Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An Attention-Enhanced MobileViT Framework for Multi-Task Skin Lesion Classification and Segmentation

Yanqiang Ge

gyq@aynu.edu.cn

Anyang Normal University

Bingqian Lu

Anyang Normal University

Haoyu Wang

Anyang Normal University

Zixu Yang

Anyang Normal University

Zhan Zhang

Anyang Normal University

Research Article

Keywords: MobileViT, skin lesion recognition, multi-task learning, CBAM, lightweight network

Posted Date: August 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10583629/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Deep learning-based automated skin lesion diagnosis has advanced substantially, but balancing diagnostic accuracy and computational efficiency remains challenging for resource-constrained primary-care settings. Heavyweight segmentation and multi-task models deliver strong performance but carry prohibitive parameter overhead for practical deployment, while existing lightweight networks lack robust spatial feature enhancement and uncertainty constraints, limiting clinical diagnostic reliability. This work proposes an attention-augmented lightweight multi-task framework based on MobileViT-XS for joint skin lesion classification and semantic segmentation. The framework adopts MobileViT-XS as the shared backbone to extract multi-scale lesion features, integrates a convolutional block attention module (CBAM) to enhance lesion-related spatial representations under dense pixel-level supervision, and constructs a dual-branch multi-task architecture forming an asymmetric task promotion paradigm favoring classification optimization. A conservative prediction strategy coupled with risk stratification is further designed to mitigate overconfident outputs for ambiguous samples and improve diagnostic safety. Experiments on the ISIC 2019 dataset demonstrate that the proposed model achieves 84.37% accuracy, 79.46% Macro-F1, 0.935 Macro-AUC, 86.37% Dice coefficient, and 78.09% mIoU with only 2.62 M parameters. The multi-task mechanism significantly boosts classification performance compared with the single-task baseline, and yields competitive results against heavyweight architectures with far fewer parameters. Visualization validates precise lesion localization and favorable model interpretability. Overall, the proposed framework strikes a well-balanced trade-off among accuracy, computational efficiency, visual interpretability and clinical prediction safety, providing a deployable lightweight solution for primary-care skin lesion screening. Prospective multi-center clinical validation is required before real-world clinical application.

1. Introduction

1.1. Background

Skin lesions are common clinical conditions that include benign nevi and malignant tumors such as melanoma and basal cell carcinoma. More than three million new cases of skin cancer are reported worldwide each year [1]. The 5-year survival rate of early melanoma exceeds 99%, whereas it decreases to approximately 30% after distant metastasis [2], emphasizing the importance of early screening. Conventional diagnosis depends heavily on dermatologist experience and is affected by inter-observer variability, with reported agreement ranging from 0.53 to 0.81 [3]. In primary-care settings, the shortage of specialized dermatological resources further limits timely and accurate screening.

Deep learning has substantially advanced automated medical image analysis [4,5]. Convolutional neural networks (CNNs) and vision Transformers (ViTs) have improved skin lesion classification and segmentation [6–10]; however, high-performance models are often computationally expensive and difficult to deploy on mobile or resource-limited devices. Lightweight CNNs are more deployment-friendly but often lack sufficient global feature modeling. Therefore, a model that integrates lightweight

deployment, global context modeling, lesion localization, and risk-aware output is needed for practical skin lesion screening.

1.2. Research gaps and motivation

This study is motivated by three key research gaps existing in current skin lesion analysis research. First, the accuracy-efficiency trade-off of lightweight models for dermoscopic lesion diagnosis remains insufficiently investigated. Although MobileViT architectures have demonstrated compelling performance and computational balance in general visual recognition tasks, their applicability, effectiveness, and task adaptability for fine-grained skin lesion classification and segmentation have not been comprehensively validated.

Second, the collaborative mechanism between classification and segmentation within lightweight multi-task frameworks remains unclear. Existing multi-task models such as MT-TransUNet have verified the feasibility of joint skin lesion classification and segmentation learning. However, their excessively large parameter overhead severely limits practical deployment on portable and low-resource clinical devices, making them unsuitable for primary-care screening scenarios. Whether lightweight backbones can achieve effective cross-task feature collaboration while maintaining low computational costs still lacks in-depth exploration.

Third, most current skin lesion diagnosis models only provide deterministic prediction results without explicitly quantifying or constraining uncertain cases. Such overconfident output mechanisms may yield unreliable predictions for ambiguous, boundary-blurred, or atypical lesion samples, potentially increasing clinical misdiagnosis risks and hindering safe artificial intelligence-assisted diagnosis [11,12].

1.3. Contributions

Targeting the above research gaps, this study constructs a lightweight, high-performance, safe, and interpretable multi-task framework for automated skin lesion diagnosis. The core contributions of this work are summarized as follows.

First, this study systematically validates the applicability and superiority of MobileViT-XS as a lightweight backbone for skin lesion analysis. Different from conventional lightweight CNNs lacking long-range spatial modeling and heavy Transformer models with high computational costs, MobileViT-XS integrates local texture extraction and global dependency modeling, which well matches the multi-scale and boundary-ambiguous characteristics of dermoscopic lesions. Extensive comparisons with EfficientNet-B0, ResNet-50, U-Net, DeepLabV3+, and GS-TransUNet demonstrate that the proposed model achieves an optimal accuracy-efficiency trade-off. With only 2.62 M parameters, our method yields superior comprehensive classification performance, effectively alleviating the unbalanced accuracy-cost problem of existing lightweight skin lesion models.

Second, this work establishes an attention-enhanced lightweight multi-task collaborative learning paradigm for joint classification and segmentation. By incorporating pixel-level segmentation

supervision to assist image-level classification optimization, the framework reveals a typical asymmetric multi-task promotion effect. Segmentation supervision significantly refines lesion spatial features and substantially improves classification performance. Benefiting from dense pixel-level supervision, the embedded CBAM module further enhances lesion-aware feature representation, overcoming the limited attention gain observed in single-task frameworks. Compared with heavy multi-task models, the proposed method achieves stable cross-task cooperation at extremely low parameter cost, greatly improving the practicability of multi-task skin lesion diagnosis for primary-care and edge-device deployment.

Third, this study introduces a conservative prediction constraint strategy and risk stratification mechanism to alleviate model overconfidence and improve diagnostic safety. By setting reasonable thresholds for confidence deviation and malignant probability screening, the model effectively identifies uncertain and hard-to-distinguish samples, thereby reducing potential clinical diagnosis risks. Combined with visualization for intuitive interpretability analysis, the proposed framework realizes accurate, stable, and explainable skin lesion screening, which further enhances the clinical credibility and practical application value of lightweight intelligent diagnosis systems.

