


HI-MViT: A lightweight model for explainable skin disease classification based on modified MobileViT

DIGITAL HEALTH
Volume 9: 1–30
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20552076231207197
journals.sagepub.com/home/dhj



Yuhan Ding^{1,2,3,†}, Zhenglin Yi^{2,4,†}, Mengjuan Li^{1,2}, Jianhong long^{1,2},
Shaorong Lei^{1,2}, Yu Guo^{1,2}, Pengju Fan^{1,2}, Chenchen Zuo^{1,2}
and Yongjie Wang^{1,2} 

Abstract

Objective: To develop an explainable lightweight skin disease high-precision classification model that can be deployed to the mobile terminal.

Methods: In this study, we present HI-MViT, a lightweight network for explainable skin disease classification based on Modified MobileViT. HI-MViT is mainly composed of ordinary convolution, Improved-MV2, MobileViT block, global pooling, and fully connected layers. Improved-MV2 uses the combination of shortcut and depth classifiable convolution to substantially decrease the amount of computation while ensuring the efficient implementation of information interaction and memory. The MobileViT block can efficiently encode local and global information. In addition, semantic feature dimensionality reduction visualization and class activation mapping visualization methods are used for HI-MViT to further understand the attention area of the model when learning skin lesion images.

Results: The International Skin Imaging Collaboration has assembled and made available the ISIC series dataset. Experiments using the HI-MViT model on the ISIC-2018 dataset achieved scores of 0.931, 0.932, 0.961, and 0.977 on F1-Score, Accuracy, Average Precision (AP), and area under the curve (AUC). Compared with the top five algorithms of ISIC-2018 Task 3, Marco's average F1-Score, AP, and AUC have increased by 6.9%, 6.8%, and 0.8% compared with the suboptimal performance model. Compared with ConvNeXt, the most competitive convolutional neural network architecture, our model is 5.0%, 3.4%, 2.3%, and 2.2% higher in F1-Score, Accuracy, AP, and AUC, respectively. The experiments on the ISIC-2017 dataset also achieved excellent results, and all indicators were better than the top five algorithms of ISIC-2017 Task 3. Using the trained model to test on the PH² dataset, an excellent performance score is obtained, which shows that it has good generalization performance.

Conclusions: The skin disease classification model HI-MViT proposed in this article shows excellent classification performance and generalization performance in experiments. It demonstrates how the classification outcomes can be applied to dermatologists' computer-assisted diagnostics, enabling medical professionals to classify various dermoscopic images more rapidly and reliably.

Keywords

Dermatology classification, HI-MViT, deep learning, lightweight model, explainable artificial intelligence

Submission date: 8 May 2023; Acceptance date: 26 September 2023

¹Department of Burns and Plastic Surgery, Xiangya Hospital, Central South University, Changsha, Hunan, China

²National Clinical Research Center for Geriatric Disorders, Xiangya Hospital, Central South University, Changsha, China

³School of Computer Science and Engineering, Central South University, Changsha, Hunan, China

⁴Departments of Urology, Xiangya Hospital, Central South University, Changsha, China

[†]These authors have contributed equally to this work as first authors.

Corresponding author:

Yongjie Wang, Xiangya Hospital Central South University, No. 87, Xiangya Road, Changsha City, Hunan 410008, China.

Email: yongjiawang@csu.edu.cn



Introduction

Dermatosis is a common disease, frequently occurring disease in medicine. The World Health Organization once said that skin diseases will be the most common disease in human history in the 21st century, with the highest morbidity rate and the highest disability rate. About 30% to 70% of people of different races and ages in the world suffer from skin diseases.¹ Dermatologists employ dermoscopy, a noninvasive skin imaging method, to aid in diagnosis.² Although qualified human experts can correctly identify a substantial proportion of dermoscopic images with an accuracy of 80%, it requires considerable time and work to do so.³ Therefore, research that uses computer-aided diagnosis methods to categorize images of skin illnesses is extremely significant.

The traditional method of classifying skin disease images is based on a series of processes including preprocessing, lesion extraction, and feature extraction.⁴ First, use effective methods such as asymmetrical, border, color, diameter Rule, Menzie's rule, or seven-point checklist to extract features from skin lesion images, and then use various machine learning methods such as eXtreme gradient boosting, decision tree, or support vector machine to classify hand-designed features.⁵ Because the traditional method relies heavily on the quantity and quality of manually designed features, it cannot be used for more types of skin lesion image classification, and cannot meet the classification requirements of higher accuracy.⁶ Compared with traditional methods, methods based on convolutional neural networks (CNN) can learn meaningful features directly from data, Esteva et al.⁷ employed a CNN framework based on Inception-V3 to train a skin disease classification model with an accuracy rate of 71.2% and verified that the algorithm was capable of categorization accuracy on par with 21 dermatologists who hold board certification. Duman et al.⁸ propose a novel ensemble method that combines various advantages of several existing CNN models to deal with large-scale imbalanced datasets. By using the weighted aggregation method, the accuracy score is improved by 5% to 10%, compared with the state-of-the-art, the average sensitivity and area under the curve (AUC) values are 0.825 and 0.923, respectively, ranking second. Compared with traditional methods, the classification of images of skin diseases has been improved using CNN-based techniques. The following difficulties still exist in the more effective categorization of skin lesions because of the uniqueness of skin disease images: (1) The dermoscopic image only contains a small portion of the skin lesion region; the majority of the space is taken up by normal tissues or other unimportant details, which could skew the findings of the recognition process. (2) The classification process is more challenging and it is challenging to acquire reliable findings due to the similarities between classes and differences

within classes of skin lesions.⁹ Figure 1 shows images of three skin diseases: melanocytic nevi, melanoma, and actinic keratosis. It is evident that while various types of dermoscopic images may share geometric shapes or colors, the same type of dermoscopic images may have stark visual contrasts, which would negatively impact the model's capacity to generalize.

For these factors, it is typically more challenging to categorize images of skin lesions than it is to classify the objects and settings in natural images. As a result, better models must be created to improve the efficacy of the classification of skin lesions. In 2020, the Google team applied the Transformer¹⁰ technology widely used in the field of natural language processing to the field of computer vision and proposed Vision Transformer¹¹ (ViT). This study has established a new standard for employing Transformer-based techniques to solve issues related to computer vision. Transformer-based techniques are currently being commonly applied in the area of medical image processing, significantly cutting down on time and labor expenses. Using the ISIC-2017 dataset,¹² Wang et al.¹³ created the O-Net algorithm to categorize dermoscopy images. However, compared with the CNN-based lightweight model, ViT still has a big gap in terms of model parameters and inference speed, and it is very difficult to implement it on the mobile terminal. Sachin et al.¹⁴ offered a lightweight general visualization Transformer for mobile devices as a solution to these problems, which is the first lightweight ViT work based on the performance of lightweight CNN networks. According to experimental findings, MobileViT outperforms MobileNetV3, CrossViT, and other nets on a wide range of tasks and datasets.

To better understand the development process and latest progress of skin lesion image classification technology, we review the relevant work of skin lesion image classification models based on deep learning^{15–28} in recent years in Table 1. It details the keywords, methods, data, modality, dermatology categories, and results of the relevant models.

Building on related work, we propose HI-MViT, a lightweight model for explainable skin disease classification and the goal is to model local and global information with fewer parameters to achieve fast and robust dermoscopic image classification performance on the mobile side. It uses MobileNet²⁹ and Vision Transformer hybrid architecture MobileViT as the basis, and innovatively designed the Improved-MV2 block to improve it, which can substantially scale back on the number of calculations and facilitate better landing on the mobile terminal. Experiments on the ISIC-2018 dataset³⁰ show that F1-Score, Accuracy, average precision (AP), and AUC have achieved excellent scores of 0.931, 0.932, 0.961, and 0.977, respectively. The following is a summary of this article's major contributions:

1. This study proposes an explainable skin disease classification lightweight model HI-MViT, which is mainly

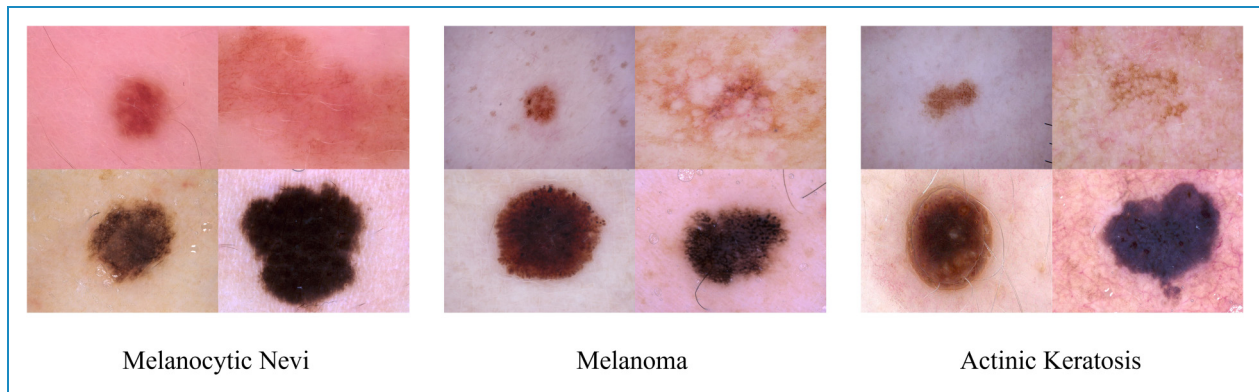


Figure 1. Dermoscopy images of melanocytic nevi, melanoma, and actinic keratosis.

composed of ordinary convolution, Improved-MV2, MobileViT block, global pooling, and fully connected layers. The network is much more concentrated on the skin lesion area thanks to the Transformer’s unique self-attention mechanism and global vision, which increases the classification accuracy of skin lesion images. The framework also inherits the lightweight and high efficiency of CNN.

2. Innovatively proposed the Improved-MV2 block. The combination of the expansion layer, projection layer, and depthwise separable convolution uses a shortcut to drastically cut back on calculation while ensuring the efficient implementation of information interaction and memory. It is more suitable for the use of lightweight models and facilitates the implementation of models on mobile terminals.
3. A large number of performance experiments and visualization experiments were conducted on the ISIC-2018 and ISIC-2017 skin lesion classification datasets released by the International Skin Imaging Collaboration to fully verify the superiority of the model. Under the same data processing and experimental conditions, HI-MViT achieved better performance than the selected existing mainstream classification models. At the same time, the trained model is tested on the PH² dataset, and excellent classification performance is obtained, which proves that it has good generalization performance and robustness.

Methods

HI-MViT structure

The structure of the HI-MViT model proposed in this article is shown in Figure 2. Its purpose is to build a lightweight interpretable deep learning model that can be deployed on the mobile terminal for accurate and fast classification of skin lesion images. HI-MViT is mainly composed of ordinary convolution, Improved-MV2, MobileViT block, global pooling, and fully connected layers. Blocks marked with a down arrow represent the need for downsampling.

MobileViT block

The MobileViT block can efficiently encode local and global information. In MobileViT block, for a given input tensor $X \in R^{H \times W \times C}$, an $n \times n$ standard convolutional layer is first used, preceded by a 1×1 convolutional layer to generate features $X_L \in R^{H \times W \times d}$. $n \times n$ convolutional layers encode local spatial information, while pointwise convolutions project tensors into d -dimensional spaces by learning linear combinations of input channels.¹⁴

We extend X_L into N non-overlapping flattened patches $X_U \in R^{P \times N \times d}$ to allow HI-MViT to acquire a global description with spatial induction bias. The relation $X_G \in R^{P \times N \times d}$ between patches is encoded by applying Transformer, as shown in equation 1, where $P = wh$, $N = HW/P$ is the number of patches, $h \leq N$, $w \leq N$ are the height and width of the patch correspondingly, $p \in \{1, \dots, P\}$.

$$X_G(p) = \text{Transformer}(X_U(p)), 1 \leq p \leq P \quad (1)$$

HI-MViT does not lose either the patch sequence or the spatial order of the pixels inside every patch. The result of folding $X_G \in R^{P \times N \times d}$ is $X_F \in R^{H \times W \times d}$. After that, X_F is pointwise convected to a low-dimensional region and joined with X using the Concat method. The local and global characteristics are then combined in the concatenated tensor using a further $n \times n$ convolutional layer.

As a result the fact that $X_U(p)$ uses convolution to encode local information for $n \times n$ areas. So every pixel in X_G can encode the details of every pixel in X for the patch at the P —position P . The global knowledge for this patch is encoded by $X_G(p)$. As a consequence, Figure 3 depicts the HI-MViT $H \times W$ ’s overall efficient receptive field.

Improved-MV2 block

The Improved-MV2 block executes a depthwise separable convolution operation, a 1×1 expansion layer to raise the total amount of channels, and a 1×1 projection layer to return the channel count to its initial size.³¹ The skip connection is established between two bottleneck layers with

Table 1. Related work on image classification models for skin diseases based on deep learning.

References	Keywords	Methods	Datasets	Modality	Classification type	Results
Carcagni et al. ¹⁵ (2019)	CNN, multilevel learning	A skin lesion classification model based on DenseNet and multilevel learning is proposed	HAM10000	Image	7 Classes (aktinic keratoses, basal cell carcinoma, benign ceratosis, dermatofibroma, melanocytic nevi, melanoma, vascular lesion)	Precision: 0.880 Recall: 0.760 F1-Score: 0.820
Jain et al. ¹⁶ (2021)	CNN, transfer learning	Analyzing the performance of skin lesion classification using different data processing and transfer learning networks	HAM10000	Image	7 Classes (aktinic keratoses, basal cell carcinoma, benign ceratosis, dermatofibroma, melanocytic nevi, melanoma, vascular lesion)	Accuracy: 0.897 Precision: 0.888 Recall: 0.896 F1-Score: 0.890
Hu et al. ¹⁷ (2022)	CNN, lightweight model, attention mechanism	An improved lightweight mobile network is proposed based on MobileNet for skin disease classification	HAM10000	Image	7 Classes (aktinic keratoses, basal cell carcinoma, benign ceratosis, dermatofibroma, melanocytic nevi, melanoma, vascular lesion)	Accuracy: 0.850 Sensitivity: 0.849 Specificity: 0.955 Precision: 0.841 F1-Score : 0.842
Muhaaba et al. ¹⁸ (2022)	CNN, lightweight model	An automated system is proposed for classifying skin diseases based on pretrained MobileNet-V2 models	Self-collected dataset	Image, metadata	7 Classes (healthy, acne vulgaris, atopic dermatitis, lichen planus, onychomycosis, tinea capitis, unknown)	Accuracy: 0.975 Sensitivity: 0.977 Precision: 0.977
Xin et al. ¹⁹ (2022)	Vision transformer, contrastive learning	A relatively simple skin cancer classification model is proposed based on the transformer framework and contrastive learning method	HAM10000, self-collected dataset	Image	(1) HAM10000 : 7 classes (aktinic keratoses, basal cell carcinoma, benign ceratosis, dermatofibroma, melanocytic nevi, melanoma, vascular lesion) (2) Self-collected dataset : 3 classes (basal cell carcinoma, malignant melanoma, squamous cell carcinoma)	(1) HAM10000 : Accuracy: 0.943 Precision: 0.941 AUC: 0.987 (2) Self-collected dataset : Accuracy: 0.941 Precision: 0.942 F1-Score: 0.941
He et al. ²⁰ (2022)	Transformer, multihead attention	A full transformer network that can learn long-range contextual information for skin lesion analysis is proposed	ISIC-2018	Image	7 classes (aktinic keratoses, basal cell carcinoma, benign ceratosis, dermatofibroma, melanocytic nevi, melanoma, vascular lesion)	Accuracy: 0.927 Sensitivity: 0.857 Specificity: 0.936 Precision: 0.621 AUC: 0.897

(continued)

Table 1. Continued.

