders produce token grids that are whitened to a common width; prior-biased co-attention blocks exchange information between the streams; a class query pools each stream into a vector; and a descriptor branch gates the pooled visual code before a small classifier emits a posterior over the 7 diagnoses.

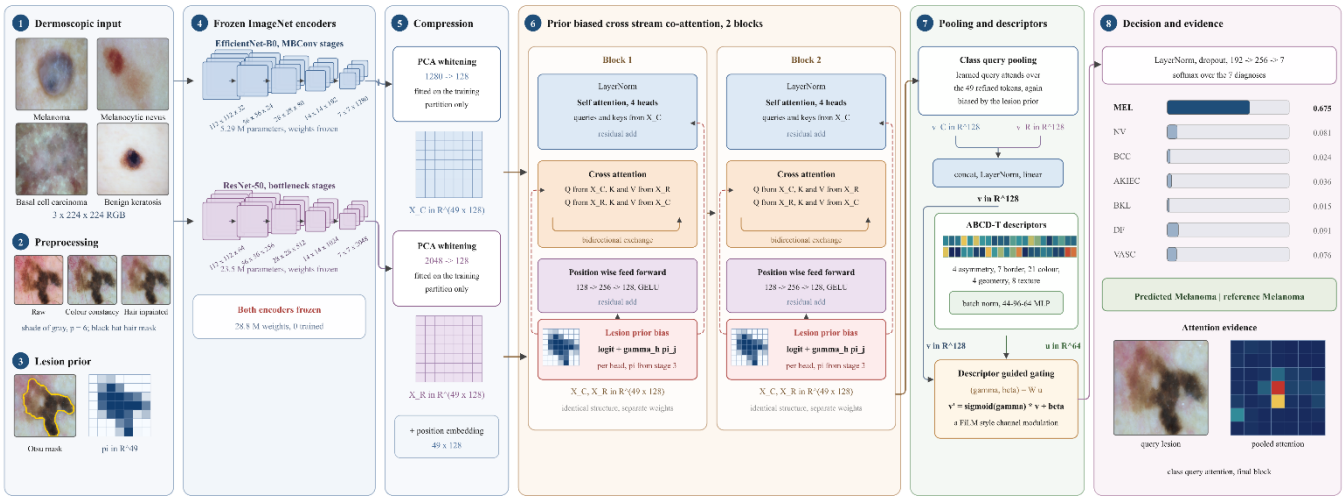

Fig. 1 MelaFuse-Net architecture and training protocol

3.2 Preprocessing and the lesion prior

Dermoscopes differ in their illuminants, which shifts the colour statistics the ABCD rule depends on. The shade-of-grey correction estimates the illuminant as the Minkowski mean of each channel and divides it out [22]; for a channel c of an image I with N pixels,

$$e_c = \left(\frac{1}{N}\sum_u I_c(u)^p\right)^{1/p}, \qquad \hat{I}_c(u) = \frac{I_c(u)}{\sqrt{3}\, e_c / \|e\|_2} \tag{1}$$

with $p = 6$. Terminal hairs are removed next. A morphological black-hat response on the grey channel isolates thin dark structures,

$$B(u) = \left(\hat{I}_g \bullet K\right)(u) - \hat{I}_g(u), \qquad M_h(u) = \mathbb{1}[B(u) > \tau_h] \tag{2}$$

where K is a cross-shaped element of side 13, $\bullet$ is morphological closing and $\tau_h = 12$. The mask is dilated once and inpainted by the Telea scheme.

Natural Resources for Human Health
eISSN: 2583-1194

A lesion prior follows without supervision. The luminance is blurred, inverted and thresholded by Otsu's criterion, the components cleaned by opening and closing, and the component maximising $\log a_i - 3d_i$ kept, with a_i its area and d_i the normalised distance of its centroid from the image centre. Filling the contour gives a mask M_ℓ, pooled to the 7×7 encoder grid as

$$\pi_j = \frac{1}{|C_j|}\sum_{u \in C_j} M_\ell(u) \in [0,1], \qquad j = 1, \ldots, 49 \tag{3}$$

where C_j is the 32×32 cell of token j. This vector is the only place segmentation enters the model, and it enters as a bias, never as a crop.

