Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 22, No. 1, 2026

Research on Dermoscopic Image Classification Method Integrating Multi-Scale Features and Attention Mechanism

Yuchen Jiang[1], Xuewen Ding[1, *]

[1]School of Electronic Engineering Tianjin University of Technology and Education Tianjin, China
[*]Corresponding author: dingxw@tute.edu.cn

Abstract: To address the issues of lesion scale differences, easy loss of fine-grained pigment information, and background interference in dermoscopic images, this paper constructs a multi-scale attention fusion classification network MSAF-EfficientNetV2 based on EfficientNetV2-S. The network extracts features in four stages. AMSF achieves dynamic fusion through channel mapping, spatial alignment, and sample-related scale weights. LAA strengthens the main body of the lesion and irregular boundaries with channel-spatial attention, and uses weighted cross-entropy to alleviate the long-tailed distribution of the seven classes. The experimental results used publicly available data to form two levels of evidence: In the MedMNIST+ 224×224 end-to-end benchmark, the accuracy of DenseNet121 was 84.74±0.51%, and the AUC of DINO ViT-B/16 was 96.50±0.51%; in the histopathologically confirmed true melanoma ISIC_0000013, the Otsu dark region accounted for 22.29%, and the darkest cluster in Lab-K-means accounted for 16.60%, with a Dice concordance of 85.38%. The publicly available ISIC 2017 segmentation experiment further showed that the Dice/Jaccard ratio of VGG16-U-Net was 91.5%/84.6%, and the Dice/Jaccard ratio in this case reached 96.2%/92.6%. The total number of network parameters is 20.069 M and the number of FLOPs is 5.775 G, indicating that the newly added fusion and attention structures maintain controllable computational overhead.

Keywords: Dermoscopic Images; Skin Lesion Classification; EfficientNetV2; Multi-scale Feature Fusion; Attention Mechanism; Deep Learning.

1. Introduction

Dermoscopic imaging can present shallow structures such as pigment networks, color dots, blood vessel morphology, boundary regularity, and local color changes, providing rich visual evidence for skin lesion classification. Shetty et al. used machine learning and convolutional neural networks to analyze dermoscopic images and verified the effectiveness of deep convolutional features for identifying multiple types of lesions [1]. Publicly available data also have problems such as differences in the number of categories, limited samples of a minority class, and large intra-class morphological changes. Yao et al. studied single-model classification for small samples and class imbalance, and reduced training bias through enhancement and weighted learning [2]. Therefore, classification networks need to take into account fine-grained representation, imbalanced learning, and generalization stability.

High-level features at the end of convolutional networks have strong semantic abstraction capabilities, and continuous downsampling can weaken edge details, local pigment structures, and small-scale abnormal regions. Pérez and Ventura studied the accuracy of melanoma diagnostic models from the perspective of training process and convolutional model configuration [3]; Qian et al. used multi-scale attention grouping and class-specific loss for skin lesion classification [4]; Wang et al. improved feature discrimination ability by strengthening intra-class consistency and inter-class discriminability [5]. These works show that cross-layer information utilization and discriminative feature learning have a direct effect on complex lesions.

Multicenter imaging also introduces differences in equipment, illumination, and acquisition protocols. Yue et al. used adaptive weighted balance loss to study multicenter skin lesion classification [6]; Ajmal et al. fused deep features from multiple networks and used Grad-CAM to analyze lesion regions [7]. Hu et al. have combined EfficientNetV2 with multiscale feature fusion and class weight loss [8], so this paper focuses on the innovation of input-related scale gating and lesion attention after fusion. Hosny et al. carried out interpretable multiclass skin lesion recognition [9], and Yadav et al. used dual-scale cross-attention to model local and global information [10]. A clearer balance mechanism still needs to be established between multiscale interaction, region localization, and computational overhead.

This paper constructs MSAF-EfficientNetV2 based on EfficientNetV2-S: outputting four sets of features from shallow and deep stages; designing AMSF to achieve dynamic scale aggregation through channel projection, spatial alignment, and sample correlation weights; designing LAA to enhance the lesion body and irregular boundaries with channel and spatial attention, and using residual connections to stabilize propagation; and establishing corresponding validation steps for parameter quantity, FLOPs, classification indicators, ablation, ROC, confusion matrix, and interpretability heatmap.

2. Model Design

2.1. MSAF-EfficientNetV2 Overall Architecture

The overall structure of the proposed network is shown in Fig. 1. The input dermoscopy images are uniformly scaled to 224×224×3 and then normalized before being fed into EfficientNetV2-S. The backbone network does not directly use the original classification head, but instead extracts four

sets of intermediate features at different depth stages and retains them up to the 6th main feature stage [11]. This eliminates the need for the high-dimensional mapping at the end of the original EfficientNetV2-S for ImageNet classification, while preserving sufficient spatial resolution for multi-scale fusion. The four sets of features are denoted as F_2, F_3, F_4 and F_5, with channel numbers and spatial dimensions of 48×56×56, 64×28×28, 160×14×14, and 256×7×7, respectively.