2. Related Work

2.1. Deep learning for skin lesion classification

CNNs are fundamental tools for skin lesion image recognition. ResNet alleviated gradient degradation in deep networks through residual connections [13], while ViT introduced self-attention into vision by modeling images as patch sequences [14]. Swin Transformer reduced the computational burden of global attention through shifted windows [15]. These models have improved recognition accuracy, but their parameter sizes and computation costs remain obstacles for lightweight deployment.

2.2. Lightweight network architectures

Lightweight networks aim to achieve competitive performance with fewer parameters and lower computational cost [16–18]. MobileNetV3 uses neural architecture search and hardware-aware optimization [19], ShuffleNetV2 improves efficiency through channel split and shuffle operations [20], and EfficientNet-Lite provides compact variants of compound-scaled CNNs [21]. MobileViT combines CNN-based local representation with Transformer-based global modeling [22], making it attractive for skin lesion analysis where local texture and global morphology are both important. The representative lightweight backbones and their ImageNet-1k performance are summarized in Table 1, which motivates the selection of MobileViT-XS as the backbone.

Table 1
Performance comparison of representative lightweight networks on ImageNet-1k.

Network	Parameters (M)	Top-1 Accuracy (%)	Characteristics
MobileNetV3-Small [12]	2.5	67.4	Pure CNN; NAS-optimized
ShuffleNetV2 0.5x [24]	1.4	60.3	Pure CNN; channel shuffle
EfficientNet-Lite0 [11]	4.7	75.1	Pure CNN; compound scaling
MobileViT-XXS [8]	1.3	69.0	CNN-ViT hybrid
MobileViT-XS [8]	2.3	74.8	CNN-ViT hybrid
MobileViT-S [8]	5.6	78.4	CNN-ViT hybrid

2.3. Attention mechanisms in medical image analysis

Attention mechanisms allow models to focus on informative regions and suppress irrelevant background. CBAM sequentially applies channel and spatial attention [23]. For skin lesions, which often have ambiguous borders and complex backgrounds, CBAM can enhance lesion-related color, texture, and boundary features. Attention U-Net further demonstrated the value of attention gates in medical image segmentation [24].

2.4. Multi-task learning in medical imaging

Multi-task learning shares feature representations across related tasks [25]. In skin lesion analysis, classification and segmentation are complementary: segmentation provides spatial supervision for lesion localization, while classification encourages discriminative semantic representations. MT-TransUNet introduced multi-task tokens for joint skin lesion classification and segmentation [26,27], but its parameter scale is unsuitable for resource-constrained deployment. Lightweight multi-task architectures therefore remain an important research direction.

Existing methods usually emphasize either accuracy or efficiency. Heavyweight models achieve high performance but require substantial computation; lightweight CNNs are efficient but provide limited global modeling; and multi-task Transformer models are accurate but large. This highlights the need for a lightweight, attention-enhanced, multi-task model with risk-aware output.

3. Methodology

3.1. Overall architecture

The proposed MobileViT-XS + MTL+CBAM framework uses MobileViT-XS as the shared feature extraction backbone. A preprocessed *256 x 256* skin lesion image is fed into the backbone to learn deep representations. Two key modifications are introduced: CBAM is inserted at the end of the backbone for feature enhancement, and the original single classification head is extended into parallel classification

and segmentation branches. The overall data flow and the relationship among the shared backbone, CBAM module, and task-specific heads are illustrated in Fig. 1.

The CBAM module follows a global enhancement strategy before the task-specific heads. The classification branch outputs probabilities for eight lesion categories, whereas the segmentation branch outputs a lesion mask. The shared backbone enables implicit knowledge transfer between classification and segmentation.

3.2. MobileViT-XS backbone

MobileViT-XS combines CNN-based local feature extraction with Transformer-based global modeling [22]. MobileViT blocks unfold local feature maps into sequences, apply Transformer encoders for global interaction, and fold the features back into two-dimensional maps. The output feature map has a size of *16 x 16 x* 320and is used by CBAM and the task-specific heads. The detailed stage-wise configuration of the MobileViT-XS backbone is listed in Table 2.

Table 2
MobileViT-XS backbone structure.

Stage	Operation	Output size	Channels	Description
Input	-	256 x 256	3	RGB skin lesion image
Stage 1	Conv 3 x 3, stride 2	128 x 128	16	Initial feature extraction
Stage 2	Inverted residual block x1	64 x 64	24	Intermediate feature extraction
Stage 3	Inverted residual block x1	32 x 32	48	Intermediate feature extraction
Stage 4	MobileViT block x1	16 x 16	96	Local-global feature fusion
Stage 5	MobileViT block x1	16 x 16	320	High-dimensional representation
Output	-	16 x 16 x 320	-	Shared feature map

3.3. CBAM attention module

Given an input feature map *F*, CBAM sequentially applies channel attention and spatial attention [23]. Channel attention identifies informative feature channels through global average pooling(GAP) and global max pooling(GMP) followed by a shared multi-layer perceptron(MLP):

$$F_{avg} = \mathrm{GAP}\left(F\right) = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W}F(:,i,j)$$

1

$$F_{max} = \mathrm{GMP}\left(F\right) = max_{i,j}F(:,i,j)$$

2

$$M_c\left(F\right) = \mathrm{sigmoid}\left(\mathrm{MLP}(\ F_{avg}) + \mathrm{MLP}(F_{max}\ )\right)$$

3

The channel-enhanced feature is $F' = M_c\left(F\right) \odot F$. The spatial attention module first aggregates *F'* across the channel dimension via pooling operations, followed by a convolutional layer to generate the spatial attention mask.

$$M_s\left(F'\right) = \mathrm{sigmoid}\left(f^{7\times7}\left(\left[\mathrm{GAP}_c(\ F'); \mathrm{GMP}_c(F'\ )\right]\right)\right)$$

4

$$F'' = M_s\left(F'\right) \odot \left(M_c(\ F) \odot F\right)$$

5

where *F ∈ R^(C×H×W)* denotes the input feature map; *C*, *H*, and *W* denote the channel number, feature-map height, and feature-map width, respectively; $F(:,i,j)$is the *C*-dimensional feature vector at spatial position (*i*, *j*); F_{avg} and F_{max} are the global average-pooled and max-pooled channel descriptors; *GAP* and *GMP* denote global average pooling and global max pooling; *MLP* is the shared multi-layer perceptron; $M_c\left(F\right)$ is the channel attention map; GAP_c and GMP_c denote average and max pooling along the channel dimension; *[;]* denotes channel-wise concatenation; $f^{7\times7}$ denotes a convolution with a 7×7 kernel; $M_s\left(F'\right)$ is the spatial attention map; $\odot$ denotes element-wise multiplication with broadcasting; and F'' is the final attention-refined feature map.