References	Keywords	Methods	Datasets	Modality	Classification type	Results
Sarker et al. ²¹ (2022)	CNN, transformer	A transformer-based skin lesion classification model is proposed	HAM10000	Image	7 classes (aktinic keratoses, basal cell carcinoma, benign ceratosis, dermatofibroma, melanocytic nevi, melanoma, vascular lesion)	Accuracy: 0.902 Precision: 0.995 Recall: 0.941 F1-Score: 0.963
Nakai et al. ²² (2022)	CNN, transformer, self-attention, position knowledge	A novel enhanced deep bottleneck transformer model is proposed for skin lesion classification	ISIC-2017, HAM10000	Image	(1) ISIC-2017 : 3 classes (melanoma, nevus, seborrheic keratosis) (2) HAM10000 : 7 classes (aktinic keratoses, basal cell carcinoma, benign ceratosis, dermatofibroma, melanocytic nevi, melanoma, vascular lesion)	(1) ISIC-2017 : Accuracy: 0.921 Sensitivity: 0.901 Specificity: 0.919 (2) HAM10000 : Accuracy: 0.958 Precision: 0.961
Aladhadh et al. ²³ (2022)	Medical vision transformer	A two-stage skin cancer classification model based on the medical vision transformer is proposed	HAM10000	Image	7 classes (aktinic keratoses, basal cell carcinoma, benign ceratosis, dermatofibroma, melanocytic nevi, melanoma, vascular lesion)	Accuracy: 0.961 Sensitivity: 0.965 Precision: 0.960 F1-Score: 0.970
Gocer ²⁴ (2023)	CNN, capsule network	A novel network including adjustable and fully convolutional capsule layers is proposed	HAM10000	Image	7 classes (aktinic keratoses, basal cell carcinoma, benign ceratosis, dermatofibroma, melanocytic nevi, melanoma, vascular lesion)	Accuracy: 0.952 Sensitivity: 0.954 Specificity: 0.992 F1-Score: 0.950
Anand et al. ²⁵ (2023)	CNN, U-Net architecture, fusion model	A fusion model that combines U-Net and convolutional neural network models is proposed	HAM10000	Image	7 classes (aktinic keratoses, basal cell carcinoma, benign ceratosis, dermatofibroma, melanocytic nevi, melanoma, vascular lesion)	Accuracy: 0.980 Sensitivity: 0.849 Specificity: 0.979 Precision: 0.885
Ayas ²⁶ (2023)	Swin transformer	A Swin transformer model is proposed for multiclass skin lesion classification	ISIC-2019	Image	9 classes (aktinic keratosis, basal cell carcinoma, benign keratosis, dermatofibroma, melanocytic nevus, melanoma, squamous cell carcinoma, vascular lesion)	Accuracy: 0.972 Sensitivity: 0.823 Specificity: 0.979

(continued)

Table 1. Continued.

References	Keywords	Methods	Datasets	Modality	Classification type	Results
Cai et al. ²⁷ (2023)	Transformer, multimodal fusion, attention mechanism	A novel multimodal transformer for processing images and metadata is proposed	ISIC-2018, self-collected dataset	Image, metadata	(1) ISIC-2018 : 7 classes (aktinic keratoses, basal cell carcinoma, benign ceratosis, dermatofibroma, melanocytic nevi, melanoma, vascular lesion) (2) Self-collected dataset : 9 classes (skin necrosis, skin defect, skin, and soft tissue infection, gangrene, sinus tract, first-degree burn, second-degree burn, third-degree burns, scar healing)	(1) ISIC-2018 : Accuracy: 0.938 Sensitivity: 0.901 Specificity: 0.984 F1-Score: 0.901 AUC: 0.993 (2) Self-collected dataset : Accuracy: 0.816 Sensitivity: 0.854 Specificity: 0.975 F1-Score: 0.820 AUC: 0.974
Mukadam and Patil ²⁸ (2023)	CNN, generative adversarial network	A skin cancer classification framework based on enhanced super-resolution generative adversarial networks and custom convolutional neural networks is proposed	HAM10000	Image	7 classes (aktinic keratoses, basal cell carcinoma, benign ceratosis, dermatofibroma, melanocytic nevi, melanoma, vascular lesion)	Protocol-I : Accuracy: 0.988 Protocol-II : Accuracy: 0.984 Protocol-III : Accuracy: 0.989

AUC: area under the curve; CNN: convolutional neural networks.

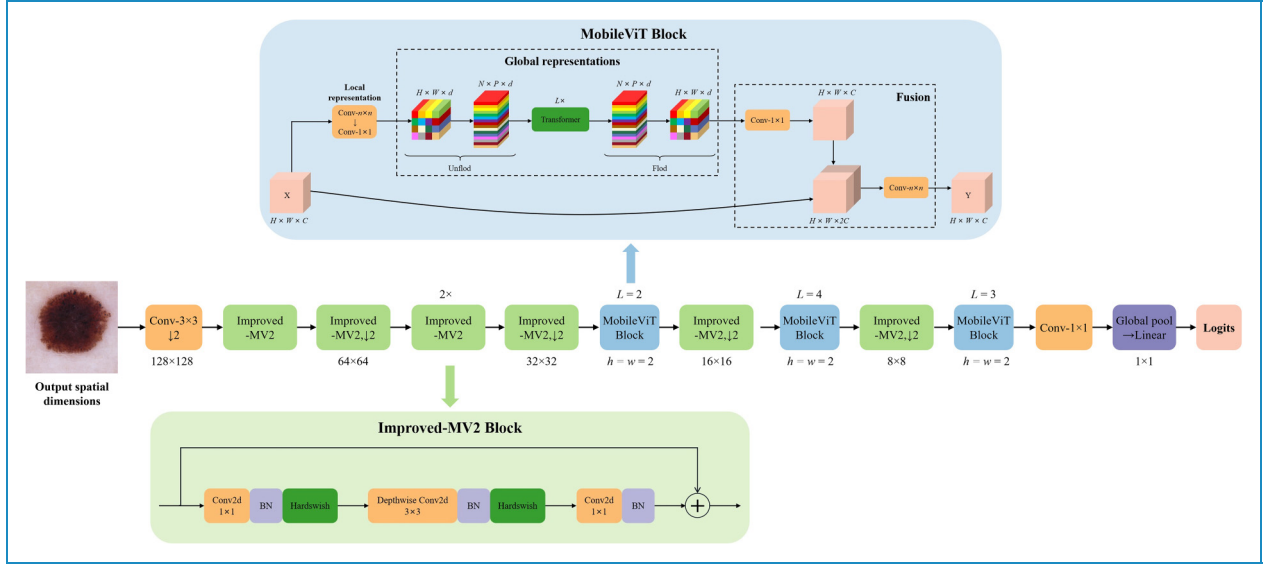


Figure 2. Network structure diagram of HI-MViT.

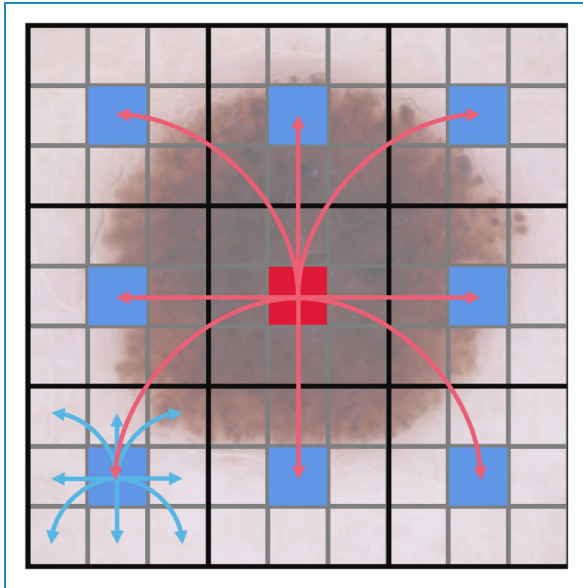


Figure 3. In MobileViT block, every pixel is aware of other pixels. Gray grids represent pixels, and black grids composed of gray grids represent patches. Through the transformer, the red pixel processes the blue pixel at the corresponding point in other patches. This enables the red pixel to encode information from every pixel in the image since the blue pixel has previously used convolution to encode data from nearby pixels.

fewer channels, which is just the opposite of the residual structure in ResNet.³² Figure 4(a) depicts the structure when stride=1 and Figure 4(b) depicts the architecture when stride=2.

Depthwise convolution and pointwise convolution are the two components of Depthwise Separable Convolution (DSC). Every convolution kernel is convolved with a

different dimension of the input feature matrix in a standard convolution. Every channel of the input feature map is given its convolution kernel using depthwise convolution, which then combines the outcomes of all the convolution kernels to get the final output. Typically, a depthwise convolution layer comes after a pointwise convolution operation. Pointwise convolution is a 1×1 convolution that executes channel merging on the feature map produced by depthwise convolution in addition to allowing users to flexibly modify the number of output channels. Therefore, DSC not only has the advantages of fewer parameters and faster calculation but also overcomes the disadvantage of lack of information interaction in group convolution.³³ Ordinary convolution is contrasted with depthwise convolution and pointwise convolution in Figure 5.

Through this modification to the feature map's channel splitting, DSC has decreased the total amount of parameters, which has improved the network's lightweight. Every spot in the spatial position of the related feature map will undergo a convolution operation under the assumption that the input feature map is $D_k \times D_k \times M$ pixels in size, the convolution kernel is $D_F \times D_F \times M$ pixels in size, and there is N convolution kernels total. Then it can be seen that a single convolution requires a total of $D_k \times D_k \times D_F \times D_F \times M$ calculations.

Therefore, for N convolutions, the total calculation amount is: $D_k \times D_k \times D_F \times D_F \times M \times N$. In the same way, it can be analyzed that the total calculation of depthwise convolution is: $D_k \times D_k \times D_F \times D_F \times M$, and the total calculation of pointwise convolution is: $M \times N \times D_k \times D_k$. So the total calculation of DSC is: $D_k \times D_k \times D_F \times D_F \times M + M \times N \times D_k \times D_k$. Then,

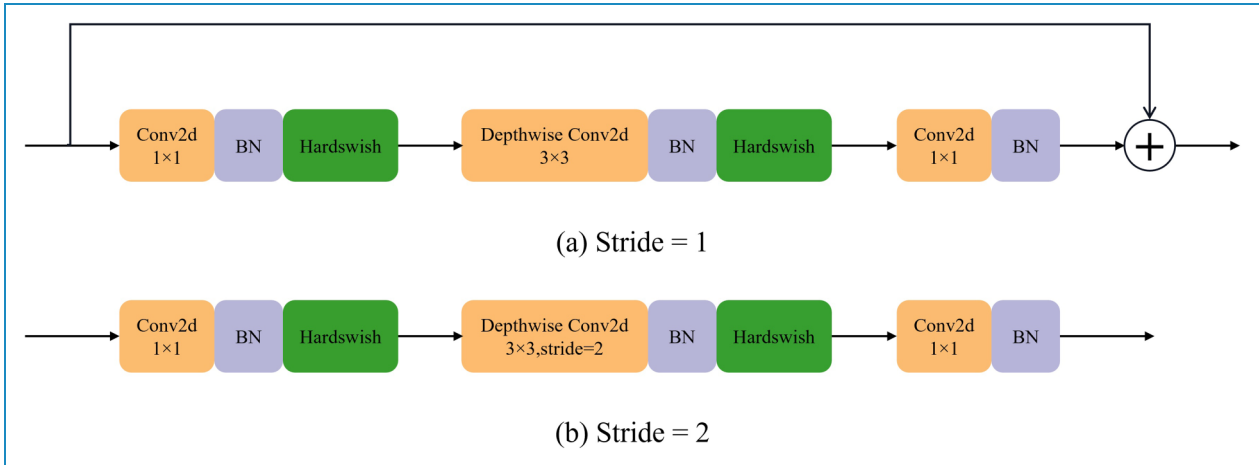


Figure 4. Improved-MV2 block structure: (a) stride = 1 and (b) stride = 2.

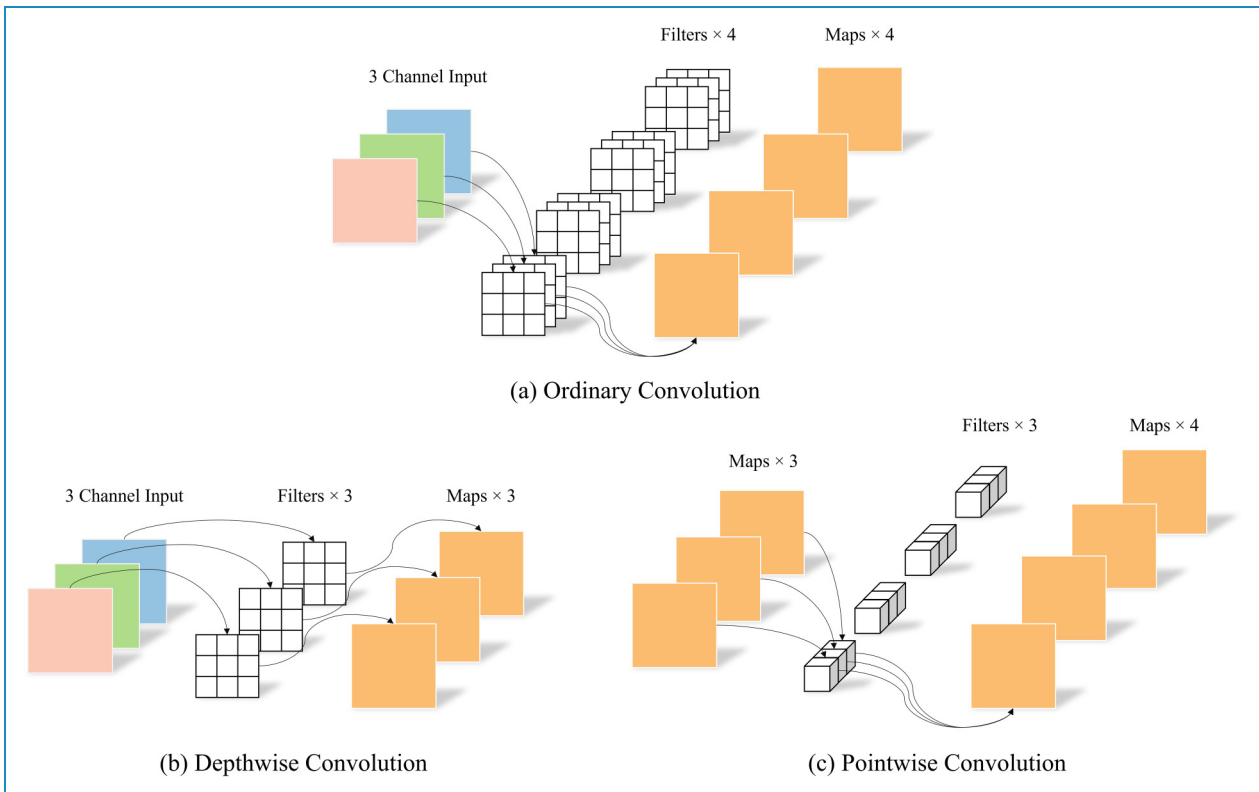


Figure 5. Comparison of ordinary convolution with depthwise convolution and pointwise convolution.

compared to ordinary convolution, the ratio of DSC calculation to ordinary convolution is: $\frac{1}{N} + \frac{1}{D_f}$. Taking this article as an example, in theory, the calculation amount of ordinary convolution is 8~9 times that of DSC, and the calculation efficiency of DSC is far better than that of ordinary convolution.