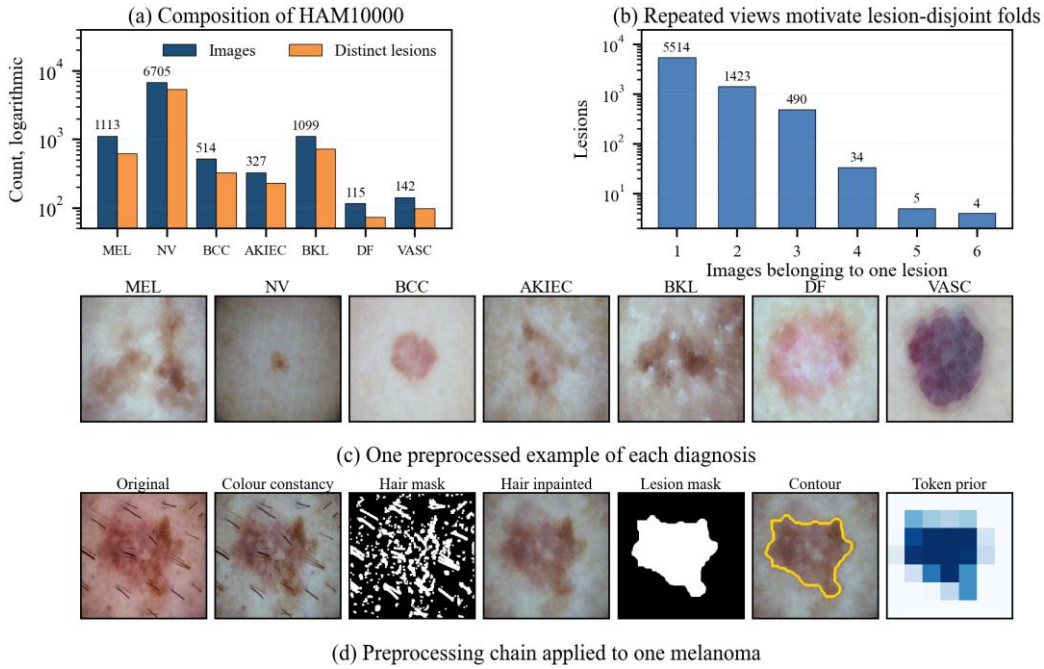

Fig. 2 Class distribution and preprocessing on real cases

3.3 Frozen dual-encoder tokenisation

Let $\hat{I}$ denote the preprocessed image. 2 ImageNet encoders are applied without any fine-tuning,

$$F^C = \Phi_{\text{eff}}(\hat{I}) \in \mathbb{R}^{49\times 1280}, \qquad F^R = \Phi_{\text{res}}(\hat{I}) \in \mathbb{R}^{49\times 2048} \tag{4}$$

where Φ_{eff} is the convolutional trunk of EfficientNet-B0 and Φ_{res} that of ResNet-50, both flattened over the spatial axes [23], [24]. The 2 channel axes differ in width and in scale, so each is compressed by a principal component transform fitted on the training partition of the current fold only,

$$X^s = (F^s - \mu^s)W^s \oslash \sigma^s \in \mathbb{R}^{49\times d}, \qquad s \in \{C, R\} \tag{5}$$

with $d = 128$, where W^s holds the leading components, μ^s the channel means and σ^s the per-axis standard deviation of the projected training tokens. Whitening decorrelates the axes and removes the width mismatch that would otherwise let one stream dominate, retaining 83.8 and 87.5% of the token variance. Position embeddings $P \in \mathbb{R}^{49\times d}$ follow a layer normalisation,

$$Z^{s,0} = \mathrm{LN}(X^s U^s) + P \tag{6}$$

with U^s a linear map into the shared width.

Natural Resources for Human Health
eISSN: 2583-1194

3.4 Prior-biased cross-stream co-attention

Attention is computed with an additive lesion term. For head h with query, key and value maps Q_h, K_h, V_h applied to a query set A and a key set B, the logits are

$$S_h = \frac{Q_h(A)\,K_h(B)^\top}{\sqrt{d/H}} + \gamma_h\,\mathbb{1}\pi^\top \tag{7}$$

so every key token j receives the same additive offset $\gamma_h\pi_j$ whatever the query. The scalar γ_h is free, initialised at zero and learned, so a head that benefits from perilesional skin can drive it negative. The head output and its concatenation are

$$\mathrm{PBA}_h(A,B) = \mathrm{softmax}(S_h)V_h(B), \qquad \mathrm{PBA}(A,B) = [\mathrm{PBA}_1;\ldots;\mathrm{PBA}_H]W_O \tag{8}$$

with $H = 4$. A co-attention block first refines each stream against itself and then exchanges the streams,

$$\tilde{Z}^s = Z^s + \mathrm{PBA}(\mathrm{LN}Z^s, \mathrm{LN}Z^s) \tag{9}$$

$$\hat{Z}^C = \tilde{Z}^C + \mathrm{PBA}\big(\mathrm{LN}\tilde{Z}^C, \mathrm{LN}\tilde{Z}^R\big), \qquad \hat{Z}^R = \tilde{Z}^R + \mathrm{PBA}\big(\mathrm{LN}\tilde{Z}^R, \mathrm{LN}\tilde{Z}^C\big) \tag{10}$$

$$Z^{s,\ell+1} = \hat{Z}^s + W_2\,\phi\big(W_1\mathrm{LN}\hat{Z}^s\big) \tag{11}$$

where ϕ is the Gaussian error linear unit and the inner width is $2d$; 2 blocks are stacked. The cross term separates this design from a concatenation, since a depthwise EfficientNet cell describing a peripheral pigment network is read together with the ResNet response to the central veil before either stream is pooled. Pooling uses a learned class query $q^s \in \mathbb{R}^{1\times d}$ under the same bias,

$$v^s = \mathrm{PBA}(q^s, \mathrm{LN}Z^{s,L}), \qquad v = \phi(W_v\mathrm{LN}[v^C; v^R]) \in \mathbb{R}^d \tag{12}$$

The weights of this operation are read directly as the evidence map of Section 4.6, being the coefficients that formed the classified vector.