This sequence of operations first learns channel-wise importance and then learns spatial importance. Therefore, CBAM refines the shared feature representation before it is passed to the downstream classification and segmentation branches. The position of the CBAM block and its channel-spatial refinement process are shown in Fig. 2.

3.4. Multi-task learning design

The CBAM-enhanced backbone is connected to a classification head and a segmentation head. The classification head applies global average pooling, dropout with *p = 0.3*, a fully connected layer, and softmax normalization:

$$p_i = \frac{\exp\left(z_i\right)}{\sum_{j=1}^{8}\exp\left(z_j\right)}, i = 1,2,\ldots,8$$

6

The segmentation head uses a lightweight decoder. The *16 x 16 x 320* feature map is upsampled to *256 x 256*, compressed with two *3 x 3* convolutional layers, and mapped to foreground and background channels by a *1 x 1* convolution:

$$s_k(x,y) = \frac{1}{1+\exp(-f_k(x,y))}, k \in \{0,1\}$$

7

where $\mathrm{z_i}$ is the logit of the *i*-th lesion class; $\mathrm{p_i}$ is the predicted probability of the *i*-th class after softmax normalization; *i = 1, 2, ..., 8* corresponds to the eight lesion categories; $f_k(x,y)$is the segmentation logit at pixel position *(x, y)* for channel *k*; *k ∈{0, 1}* denotes background and lesion foreground channels; and $s_k(x,y)$ is the corresponding pixel-wise probability after sigmoid activation.

Equations (6) and (7) define the outputs of the two task-specific heads. The classification head converts global lesion features into class probabilities, whereas the segmentation head converts dense features into pixel-level foreground/background probabilities.

3.5. Joint loss function

The total loss is defined as:

$$\mathscr{L}_{\text{total}} = \lambda_{cls}\mathscr{L}_{cls} + \lambda_{seg}\mathscr{L}_{seg}$$

8

where λ_{cls} and λ_{seg} are both set to 1.5. The classification loss is a weighted cross-entropy loss to alleviate class imbalance: The resulting class distribution and inverse-frequency weights are reported in Table 3.

$$\mathscr{L}_{cls} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} w_c y_{i,c} \log\left(p_{i,c}\right)$$

9

$$\mathrm{w}_c = \frac{N}{C \cdot n_c}$$

10

Table 3
ISIC 2019 class distribution and loss weights.

Class	Lesion type	Samples	Proportion (%)	Weight $\varvec{w}_{\varvec{c}}$
AK	Actinic keratosis	867	3.42	3.65
BCC	Basal cell carcinoma	3323	13.12	0.95
BKL	Benign keratosis-like lesion	2624	10.36	1.21
DF	Dermatofibroma	239	0.94	13.25
MEL	Melanoma	4522	17.85	0.70
NV	Melanocytic nevus	12875	50.83	0.25
SCC	Squamous cell carcinoma	628	2.48	5.04
VASC	Vascular lesion	253	1.00	12.52

The segmentation loss combines Dice loss and binary cross-entropy loss:

$$\mathscr{L}_{seg} = \mathscr{L}_{Dice} + \mathscr{L}_{BCE}$$

11

$$\mathscr{L}_{Dice} = 1 - \frac{2\sum_i s_{1,i} g_i + \epsilon}{\sum_i s_{1,i} + \sum_i g_i + \epsilon}$$

12

$$\mathscr{L}_{BCE} = -\frac{1}{M}\sum_{i=1}^{M} [g_i \log(s_{1,i}) + (1 - g_i)\log(1 - s_{1,i})]$$

13

where $\mathscr{L}_{\text{total}}$ is the total multi-task loss; $\mathscr{L}_{cls}$ and $\mathscr{L}_{seg}$ denote the classification and segmentation losses; λ_{cls} and λ_{seg} are their loss weights; *N* is the batch size; *C = 8* is the number of lesion classes; $y_{i,c}$ is the one-hot ground-truth label for sample *i* and class *c*; $p_{i,c}$ is the predicted class probability; w_c is the class-specific loss weight; n_c is the number of training samples in class *c*; $s_{1,i}$ is the predicted foreground probability of pixel *i* ; g_i is the ground-truth binary mask label; *ε* is a smoothing constant; and *M* is the number of pixels in an image.

The weighted classification loss encourages the model to pay more attention to minority lesion classes, which is consistent with imbalance-aware learning strategies such as focal loss, class-balanced reweighting, and logit adjustment [28–30], whereas the segmentation loss jointly optimizes region overlap and pixel-wise classification accuracy. The total loss therefore couples semantic discrimination with lesion localization.

3.6. Inference workflow and risk assessment

The inference workflow includes image preprocessing, model inference, post-processing, and risk output. The input image is resized to *256 x 256*, normalized, and standardized using ImageNet statistics. The model outputs the classification probability vector and segmentation mask in a single forward pass. The mask is binarized using a threshold of 0.5, geometric lesion indicators are calculated, and classification row-normalized confusion matrix are generated from the classification branch.

A conservative output strategy is used because model confidence is not always reliable. Two thresholds are defined before formalizing the decision rule: a minimum confidence threshold θ_{conf}= 0.5 and a minimum margin threshold θ_{diff}= 0.15. After inference, the largest class probability p_{max} and the second-largest probability p_{second} are extracted, and their margin is computed as $\delta = p_{max} - p_{second}$. The sample is assigned to the review pathway when either the confidence or the class margin is insufficient:

$$\text{output} = \begin{cases} \text{'Needs review'}, & \text{if } p_{max} < \theta_{conf} \text{ or } \delta < \theta_{diff} \\ \text{class}_{pred}, & \text{otherwise} \end{cases}$$

14

If the sample does not trigger the review condition, the malignant-class probability is calculated as:

$$p_{malignant} = p_{MEL} + p_{BCC} + p_{SCC} + p_{AK}$$

15

where class_{pred} is the predicted class label when the review condition is not triggered; and $p_{malignant}$ denotes the summed probability of malignant or potentially malignant categories.