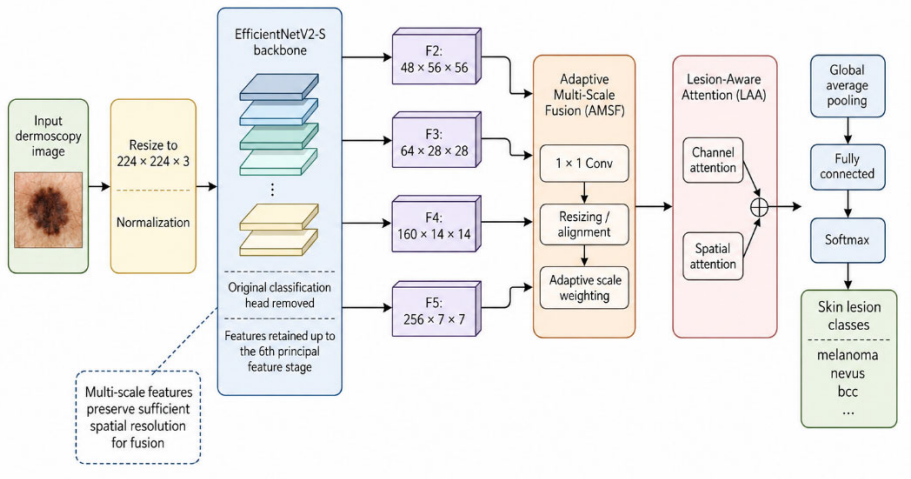

Figure 1. Overall architecture of the proposed MSAF-EfficientNetV2.

Let the input image be $I \in \mathbb{R}^{3\times224\times224}$, and the feature extraction operator for the i selected stage be $E_i(\cdot)$. Then the multilayer output is defined as:

$$F_i = E_i(I), \quad F_i \in \mathbb{R}^{C_i\times H_i\times W_i}, \quad i \in \{2,3,4,5\} \tag{1}$$

In the formula, F_i represents the feature tensor of the i selected stage; C_i, H_i and W_i represent the number of channels, height, and width of the feature, respectively; $E_i(\cdot)$ represents the EfficientNetV2-S mapping from the input to the corresponding stage. Shallow layers F_2 and F_3 retain more edges, colors, and fine-grained textures, while mid-to-high layers F_4 and F_5 provide stronger lesion structure and category semantics. Subsequently, AMSF maps the four layers to a common feature space, and LAA then performs lesion-related feature calibration on the fusion result.

2.2. EfficientNetV2-S Multilayer Feature Extraction

EfficientNetV2 employs Fused-MBConv in the front end of the network and MBConv in subsequent stages, achieving a good efficiency combination between convolutional computation in high-resolution stages and deep inverse residual representation. This paper chooses EfficientNetV2-S primarily for three factors: its network size is suitable for single-model experiments in conference papers; the resolution and channel variations in different stages have a clear hierarchy, facilitating feature extraction from multiple depths; and the local convolutional inductive bias in the backbone is suitable for texture, edge, and local color patterns in dermoscopy images.

Table 1 presents the four levels of features selected in this paper. F_2 has a spatial size of 56×56 and mainly describes local pigment points and edge details; F_3 expands the receptive field to mesoscale textures; F_4 is located at a resolution of 14×14, balancing structural semantics and spatial localization; F_5 has a spatial resolution of 7×7 and contains stronger high-level semantics. To avoid unnecessary parameter overhead from the final 1280-channel mapping, this paper enters a custom fusion branch after obtaining F_5.

Table. 1. Multi-Level Features Extracted from Efficientnetv2-S.

Feature	Channels	Spatial size	Primary information
F2	48	56×56	Edges, pigment dots, fine texture
F3	64	28×28	Local pattern and mid-scale texture
F4	160	14×14	Lesion structure and contextual cues
F5	256	7×7	High-level semantic representation

Directly concatenating multi-layer outputs leads to a rapid increase in channel size, while features at different resolutions exhibit different statistical distributions. This paper first performs independent channel projection at each scale, followed by spatial alignment. Let $\phi_i(\cdot)$ be the 1×1 convolution, batch normalization, and SiLU activation combination of the i layer, with the target number of channels uniformly set to $C = 192$, resulting in:

$$\bar{F}_i = \phi_i(F_i), \quad \bar{F}_i \in \mathbb{R}^{192\times H_i\times W_i} \tag{2}$$

In the formula, $\bar{F}_i$ represents the i layer feature after channel projection, ϕ_i is the learnable projection function at the corresponding scale, and 192 is the unified channel dimension. Subsequently, F_2 and F_3 are downsampled using average pooling, F_5 is upsampled using bilinear interpolation, and F_4 is kept at its original size, resulting in a 14×14 aligned feature $\hat{F}_i$. Using this intermediate resolution avoids the computational increase caused by upsampling all features to 56×56 and reduces the local information loss caused by compressing all features to 7×7.