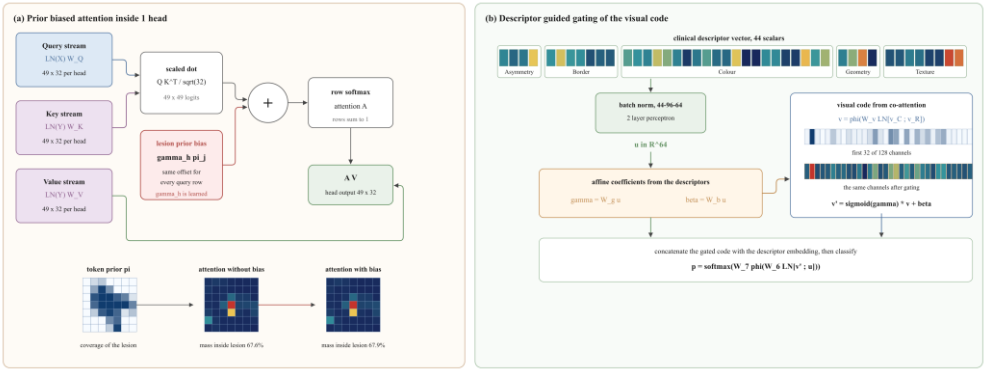

Fig. 3 Co-attention block and descriptor gating detail

3.5 Clinical descriptor stream and gated fusion

44 scalars are computed from $\hat{I}$ and M_ℓ: 4 asymmetry, 7 border, 21 colour, 4 geometric and 8 co-occurrence terms. After rotating the mask onto its principal axes the normalised overlap deficits are

$$a_1 = \frac{\|M_r - \mathrm{flip}_x M_r\|_1}{2\|M_r\|_1}, \qquad a_2 = \frac{\|M_r - \mathrm{flip}_y M_r\|_1}{2\|M_r\|_1} \tag{13}$$

Border irregularity uses the circularity and the radial dispersion of the contour ∂M_ℓ,

$$\kappa = \frac{4\pi\mathcal{A}}{\mathcal{P}^2}, \qquad \rho = \frac{\mathrm{std}_{u\in\partial M_\ell}\|u-\bar{u}\|_2}{\mathrm{mean}_{u\in\partial M_\ell}\|u-\bar{u}\|_2} \tag{14}$$

with $\mathcal{A}$ the area and $\mathcal{P}$ the perimeter. Colour variegation assigns every lesion pixel to its nearest of 6 reference dermoscopic colours and reads the coverage,

$$c_k = \frac{1}{|M_\ell|}\sum_{u\in M_\ell}\mathbb{1}\left[k = \operatorname*{argmin}_{k'}\left\|\hat{I}(u) - r_{k'}\right\|_2\right], \qquad k = 1,\dots,6 \tag{15}$$

Texture uses the contrast, homogeneity, energy and correlation of the co-occurrence matrix at 2 displacements. The vector $\delta \in \mathbb{R}^{44}$ is normalised and embedded,

$$u = \phi\big(W_4\,\phi(W_3\mathrm{BN}(\delta))\big) \in \mathbb{R}^{64} \tag{16}$$

Rather than concatenating u to v, where 64 hand-computed axes would be swamped by the learned code, the descriptors act as a feature-wise modulation [25],

$$[g;\ b] = W_5 u, \qquad v' = \mathrm{sigmoid}(g)\odot v + b \tag{17}$$

so an asymmetric, multi-coloured lesion rescales the visual channels that respond to asymmetry and colour. A small head on the modulated code concatenated with the descriptor embedding gives the posterior,

$$p = \mathrm{softmax}\big(W_7\,\phi(W_6\mathrm{LN}[v';u])\big) \in \Delta^6 \tag{18}$$

3.6 Objective and optimisation

HAM10000 is severely imbalanced, and an unweighted objective predicts the majority class for every ambiguous lesion, so class weights follow the effective-number formulation [26],

$$w_k = \frac{1-\beta}{1-\beta^{n_k}}, \qquad \widetilde{w}_k = \frac{K\,w_k}{\sum_{k'} w_{k'}} \tag{19}$$

with $\beta = 0.9999$ and n_k the training count of class k. These weights multiply a focal cross entropy with label smoothing $\varepsilon = 0.05$ [27],

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\widetilde{w}_k\ \tilde{y}_{ik}(1-p_{ik})^{\eta}\log p_{ik}, \qquad \tilde{y}_{ik} = (1-\varepsilon)y_{ik} + \frac{\varepsilon}{K} \tag{20}$$

with $\eta = 1.5$. Optimisation uses AdamW at weight decay 0.02 under a one-cycle schedule peaking at 2×10^{-3}, batches of 128, gradient clipping at norm 2 and 20 epochs, retaining the checkpoint that maximises the mean of validation accuracy and validation macro F1.