Eq. (14) implements the safety-first review rule, whereas Eq. (15) computes the cumulative malignant probability used for subsequent risk stratification. Low confidence or a small margin between the top two classes indicates that the model lacks sufficient evidence for a deterministic output, so the sample is flagged for expert review.

Risk is assigned as high if class_{pred} belongs to the malignant set, medium if $p_{malignant}$ >= 0.25, and classified as low risk otherwise.This rule follows a safety-oriented principle: malignant lesions should not be under-reported. Table 4 provides the complete mapping from model uncertainty and malignant-class probability to the final risk category.

Table 4
Conservative output and risk stratification logic.

Condition	Output	Interpretation
p_{max} < 0.50	Needs review	Overall model confidence is insufficient
δ < 0.15	Needs review	The model is uncertain between classes
class_{pred} is malignant	High risk	Predicted malignant lesion
class_{pred} is benign and $p_{malignant}$ >= 0.25	Medium risk	Potential malignancy remains
class_{pred} is benign and $p_{malignant}$ < 0.25	Low risk	Low malignant probability

4. Experiments and Results

4.1. Dataset and preprocessing

The ISIC 2019 public dataset [31–33] was used for classification training and validation, and segmentation masks from ISIC 2018 [34] were matched for joint supervision. ISIC 2019 contains 25,331 images from eight lesion categories: AK, BCC, BKL, DF, MEL, NV, SCC, and VASC. A total of 772 samples could be matched with segmentation masks. Related patient-centric dermoscopic datasets with metadata further support external validation under more clinically realistic settings [35]. Data cleaning verified image paths and label validity. The dataset was split into training and validation sets at an 8:2 ratio using stratified random sampling.

All images were resized to 256 x 256 pixels, normalized by 255, and standardized using ImageNet means [0.485, 0.456, 0.406] and standard deviations [0.229, 0.224, 0.225]. Data augmentation was applied during training to improve generalization, consistent with mixing- and policy-based augmentation strategies used to improve robustness [36–38]. The specific augmentation operations and probabilities are summarized in Table 5.

Table 5
Data augmentation settings.

Augmentation	Parameters	Probability	Purpose
Random horizontal flip	-	0.5	Left-right mirroring
Random vertical flip	-	0.5	Top-bottom mirroring
Random rotation	[-15, 15] degrees	0.5	Different shooting angles
Random resized crop	Scale [0.8, 1.2], crop to 256 x 256	1.0	Different shooting distances
Color jitter	Brightness 0.2, contrast 0.2, saturation 0.2	0.5	Different illumination conditions

4.2. Experimental environment and settings

The hardware and software environment used for training and evaluation is summarized in Table 6.

Table 6
Experimental environment.

Component	Version/Model	Description
Python	3.9+	Core development language
PyTorch	>= 2.0.0	Deep learning framework
torchvision	>= 0.15.0	Vision processing library
timm	0.9.12	Pretrained model library
OpenCV	4.8.1.78	Image preprocessing
CUDA	11.8	GPU acceleration
GPU	NVIDIA RTX 4090	Training hardware with 24 GB memory

Experiments were designed for both eight-class classification and lesion segmentation. Ablation studies evaluated the effects of CBAM and multi-task learning, and comparative experiments included five baselines.

4.3. Evaluation metrics

Classification performance was evaluated using accuracy, Macro-F1, sensitivity and specificity. Segmentation performance was evaluated using Dice coefficient and mean intersection over union (*mIoU*). These metrics provide complementary perspectives on overall accuracy, class imbalance, malignant-lesion detection, boundary quality, and region overlap.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

16

$$\text{Macro-F1} = \frac{1}{C}\sum_{i=1}^{C} F1_i, \text{where } F1_i = \frac{2P_i R_i}{P_i + R_i}$$

17

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

18

$$\text{Specificity} = \frac{TN}{TN + FP}$$

19

$$\text{Dice} = \frac{2|A \cap B|}{|A| + |B|}$$

20

$$\text{mIoU} = \frac{1}{N}\sum_{i=1}^{N} \frac{|A_i \cap B_i|}{|A_i \cup B_i|}$$

21

where *TP*, *TN*, *FP*, and *FN* denote true positives, true negatives, false positives, and false negatives; P_i and R_i are the precision and recall of class *i*; $F1_i$ is the class-specific F1 score; *C* is the number of classes; *A* and *B* denote the predicted and ground-truth foreground masks; |·| denotes the number of pixels in a set; and *N* is the number of evaluated samples for *mIoU* calculation.

These metrics were selected because the dataset is imbalanced and the task involves both image-level diagnosis and pixel-level lesion localization. Accuracy reflects overall correctness, *Macro-F1* reduces dominance by majority classes, sensitivity emphasizes malignant-lesion detection, specificity measures correct rejection of negative cases, and Dice/*mIoU* quantify segmentation overlap.

4.4. Ablation study

The ablation experimental results for the classification task are summarized in Table 7. Compared with the MobileViT-XS single-task baseline model, simply embedding the CBAM attention module yielded marginal performance gains: the overall accuracy only rose slightly from 72.19% to 75.09%, with limited promotion on Macro-F1, Sensitivity and Specificity metrics. By contrast, introducing multi-task learning

(MTL) that couples classification and lesion segmentation brought a substantial accuracy improvement up to 83.66%, which verifies that pixel-wise segmentation supervision can effectively excavate lesion-associated feature information and optimize the representation ability of the backbone for skin lesion classification. On the basis of the multi-task framework, further embedding the CBAM module to strengthen channel-wise lesion feature aggregation achieved the optimal comprehensive performance: the accuracy reached 84.37%, Macro-F1 was elevated to 79.46%, and Specificity also attained the highest value of 97.22%. This phenomenon demonstrates that the feature enhancement effect brought by the CBAM attention mechanism can be fully stimulated under the constraint of joint segmentation-classification multi-task supervision.

Table 7
Classification ablation results.