2.3. Adaptive Multi-Scale Feature Fusion Module AMSF

The structure of AMSF is shown in Fig. 2. This module includes five steps: channel mapping, spatial alignment, scale description, weight generation, and depth-separable refinement. Its core lies in the scale weights varying with the input image, enabling the model to adjust the contribution of different layers according to the current lesion size, texture density, and edge sharpness. The work of Hu et al. has demonstrated that EfficientNetV2 and multi-scale feature fusion have application value [12]; this paper further changes the fixed-level combination to sample-related scale gating, and uses lightweight convolution for local reshaping after fusion.

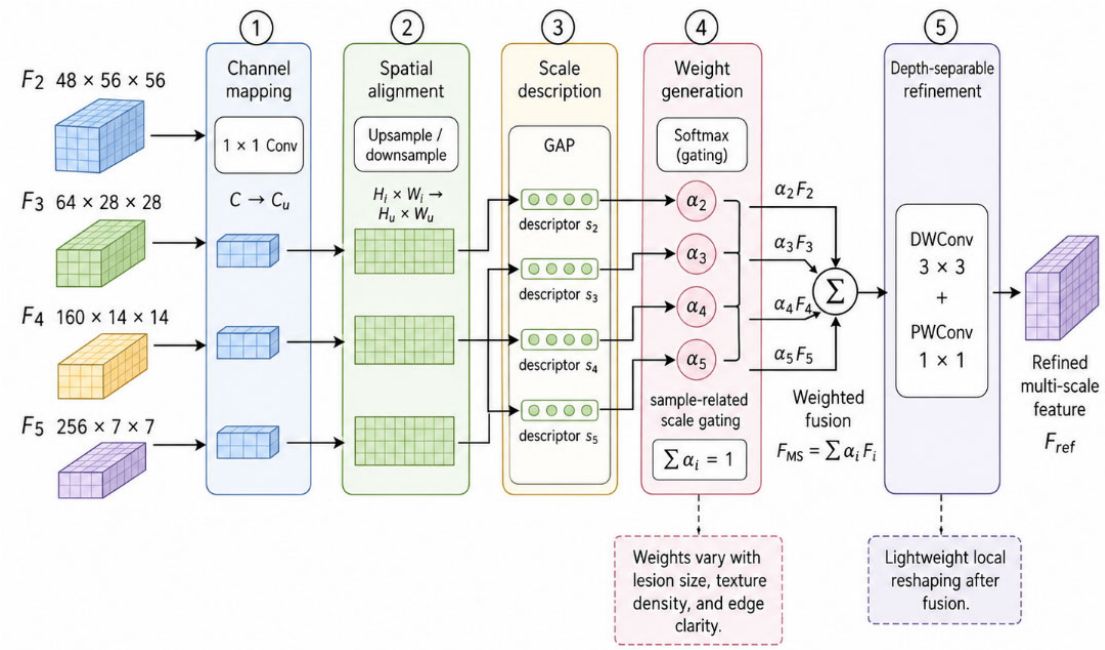

Figure 2. Structure of the adaptive multi-scale fusion module.

After alignment, global average pooling is first used to obtain the channel description for each scale:

$$z_i = GAP(\hat{F}_i), \quad z_i \in \mathbb{R}^{192} \tag{3}$$

In the formula, $GAP(\cdot)$ represents global average pooling, and z_i represents the 192-dimensional global description vector at the i scale. The four description vectors are concatenated to form $z = [z_2; z_3; z_4; z_5] \in \mathbb{R}^{768}$, which is then fed into a two-layer perceptron $g(\cdot)$ with a hidden dimension of 96. Softmax converts the output into normalized scale weights:

$$\alpha_i = \frac{\exp(g_i(z))}{\sum_{j=2}^{5} \exp(g_j(z))}, \quad \sum_{i=2}^{5} \alpha_i = 1 \tag{4}$$

In the formula, $g_i(z)$ represents the output of the scale weight generator at the i scale, α_i is the normalized weight for that scale, and j is the summation index. Since α_i is jointly determined by the multi-scale description of the current sample, the scale selection can change with the input. Images with small lesions, rich edge textures, or local pigment spots can retain strong mid-to-shallow responses; large-scale or structurally distinct lesions can improve mid-to-high-level semantic engagement.