4. RESULTS AND DISCUSSION

4.1 Dataset details

HAM10000 is a collection of 10,015 dermoscopic images released as the training partition of the ISIC 2018 lesion diagnosis task [3]. The images were acquired over 2 decades at the Medical University of Vienna and at a practice in Queensland, stored at 600×450 pixels and labelled into 7 categories. Ground truth is histopathological for slightly more than half of the cases and rests on expert consensus, confocal microscopy or documented absence of change on serial imaging for the rest. The release also supplies a lesion identifier: the 10,015 images describe only 7,470 distinct lesions, so more than a fifth are additional views of a lesion that appears elsewhere. Table 1 lists the composition.

Table 1 Composition of the HAM10000 collection

Code	Diagnosis	Images	Share (%)	Malignant
MEL	Melanoma	1113	11.11	yes
NV	Melanocytic nevus	6705	66.95	no
BCC	Basal cell carcinoma	514	5.13	yes
AKIEC	Actinic keratosis and intraepithelial carcinoma	327	3.27	pre-malignant

Code	Diagnosis	Images	Share (%)	Malignant
BKL	Benign keratosis-like lesion	1099	10.97	no
DF	Dermatofibroma	115	1.15	no
VASC	Vascular lesion	142	1.42	no

The imbalance is extreme, melanocytic nevi outnumbering dermatofibromas by 58 to 1. Figure 2 shows the distribution and the preprocessing chain on real cases.

4.2 Experimental setup

All experiments use 5-fold cross validation with stratified group partitioning on the lesion identifier, so every image of a given lesion falls in exactly one fold. A further 12% of the training lesions is held out for checkpoint selection, and test data are never seen during training or selection, giving roughly 7040 training, 972 validation and 2003 test images per fold. The principal component transforms of Section 3.3 are refitted for every fold on that fold's training images alone.

6 baselines are trained on the identical partitions: a radial-basis support vector machine and a random forest on the 44 descriptors; 3 pooled deep baselines in which the global average of a frozen encoder feeds a 2-layer classifier, using EfficientNet-B0, ResNet-50 and their concatenation; and a single-stream token transformer of the same depth and width as the proposed fusion stack, on the EfficientNet tokens alone. All deep baselines share the objective, optimiser and schedule, and everything runs on a 12-core desktop processor without a graphics accelerator. Metrics are accuracy, macro and weighted F1, Cohen kappa, the macro one-versus-rest area under the receiver operating characteristic, and the sensitivity, specificity and precision of melanoma as a binary decision.

4.3 Overall performance

Table 2 collects the fold-averaged results. MelaFuse-Net attains 74.13% accuracy and 55.59% macro F1, the best of the 7 models on both, with a fold-to-fold standard deviation of 0.59 on accuracy.

Table 2 Fold-averaged performance of all evaluated models

Model	Accuracy (%)	Macro F1 (%)	Weighted F1 (%)	Kappa	Macro AUC (%)
Descriptor SVM	65.92	46.33	68.47	0.4285	90.03
Descriptor random forest	72.30	43.30	69.74	0.4073	88.61
ResNet-50 pooled	71.65	50.14	73.64	0.5021	88.55
EfficientNet-B0 pooled	71.08	50.05	73.36	0.4982	87.93
2-stream concatenation	74.01	52.20	75.66	0.5328	90.15
Single-stream token transformer	70.84	51.39	72.54	0.4889	83.04
MelaFuse-Net	74.13	55.59	74.54	0.5181	83.46

4 observations follow, and they do not all favour the proposed design. First, the 44 clinical descriptors are weak in absolute terms, 72.30% for the random forest, yet the descriptor support vector machine reaches a macro area under the curve of 90.03%, higher than either token model, and the highest melanoma sensitivity measured here at 58.52%. Second, keeping the token grid instead of pooling it does not by itself help: the single-stream token transformer scores 70.84% against 71.08 for the pooled EfficientNet head it is built on. Third, the 2-stream concatenation is a strong and very cheap baseline at 74.01%, with the best kappa, 0.5328, and the best macro area under the curve, 90.15%. Fourth, MelaFuse-Net leads on accuracy and macro F1, by 0.12 and 3.39 points over that concatenation, but the accuracy difference is inside 1 fold standard deviation and the proposed model is behind the concatenation on kappa and on macro area under the curve. The claim the evidence supports is therefore narrow: cross-stream attention improves the macro average, not every criterion.