Architecture	Accuracy (%)	Macro-F1 (%)	Sensitivity (%)	Specificity (%)	Parameters (M)
MobileViT-XS baseline	72.19	73.64	79.84	95.70	1.94
MobileViT-XS+CBAM	75.09	74.88	78.39	96.12	1.95
MobileViT-XS + MTL	83.66	76.3	79.96	97.14	2.60
ours	84.37	79.46	79.68	97.22	2.62

The ablation results for lesion segmentation are presented in Table 8. Under the single-task segmentation setting, integrating the CBAM attention module achieved marginal performance improvements, with the Dice coefficient rising slightly from 90.45% to 90.50% and mIoU increasing from 84.00% to 84.05%. When introducing multi-task learning (MTL) for joint classification and segmentation, the segmentation metrics declined noticeably: the Dice coefficient dropped to 86.23% and mIoU decreased to 77.77%. This phenomenon can be attributed to task competition, where the classification task dominates gradient optimization and impairs segmentation feature learning. On this multi-task basis, further embedding CBAM yielded moderate gains, lifting Dice to 86.37% and mIoU to 78.09%. This observation demonstrates that the attention mechanism can partially alleviate the adverse impacts of task conflict in multi-task optimization.

Table 8
Segmentation ablation results.

Architecture	Dice(%)	mIoU(%)	Parameters (M)	Inference (ms)
MobileViT-XS segmentation	90.45	84.00	2.82	935
MobileViT-XS+CBAM	90.50	84.05	2.82	804
MobileViT-XS + MTL	86.23	77.77	2.60	758
ours	86.37	78.09	2.62	735

4.5. Comparison with existing methods and computational complexity

Table 9 reports the classification performance of different models. The proposed method obtains an overall accuracy of 84.37%, along with a Macro-F1 score of 79.46%, which surpasses all comparative baselines. Specifically, our method outperforms EfficientNet-B0 by 2.74 percentage points in accuracy and achieves better Macro-F1. ResNet-50 and GS-TransUNet exhibit inferior classification results on this skin lesion dataset, with substantially lower Accuracy and Macro-F1 values.

Table 9
Classification comparison with existing methods.

Method	Accuracy (%)	Macro-F1 (%)	Sensitivity (%)	Specificity (%)
EfficientNet-B0	81.63	78.45	82.97	97.09
ResNet-50	62.54	58.43	74.05	94.55
GS-TransUNet	63.35	48.94	52.39	94.14
ours	84.37	79.46	79.68	97.22

Table 10 presents the segmentation performance of competing methods. The proposed multi-task model achieves a Dice coefficient of 86.37% and an mIoU of 78.09%. Compared with single-task segmentation baselines, U-Net [39] and DeepLabV3 + obtain superior segmentation metrics, while GS-TransUNet yields severely degraded performance on the skin lesion dataset. Although our segmentation accuracy is slightly lower than the above two single-task segmentation networks, the proposed model simultaneously fulfills skin lesion classification and segmentation tasks within only 2.62M parameters. Such accuracy-computation trade-off makes it suitable for lightweight portable skin disease screening systems.

Table 10
Segmentation comparison with existing methods.

Method	Dice(%)	mIoU(%)	Inference (ms)
U-Net	0.8961	0.8274	725
DeepLabV3+	0.9009	0.8335	654
GS-TransUNet	0.5521	0.4026	265
ours	0.8637	0.7809	735

4.6. Visualization and interpretability analysis

Visualization results provide qualitative evidence for the proposed method. The confusion matrix showed that the model achieves the highest recall on NV (90.0%), while SCC exhibits the lowest recall (66.7%) owing to frequent confusion with other keratinocytic lesions. It is noteworthy that 17.9% of

malignant MEL samples are misclassified as benign BKL, indicating potential risks of missed clinical diagnosis. The class-wise confusion pattern is shown in Fig. 3.

Once the user uploads an image, the intelligent diagnosis pipeline is activated. The interface displays core outputs, including diagnostic conclusions, risk grades, and confidence scores. The original image, segmentation mask, Grad-CAM heatmap[40], and overlay visualization are presented simultaneously. Supplementary quantitative information, such as lesion size, ABCDE score, color distribution, and an eight-class probability chart, is provided to facilitate intuitive interpretation of the diagnostic evidence. The diagnosis result interface is shown in Fig. 4.

5. Discussion

5.1. Interpretation of results

The experiments verify that MobileViT-XS serves as an effective lightweight backbone for skin lesion analysis. The proposed model achieves an accuracy of 84.37% and a Macro-F1 of 79.46% with only 2.62M parameters. Compared with classic convolutional networks, our method outperforms ResNet-50 by a large margin in classification metrics, while consuming merely around one-tenth of its parameters. It also surpasses EfficientNet-B0, another widely adopted lightweight baseline, with consistent improvements in Accuracy, Macro-F1 and Specificity. Moreover, the recently proposed GS-TransUNet yields poor classification performance on the ISIC 2019 dataset, demonstrating that Transformer-based architectures designed for CT organ segmentation lack sufficient adaptability to dermoscopic skin lesion images with complex textures and ambiguous lesion boundaries.

Multi-task learning brings prominent improvements to classification accuracy, since pixel-wise segmentation supervision drives the shared backbone to capture lesion-related spatial features. Nevertheless, multi-task optimization introduces task competition. Compared with dedicated single-task segmentation models including U-Net and DeepLabV3+, our multi-task method achieves slightly lower Dice and mIoU. It is worth noting that U-Net and DeepLabV3 + are trained solely for segmentation, whereas our unified network simultaneously completes lesion classification and segmentation. In contrast, GS-TransUNet exhibits severe performance degradation on skin lesion segmentation, which further proves its limited generalization ability for dermoscopic data.

The performance gain introduced by CBAM is conditioned on the task paradigm. When deployed in the single-task framework for classification, CBAM contributes marginal performance promotion. A plausible explanation is that image-level classification supervision offers insufficient spatial guidance for attention learning. In contrast, under the multi-task joint learning paradigm, segmentation supervision provides dense spatial cues to optimize the attention module, leading to a more notable improvement in Macro-F1. This observation reveals that attention mechanisms can exert greater feature enhancement capability when combined with pixel-level dense supervision. Benefiting from the combination of

MobileViT-XS backbone, multi-task learning and CBAM attention, the proposed lightweight model achieves competitive comprehensive performance, striking a favorable balance between diagnostic accuracy and computational cost for portable skin lesion screening systems.

5.2. Strengths

The proposed framework possesses four main advantages. First, it achieves a favorable balance between prediction accuracy and computational efficiency, enabling deployment on consumer-grade edge hardware. Second, it leverages collaborative learning between classification and segmentation tasks to boost skin lesion recognition performance. Third, the conservative inference strategy mitigates overconfident predictions for ambiguous samples, a well-known challenge widely discussed in existing research on model calibration and uncertainty estimation [11,12]. Fourth, visualization endows the model with interpretability, which can assist clinicians in comprehending the reasoning behind model outputs.