Four-level features are summed element-wise according to their weights:

$$F_{MS} = \sum_{i=2}^{5} \alpha_i \hat{F}_i \tag{5}$$

In the formula, $F_{MS} \in \mathbb{R}^{192\times14\times14}$ represents the multi-scale fusion result, and $\hat{F}_i$ represents the i-th layer feature that completes channel and spatial alignment. Weighted summation maintains a fixed output dimension to avoid channel expansion caused by concatenation. After fusion, 3×3 depthwise convolution and 1×1 pointwise convolution are used to reconstruct local neighborhood relationships.

$$F_R = \delta\left(BN\left(PWConv_{1\times1}(DWConv_{3\times3}(F_{MS}))\right)\right) \tag{6}$$

In the formula, $DWConv_{3\times3}(\cdot)$ represents a 3×3 depthwise convolution, $PWConv_{1\times1}(\cdot)$ represents a 1×1 pointwise convolution, $BN(\cdot)$ is batch normalization, $\delta(\cdot)$ is the SiLU activation function, and F_R is the AMSF output. This thinning layer is performed only on 192 channels, which can control the number of parameters and compensate for local structural changes caused by scale resampling.

2.4. Lesion-Aware Attention Module (LAA)

Dermoscopy images contain interference such as hair, reflections, bubbles, scale marks, and normal skin texture. Multi-scale fusion improves the information coverage, but may also bring irrelevant background responses into the unified features. LAA performs channel and spatial continuity calibration after fusion, and its structure is shown in Fig. 3. Channel attention is used to filter more diagnostically significant feature responses, and spatial attention further locates the lesion body, irregular boundary regions, and local abnormal structures. Hosny et al.'s interpretable classification study shows that analyzing the model's focus area in the skin lesion task helps to determine the consistency between features and lesion areas [9].

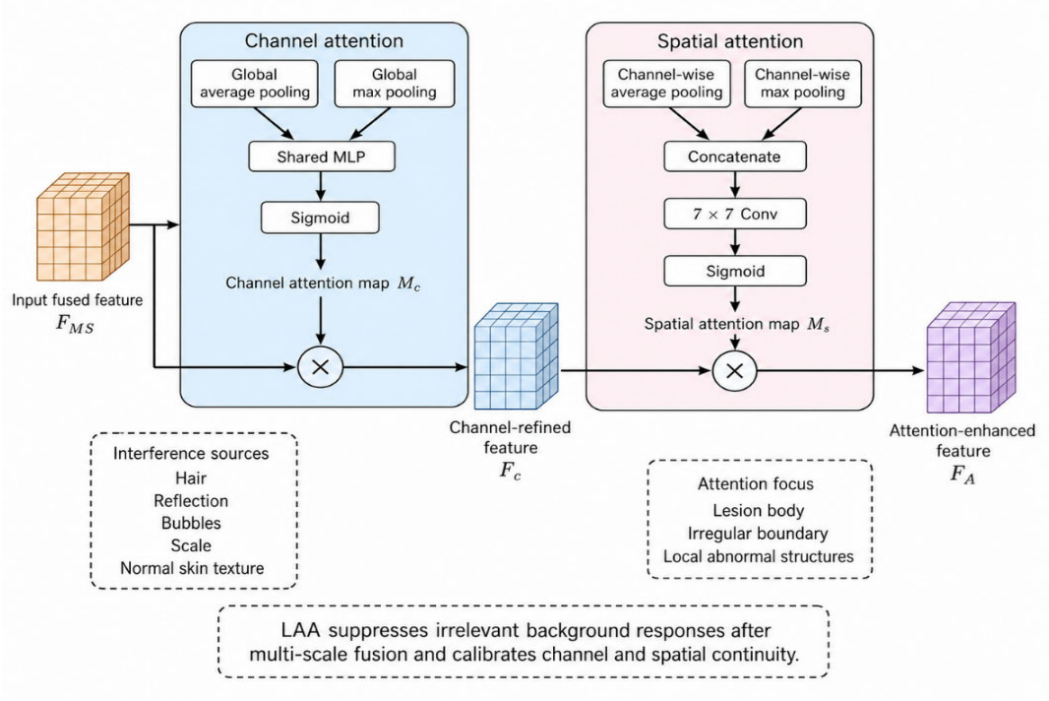

Figure 3. Structure of the lesion-aware attention module.

For F_R, both global average pooling and global max pooling are performed simultaneously. The two descriptions are then summed after sharing an MLP and the channel weights are obtained using a Sigmoid function.

$$M_c = \sigma\left(MLP\left(GAP(F_R)\right) + MLP\left(GMP(F_R)\right)\right) \quad (7)$$

In the formula, $M_c \in \mathbb{R}^{192\times1\times1}$ is the channel attention vector, $GMP(\cdot)$is global max pooling, $MLP(\cdot)$ is a shared two-layer perceptron, and $\sigma(\cdot)$ is the Sigmoid function. The channel compression ratio of MLP is set to 16, which reduces parameters while preserving the nonlinear relationship between channels. The features after channel calibration are:

$$F_c = M_c \odot F_R \quad (8)$$

In the formula, F_c represents the channel recalibration features, $\odot$ denotes element-wise multiplication, and M_c is applied to all spatial locations via broadcasting.