Improved-MV2 additionally permits memory-efficient implementation, which is crucial for the use of lightweight

models on mobile terminals. Create a directed acyclic computational hypergraph G with nodes representing tensors of intermediate calculations and edges representing operations to use a typical efficient inference implementation in PyTorch. The overall amount of tensors that must be maintained in memory is kept as low as possible by scheduling computations accordingly. As a whole, all reasonable calculation sequences are searched and the smallest one

Table 2. Specific network parameter configuration information of the HI-MViT model.

Layer	Output size	Output stride	Repeat	Output channels
Image	256 × 256	1	-	-
Conv-3 × 3, ↓2	128 × 128	2	1	16
Improved-MV2			1	32
Improved-MV2, ↓2	64 × 64	4	1	64
Improved-MV2			2	64
Improved-MV2, ↓2	32 × 32	8	1	96
MobileViT block ($L = 2$)			1	96($d = 144$)
Improved-MV2, ↓2	16 × 16	16	1	128
MobileViT block ($L = 4$)			1	128($d = 192$)
Improved-MV2, ↓2	8 × 8	32	1	160
MobileViT block ($L = 3$)			1	160($d = 240$)
Conv-1 × 1			1	640
Global pool	1 × 1	256	1	-
Linear			1	1000
Network Parameters	-	-	-	5.6M

is selected. The specific operation process is shown in equation 2:

$$M(G) = \min_{\pi \in \Sigma(G)} \max_{i \in 1 \dots n} \left[\sum_{A \in R(i, \pi, G)} |A| \right] + \text{size}(\pi_i) \quad (2)$$

where $|A|$ is the size of tensor A and $R(i, \pi, G)$ a sequence of intermediary tensors linked to any of the $(\pi_i \dots \pi_n)$ nodes. The total quantity of RAM needed for internal storage throughout the process is $\text{size}(i)$. There is just one atypical order of computation that makes calculating the memory needed for inference on graph G simpler for graphs with only ordinary parallel structures (such as residual components), as shown in equation 3:

$$M(G) = \max_{op \in G} \left[\sum_{A \in op_{\text{inp}}} |A| + \sum_{A \in op_{\text{out}}} |B| + |op| \right] \quad (3)$$

Taking the Bottleneck Residual block as an example, the operation $F(x)$ can be expressed as a combination of three operations, as shown in equation 4:

$$F(x) = [A \cdot N \cdot B]x \quad (4)$$

where $A: R^{s \times s \times k} \rightarrow R^{s \times s \times n}$ is the linear transformation, $N: R^{s \times s \times n} \rightarrow R^{s \times s \times n}$ the nonlinear transformation of each

channel, $B: R^{s \times s \times n} \rightarrow R^{s \times s \times k'}$ and is the linear transformation of the other output x is the input, s the stride, k the size of the convolution kernel, and n the quantity of channels.

For HI-MViT, $N = \text{Hardswish} \cdot \text{dwise} \cdot \text{Hardswish}$, the result is applied to each channel transform. Assuming that the input size is dominated by $|x|$ and the output size by $|y|$, then the memory required for $F(x)$ computation can be reduced to as shown in equation 5:

$$|s^2 k| + |s^2 2k'| + O(\max(s^2, s'^2)) \quad (5)$$

In Improved-MV2 we used Hardswish as the activation function, replacing rectified linear unit 6 (ReLU6)³⁴ in the original version. Due to the large number of channels in the middle two layers, using Hardswish will not cause too much information loss, and Hardswish has multiple advantages in the use of lightweight models. The majority of hardware and software architectures offer ReLU implementations that are optimized. Second, by operating in quantized mode, it prevents any potential loss of numerical accuracy brought on by various implementations of approximative sigmoid forms.³⁵ The Hardswish activation function can also be implemented as a divided function to

decrease the amount of memory allocation, which will significantly lower the latency cost and make it easier to deploy HI-MViT in the future on mobile devices. Equation 6 defines the Hardswish activation function, where x is the input variable.

$$\text{Hardswish}[x] = x \frac{\text{ReLU6}(x+3)}{6} \quad (6)$$

HI-MViT parameter

Table 2 displays the details of the configuration of the HI-MViT model's specific network parameters, where d stands for the input dimension of the transformer layer in the MobileViT block.

Evaluation metrics

Precision, Recall, Accuracy, F1-Score, AP, and AUC were chosen as the network assessment metrics to thoroughly assess the efficacy of the HI-MViT classification algorithm,³⁶ and the definitions are shown in equations (7) to (12):

$$\text{Precision} = \frac{TP}{TP + FP} \quad (7)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (8)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + PN} \quad (9)$$

$$\text{F1-Score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (10)$$

$$\text{AP} = \int_0^1 p(r) d(r) \quad (11)$$

$$\text{AUC} = \int_0^1 t(f) d(f) \quad (12)$$

Among them, TP , TN , FP , and FN are the amount of true positive, true negative, false positive, and false negative samples, respectively, t and f are the true positive rate and false positive rate, p is Precision, and r is Recall.

The percentage of the real positive class (TP) compared to the total positive classes ($TP + FP$) which are considered to be positive is known as Precision. The percentage of all true positive classes ($TP + FN$) that are considered to be positive classes is known as Recall (TP). The percentage of all accurate assessment information ($TP + TN$) to the whole is the accuracy score. While evaluating the correctness of its performance, the F1-Score assigns equal weight to the Precision and Recall scores. The Precision-Recall (PR) curve has Recall on the abscissa and Precision on the ordinate, and its area is denoted by the symbol AP. It can be regarded as a key indicator for

assessing the overall performance of the model because it also considers the model's Precision and Recall. The false positive rate (FPR) is on the abscissa and the true positive rate (TPR) is on the ordinate of the receiver operating characteristic (ROC) curve, which is referred to as AUC. It served as the benchmark for evaluation in the ISIC-2017 skin disease categorization challenge and carefully assessed the model's sensitivity and specificity.³⁷

Results

Datasets

This is a diagnostic study aimed at developing and validating the diagnostic accuracy of our proposed model. In this article, we conduct a series of experiments on the HI-MViT model on two skin lesion classification challenge datasets (i.e. ISIC-2017 and ISIC-2018 datasets) and a publicly available dermoscopic image dataset (i.e. PH² dataset). Among them, the PH² dataset is only used in the test, mainly for the evaluation of the generalization performance of the model. The 10,015 dermoscopic images for the HAM10000 dataset were collected over a 20-year period from 2 different sites: Dermatology at the Medical University of Vienna, Austria, and the Cliff Rosendahl Skin Cancer Clinic in Queensland, Australia. Inclusion criteria: (1) included dermoscopic images were of sufficient quality for analysis and diagnosis and (2) ensure that each image has appropriate clinical diagnostic labeling, as well as labeling of dermoscopic features (e.g. pigment grid, structure, blood vessels, etc.). Exclusion criteria: (1) exclusion of poor-quality images, which may include blurred, overexposed, or underexposed images and (2) unmarked or poorly marked images are excluded.

The ISIC series datasets are aggregated and published by the International Skin Imaging Collaboration. The ISIC-2017 classification dataset consists of 2750 dermoscopic images, which are divided into 3 different categories: melanoma, seborrheic keratosis, and nevus. The HAM10000 dataset, also known as the ISIC-2018 skin lesion classification dataset, has 10,015 dermoscopic images with a size of 600×450 pixels that span 7 main categories of skin illnesses. The PH² dataset is released by the Dermatology Service of Pedro Hispano Hospital, which contains 200 dermatoscopy images with a resolution of 768×560 pixels, consisting of 80 nevus, 80 atypical nevus, and 40 melanomas images. In the ISIC-2018 dataset, we divided the dataset into training set, verification set, and test set according to the ratio of 6:3:1. Then, data enhancement is performed on the divided training set, and the data processing process is strictly controlled to prevent data leakage. While on the ISIC-2017 dataset, we use the division ratio specified in the official challenge. Table 3 displays the dataset's precise distribution of data.

Table 3. The specific data distribution of ISIC-2017, ISIC-2018, and PH² classification datasets.

Dataset	Disease	Training	Validation	Test
ISIC-2017	Melanoma	374	30	117
	Seborrheic keratosis	254	42	90
	Nevus	1372	78	393
ISIC-2018	Actinic keratosis (AKIEC)	197	98	32
	Basal cell carcinoma (BCC)	309	154	51
	Benign keratosis (BKL)	660	330	109
	Dermatofibroma (DF)	70	34	11
	Melanoma (MEL)	668	334	111
	Melanocytic nevi (NV)	4024	2011	670
	Vascular skin lesion (VASC)	86	42	14
PH ²	Nevus	-	-	80
	Atypical nevus	-	-	80
	Melanoma	-	-	40

When preprocessing the dermoscopic images in the dataset, we first resize the images to 256×256 and use CenterCrop to return a center-cropped image.³⁸ Then a variety of data enhancement methods are used, including random rotation ($[-90^\circ, +90^\circ]$), random vertical flip, random horizontal flip, adding noise, adjusting contrast, and normalization to prevent excessive model training. As it is known, deep networks are data-hungry and a lot of augmentation methods have been applied with dermoscopy images to increase reliability and robustness.^{39–41} Therefore, an increased number of data has been used in this work to improve the performance of the proposed approach. Specifically, we use the min–max normalization method to adjust the pixel values of the image to the range $[0,1]$. Intensities in medical images are usually inhomogeneous and significantly affect the performance of automated image analysis techniques. Although various normalization algorithms with different image types have been applied to obtain high performance in the literature,^{42,43} they can cause increased computational costs. At the same time, we also randomly add Gaussian noise and Salt noise to the image to achieve data enhancement. Although generally Gaussian or speckle type of noise occurs in dermoscopy images, they may include different types of noise that can

be caused by different reasons such as imaging techniques or environmental factors.⁴⁴

Figure 6 shows schematic diagrams of images of different categories in the ISIC-2017, ISIC-2018, and PH² classification datasets.

Implementation details

The deep learning framework PyTorch 1.12.1 was employed to implement each experiment in this article. The hardware device used in the experiment is a computer equipped with four GeForce RTX 2080 Ti GPUs and 64GB of memory, and the operating system is Ubuntu 20.04.3. After a series of detailed comparative experiments, this article selects the AdamW⁴⁵ adaptive learning rate optimization algorithm as the optimizer and the loss function selects the cross entropy function. The weight decay is adjusted to $1E-2$, the batch size is 16, and the epoch is 50. The starting learning rate has been set to 0.0002. On the ImageNet 21 K dataset,²⁷ the network is pretrained in this study, and the pretraining variables are obtained. After loading it into the HI-MViT model, transfer learning is performed to ensure the best results for skin disease image classification. Specifically, first, migrate the pre-trained parameters obtained on ImageNet, modify the number of categories of the classification head, and then use the dermoscopic image to fine-tune some parameters of the model until convergence.

To fairly compare the performance of different methods and reduce the impact of chance and error, we conducted a five-fold cross-validation on ISIC-2017 and ISIC-2018, and showed the average performance and standard deviation.

Comparative experiment

We conducted comparative experiments on the ISIC-2018 dataset to explore the hyperparameter selection in model design and verify the effectiveness of Improved-MV2.

Hyperparameter optimization. We designed comparative experiments to select training hyperparameters, including optimizers and loss functions, for the HI-MViT skin lesion image classification model. In the comparison experiment, the optimizer selected the current mainstream and cutting-edge adaptive learning rate optimization algorithms: NAdam, AdaMax,⁴⁶ and AdamW, and the loss function selected FocalLoss,⁴⁷ PolyLoss,⁴⁸ and Cross Entropy Loss function. The classification results of the HI-MViT structure employing various optimizers and loss functions are shown in Table 4.

Table 4 shows that the HI-MViT model performs best when AdamW is employed as the optimizer and Cross Entropy Loss is utilized as the loss function. The scores of Precision, Recall, F1-Score, Accuracy, AP, and AUC are respectively 0.931, 0.932, 0.931, 0.932, 0.961, and

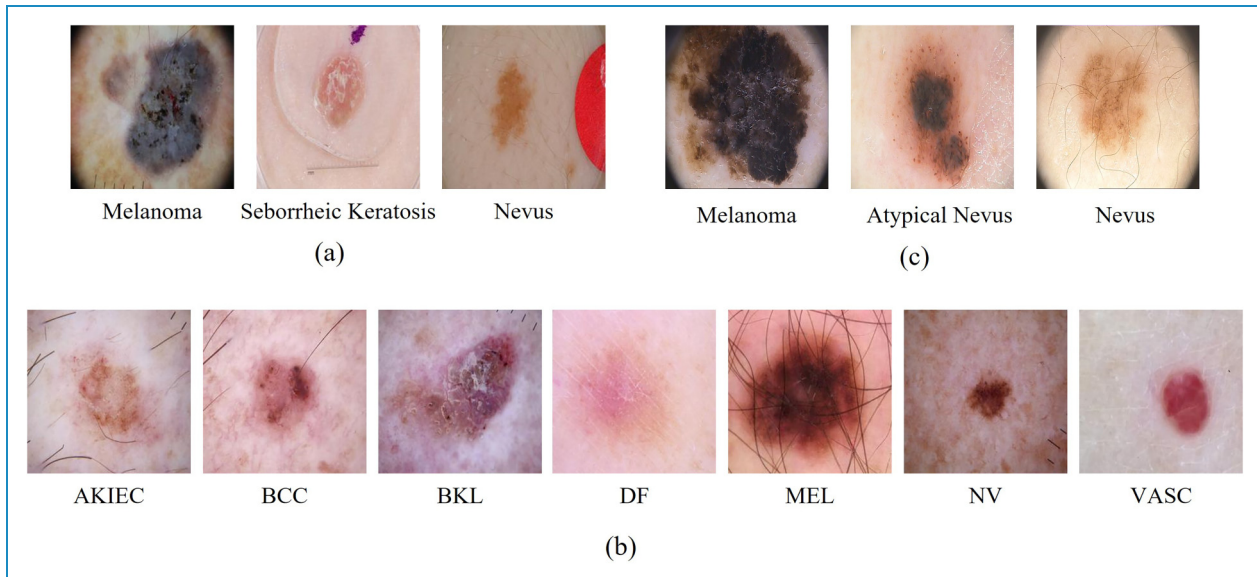


Figure 6. Schematic diagram of images of different categories in (a) ISIC-2017, (b) ISIC-2018, and (c) PH² classification datasets.