5.3. Limitations

Several limitations of this study should be acknowledged. First, the model is trained exclusively on public ISIC dermoscopic datasets, where images are collected under standardized and relatively controlled imaging environments. Consequently, domain shift may arise when the model is generalized to real-world clinical scenarios, including primary-care examinations and informal self-screening images with unconstrained lighting, magnification, and shooting angles. Previous external evaluations have demonstrated performance discrepancies across different skin types and demographic distributions, emphasizing the necessity of more diverse and representative validation cohorts [41,42]. In particular, the current dataset lacks sufficient samples from Asian populations, while minority lesion categories (DF and VASC) consistently suffer from insufficient feature learning and relatively poor classification accuracy. In addition, the fixed loss weighting strategy adopted in the multi-task training pipeline leads to unavoidable task competition, which partially degrades segmentation performance. This indicates that dynamically adaptive task-balancing mechanisms should be explored to further optimize multi-task collaboration. Furthermore, the conservative prediction thresholds, including θ_{conf}=0.50, θ_{diff}=0.15, and $p_{malignant}$ =0.25, are empirically predefined in this study and require further statistical calibration based on large-scale clinical data. Finally, all experiments in this work are conducted under retrospective dataset-based evaluation protocols. Therefore, prospective clinical validation is still essential before reliable practical deployment can be achieved.

5.4. Clinical applicability

The proposed method is intended as an auxiliary screening tool for primary-care settings rather than an independent diagnostic system. All high-risk or uncertain outputs should be reviewed by qualified dermatologists. Clinical translation requires staged validation, including retrospective multi-center testing, prospective single-center trials, and prospective multi-center studies [43–45]. Ethical and regulatory issues such as informed consent, data privacy, responsibility attribution, fairness, and medical-device registration should also be addressed.

6. Conclusion and Future Work

6.1. Conclusion

This study proposes an attention-enhanced lightweight MobileViT-XS framework to simultaneously tackle skin lesion classification and semantic segmentation tasks. Built upon the compact MobileViT-XS backbone embedded with the CBAM attention module, the designed multi-task model achieves favorable comprehensive performance with merely 2.62M parameters. On the ISIC 2019 dataset, the model reaches a classification accuracy of 84.37%, a Macro-F1 of 79.46%, and attains a Dice coefficient of 86.37% and an mIoU of 78.09% for lesion segmentation. Ablation experiments uncover an asymmetric interaction between the two tasks during joint optimization. Pixel-level segmentation supervision significantly facilitates the learning of lesion-related spatial features and greatly boosts classification performance. Nevertheless, task competition emerges under multi-task training, leading to a moderate decline in segmentation metrics compared with dedicated single-task segmentation models. Further experiments validate that the CBAM attention mechanism delivers limited performance gains within the single-task paradigm. When combined with multi-task learning, dense segmentation supervision guides attention weight optimization, enabling CBAM to strengthen discriminative spatial features and bring consistent performance improvements.

Furthermore, this work introduces a conservative inference strategy and risk stratification mechanism, which imposes constraints on model outputs to reduce hazardous overconfident predictions for ambiguous dermoscopic samples. Combined with visual explanations that improve model interpretability, the proposed integrated framework provides a practical and lightweight technical solution for auxiliary skin lesion screening, which has promising application prospects in resource-limited primary-care scenarios and portable edge devices.

6.2. Future work

Future work will focus on three directions. First, multi-center and multi-device clinical datasets should be collected to improve domain generalization and demographic representativeness. Second, dynamic multi-task loss balancing, focal loss, synthetic augmentation, uncertainty estimation, calibration assessment, and more advanced interpretability methods should be explored [11,12,28–30,36–38,40,46]. Third, mobile and edge deployment should be implemented using ONNX or TFLite, followed by prospective clinical validation to evaluate real-world effectiveness and safety.

Declarations

Funding

This work was supported by the Henan Provincial Science and Technology Research Project (Grant No. 242102210116) .

Author Contribution

Y.G. conceived and supervised the project, conducted benchmark evaluations, and drafted the manuscript. B.L. carried out the technical validation experiments. Z.Z. participated in data collection and offered guidance on image processing. H.W. and Z.Y. assisted with data collection. All authors reviewed and approved the final manuscript.

Data Availability

The datasets used in this study are publicly available in online repositories. The skin image dataset is available from the International Skin Imaging Collaboration at: https://challenge.isic-archive.com/data/#2019,Segmentation mask images are available from the International Skin Imaging Collaboration at:https://challenge.isic-archive.com/data/#2018.These datasets can be accessed through their respective repositories subject to the terms and conditions of each platform.

References

1. World Health Organization. Radiation: Ultraviolet (UV) radiation and skin cancer. Geneva: World Health Organization, 2017. Available from: https://www.who.int/news-room/questions-and-answers/item/radiation-ultraviolet-(uv)-radiation-and-skin-cancer.
2. Siegel RL, Giaquinto AN, Jemal A. Cancer statistics, 2024. CA: A Cancer Journal for Clinicians, 2024, 74(1): 12–49. DOI: 10.3322/caac.21820.
3. Kassem MA, Hosny KM, Damasevicius R, et al. Machine learning and deep learning methods for skin lesion classification and diagnosis: A systematic review. Diagnostics, 2021, 11(8): 1390. DOI: 10.3390/diagnostics11081390.
4. Liu T, Liu J, Ma J, et al. Application progress of deep learning in skin cancer lesion classification and diagnosis. Computer Systems & Applications, 2024, 33(12): 1–15.
5. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Medical Image Analysis, 2017, 42: 60–88. DOI: 10.1016/j.media.2017.07.005.
6. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 2017, 542(7639): 115–118. DOI: 10.1038/nature21056.
7. Zafar M, Sharif MI, Sharif MI, et al. Skin lesion analysis and cancer detection based on machine/deep learning techniques: A comprehensive survey. Life, 2023, 13(1): 146. DOI: 10.3390/life13010146.
8. Brinker TJ, Hekler A, Utikal JS, et al. Skin cancer classification using convolutional neural networks: Systematic review. Journal of Medical Internet Research, 2018, 20(10): e11936. DOI: 10.2196/11936.
9. Haenssle HA, Fink C, Schneiderbauer R, et al. Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison

to 58 dermatologists. Annals of Oncology, 2018, 29(8): 1836–1842. DOI: 10.1093/annonc/mdy166.
10. Tschandl P, Rinner C, Apalla Z, et al. Human-computer collaboration for skin cancer recognition. Nature Medicine, 2020, 26(8): 1229–1234. DOI: 10.1038/s41591-020-0942-0.
11. Guo C, Pleiss G, Sun Y, et al. On calibration of modern neural networks. ICML, 2017: 1321–1330.
12. Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. NeurIPS, 2017: 6402–6413.
13. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. CVPR, 2016: 770–778. DOI: 10.1109/CVPR.2016.90.
14. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16 x 16 words: Transformers for image recognition at scale. ICLR, 2021.
15. Liu Z, Lin Y, Cao Y, et al. Swin Transformer: Hierarchical vision transformer using shifted windows. ICCV, 2021: 10012–10022. DOI: 10.1109/ICCV48922.2021.00986.
16. Sandler M, Howard A, Zhu M, et al. MobileNetV2: Inverted residuals and linear bottlenecks. CVPR, 2018: 4510–4520. DOI: 10.1109/CVPR.2018.00474.
17. Chollet F. Xception: Deep learning with depthwise separable convolutions. CVPR, 2017: 1251–1258. DOI: 10.1109/CVPR.2017.195.
18. Radosavovic I, Kosaraju RP, Girshick R, et al. Designing network design spaces. CVPR, 2020: 10428–10436. DOI: 10.1109/CVPR42600.2020.01044.
19. Howard A, Sandler M, Chu G, et al. Searching for MobileNetV3. ICCV, 2019: 1314–1324. DOI: 10.1109/ICCV.2019.00140.
20. Ma N, Zhang X, Zheng HT, et al. ShuffleNet V2: Practical guidelines for efficient CNN architecture design. ECCV, 2018: 116–131. DOI: 10.1007/978-3-030-01264-9_8.
21. Tan M, Le QV. EfficientNet: Rethinking model scaling for convolutional neural networks. ICML, 2019: 6105–6114.
22. Mehta S, Rastegari M. MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer. ICLR, 2022.
23. Woo S, Park J, Lee JY, et al. CBAM: Convolutional block attention module. ECCV, 2018: 3–19. DOI: 10.1007/978-3-030-01234-2_1.
24. Oktay O, Schlemper J, Folgoc LL, et al. Attention U-Net: Learning where to look for the pancreas. arXiv:1804.03999, 2018.
25. Kendall A, Gal Y, Cipolla R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. CVPR, 2018: 7482–7491. DOI: 10.1109/CVPR.2018.00781.
26. Chen J, Chen J, Zhou Z, et al. MT-TransUNet: Mediating multi-task tokens in transformers for skin lesion segmentation and classification. arXiv:2112.01767, 2021.
27. Yang H, Bai Z. CoT-TransUNet: A lightweight context Transformer network for medical image segmentation. Computer Engineering and Applications, 2023, 59(03): 218–225.

28. Lin TY, Goyal P, Girshick R, et al. Focal loss for dense object detection. ICCV, 2017: 2980–2988. DOI: 10.1109/ICCV.2017.324.
29. Cui Y, Jia M, Lin TY, et al. Class-balanced loss based on effective number of samples. CVPR, 2019: 9268–9277. DOI: 10.1109/CVPR.2019.00949.
30. Menon AK, Jayasumana S, Rawat AS, et al. Long-tail learning via logit adjustment. ICLR, 2021.
31. Codella NCF, Gutman D, Celebi ME, et al. Skin lesion analysis toward melanoma detection: A challenge at ISBI, hosted by ISIC. ISBI, 2018. DOI: 10.1109/ISBI.2018.8363547.
32. Tschandl P, Rosendahl C, Kittler H. The HAM10000 dataset: A large collection of multi-source dermatoscopic images. Scientific Data, 2018, 5: 180161.
33. Hernández-Pérez C, Combalia M, Podlipnik S, Codella NC, Rotemberg V, Halpern AC, Reiter O, Carrera C, Barreiro A, Helba B, Puig S, Vilaplana V, Malvehy J. BCN20000: Dermoscopic lesions in the wild. Scientific Data. 2024 Jun 17;11(1):641. DOI: 10.1038/s41597-024-03387-w.
34. Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, Allan Halpern: "Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC)", 2018; https://arxiv.org/abs/1902.03368
35. Rotemberg V, Kurtansky N, Betz-Stablein B, et al. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. Scientific Data, 2021, 8: 34. DOI: 10.1038/s41597-021-00815-z.
36. Zhang H, Cisse M, Dauphin YN, et al. mixup: Beyond empirical risk minimization. ICLR, 2018.
37. Cubuk ED, Zoph B, Mane D, et al. AutoAugment: Learning augmentation strategies from data. CVPR, 2019: 113–123.
38. Yun S, Han D, Oh SJ, et al. CutMix: Regularization strategy to train strong classifiers with localizable features. ICCV, 2019: 6023–6032.
39. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. MICCAI, 2015. DOI: 10.1007/978-3-319-24574-4_28.
40. Selvaraju RR, Cogswell M, Das A, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization. ICCV, 2017: 618–626. DOI: 10.1109/ICCV.2017.74.
41. Groh M, Harris C, Soenksen LR, et al. Evaluating deep neural networks trained on clinical images in dermatology with the Fitzpatrick 17k dataset. CVPR Workshops, 2021: 1820–1828. DOI: 10.1109/CVPRW53098.2021.00201.
42. Daneshjou R, Vodrahalli K, Novoa RA, et al. Disparities in dermatology AI performance on a diverse, curated clinical image set. Science Advances, 2022, 8(31): eabq6147. DOI: 10.1126/sciadv.abq6147.
43. Heinlein L, Maron RC, Hekler A, et al. Clinical melanoma diagnosis with artificial intelligence: Insights from a prospective multicenter study. arXiv:2401.14193, 2024.
44. Kelly CJ, Karthikesalingam A, Suleyman M, et al. Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine, 2019, 17: 195. DOI: 10.1186/s12916-019-1426-2.

45. Liu X, Rivera SC, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. Nature Medicine, 2020, 26(9): 1364–1374. DOI: 10.1038/s41591-020-1034-x.
46. Chen Z, Badrinarayanan V, Lee CY, Rabinovich A. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. ICML, 2018: 794–803.