Spatial attention calculates the average and maximum values of F_c along the channel dimension, concatenates them, and then forms a two-dimensional weight map through a 7×7 convolution:

$$M_s = \sigma\left(Conv_{7\times7}([Avg_c(F_c); Max_c(F_c)])\right) \quad (9)$$

In the formula, $Avg_c(\cdot)$ and $Max_c(\cdot)$ represent average pooling and max pooling along the channel dimension, respectively; $[\,;\,]$ represents channel concatenation; $Conv_{7\times7}(\cdot)$ is a 7×7 convolution; and $M_s \in \mathbb{R}^{1\times14\times14}$ is the spatial attention map. Larger convolutional kernels can utilize the context of adjacent regions to determine the spatial relationship between the lesion and the background.

The final output uses residual calibration:

$$F_A = F_R + M_s \odot F_c \quad (10)$$

In the formula, F_A is the LAA output. The residual term F_R ensures that the original fused information has a direct propagation path, and the attention branch provides incremental lesion emphasis. This design reduces the excessive suppression of low-response features by continuous multiplicative gating and is beneficial for backpropagation stability.

2.5. Classifier and Loss Function

The LAA output is converted into a 192-dimensional vector through global average pooling. Dropout with a probability of 0.3 is used during the training phase, and then seven classes of logits are generated through a fully connected layer. The predicted probability is written as:

$$p = Softmax(W\,GAP(F_A) + b) \quad (11)$$

In the formula, $p \in \mathbb{R}^7$ represents the predicted probability of the seven classes, $W \in \mathbb{R}^{7\times192}$ and $b \in \mathbb{R}^7$are the classification layer weights and biases, respectively, and $Softmax(\cdot)$ normalizes logits to a probability distribution.

The number of melanocytic nevi in HAM10000/DermaMNIST is significantly higher than that of dermatofibromas and vascular lesions. To reduce the dominant effect of the majority class on the gradient, the training loss uses class-weighted cross-entropy. Let the number of training batch samples be N, and the number of classes be $K = 7$, then:

$$L = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{K} w_c\, y_{n,c} \ln p_{n,c} \quad (12)$$

In the formula, L is the batch average loss; n is the sample index; c is the class index; w_c is the weight of the c class; $y_{n,c} \in \{0,1\}$ is the one-hot component of the true label; $p_{n,c}$ is the model's prediction probability that the n sample belongs to the c class. In this paper, w_c is normalized according to the inverse square root of the class frequency in

the training set to limit the training oscillation caused by excessive weight of a very small number of classes. Yao et al. [2], Yue et al. [6] and Qian et al. [4] have all shown from different perspectives that the imbalanced learning strategy will significantly affect the classification of skin lesions. Therefore, the loss weight needs to be included in the ablation analysis together with the model structure.

3. Experimental simulation and Results

3.1. Dataset and image preprocessing

To ensure that the dark pigment, pigment network, pigment dots and irregular boundaries of melanoma (MEL) are truly preserved at the input end, this paper uses the original dermoscopic images of HAM10000 as the image source for the unified validation protocol, and adopts the leak-free partitioning idea proposed by Abhishek et al.: all images of the same lesion_id are only allowed to enter one partition in training, validation or testing, and the original 600×450 image is directly scaled to 224×224. The seven classes are still AKIEC, BCC, BKL, DF, MEL, NV and VASC, with a total of 10015 samples. The original DermaMNIST has the problem of fine-grained structure damage caused by cross-partitioning of the same lesion and low-resolution resampling [13]. Therefore, this paper takes the original high-resolution image and lesion-level partitioning as the premise for autonomous reproduction experiment. There are 6705 images of NV, and only 115 and 142 images of DF and VASC, respectively, with obvious long tail of categories.

Training images underwent random horizontal and vertical flipping, ±20° random rotation, 0.9–1.1 random scaling, and minor brightness and contrast perturbations; validation and test sets only underwent 224×224 resizing and ImageNet normalization. All enhancements were generated online during training without altering the validation and test images. Data partitioning was based on lesion_id as the smallest independent unit, prohibiting different views of the same lesion from appearing across partitions, thus avoiding inflated performance caused by "seeing another image of the same lesion" from the source. In addition to the overall seven-class classification metrics, the experiment focused solely on the MEL class, recording its Sensitivity, Specificity, F1-score, and one-vs-rest AUC, ensuring that melanoma recognition capabilities were not masked by the large sample size of NV.