Table 4. Classification performance of the HI-MViT model using different optimizers and loss functions.

Optimizer	Loss function	Precision	Recall	F1-Score	Accuracy	AP	AUC
NAdam	FocalLoss	0.643 ± 0.001	0.704 ± 0.007	0.634 ± 0.009	0.704 ± 0.007	0.728 ± 0.007	0.870 ± 0.013
	PolyLoss	0.773 ± 0.007	0.796 ± 0.011	0.768 ± 0.001	0.796 ± 0.011	0.845 ± 0.012	0.932 ± 0.006
	Cross entropy loss	0.754 ± 0.012	0.779 ± 0.013	0.744 ± 0.010	0.779 ± 0.013	0.823 ± 0.011	0.915 ± 0.003
AdaMax	FocalLoss	0.566 ± 0.011	0.699 ± 0.010	0.624 ± 0.003	0.699 ± 0.010	0.724 ± 0.011	0.879 ± 0.006
	PolyLoss	0.793 ± 0.012	0.812 ± 0.007	0.797 ± 0.006	0.812 ± 0.007	0.870 ± 0.004	0.939 ± 0.009
	Cross entropy loss	0.776 ± 0.005	0.797 ± 0.006	0.774 ± 0.010	0.797 ± 0.006	0.856 ± 0.002	0.927 ± 0.008
AdamW	FocalLoss	0.873 ± 0.001	0.877 ± 0.012	0.872 ± 0.002	0.877 ± 0.012	0.930 ± 0.007	0.968 ± 0.003
	PolyLoss	0.868 ± 0.002	0.878 ± 0.005	0.868 ± 0.003	0.878 ± 0.005	0.920 ± 0.010	0.965 ± 0.002
	Cross entropy loss	0.931 ± 0.003	0.932 ± 0.002	0.931 ± 0.005	0.932 ± 0.002	0.961 ± 0.004	0.977 ± 0.001

AUC: area under the curve.

0.977, respectively. Compared with the AdamW + FocalLoss combination and AdamW + PolyLoss combination with the second-best performance, the indicators have increased by 5.8%, 5.4%, 5.9%, 5.4%, 3.1%, and 0.9%, respectively.

To more intuitively understand the changes in indicators during training and verification, Figure 7 depicts the accuracy and loss values' tendency as the number of model training and verification steps grows for various optimizer and loss function configurations. Figure 7 shows that the optimal performance combination AdamW + Cross Entropy Loss can learn the features contained in the

images in the dataset in fewer epochs during training and verification. And it converges quickly under the premise of ensuring stability.

Improved-MV2 validity verification. We performed comparison studies on the Improved-MV2 block in the HI-MViT architecture suggested in this article to confirm its efficacy. We set the activation functions in the original MV2 block to ReLU6, Gaussian error linear unit (GELU),⁴⁹ exponential linear unit (ELU),⁵⁰ leaky rectified linear unit (Leaky ReLU),⁵¹ parametric rectified linear unit (PReLU)⁵² respectively. The pooling layer of

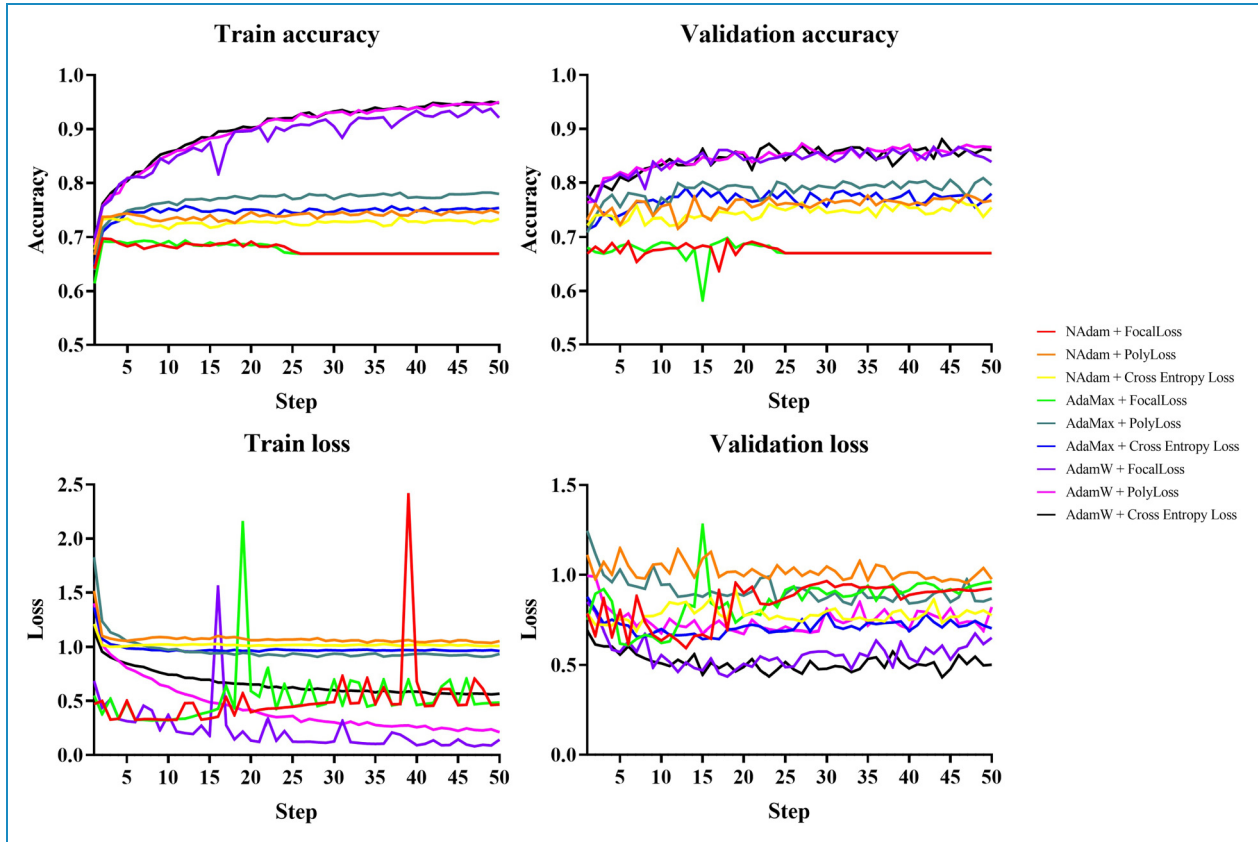


Figure 7. Accuracy and loss value changes of different optimizers and loss function combinations during training and verification.

the model is divided into two methods of global average pooling⁵³ and global maximum pooling⁵⁴ for discussion and experimentation. The comparative experimental findings on the ISIC-2018 classification dataset are displayed in Table 5.

Table 5's results indicate that HI-MViT performs highest in the experiment when the pooling layer is adjusted to Global Average Pooling, with grades of 0.931 for F1-Score, 0.932 for Accuracy, 0.961 for AP, and 0.977 for AUC. Compared with the original MobileViT model (using ReLU6), the improvements are 5.1%, 5%, 7.4%, and 2.5%, respectively. Compared with the suboptimal model in the performance of each indicator, it has increased by 4.9%, 4.5%, 3.1%, and 0.8%, respectively. When HI-MViT uses global average pooling, F1-Score, Accuracy, AP, and AUC are increased by 1.1%, 1.2%, 1.3%, and 0.7%, respectively, compared with the case of using global maximum pooling. It can be known from the above conclusions that the innovative design of the Improved-MV2 block is very effective in the HI-MViT model.

Classification results on the ISIC-2018 dataset

Compared with mainstream algorithms. The comparative evaluation between this model and the currently mainstream

methods on the ISIC-2018 classification dataset is shown in Table 6, including ResNeXt,⁵⁵ ShuffleNet,⁵⁶ MnasNet,⁵⁷ MobileNet,³¹ MobileOne,⁵⁸ ConvNeXt,⁵⁹ Vision Transformer,¹¹ Swin Transformer,⁶⁰ MetaFormer,⁶¹ EfficientFormer,⁶² MaxViT,⁶³ and FasterViT.⁶⁴

As can be seen from the results in Table 6, compared with the mainstream classification models based on CNN or Transformer selected in the experiment, our scheme performs better in various evaluation indicators. Compared with ConvNeXt, the most competitive CNN architecture, our model is 5.0%, 3.4%, 2.3% and 2.2% higher in F1-Score, Accuracy, AP and AUC, respectively. Compared with FasterViT, which has the highest AUC score in the Transformer architecture, the F1-Score, Accuracy, and AP of HI-MViT have increased by 5.2%, 5.0%, and 3.4%, respectively. The experimental results fully demonstrate that the method proposed in this article has excellent performance in the comprehensive performance of skin lesion classification.

Compared with the top algorithms in the ISIC-2018 classification challenge. The top five algorithms on ISIC-2018 Task 3 are contrasted with the results of the HI-MViT model provided in this article in Table 7. These algorithms do not use additional data. The deep learning strategies adopted by the top

Table 5. Validation of the Improved-MV2 block on the ISIC-2018 classification dataset.

Pooling	Methods	Precision	Recall	F1-Score	Accuracy	AP	AUC
Global average pooling	ReLU6	0.889 ± 0.008	0.882 ± 0.011	0.88 ± 0.006	0.882 ± 0.011	0.887 ± 0.015	0.952 ± 0.008
	GELU	0.884 ± 0.012	0.883 ± 0.008	0.877 ± 0.012	0.883 ± 0.008	0.905 ± 0.013	0.946 ± 0.014
	ELU	0.882 ± 0.009	0.887 ± 0.014	0.882 ± 0.010	0.887 ± 0.014	0.93 ± 0.010	0.969 ± 0.004
	Leaky ReLU	0.874 ± 0.009	0.876 ± 0.003	0.87 ± 0.003	0.876 ± 0.003	0.914 ± 0.006	0.961 ± 0.011
	PReLU	0.865 ± 0.001	0.865 ± 0.005	0.858 ± 0.013	0.865 ± 0.005	0.906 ± 0.006	0.951 ± 0.010
	HI-MViT(Ours)	0.931 ± 0.003	0.932 ± 0.002	0.931 ± 0.005	0.932 ± 0.002	0.961 ± 0.004	0.977 ± 0.001
Global max pooling	ReLU6	0.875 ± 0.013	0.876 ± 0.013	0.87 ± 0.011	0.876 ± 0.013	0.914 ± 0.002	0.955 ± 0.009
	GELU	0.876 ± 0.015	0.878 ± 0.003	0.874 ± 0.003	0.878 ± 0.003	0.929 ± 0.002	0.968 ± 0.008
	ELU	0.873 ± 0.014	0.876 ± 0.012	0.872 ± 0.007	0.876 ± 0.012	0.917 ± 0.009	0.959 ± 0.003
	Leaky ReLU	0.874 ± 0.008	0.878 ± 0.006	0.873 ± 0.012	0.878 ± 0.006	0.923 ± 0.008	0.958 ± 0.004
	PReLU	0.870 ± 0.016	0.874 ± 0.013	0.871 ± 0.004	0.874 ± 0.013	0.91 ± 0.006	0.958 ± 0.013
	HI-MViT(Ours)	0.922 ± 0.003	0.920 ± 0.005	0.920 ± 0.002	0.920 ± 0.005	0.948 ± 0.004	0.970 ± 0.006

AUC: area under the curve; ELU: exponential linear unit; GELU: Gaussian error linear unit; PReLU: parametric rectified linear unit.

five teams are the deep neural network (DCNN) integrated model⁶⁵ (Team-1), CNN-based model⁶⁶ (Team-2), Xception + DenseNet121 hybrid model⁶⁷ (Team-3), Inceptionv4 + ResNet-152 + DenseNet-161 Hybrid model⁶⁸ (Team-4), and improved DCNN model⁶⁹ (Team-5).

Table 7 shows that, except for BCC and NV, the HI-MViT framework yields the best results for all evaluation markers across the five categories. The AUC score on BCC is only 0.1% lower than the best performance, which is 0.991, but it improves the best performance obtained by the model of Team-1 by 2.3% and 1.6% on F1-Score and AP. Although only 0.3% below the best performance in NV's AP score, 0.978, HI-MViT improves the best performance obtained by the model of Team-1 F1-Score and AP by 2.1% and 0.3%. Taking AKIEC as an example, compared with the top five algorithms, HI-MViT performed suboptimally in terms of various indicators. F1-Score, AP, and AUC increased by 18.9%, 16.6%, and 0.9%, respectively, achieving 0.939, 0.993, and 1 good results. Macro-average can more comprehensively measure the classification performance of the model in each category. The scores of F1-Score, AP, and AUC of HI-MViT are 0.899, 0.960, and 0.986, respectively. Compared with the models of Team-1 and Team-2 with suboptimal performance, they have improved by 6.9%, 6.8%, and 0.8%.

Figure 8 shows the grouped histogram of the evaluation indicators of the HI-MViT model on the ISIC-2018 test set.

The model performs optimally overall for classifying images of different skin diseases, especially the AUC, which is the gold standard for evaluation, and the scores of each category are close to full marks.

To understand the classification effect of HI-MViT for different categories in the ISIC-2018 dataset more clearly and intuitively, Figure 9 shows the confusion matrix of HI-MViT on the test set. By observing Figure 9, it can be found that there are relatively more misjudgments in the NV and MEL categories, which may be caused by the high similarity images between the NV and MEL categories.

The PR and ROC curves of HI-MViT on the test set are displayed in Figure 10. AP is a representation of the region beneath the PR curve. The PR curve implies improved model performance when it is more inclined to the upper right. The square value under the ROC curve, or AUC, is a measure of how well the algorithm generalizes. The nearer the ROC curve is to the (0,1) point, the greater. It can be observed that in the PR curve, the AP values of AKIEC, DF, and VASC are the highest, which are 0.993, 0.992, and 1, respectively. In the ROC curve, the curves of AKIEC, BCC, DF, and VASC are all very close to the (0,1) point, and the AUC values are 1, 0.991, 1, and 1, respectively.

Compared with the latest skin lesion classification models. To better represent the good performance of HI-MViT in skin

Table 6. Performance comparison between the method in this article and the mainstream models on the ISIC-2018 dataset.