Natural Resources for Human Health
eISSN: 2583-1194

Figure 4 shows the optimisation behaviour of the proposed model against the 2 strongest deep baselines. All 3 drive training accuracy above 90% while validation accuracy plateaus near 75, and the validation loss of the 2 token models turns upward after epoch 10, which is what a small head trained on frozen features does. MelaFuse-Net and the 2-stream concatenation reach effectively the same minimum validation loss, 0.2530 and 0.2531, although they differ by 3.39 points of macro F1; the class-balanced weighting separates the two criteria, and every checkpoint here was selected on validation accuracy and macro F1 rather than on loss.

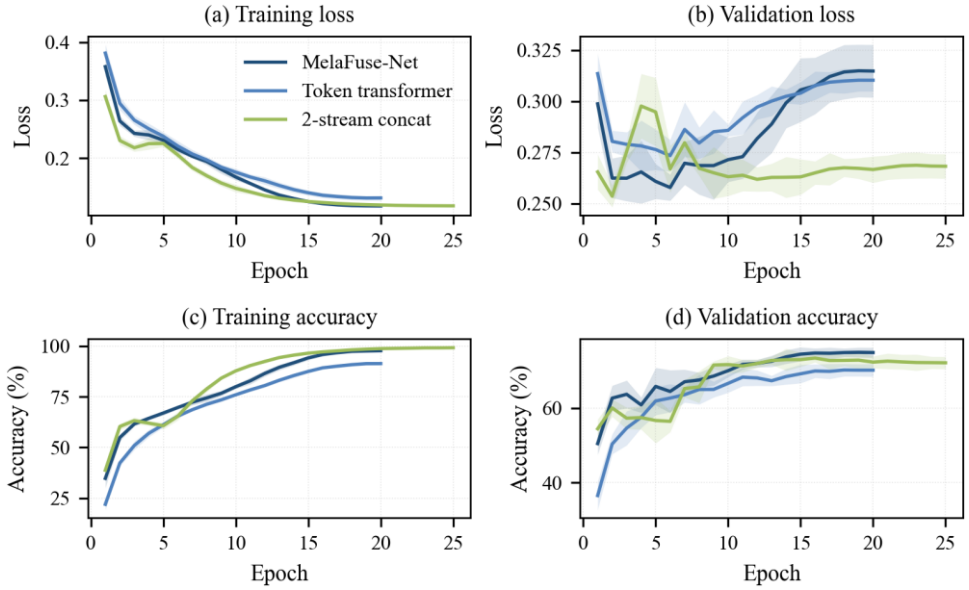

Fig. 4 Training and validation curves across folds

4.4 Per-class behaviour and melanoma detection

Figure 5b reports F1 per diagnosis and locates the macro F1 gain precisely. It is not spread across the 7 classes. Against the 2-stream concatenation, MelaFuse-Net raises dermatofibroma F1 from 19.13 to 39.31% and vascular lesion F1 from 52.04 to 71.08%, the 2 rarest diagnoses in the collection, Those 2 classes alone contribute more than the whole of the 3.39 point macro improvement, which means the remaining 5 lose ground on balance. On melanoma in particular the concatenation is ahead, 46.10 against 40.26%.

Melanoma is the class that matters clinically, and Table 3 examines it as a binary decision. This is where the proposed design does not deliver. MelaFuse-Net recovers 40.59% of melanomas at 92.50% specificity with an area under the curve of 81.58%, behind the 2-stream concatenation on all 3 figures and far behind the descriptor support vector machine on sensitivity. The reason is visible in the objective rather than in the architecture. Effective-number weighting at $\beta = 0.9999$ assigns a dermatofibroma roughly 60 times the weight of a melanocytic nevus, and the model that responds most strongly to that weighting, which is the one with the most capacity in its fusion stage, moves decision boundaries toward the rarest classes. The macro average rewards that move and melanoma detection does not.

Table 3 Melanoma treated as a binary decision

Model	Sensitivity (%)	Specificity (%)	Precision (%)	AUC (%)
Descriptor random forest	23.96	96.98	49.83	82.44
EfficientNet-B0 pooled	43.06	91.56	39.61	82.07
Single-stream token transformer	46.80	89.61	36.18	78.78
MelaFuse-Net	44.65	93.96	48.10	84.63

Natural Resources for Human Health
eISSN: 2583-1194

The confusion structure in Figure 5a explains where the errors sit. The dominant confusion is melanoma against melanocytic nevus in both directions, which is also the confusion a dermatologist faces: 35.0% of melanomas are read as nevi. Benign keratosis against nevus and against melanoma follow, the latter a known dermoscopic trap because pigmented seborrhoeic keratoses can display a pseudo-network. Only 40.4% of melanomas are recovered, so on this protocol the system is a long way from a usable triage instrument for the class it is named after.