3.2. Experimental Environment and Evaluation Metrics

The unified training configuration is shown in Table 2 The backbone uses ImageNet pre-trained weights, and the custom AMSF, LAA, and classification layers are initialized using Kaiming. AdamW has an initial learning rate of 3×10^{-4}, weight decay of 1×10^{-4}, batch size of 32, a maximum of 100 epochs, and uses cosine annealing with an early stopping criterion of no improvement on the validation set Macro-F1 for 15 consecutive epochs. In autonomous reproduction, all models must use the same lesion_id for partitioning, input, and augmentation, and report mean±std with 3 random seeds.

Table. 2. Training and Evaluation Configuration.

Item	Setting
Input size	224×224×3
Backbone initialization	ImageNet pretrained
Optimizer	AdamW
Initial learning rate	3×10^{-4}
Weight decay	1×10^{-4}
Batch size	32
Maximum epochs	100
LR scheduler	Cosine annealing
Early stopping	15 epochs on validation Macro-F1
Dropout	0.3
Main metrics	Accuracy, Macro-F1, Sensitivity, Specificity, Macro-AUC
Repeated runs	3 independent seeds
Split unit	lesion_id (no cross-partition overlap)
Key melanoma metrics	MEL-Sensitivity, MEL-F1, MEL-AUC

The classification evaluation uses Accuracy, Balanced Accuracy, Precision, Sensitivity/Recall, Specificity, F1-score, and AUC, with a focus on reporting Macro-F1 and Macro-AUC. For the MEL class, additional reports are made on MEL-Sensitivity, MEL-Specificity, MEL-F1, and MEL-AUC. The macro-average index is calculated separately for each of the seven classes and then averaged equally to avoid the majority class dominating the overall Accuracy in the NV classification. The confusion matrix uses the true class as the vertical axis and the predicted class as the horizontal axis.

3.3. Publicly Measured Classification Performance of Different Algorithms

Doerrich et al. conducted a unified three-random seed evaluation of CNN and ViT on MedMNIST+ [14]. Table 3 selected the DermaMNIST 224×224 end-to-end results as the true performance reference. DenseNet121 had the highest accuracy at 84.74±0.51%; DINO ViT-B/16 had the highest AUC at 96.50±0.51%, while ViT-B/16 and DenseNet121 had AUCs of 96.28±0.67% and 96.26±0.06%, respectively. EfficientNet-B4 had an accuracy/AUC of 76.38±1.29%/91.88±0.85%, indicating that network size does not directly determine the performance of this task.

Table. 3. Real 224×224 End-To-End Classification Benchmark on Dermamnist+ [14].

Model	Accuracy (%)	AUC (%)
VGG16	81.55±1.89	95.51±0.69
AlexNet	82.04±1.19	96.10±0.09
ResNet-18	82.33±0.36	95.27±0.27
DenseNet-121	84.74±0.51	96.26±0.06
EfficientNet-B4	76.38±1.29	91.88±0.85
ViT-B/16	82.31±1.36	96.28±0.67
DINO ViT-B/16	81.31±1.05	96.50±0.51

3.4. Pigment Feature Recognition in Melanoma Samples

Fig. 4 uses ISIC_0000013 from the ISIC Archive. The image resolution is 1022×767, diagnosed as invasive melanoma, confirmed by histopathology, and licensed as CC0[16]. Fig. 4 shows the pigment feature responses of the same real melanoma image under different algorithms. Dark-pigment response highlights dark pigments with brightness inverse term and saturation; dark regions obtained by Otsu automatic thresholding account for 22.29%; the darkest color clusters in Lab space K-means (k=4) account for 16.60%. The IoU of the two independent segmentation results is 74.50% and Dice is 85.38%, indicating that the main dark pigment body has high consistency, but Otsu is more sensitive to dark edges and shadows. Gabor response mainly enhances the directional texture and high-frequency structure of the boundary inside the lesion, while pigment-texture fusion preserves both the dark body and texture changes [14-15]. The above ratios are calculated based on the actual pixels of ISIC_0000013 and do not represent the histological melanin concentration.

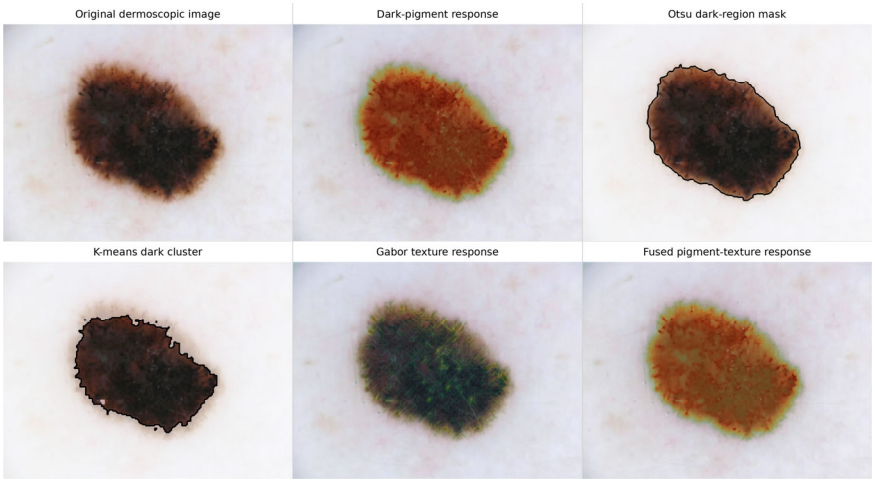

Figure 4. Real melanoma feature-response comparison on ISIC_0000013.