Methods	Networks	Precision	Recall	F1-Score	Accuracy	AP	AUC	Params
CNN	ResNeXt	0.837 ± 0.007	0.844 ± 0.006	0.839 ± 0.013	0.844 ± 0.006	0.896 ± 0.003	0.959 ± 0.007	23.0M
	ShuffleNet	0.869 ± 0.006	0.869 ± 0.002	0.867 ± 0.011	0.869 ± 0.002	0.925 ± 0.004	0.969 ± 0.003	1.3M
	MnasNet	0.862 ± 0.002	0.864 ± 0.015	0.860 ± 0.011	0.864 ± 0.015	0.932 ± 0.013	0.974 ± 0.002	3.1M
	MobileNet	0.864 ± 0.006	0.867 ± 0.013	0.865 ± 0.011	0.867 ± 0.013	0.922 ± 0.013	0.967 ± 0.001	5.0M
	MobileOne	0.857 ± 0.011	0.881 ± 0.002	0.868 ± 0.009	0.881 ± 0.002	0.925 ± 0.003	0.952 ± 0.003	6.5M
	ConvNeXt	0.868 ± 0.004	0.898 ± 0.007	0.881 ± 0.007	0.898 ± 0.007	0.938 ± 0.008	0.955 ± 0.005	27.8M
Transformer	Vision Transformer	0.679 ± 0.001	0.742 ± 0.005	0.696 ± 0.013	0.742 ± 0.005	0.795 ± 0.004	0.907 ± 0.002	86.2M
	Swin Transformer	0.823 ± 0.002	0.861 ± 0.014	0.839 ± 0.003	0.861 ± 0.014	0.904 ± 0.006	0.937 ± 0.008	27.5M
	MetaFormer	0.855 ± 0.005	0.898 ± 0.007	0.874 ± 0.013	0.898 ± 0.007	0.931 ± 0.015	0.955 ± 0.003	11.4M
	EfficientFormer	0.857 ± 0.004	0.881 ± 0.008	0.868 ± 0.006	0.881 ± 0.008	0.925 ± 0.002	0.952 ± 0.003	11.3M
	MaxViT	0.820 ± 0.001	0.833 ± 0.014	0.823 ± 0.010	0.833 ± 0.014	0.884 ± 0.005	0.953 ± 0.011	28.5M
	FasterViT	0.878 ± 0.015	0.882 ± 0.005	0.879 ± 0.014	0.882 ± 0.005	0.927 ± 0.011	0.969 ± 0.005	31.4M
Ours	HI-MViT	0.931 ± 0.003	0.932 ± 0.002	0.931 ± 0.005	0.932 ± 0.002	0.961 ± 0.004	0.977 ± 0.001	4.9M

AUC: area under the curve; CNN: convolutional neural networks.

lesion classification, we selected six latest networks^{70–75} specifically designed for skin lesion classification for comparison, and the results are shown in Table 8. As can be seen, Table 8 covers CNN-based, Transformer-based, and lightweight networks, and in comparison, our method has a better accuracy score on the ISIC-2018 dataset.

Visualization of the learned embeddings and CAMs. To enable the model to concentrate more on the skin lesion area and enhance the classification effect, HI-MViT employs the Transformer-based multihead self-attention method to acquire the global representation of the image. To verify this, we used two methods to visually compare the ConvNeXt and HI-MViT: one is to visualize the dimensionality reduction of semantic features in the test set, and the other is to visualize the Class Activation Map (CAM).

First, we use a linear dimensionality reduction method (PCA⁷⁶) and two nonlinear dimensionality reduction methods (t-SNE⁷⁷ and UMAP⁷⁸) to visualize the features learned by a specific layer of the model. Figure 11 shows the results of ConvNeXt and HI-MViT using PCA, t-SNE, and UMAP on the ISIC-2018 test set for dimensionality reduction visualization, which can reflect the ability of the model to distinguish different types of samples. It can be seen from Figure 10 that among the three dimensionality

reduction visualization algorithms, the visualization results of HI-MViT are better than ConvNeXt, with smaller intraclass distances, larger interclass distances, clear cluster boundaries, and fewer mixed samples. This reflects that HI-MViT has a stronger ability to discriminate images of skin lesions.

CAM visualizes which pixels of an image a neural network pays attention to when predicting a class. In this article, six different CAM methods are used to generate class activation heatmaps for skin disease images, including GradCAM,⁷⁹ Guided Grad-CAM, GradCAM++,⁸⁰ AblationCAM,⁸¹ ScoreCAM,⁸² and EigenCAM.⁸³ Among them, Guided Grad-CAM can generate high-resolution fine-grained thermal maps. Figure 12 shows the class activation heatmap obtained by HI-MViT on different categories of skin disease images in the ISIC-2018 dataset. It can be observed that there are differences in the specific subregions and concentrations that different CAM algorithms focus on HI-MViT, but they all focus on the attention region representing the location of the skin lesion when performing classification, rather than the surrounding normal skin tissue and hair tissue.

Visualization of effective receptive fields and attention maps. Different from traditional CNN’s process of gradually expanding the receptive field by using larger convolution

Table 7. Performance comparison between the method in this article and the top five algorithms of the ISIC-2018 classification challenge.

Methods	AKIEC			BCC			BKL			DF		
	F1-Score	AP	AUC	F1-Score	AP	AUC	F1-Score	AP	AUC	F1-Score	AP	AUC
Team-1 (Zhuang et al. ⁶⁵)	0.736	0.819	0.988	0.862	0.933	0.992	0.833	0.899	0.969	0.835	0.871	0.982
Team-2 (Li and Li ⁶⁶)	0.747	0.827	0.991	0.822	0.919	0.992	0.852	0.927	0.977	0.800	0.849	0.980
Team-3 (Amro et al. ⁶⁷)	0.750	0.807	0.979	0.769	0.845	0.977	0.773	0.818	0.937	0.736	0.818	0.976
Team-4 (Bissoto et al. ⁶⁸)	0.651	0.741	0.983	0.793	0.901	0.991	0.763	0.853	0.958	0.840	0.881	0.987
Team-5 (Pan and Xia ⁶⁹)	0.688	0.484	0.866	0.802	0.655	0.906	0.772	0.629	0.868	0.824	0.685	0.896
HI-MViT(Ours)	0.939	0.993	1	0.885	0.948	0.991	0.885	0.940	0.977	0.909	0.992	1
Methods	MEL			NV			VASC			Macro-average		
	F1-Score	AP	AUC	F1-Score	AP	AUC	F1-Score	AP	AUC	F1-Score	AP	AUC
Team-1 (Zhuang et al. ⁶⁵)	0.750	0.821	0.945	0.934	0.977	0.974	0.857	0.919	0.998	0.830	0.891	0.978
Team-2 (Li and Li ⁶⁶)	0.740	0.814	0.949	0.939	0.981	0.974	0.896	0.928	0.998	0.828	0.892	0.980
Team-3 (Amro et al. ⁶⁷)	0.525	0.638	0.883	0.872	0.979	0.969	0.824	0.929	0.998	0.750	0.833	0.960
Team-4 (Bissoto et al. ⁶⁸)	0.589	0.715	0.934	0.895	0.977	0.965	0.844	0.912	0.997	0.768	0.854	0.974
Team-5 (Pan and Xia ⁶⁹)	0.683	0.505	0.813	0.923	0.896	0.902	0.831	0.700	0.885	0.789	0.650	0.876
HI-MViT(Ours)	0.788	0.871	0.960	0.965	0.978	0.977	0.923	1	1	0.899	0.960	0.986

AUC: area under the curve; BCC: basal cell carcinoma; CNN: convolutional neural networks; DF: dermatofibroma; MEL: melanoma.

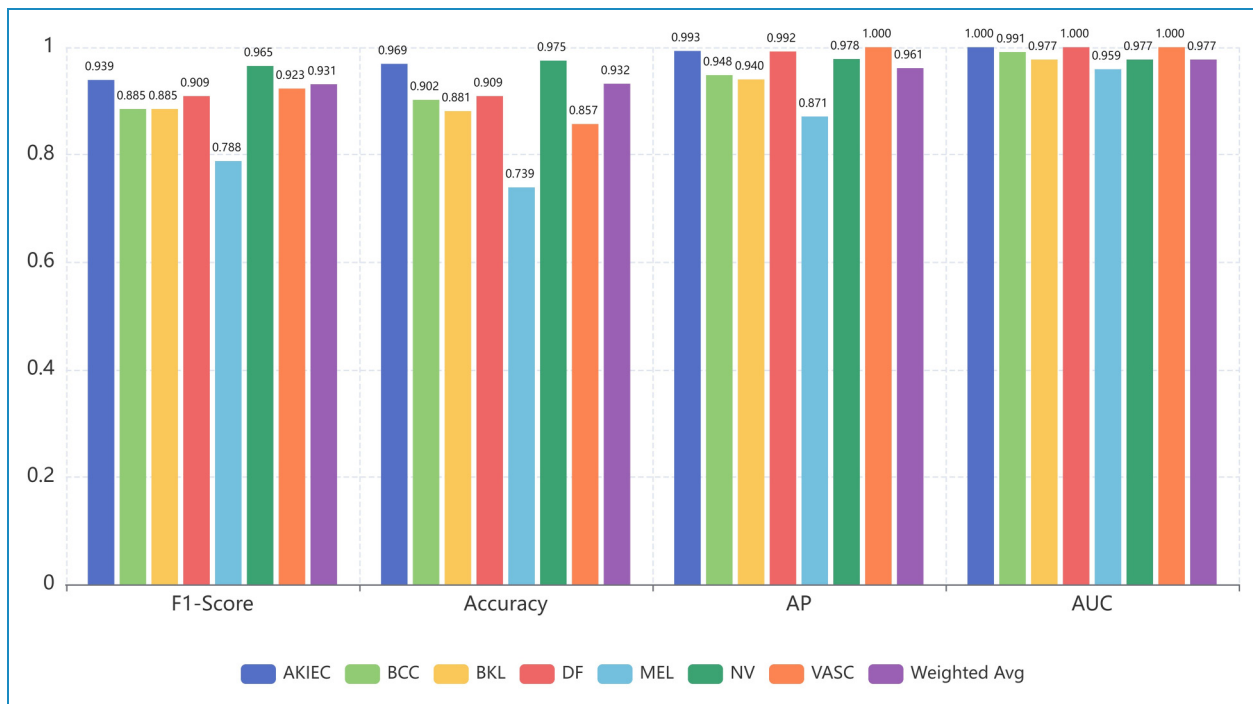


Figure 8. Grouped histogram of evaluation indicators of HI-MViT model on ISIC-2018 test set.

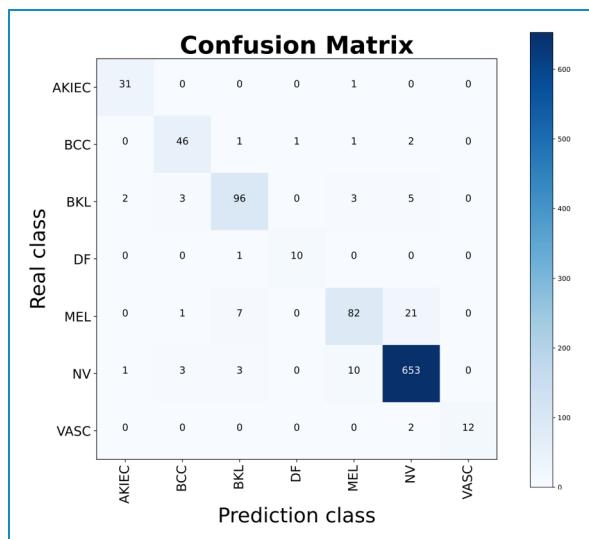


Figure 9. Confusion matrix of HI-MViT on the ISIC-2018 test set.

kernels and deeper convolutions, the Transformer-based HI-MViT model we propose can obtain global representation earlier by modeling long-distance dependencies. We visualize the average effective receptive field at different stages of the ConvNeXt and HI-MViT models,⁸⁴ and the results are shown in Figure 13. It can be observed that compared with the convolutional neural network, the global information interaction capability of HI-MViT can help the feature extractor to quickly establish a global receptive field, thereby achieving faster context understanding.

To better understand the change of attention of HI-MViT in the process of classifying skin lesion images, we visualized the attention maps of different stages⁸⁵ of the model on ISIC-2018, as shown in Figure 14. It can be observed that HI-MViT can model the global information through the self-attention mechanism, thereby improving the ability to identify skin lesions from the global receptive field. At the same time, with the deepening of the module, the attention will gradually focus on the skin lesion area, so the lesion type of the skin lesion area can be more accurately identified.

Classification results on the ISIC-2017 dataset

Compared with mainstream algorithms. Table 9 shows the performance comparison between HI-MViT and mainstream classification models for skin lesion classification on the ISIC-2017 dataset. It can be seen that HI-MViT performs better in all indicators than other comparison models in Table 9. Compared with ConvNeXt, the master of the convolutional neural network, our method improves F1-Score, Accuracy, AP, and AUC by 7.6%, 7.4%, 8.8%, and 5.0%, respectively. In the Transformer architecture, EfficientFormer, which is also a lightweight network, is considered to be the most competitive method. Compared with its scores in F1-Score and AUC, HI-MViT increased by 4.3% and 1.9%, respectively, which reflects the strong performance of HI-MViT in skin lesion classification.

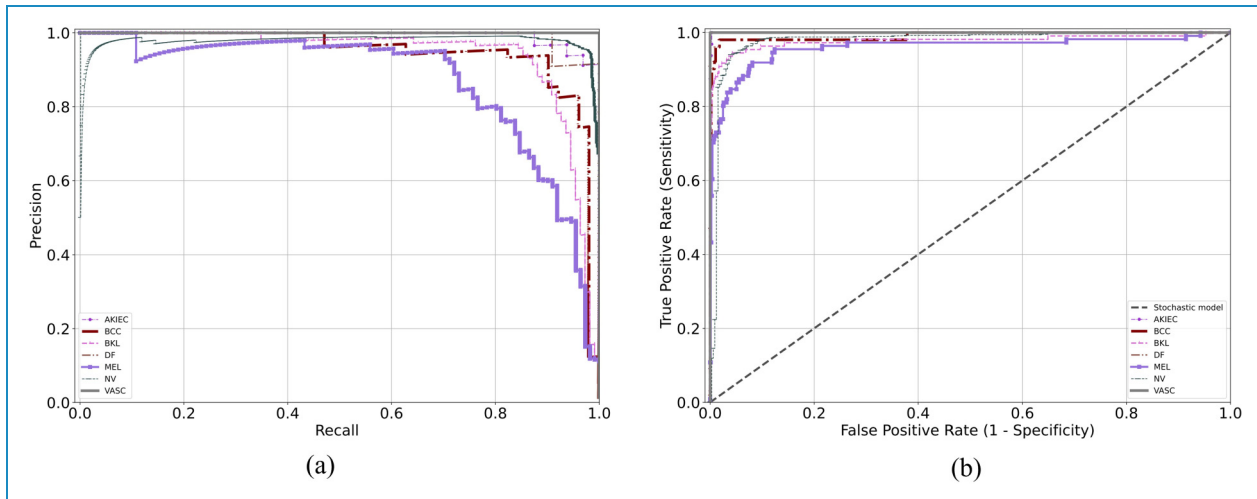


Figure 10. (a) Precision-recall curve and (b) ROC curve of HI-MViT on the ISIC-2018 test set. ROC: receiver operating characteristic.

Compared with the top algorithms in the ISIC-2017 classification challenge. The effectiveness of the technique used in this article is contrasted with the top 5 algorithms used in ISIC-2017 Task 3 in Table 10. The deep learning strategies adopted by the top five teams are the ResNet ensemble model⁸⁶ (Team-1), CNN-based model⁸⁷ (Team-2), transfer learning-based model⁸⁸ (Team-3), deep residual network model⁸⁹ (Team-4), and multitask deep learning model⁹⁰ (Team-5). The strategies outlined in this study have produced the best outcomes across all categories of evaluation metrics.