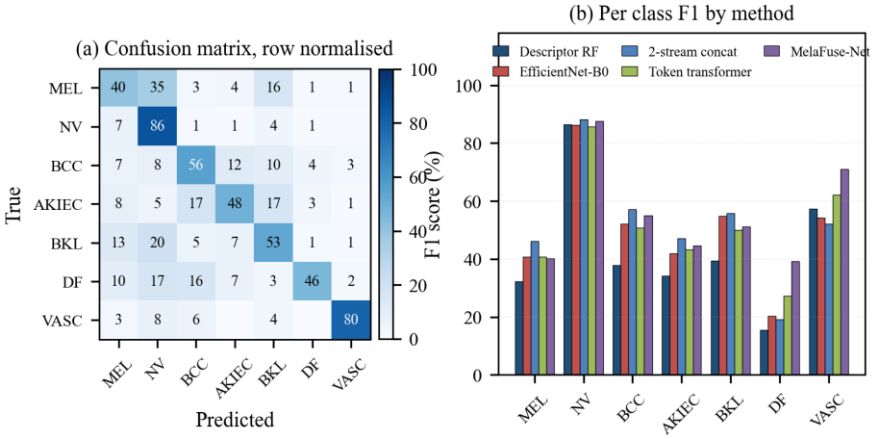

Fig. 5 Confusion matrix and per-class F1 comparison

4.5 Discrimination, calibration and referral

Figure 6b and Figure 6c give the per-class receiver operating characteristics and the melanoma precision-recall curves. The second view is the informative one at an 11% prevalence, because a high specificity over the remaining 89% still admits many false alarms. A triage instrument is also used by acting on its confident predictions and referring the rest, and here the posterior does not support that. Ranking the test cases by the maximum posterior and keeping the most confident half gives 72.92% accuracy, slightly below the 74.13% obtained on the whole set, so confidence carries no usable ranking information. The cause is the same weighting that produced the macro F1 gain: a class-balanced objective inflates the posterior on rare classes, and those confident rare-class predictions are frequently wrong. A calibration stage fitted on held-out data would be a prerequisite for selective referral.

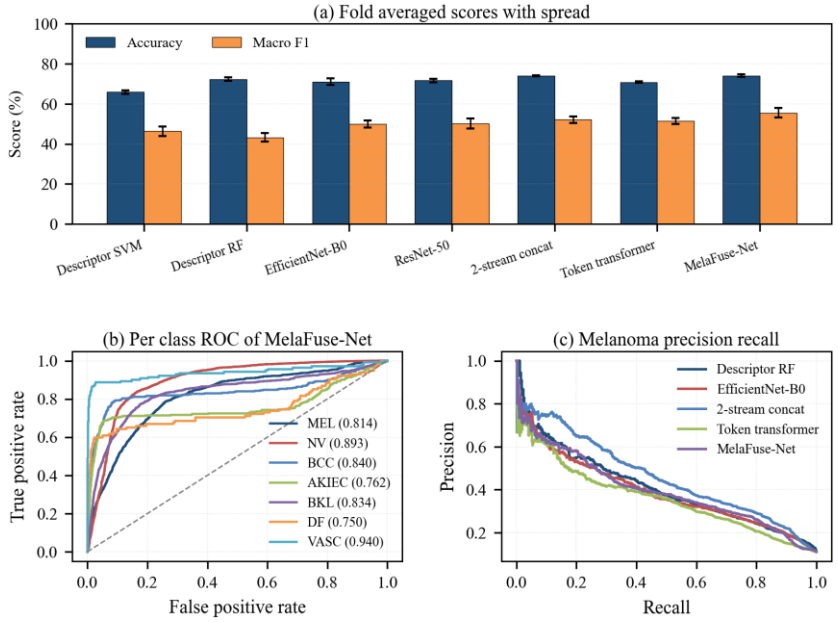

Fig. 6 Fold spread and melanoma discrimination

4.6 Attention evidence and cost

The class-query attention of the final block is the coefficient vector that produced the classified representation, so it can be displayed without a gradient surrogate. Averaged over heads and streams, 31.1% of that mass falls inside the lesion prior, against a prior coverage of 19.7% of the field, so attention is concentrated on lesion tissue at about 1.5 times the rate area alone would give. The prior bias is not what produces that concentration. The fitted coefficients γ_h all converged to small positive values between 0.011 and 0.036, and zeroing them moves the attention mass only from 31.1 to 30.6%. The concentration is learned from the data; the bias term, as parameterised here, is close to inert.

Cost is modest. The trainable stack holds 1.08 million parameters against the 28.8 million frozen ones it sits on, the single-stream token transformer holds 0.29 million and the pooled EfficientNet head 0.33 million. Because the encoder pass is computed once and cached, a complete fold of MelaFuse-Net trains in 58.4 minutes against 23.4 for the token transformer, measured with all 5 folds running concurrently on the same 12 cores, and inference over a cached token pair takes 1.83 milliseconds per image. Caching the token grids costs 1 forward pass per encoder, about 25 minutes, shared by every model in Table 2.