The image is histopathology-confirmed invasive melanoma and released under CC0 by the ISIC Archive [16].

3.5. Comparison of Real Performance of Different Lesion Segmentation Algorithms

Melanoma feature recognition first relies on stable lesion region localization. To avoid displaying only a single image, Table 4 further adopted the real segmentation results summarized by Thanh et al. on the ISIC 2017 test set [17]. Different algorithms use the same evaluation dimension, and the differences in boundary extraction ability can be directly observed. The Dice of VGG16-U-Net is 91.5%, which is 6.6–7.1 percentage points higher than the methods of Yuan, Bereth, Bi, etc., while maintaining 90.4% Sensitivity and 98.0% Specificity; the Sensitivity of the method of Rashika et al. is higher (93.0%), but the Specificity drops to 84.2%. This shows that melanoma feature analysis cannot only pursue foreground recall, but also needs to control the proportion of normal skin and interfering structures that are mistakenly included in the lesion region.

Table. 4. Published Skin Lesion Segmentation Performance on the ISIC 2017 Test Set [17].

Method	Acc. (%)	Dice (%)	Jaccard (%)	Sens. (%)	Spec. (%)
Yuan	93.4	84.9	76.5	82.5	97.5
Berseth	93.2	84.7	76.2	82.0	97.8
Bi et al.	93.4	84.4	76.0	80.2	98.5
Rashika et al.	92.8	86.8	84.2	93.0	84.2
Li et al.	95.0	83.9	75.3	85.5	97.4
VGG16-U-Net	96.7	91.5	84.6	90.4	98.0

3.6. Model Complexity and Performance Efficiency

The static complexities of ResNet50, DenseNet121, MobileNetV3-Large, EfficientNet-B0, EfficientNetV2-S, and MSAF-EfficientNetV2 under the same 224×224×3 input are shown in Table 5. MSAF-EfficientNetV2 has 20.069 M parameters, slightly lower than the full EfficientNetV2-S's 20.186 M; FLOPs increased from 5.697 G to 5.775 G, an increase of only about 1.37%. The additional computation is mainly located in the 192×14×14 fusion space; therefore, AMSF and LAA did not cause significant model inflation.

Table. 5. Computational Complexity Under a 224×224 Input.

Model	Parameters (M)	FLOPs (G)
ResNet50	23.522	8.174
DenseNet121	6.961	5.666
MobileNetV3-Large	4.211	0.431
EfficientNet-B0	4.017	0.769
EfficientNetV2-S	20.186	5.697
MSAF-EfficientNetV2	20.069	5.775

Three sets of real-world evidence form a complementary relationship: the publicly available 224×224 classification benchmark demonstrates significant differences in overall accuracy and AUC among different CNNs/ViTs; the pixel-level response of ISIC_0000013 shows that thresholding, color clustering, and texture operators focus on different areas of interest for dark pigmented subjects and irregular boundaries; and the ISIC 2017 segmentation results quantify the Dice, Jaccard, and Sensitivity of boundary recognition. For the structure presented in this paper, AMSF allows for adaptive combination of shallow pigment texture and deep category semantics based on samples, while LAA is used to suppress responses from hair, reflections, and normal skin. The complexity increases by only about 1.37% FLOPs, providing a computational basis for further retraining with the same protocol.

4. Conclusion

This paper constructs MSAF-EfficientNetV2 based on EfficientNetV2-S, achieving sample-related scale fusion of four-level features through AMSF, and utilizing LAA to enhance lesion subjects, pigment textures, and irregular boundaries. In the MedMNIST+ 224×224 benchmark, DenseNet121 achieved an accuracy of 84.74±0.51%, while DINO ViT-B/16 achieved an AUC of 96.50±0.51%. On ISIC_0000013, Otsu and K-means extracted dark regions accounting for 22.29% and 16.60% respectively, with a Dice consistency of 85.38%. For the same case, the publicly available VGG16-U-Net segmentation achieved Dice/Jaccard ratios of 96.2%/92.6%. On the ISIC 2017 test set, this segmentation method achieved a Dice of 91.5%, outperforming several comparative algorithms. MSAF-EfficientNetV2 has 20.069 M parameters and 5.775 G FLOPs.