Figure 15 is a grouped histogram of the evaluation indicators of the HI-MViT model on the ISIC-2017 test set. The model's performance on every indication of each class can be observed to be fairly balanced, and its total performance is quite outstanding. At the same time, the weighted average AUC score as the ISIC-2017 challenge evaluation standard is 0.962. The Accuracy, AP, and AUC scores on the seborrheic keratosis category achieved good results of 0.933, 0.968, and 0.988, respectively.

The confusion matrix for the algorithm on the ISIC-2017 test set is shown in Figure 16, and the deviation between the real category and the predicted category can be seen. It can be seen that a part of nevus is misjudged as melanoma because the division of nevus species does not exist in the original dataset. To fully exploit the training potential of the dataset, we generally divide all the images not marked as melanoma and seborrheic keratosis into nevus for classification. Certain nevus images may be incorrectly assessed due to the similarities between classes and the variations within classes.

Figure 17 is the PR and ROC curves of the network in this article on the ISIC-2017 test set. It can be seen that in the PR curve, the overall trend of all categories is close to

(1,1), which shows that the model is well-balanced between the two indicators of Precision and Recall. In the ROC curve, the inflection points of the three classifications are very close to (0,1), and the outstanding overall effectiveness of the model may be assessed based on the area under the curve, which is very near 1.

Compared with the latest skin lesion classification models.

Table 11 shows the accuracy scores of six networks specifically designed for skin lesion classification^{13,91–95} compared to our method on the ISIC-2017 dataset. Table 11 contains a variety of methods based on EfficientNet B3, CNN, Transformer, metaheuristic optimization algorithms, and noise label correction methods. It can be seen that our lightweight classification model HI-MViT obtains better accuracy performance.

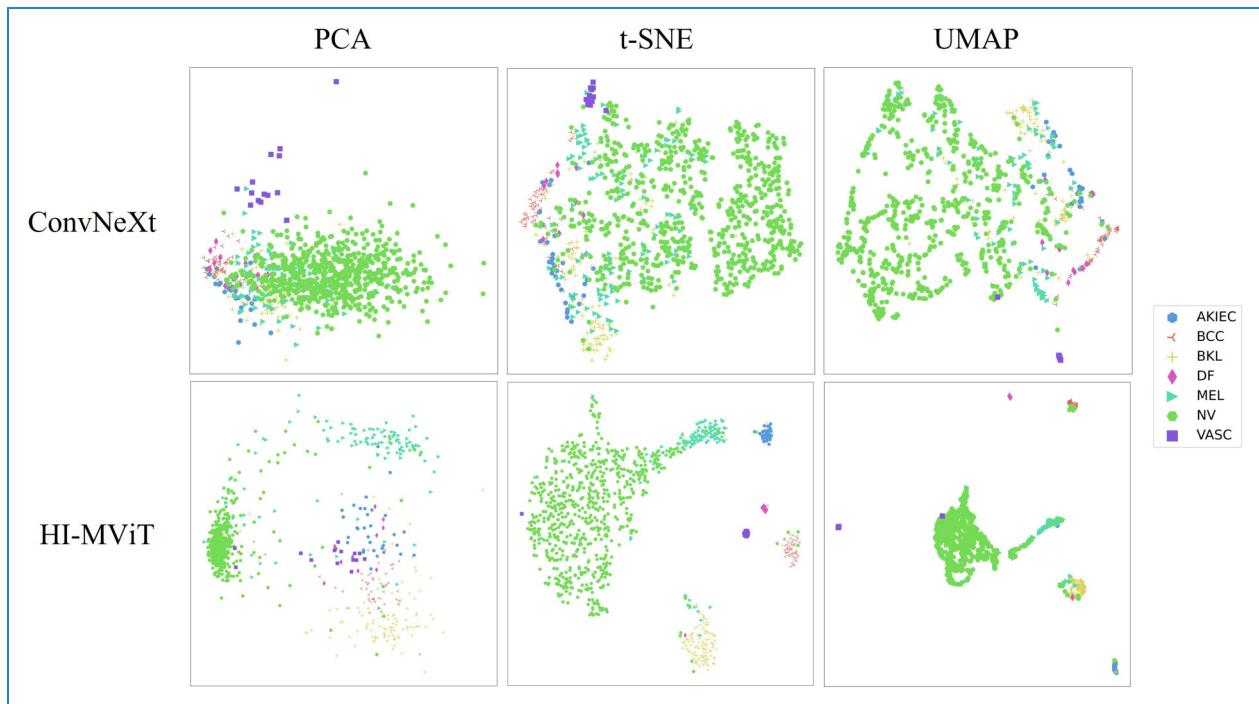
Visualization of the learned embeddings and CAMs. Figure 18 compares the results of semantic feature extraction and dimension reduction visualization of HI-MViT and ConvNeXt on the ISIC-2017 test set. It can be seen in the three dimensionality reduction visualization algorithms that there is the existence of the mixing of data points of different categories in ConvNeXt. The intercluster distribution distance of HI-MViT is far, the intracluster distribution is similar, the cluster boundary is clear, and the data is less mixed.

The CAM visualization results shown in Figure 19 allow us to more intuitively understand which pixels HI-MViT pays more attention to as the basis for classification when classifying the ISIC-2017 dataset. It can be seen that HI-MViT mainly focuses on the skin lesion area in the image, and does not pay too much attention to the surrounding skin tissue, which is also a prerequisite for ensuring accurate classification results.

Table 8. Performance comparison between the method in this article and the latest skin lesion classification models on the ISIC-2018 dataset.

References	Methods	Accuracy
Alam et al. ⁷⁰ (2022)	A feature extraction block based on CNN to construct stepwise feature tracking is used for the classification of skin disease images	0.906
Anand et al. ⁷¹ (2022)	A deep learning-based model was used to identify benign and malignant stages of skin cancer using the concepts of transfer learning methods	0.891
Popescu et al. ⁷² (2022)	A system based on the collective intelligence of nine CNNs is proposed for detecting and classifying skin lesions	0.867
Tada and Han ⁷³ (2023)	A hybrid skin lesion classification model is constructed by combining convolution operations and self-attention structures	0.911
Durães and Véstias ⁷⁴ (2023)	A low-cost intelligent embedded system for skin cancer classification based on cascade inference technology and Vitis-AI	0.870
Li et al. ⁷⁵ (2023)	An auxiliary diagnosis method for skin lesions based on enhanced MobileNet model	0.926
HI-MViT(Ours)	A Lightweight Model for Explainable Skin Disease Classification Based on Modified MobileViT	0.932

CNN: convolutional neural networks.

**Figure 11.** Dimensionality reduction visualization results of ConvNeXt and HI-MViT on the ISIC-2018 test set.

Validate generalization performance on PH² dataset

The generalization ability reflects the ability of the model to judge unknown data. A model with good generalization ability can make correct judgments when the data fluctuates. We used the model trained on ISIC-2017 to test on a

skin disease dataset PH² with an unknown distribution to verify the generalization effect of HI-MViT. Table 12 shows the test results of different mainstream classification models on the PH² dataset.

Compared with the most competitive ConvNeXt and EfficientFormer, our method improves Accuracy by 3.9%

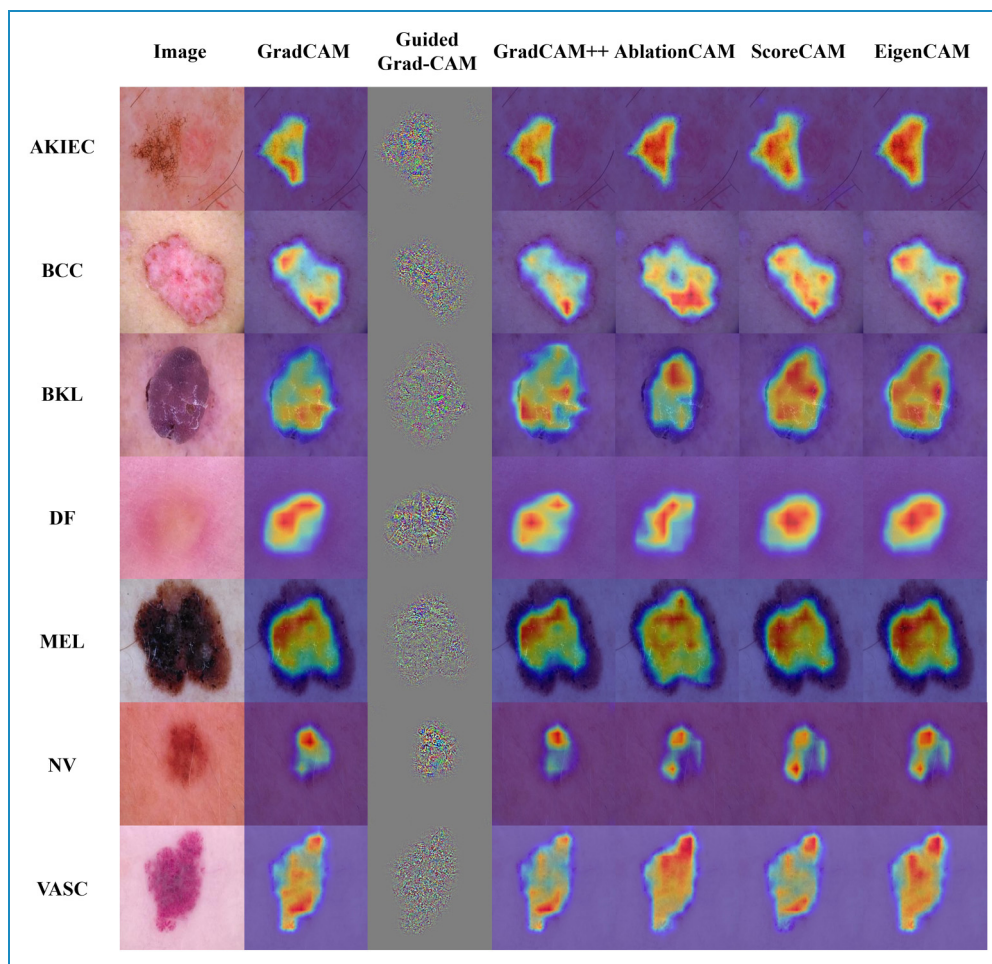


Figure 12. HI-MViT class activation heatmap on the ISIC-2018 dataset.

and 1.7%, and AUC by 3.5% and 2.4%, respectively. Therefore, according to the experimental results shown in Table 12, it can be shown that HI-MViT has good model generalization performance and robustness, and can still maintain good classification performance in the face of unknown data and parameters.

Discussions

Using dermoscopic images, our technique successfully achieves high classification accuracy of skin lesion classifications. The F1-Score, Accuracy, AP, and AUC scores for experiments utilizing the HI-MViT skin disease image classification algorithm were 0.931, 0.932, 0.961, and 0.977, respectively. Compared with the top five algorithms of ISIC-2018 Task 3, the scores of Marco's average F1-Score, AP, and AUC of HI-MViT are 0.899, 0.960, and 0.986, respectively. The improvements were 6.9%, 6.8%, and 0.8% when compared to the second-best-performing model, respectively. This demonstrates the effectiveness of the skin disease classification

methodology used in this study. Compared with FasterViT, which has the highest AUC score in the Transformer architecture, the F1-Score, Accuracy, and AP of HI-MViT have increased by 5.2%, 5.0%, and 3.4%, respectively. It can be seen that the HI-MViT findings have lower intraclass distances and higher interclass distances compared to ConvNeXt using the dimensionality reduction visualization method. This reflects that HI-MViT has a stronger ability to discriminate images of skin lesions. While performing classification findings, the model appears to focus more on the lesion site than on irrelevant skin tissue, according to explainable class-activation heatmaps. We also visualized the effective receptive field and attention map at different stages of HI-MViT, to more intuitively recognize the ability of HI-MViT to model local and global information. The performance experiment on the ISIC-2017 dataset also achieved excellent results, and the results of 0.898, 0.895, 0.947, and 0.962 were obtained on F1-Score, Accuracy, AP, and AUC, respectively. All indicators are better than the top five algorithms of ISIC-2017 Task 3. Using the

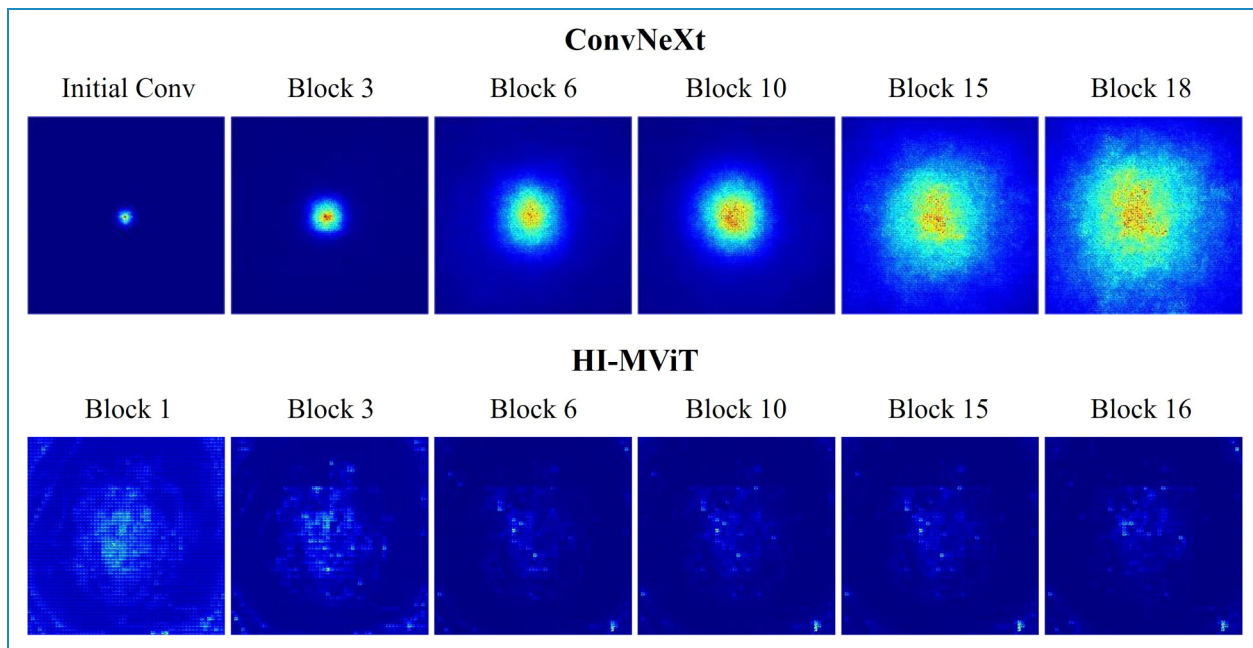


Figure 13. Effective receptive field visualization at different stages of ConvNeXt and HI-MViT.