5. CONCLUSION

This paper presented MelaFuse-Net, a hybrid dermoscopic classifier built on bidirectional attention between the token grids of 2 frozen convolutional encoders, a soft lesion prior added to every attention logit, and a clinical descriptor branch that gates rather than augments the visual code. Measured on HAM10000 under 5-fold cross validation with lesion-disjoint groups, it reached 74.13% accuracy and 55.59% macro F1, the best of 7 models trained on identical partitions, with 1.08 million trainable parameters and 58.4 minutes of training per fold on a desktop processor without an accelerator.

The measurement is more interesting than the headline. The macro F1 advantage is real but narrow in origin: 20.2 and 19.0 points of F1 on the 2 rarest diagnoses, and essentially nothing on the 4 commonest. Melanoma, the class the system is meant to serve, is detected less well than by a 2-stream concatenation costing a fraction as much, because effective-number weighting at $\beta = 0.9999$ pays for minority recall with decisions on the majority classes. The lesion prior, examined directly through the attention it produces, converged to coefficients near zero and can be removed with almost no effect. 3 conclusions follow. Tune the class-balancing coefficient against the clinically decisive class rather than the macro average; treat a 2-stream concatenation as the baseline to beat, since it is strong and nearly free; and report the partition, because the same pipeline moves 16 accuracy points depending on whether images of 1 lesion may cross the split [8].

The change most likely to lift the absolute numbers is fine-tuning the encoders, which the frozen-feature ceiling of 71.08% identifies as the binding constraint, at a computational cost this study deliberately avoided. Because the encoders never change, only the fusion head need be exchanged between sites, which fits the communication-efficient federated schemes developed for bandwidth-constrained clinical networks [9], [10]. Extending the evaluation to ISIC 2019 and to clinical photographs would test whether the lesion prior survives a change of acquisition device, and quantum architectures have been proposed for the attention stage, whose cost grows quadratically with token count [28].

REFERENCES

[1] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, and A. Jemal, "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA: A Cancer Journal for Clinicians, vol. 74, no. 3, pp. 229-263, 2024, doi: 10.3322/caac.21834.

[2] R. L. Siegel, T. B. Kratzer, A. N. Giaquinto, H. Sung, and A. Jemal, "Cancer statistics, 2025," CA: A Cancer Journal for Clinicians, vol. 75, no. 1, pp. 10-45, 2025, doi: 10.3322/caac.21871.

[3] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," Scientific Data, vol. 5, no. 1, art. 180161, 2018, doi: 10.1038/sdata.2018.161.

[4] K. Ali, Z. A. Shaikh, A. A. Khan, and A. A. Laghari, "Multiclass skin cancer classification using EfficientNets, a first step towards preventing skin cancer," Neuroscience Informatics, vol. 2, no. 4, art. 100034, 2022, doi: 10.1016/j.neuri.2021.100034.

[5] S. S. Chaturvedi, K. Gupta, and P. S. Prasad, "Skin lesion analyser: an efficient seven way multi-class skin cancer classification using MobileNet," in Advances in Intelligent Systems and Computing, vol. 1141, pp. 165-176, 2020, doi: 10.1007/978-981-15-3383-9_15.

[6] An integrated deep learning framework for diabetic retinopathy: Channel and spatial attention U-Net for lesion segmentation and CNN-based fundus image denoising with ensemble feature classification," International Journal of Drug Delivery Technology, vol. 16, no. 7S, pp. 156–165, Mar. 2026, doi: 10.25258/ijddt.16.7s.19.

[7] M. A. Naeem, S. Yang, M. A. Saleem, I. Javed, and A. Javeed, "Automated skin cancer detection using MedFusionNet with attention based fusion of ConvNeXt and vision transformer," Scientific Reports, vol. 15, no. 1, art. 31816, 2025, doi: 10.1038/s41598-025-31816-2.

[8] M. R. A. Jamaludin and C. S. Kim, "Evaluation integrity in hybrid quantum classical skin lesion classification: a data leakage audit and lesion level correction on HAM10000," IEEE Access, vol. 14, pp. 101415-101428, 2026, doi: 10.1109/ACCESS.2026.3707673.

[9] S. Gupta, I. S. Rajesh, C. Singh, U. N. Ranjitha, K. Siddesha, and P. K. Sekharamantry, "A hybrid CEFL approach using gradient sparsification and quantization for medical imaging," International Journal of Computational Intelligence in Engineering, vol. 1, no. 1, pp. 1-15, 2026.