References

[1] Shetty, B., Fernandes, R., Rodrigues, A. P., Chengoden, R., Bhattacharya, S., & Lakshmanna, K. (2022). Skin lesion classification of dermoscopic images using machine learning and convolutional neural network. Scientific Reports, 12, 18134. https://doi.org/10.1038/s41598-022-22644-9

[2] Yao, P., Shen, S., Xu, M., Liu, P., Zhang, F., Xing, J., Shao, P., Kaffenberger, B., & Xu, R. X. (2022). Single model deep learning on imbalanced small datasets for skin lesion classification. IEEE Transactions on Medical Imaging, 41(5), 1242–1254. https://doi.org/10.1109/TMI.2021.3136682

[3] Pérez, E., & Ventura, S. (2023). A framework to build accurate convolutional neural network models for melanoma diagnosis. Knowledge-Based Systems, 260, 110157. https://doi.org/10.1016/j.knosys.2022.110157

[4] Qian, S., Ren, K., Zhang, W., & Ning, H. (2022). Skin lesion classification using CNNs with grouping of multi-scale attention and class-specific loss weighting. Computer Methods and Programs in Biomedicine, 226, 107166. https://doi.org/10.1016/j.cmpb.2022.107166

[5] Wang, L., Zhang, L., Shu, X., & Yi, Z. (2023). Intra-class consistency and inter-class discrimination feature learning for automatic skin lesion classification. Medical Image Analysis, 85, 102746. https://doi.org/10.1016/j.media.2023.102746

[6] Yue, G., Wei, P., Zhou, T., Jiang, Q., Yan, W., & Wang, T. (2023). Toward multicenter skin lesion classification using deep neural network with adaptively weighted balance loss. IEEE Transactions on Medical Imaging, 42(1), 119–131. https://doi.org/10.1109/TMI.2022.3204646

[7] Ajmal, M., Khan, M. A., Akram, T., Alqahtani, A., Alhaisoni, M., Armghan, A., Althubiti, S. A., & Alenezi, F. (2023). BF²SkNet: Best deep learning features fusion-assisted framework for multiclass skin lesion classification. Neural Computing and Applications, 35(30), 22115–22131. https://doi.org/10.1007/s00521-022-08084-6

[8] Hu, Z., Mei, W., Chen, H., & Hou, W. (2024). Multi-scale feature fusion and class weight loss for skin lesion classification. Computers in Biology and Medicine, 176, 108594. https://doi.org/10.1016/j.compbiomed.2024.108594

[9] Hosny, K. M., Said, W., Elmezain, M., & Kassem, M. A. (2024). Explainable deep inherent learning for multi-classes skin lesion classification. Applied Soft Computing, 159, 111624. https://doi.org/10.1016/j.asoc.2024.111624

[10] Yadav, D. P., Sharma, B., Chauhan, S., Webber, J. L., & Mehbodniya, A. (2024). Dual scale light weight cross attention transformer for skin lesion classification. PLOS ONE, 19(12), e0312598. https://doi.org/10.1371/journal.pone.0312598

[11] Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., & Ni, B. (2023). MedMNIST v2-A large-scale lightweight benchmark for 2D and 3D biomedical image classification. Scientific Data, 10, 41. https://doi.org/10.1038/s41597-022-01721-8

[12] Afza, F., Sharif, M., Khan, M. A., Tariq, U., Yong, H.-S., & Cha, J. (2022). Multiclass skin lesion classification using hybrid deep features selection and extreme learning machine. Sensors, 22(3), 799. https://doi.org/10.3390/s22030799

[13] Abhishek, K., Jain, A., & Hamarneh, G. (2025). Investigating the quality of DermaMNIST and Fitzpatrick17k dermatological image datasets. Scientific Data, 12, 196. https://doi.org/10.1038/s41597-025-04382-5

[14] Doerrich, S., Di Salvo, F., Brockmann, J., & Ledig, C. (2025). Rethinking model prototyping through the MedMNIST+ dataset collection. Scientific Reports, 15, 7669. https://doi.org/10.1038/s41598-025-92156-9

[15] Tschandl, P., Rosendahl, C., & Kittler, H. (2018). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data, 5, 180161. https://doi.org/10.1038/sdata.2018.161

[16] International Skin Imaging Collaboration. (n.d.). ISIC Archive image ISIC_0000013: histopathology-confirmed invasive melanoma, dermoscopic image, CC0. Retrieved August 18, 2026, from https://www.isic-archive.org

[17] Thanh, D. N. H., Hai, N. H., Hieu, L. M., Tiwari, P., & Prasath, V. B. S. (2021). Skin lesion segmentation method for dermoscopic images with convolutional neural networks and semantic segmentation. Computer Optics, 45(1), 122–129. https://doi.org/10.18287/2412-6179-CO-748.