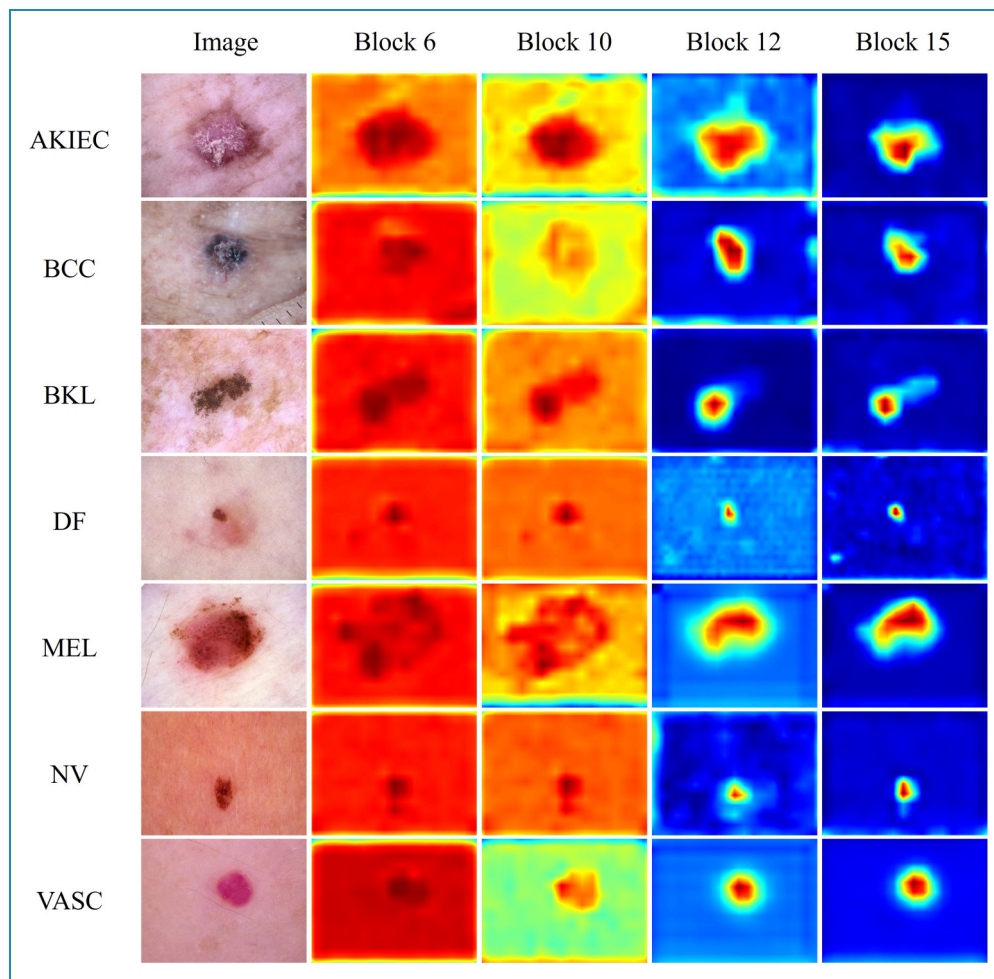


Figure 14. Visualization of HI-MViT's attention map at different stages on ISIC-2018.

Table 9. Performance comparison between the method in this article and the mainstream models on the ISIC-2017 dataset.

Methods	Networks	Precision	Recall	F1-Score	Accuracy	AP	AUC
CNN	ResNeXt	0.717 ± 0.002	0.728 ± 0.009	0.711 ± 0.003	0.728 ± 0.009	0.788 ± 0.012	0.842 ± 0.005
	ShuffleNet	0.763 ± 0.012	0.763 ± 0.011	0.759 ± 0.011	0.763 ± 0.011	0.814 ± 0.002	0.855 ± 0.013
	MnasNet	0.757 ± 0.007	0.772 ± 0.014	0.756 ± 0.007	0.772 ± 0.014	0.809 ± 0.007	0.853 ± 0.002
	MobileNet	0.752 ± 0.006	0.74 ± 0.003	0.745 ± 0.003	0.74 ± 0.003	0.801 ± 0.006	0.856 ± 0.010
	MobileOne	0.762 ± 0.006	0.79 ± 0.005	0.764 ± 0.012	0.79 ± 0.005	0.820 ± 0.010	0.878 ± 0.006
	ConvNeXt	0.828 ± 0.006	0.821 ± 0.006	0.822 ± 0.007	0.821 ± 0.006	0.859 ± 0.013	0.912 ± 0.013
Transformer	Vision Transformer	0.648 ± 0.010	0.688 ± 0.004	0.628 ± 0.014	0.688 ± 0.004	0.705 ± 0.010	0.756 ± 0.013
	Swin Transformer	0.781 ± 0.006	0.826 ± 0.014	0.770 ± 0.005	0.826 ± 0.014	0.814 ± 0.010	0.892 ± 0.001
	MetaFormer	0.780 ± 0.005	0.814 ± 0.005	0.785 ± 0.005	0.814 ± 0.005	0.866 ± 0.001	0.922 ± 0.010
	EfficientFormer	0.842 ± 0.003	0.892 ± 0.002	0.855 ± 0.012	0.892 ± 0.002	0.939 ± 0.004	0.943 ± 0.005
	MaxViT	0.811 ± 0.014	0.815 ± 0.008	0.805 ± 0.009	0.815 ± 0.008	0.858 ± 0.001	0.913 ± 0.011
	FasterViT	0.847 ± 0.011	0.844 ± 0.005	0.836 ± 0.009	0.844 ± 0.005	0.903 ± 0.014	0.949 ± 0.005
Ours	HI-MViT	0.908 ± 0.001	0.895 ± 0.003	0.898 ± 0.002	0.895 ± 0.003	0.947 ± 0.002	0.962 ± 0.004

AUC: area under the curve; CNN: convolutional neural networks.

Table 10. Performance comparison between the method in this article and the top five models of ISIC-2017 Task3.

Methods	Melanoma			Seborrheic keratosis			Macro-average		
	F1-Score	AP	AUC	F1-Score	AP	AUC	F1-Score	AP	AUC
Team-1 (Matsunaga et al. ⁸⁶)	0.625	0.711	0.868	0.599	0.790	0.953	0.612	0.750	0.911
Team-2 (Díaz ⁸⁷)	0.185	0.656	0.856	0.299	0.840	0.965	0.242	0.748	0.910
Team-3 (Menegola et al. ⁸⁸)	0.624	0.716	0.874	0.504	0.791	0.943	0.564	0.754	0.908
Team-4 (Bi et al. ⁸⁹)	0.541	0.695	0.870	0.684	0.771	0.921	0.612	0.733	0.896
Team-5 (Yang et al. ⁹⁰)	0.500	0.526	0.830	0.716	0.809	0.942	0.608	0.667	0.886
HI-MViT(Ours)	0.798	0.826	0.949	0.903	0.968	0.988	0.876	0.924	0.966

AUC: area under the curve.

trained model to test on the PH² dataset, compared with the most competitive ConvNeXt and EfficientFormer, HI-MViT has improved Accuracy by 3.9% and 1.7%, and AUC has increased by 3.5% and 2.4%. This shows that HI-MViT has good generalization enough to face changes in data and parameters. Dermatologists can use the

HI-MViT skin disease classification model's classification result as an evaluation index to aid in their diagnosis, which will enable them to more rapidly and accurately identify various dermoscopic images.

The International Skin Digital Imaging Society sponsors ISIC, a global organization devoted to skin cancer diagnosis

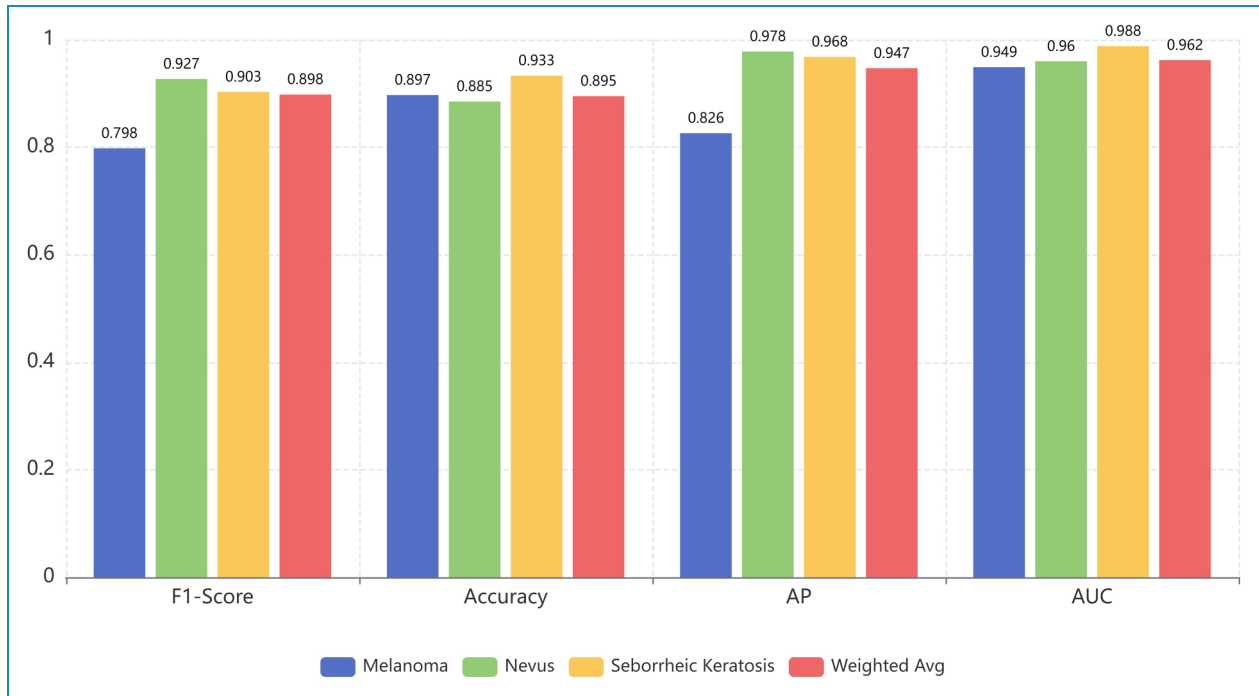


Figure 15. Grouped histogram of evaluation indicators of HI-MViT model on ISIC-2017 test set.

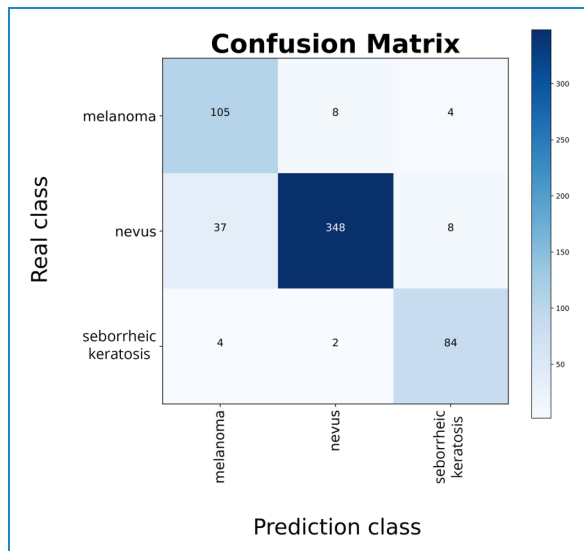


Figure 16. Confusion matrix of HI-MViT on the ISIC-2017 test set.

(ISDIS). It sponsors the ISIC-2017 and ISIC-2018 Challenge, which focuses on the study and diagnosis of skin lesions. Task 3 is skin disease classification.⁹⁶ For the categorization of skin diseases, convolutional neural network-based structures have been employed in the majority of the studies that have been published to date. Alwakid et al.⁹⁷ classified skin lesions in images by using a convolutional neural network system and a modified version of ResNet50. The HAM10000 dataset's seven skin cancer

cases with uneven samples were used in the analysis. The F1-Score is 0.86, the average accuracy is 0.86, the precision is 0.84, the recall rate is 0.86, and so on. Ali et al.⁹⁸ applied transfer learning to pretrained ImageNet weights and improved convolutional neural networks to train EfficientNets B0-B7 on the HAM10000 dataset. With an F1-Score of 87% and a Top-1 accuracy of 87.91%, EfficientNet B4 was the best model. Chaturvedi et al.⁹⁹ employed a MobileNet-based model that had been pre-trained on approximately 1.28 million images from the 2014 ImageNet Competition and then was refined via transfer learning on 10,015 dermoscopic images from the HAM10,000 dataset. The precision, recall, and F1-Score weighted average scores for the model employed in this study are 89%, 83%, and 83%, respectively, with an overall accuracy rate of 83.1% for the 7 classes in the dataset. The method in this article is superior to all the abovementioned models using the ISIC-2018 Task 3 dataset, with scores of 0.931, 0.932, 0.961, and 0.977 on F1-Score, Accuracy, AP, and AUC, respectively. Compared with the best-performing EfficientNet B4 in the above model, this article improves F1-Score and Accuracy by 6.1% and 5.29%, respectively.

Most significantly, this work makes use of the MobileViT-based explainable lightweight skin disease classification algorithm. Unlike earlier CNN-based techniques, HI-MViT combines the strengths of the Transformer and CNN: Transformer may offer spatial inductive bias, allowing it to eliminate positional bias. Also, the addition of CNN can hasten network convergence and improve the

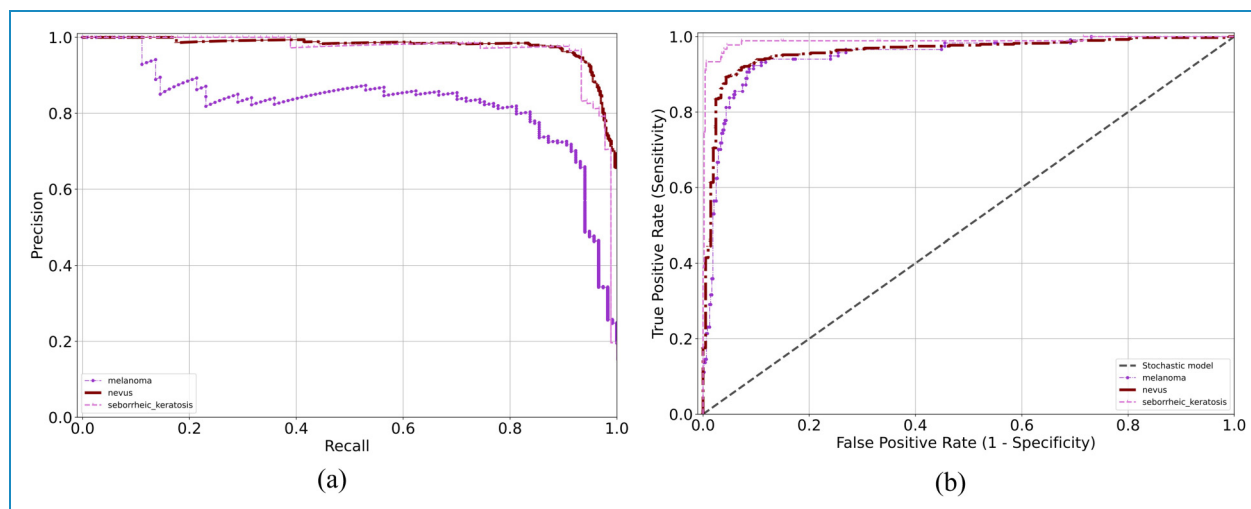


Figure 17. (a) Precision-recall curve and (b) ROC curve of HI-MViT on the ISIC-2017 test set. ROC: receiver operating characteristic.

Table 11. Performance comparison between the method in this article and the latest skin lesion classification models on the ISIC-2017 dataset.