Figures

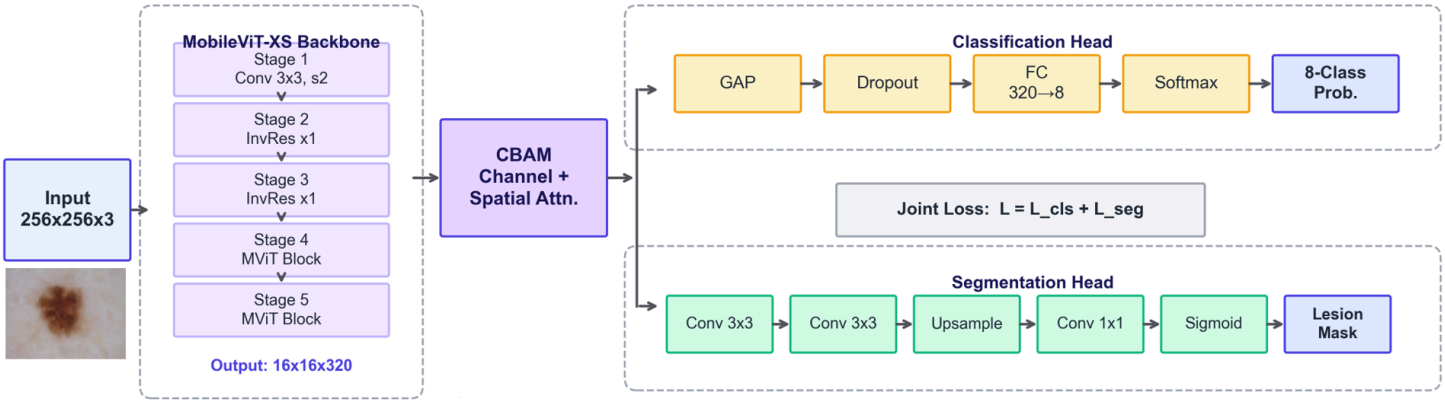


Figure 1

Overall architecture of the proposed framework.

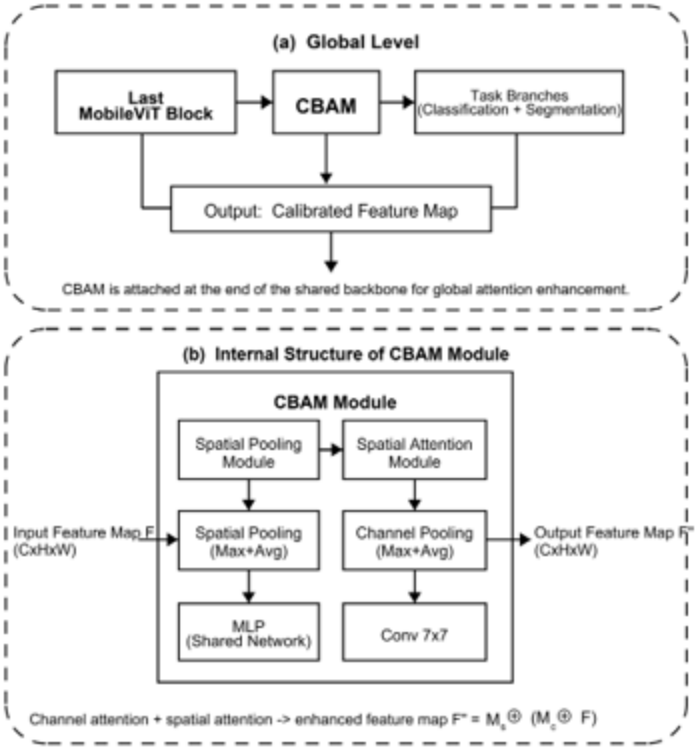


Figure 2

CBAM integration strategy.

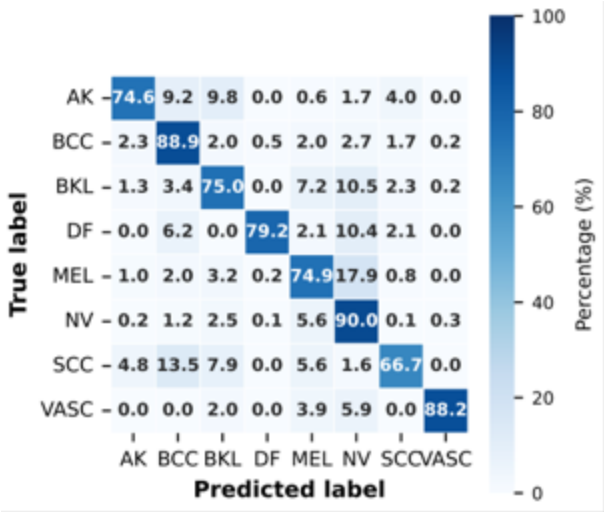


Figure 3

Classification Row-normalized confusion matrix

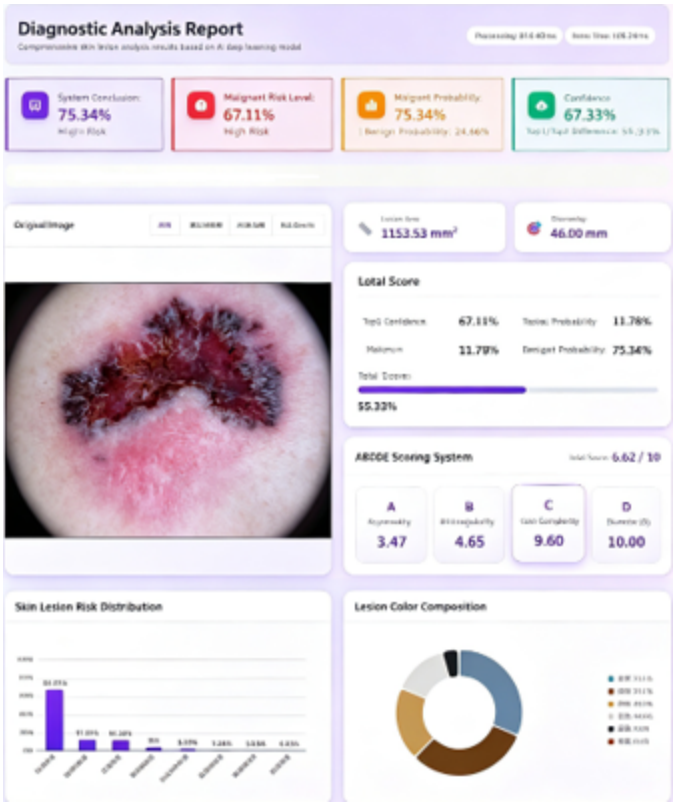


Figure 4

Diagnostic Result Interface