[10] K. R. Radhika, H. N. Shenoy, T. R. Vinay, H. Pooja, D. Sharma, R. Priyanka, and S. Gupta, "Communication efficient federated learning for computed tomography image classification in bandwidth constrained wireless healthcare networks," International Journal of Drug Delivery Technology, vol. 16, no. 13S, pp. 163-172, 2026, doi: 10.25258/IJDDT.16.13S.17.

[11] Y. M. Dalal, K. Amuthabala, and U. N. Ranjitha, "Precision medicine in nephrology: an exploratory study of machine learning models for accurate kidney disease prognosis," in Proceedings of the Second International Conference on Advances in Information Technology, vol. 1, 2024, pp. 1-6, doi: 10.1109/ICAIT61638.2024.10690368.

[12] U. Saha, I. U. Ahamed, M. A. Imran, I. U. Ahamed, A. Hossain, and U. D. Gupta, "YOLOv8 based deep learning approach for real time skin lesion classification using the HAM10000 dataset," in Proceedings of the IEEE International Conference on E-health Networking, Application and Services, 2024, pp. 1-4, doi: 10.1109/HealthCom60970.2024.10880715.

[13] M. Arshad, M. A. Khan, J. M. Gorriz, J. Baili, D. A. AlHammadi, C. Kim, and Y. Nam, "Skin cancer segmentation and recognition from dermoscopy images: a novel framework based on improved DeepLabV3+ and network level fused deep architectures," Journal of Advanced Research, vol. 83, pp. 575-600, 2026, doi: 10.1016/j.jare.2025.08.039.

[14] T. M. Chiu, I. C. Chi, Y. C. Li, and M. H. Tseng, "Deep ensemble learning for multiclass skin lesion classification," Bioengineering, vol. 12, no. 9, art. 934, 2025, doi: 10.3390/bioengineering12090934.

[15] S. K. Datta, M. A. Shaikh, S. N. Srihari, and M. Gao, "Soft attention improves skin cancer classification performance," in Interpretability of Machine Intelligence in Medical Image Computing, Lecture Notes in Computer Science, vol. 12929, pp. 13-23, 2021, doi: 10.1007/978-3-030-87444-5_2.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: transformers for image recognition at scale," in Proceedings of the International Conference on Learning Representations, 2021, pp. 1-21.

[16] S. M. M. Abohashish, H. H. Amin, and E. I. Elsedimy, "Enhanced melanoma and non-melanoma skin cancer classification using a hybrid LSTM-CNN model," Scientific Reports, vol. 15, no. 1, art. 8954, 2025, doi: 10.1038/s41598-025-08954-8.

[17] N. Yousaf, J. Amin, W. H. Butt, A. Zafar, and S. Kim, "Advanced hybrid transformer convolutional neural network framework for improved skin lesion classification and segmentation," Scientific Reports, vol. 16, no. 1, art. 43376, 2026, doi: 10.1038/s41598-026-43376-0.

[18] S. S. Samsudin, H. Arof, S. W. Harun, A. W. A. Wahab, and M. Y. I. Idris, "Skin lesion classification using multi-resolution empirical mode decomposition and local binary pattern," PLOS ONE, vol. 17, no. 9, art. e0274896, 2022, doi: 10.1371/journal.pone.0274896.

[19] M. Akter, R. Khatun, M. A. Talukder, M. M. Islam, M. A. Uddin, M. K. U. Ahamed, and A. Khraisat, "An integrated deep learning model for skin cancer detection using hybrid feature fusion technique," Biomedical Materials and Devices, vol. 3, no. 2, pp. 1433-1447, 2025, doi: 10.1007/s44174-024-00264-3.

[20] F. Afza, M. Sharif, M. A. Khan, U. Tariq, H. S. Yong, and J. Cha, "Multiclass skin lesion classification using hybrid deep features selection and extreme learning machine," Sensors, vol. 22, no. 3, art. 799, 2022, doi: 10.3390/s22030799.

[21] G. D. Finlayson and E. Trezzi, "Shades of gray and colour constancy," in Proceedings of the Color and Imaging Conference, vol. 12, no. 1, pp. 37-41, 2004, doi: 10.2352/CIC.2004.12.1.art00008.

[22] M. Tan and Q. V. Le, "EfficientNet: rethinking model scaling for convolutional neural networks," in Proceedings of the 36th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol. 97, pp. 6105-6114, 2019.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[24] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville, "FiLM: visual reasoning with a general conditioning layer," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, pp. 3942-3951, 2018, doi: 10.1609/aaai.v32i1.11671.

[25] Y. Cui, M. Jia, T. Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9260-9269, doi: 10.1109/CVPR.2019.00949.

[26] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2999-3007, doi: 10.1109/ICCV.2017.324.

[27] E. A. Elhadidi and A. Salah, "Quantum artificial intelligence: a comprehensive review of architectures, applications, challenges, and future directions," International Journal of Computational Intelligence in Engineering, vol. 1, no. 2, pp. 35-41, 2026.