References	Methods	Accuracy
Salian and Sawarkar ⁹¹ (2022)	Classification of malignant skin lesions using the concept of fine-tuned transfer learning based on the improved and fine-tuned EfficientNet B3 model	0.871
Wu et al. ⁹² (2022)	Skin lesion classification model based on deep convolutional neural network and transfer learning	0.867
Wang et al. ¹³ (2022)	It combines the advantages of CNN and transformer and fully utilizes global and local information to improve medical image segmentation and classification	0.872
Golnoori et al. ⁹³ (2023)	Improving the performance of skin lesion classification systems by optimizing the hyperparameters and architecture of deep neural networks using metaheuristic optimization algorithms	0.816
Kim et al. ⁹⁴ (2023)	A new paradigm based on deep learning is proposed, allowing the extraction of fine-grained differences between skin lesions on a pixel basis to achieve high-precision classification of skin lesions	0.878
Zhu et al. ⁹⁵ (2023)	A noise label correction method is proposed to deal with noisy labels in medical image datasets to improve classification performance	0.872
HI-MViT(Ours)	A Lightweight Model for Explainable Skin Disease Classification Based on Modified MobileViT	0.895

CNN: convolutional neural networks.

stability of the structure training process. The Transformer's self-attention mechanism can efficiently gather global information, and numerous heads can map it to various areas, strengthening the model's capacity for expression.¹⁰⁰

Table 4's hyperparameter comparison experiment findings show that each combination that chooses AdamW as

the optimizer performs much better than the other unselected combinations. This is because AdamW, as a variant of Adam, can automatically adjust the learning rate without requiring too many parameter adjustments, reducing redundancy. It can also automatically adjust the weight decay coefficient to make the model more stable and avoid overfitting. In terms of the loss function,

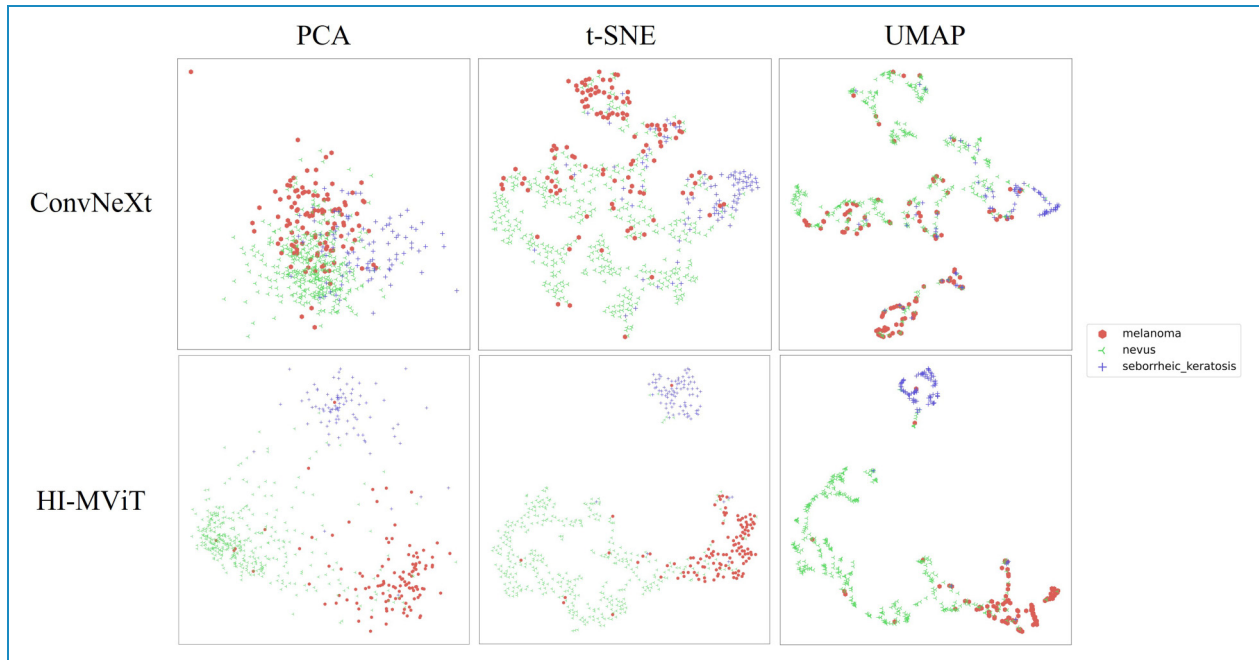


Figure 18. Dimensionality reduction visualization results of ConvNeXt and HI-MViT on the ISIC-2017 test set.

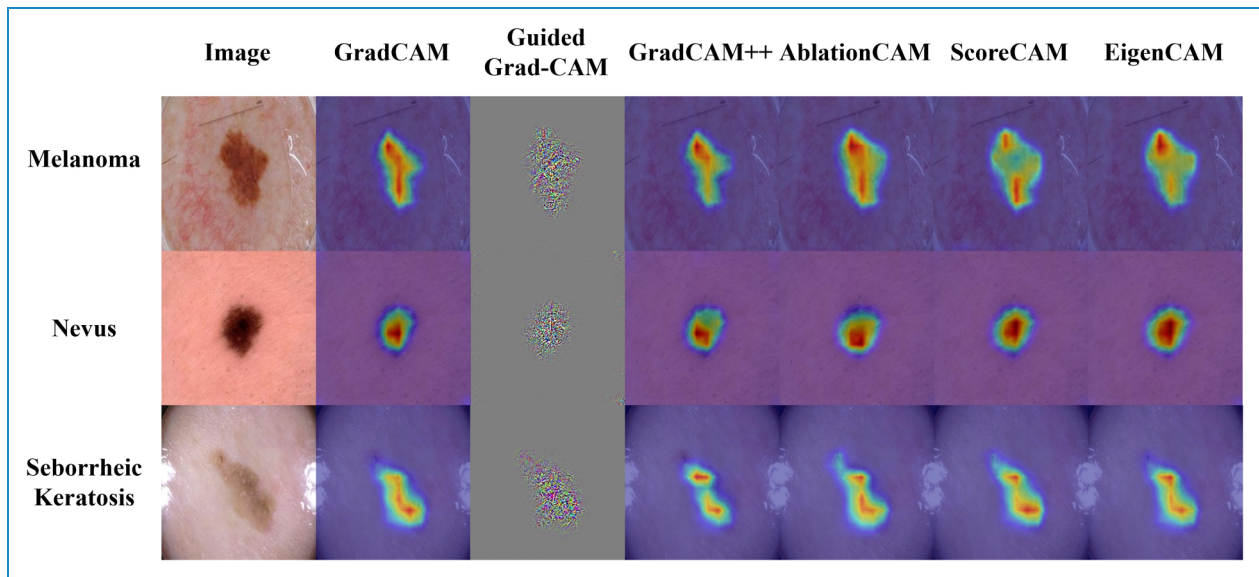


Figure 19. HI-MViT class activation heatmap on the ISIC-2017 dataset.

FocalLoss performs relatively poorly. The reason is that FocalLoss is very susceptible to noise interference, and since it values difficult samples more, wrong samples in the data may mislead the direction of model learning.¹⁰¹ However, finding the ideal collection of hyperparameters for a given dataset is challenging due to the large adjustment costs of its hyperparameters.

The efficacy verification findings for the Improved-MV2 block on the ISIC-2018 classification dataset are displayed in Table 5. Compared with global max pooling, global

average pooling performs better on the same model. This is because there is no need to optimize parameters and fixed input size in global average pooling, which can effectively avoid overfitting this layer. The spatial data is summarized using the global average pooling, which is more resistant to the spatial modification of the input.

After choosing the ISIC-2018 dataset's performance comparison of modern mainstream algorithms in Table 6, ViT performed poorly, even worse than some lightweight models whose parameters are much smaller than it. This

Table 12. The performance results of the trained model tested on the PH² dataset.

Methods	Networks	Precision	Recall	F1-Score	Accuracy	AP	AUC
CNN	ResNeXt	0.781	0.798	0.789	0.798	0.829	0.827
	ShuffleNet	0.749	0.759	0.747	0.759	0.834	0.876
	MnasNet	0.809	0.777	0.792	0.777	0.855	0.859
	MobileNet	0.822	0.902	0.857	0.902	0.896	0.883
	MobileOne	0.820	0.885	0.846	0.885	0.901	0.896
	ConvNeXt	0.817	0.875	0.839	0.875	0.909	0.928
Transformer	Vision Transformer	0.807	0.860	0.833	0.860	0.902	0.840
	Swin Transformer	0.811	0.857	0.827	0.857	0.898	0.899
	MetaFormer	0.872	0.878	0.873	0.878	0.938	0.926
	EfficientFormer	0.874	0.897	0.883	0.897	0.917	0.939
	MaxViT	0.831	0.907	0.862	0.907	0.925	0.922
	FasterViT	0.844	0.897	0.870	0.897	0.906	0.910
Ours	HI-MViT	0.885	0.914	0.897	0.914	0.942	0.963

AUC: area under the curve; CNN: convolutional neural networks.

is mainly due to the following reasons: (1) large data demand: Training Vision Transformer with less than 100 million images cannot get an optimal solution. It is hard to compile a dataset of more than 100 million images in the medical field of image analysis; occasionally, a medical image dataset only contains thousands or perhaps a few hundred images. (2) A large number of calculations and parameters: the global attention mechanism has a large number of calculations, and the calculation cost squared with the input length greatly limits its application on high-resolution input. (3) The number of stacked layers is limited, and there is an excessive smoothing problem. (4) The ViT model's training procedure is unstable and parameter sensitive.¹⁰²

From the confusion matrix in Figure 9, it can be found that different categories will misjudge each other. This is because, as shown in Figure 1, there are significant inter-class similarities and intraclass differences in skin disease images, making it difficult to distinguish. Even dermatologists often require an extensive histopathological examination to draw concrete conclusions at the time of diagnosis.¹⁰³ On the other hand, due to the problem of unbalanced samples in the dataset, the number of samples of benign lesions is often dozens of times that of malignant lesions, which makes it easy for us to think of expanding the

data to improve the classification performance of model.¹⁰⁴ To verify this conjecture, this article uses traditional methods such as flipping, rotating, and adding noise, to amplify the data. Among them, the samples of AKIEC, BCC, DF, and VASC are increased by 10 times, the samples of MEL and BKL are increased by 4 times, and the samples of data enhancement are resized as the training set. Surprisingly, the effect of the model trained using the dataset after data augmentation is even worse, and it does not achieve the effect of enhancing the classification ability of the model. This is due to the blind pursuit of the balance between classes and the abuse of data enhancement methods will introduce a large amount of additional noise, resulting in excessive differences between the training set and the test set. More specifically, incorrect data augmentation causes the original distribution of data to change, and the similarity of data distribution before and after data augmentation is low.¹⁰⁵ Therefore, before performing data amplification, it is necessary to ensure that the distribution of the data does not change significantly before and after the amplification, and at the same time ensure the quality of the new image generated by the data amplification.

Our research still has certain limitations. First always foremost, medical ethics mandate thorough testing of any

new technology's effectiveness and security in clinical settings. Medical artificial intelligence that uses dermoscopic information to diagnose diseases carries some risks. Only dermatologists should use the study's findings as a guide. Secondly, to strengthen the model's capacity for generalization, our research solely relies on dermoscopic images and fails to incorporate any other categories of medical indicators into the model's design or training, including the patient's age, race, or location of the disease.¹⁰⁶ Third, there are very few case samples from people of color in the existing skin disease dataset, and the deep learning algorithm trained with the skin disease data of white people may not be able to better diagnose people of color.¹⁰⁷ Fourth, dermoscopic image-based deep learning approaches for skin disease diagnosis rely on their databases or public databases and lack external validation with a large number of samples, which is also the path of our future study. In light of the aforementioned issues, we will continue working with the Xiangya Hospital Dermatology Department to produce a dataset of dermoscopic images that contains more members of the yellow race, more skin conditions, and more metadata. And test our model in a real medical setting, taking into account the interobserver variance. At the same time, we plan to further optimize the HI-MViT model and deploy it to the mobile terminal, so that patients can easily and quickly identify whether the lesion area is malignant or not through the smart terminal device. Combined with the online diagnosis of dermatologists, it meets the public's requirements for timeliness and convenience of diagnosis.

Conclusions

In this study, we propose HI-MViT, a lightweight model for explainable skin disease classification based on Modified MobileViT. HI-MViT is mainly composed of ordinary convolution, Improved-MV2, MobileViT block, global pooling, and fully connected layers. Improved-MV2 uses the combination of shortcut and depth classifiable convolution to substantially scale back on the number of calculations while ensuring the efficient implementation of information interaction and memory. The MobileViT block can efficiently encode local and global information. In addition, HI-MViT uses semantic feature dimension reduction visualization and class activation mapping visualization methods to further understand the model's attention area when learning skin lesion images, and better improve the classification effect. We evaluate the performance of HI-MViT on ISIC-2017 and ISIC-2018 datasets. The results show that HI-MViT achieves superior performance scores on the ISIC-2017 and ISIC-2018 datasets compared to the comparison models. At the same time, the performance of testing on the PH² dataset using the trained model is also very good, reflecting the good generalization performance of HI-MViT. In addition, the results of

comparative experiments verify the effectiveness of the Improved-MV2 module in HI-MViT. As a future work, the performance of the proposed HI-MViT method can be compared with the performance of a capsule network because capsule-based networks can preserve spatial relationships of learned features and have been used in recently published works for image classification.¹⁰⁸ In the future, we plan to further cooperate with the Dermatology Department of Xiangya Hospital to build a more diversified image database of skin lesions and deploy HI-MViT to mobile devices to better provide patients with convenient medical services.

Acknowledgements: We would like to thank the ISIC-2017 and ISIC-2018 databases for providing valuable data.

Contributorship: YD, ZY, and YW conceived and supervised the study. ML, JL, and SL contributed to data collection and assembly. ZY, YD, and YG performed data analysis and interpretation. YD, PF, and CZ performed software, visualization, and validation. All authors contributed to writing the manuscript. All authors reviewed and approved the final manuscript.

Declaration of conflicting interests: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding: The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the grant from Hunan Provincial Natural Science Foundation of China (2021JJ41026).

Ethical approval: This study was an analysis of third-party de-identified publicly available databases with pre-existing ethical review board approval. All data were fully anonymized. All participants signed informed consent.

Guarantor: YJW.

ORCID iD: Yongjie Wang  <https://orcid.org/0009-0005-5846-9092>

References

1. Hay RJ, Johns NE, Williams HC, et al. The global burden of skin disease in 2010: an analysis of the prevalence and impact of skin conditions. *J Invest Dermatol* 2014; 134: 1527–1534.
2. Binder M, Puespoeck-Schwarz M, Steiner A, et al. Epiluminescence microscopy of small pigmented skin lesions: short-term formal training improves the diagnostic performance of dermatologists. *J Am Acad Dermatol* 1997; 36: 197–202.
3. Kittler H, Pehamberger H, Wolff K, et al. Diagnostic accuracy of dermoscopy. *Lancet Oncol* 2002; 3: 159–165.

4. Goceri E. Automated skin cancer detection: where we are and the way to the future. In: 2021 44th International Conference on Telecommunications and Signal Processing (TSP), pp.48–51.
5. Wu Y, Chen B, Zeng A, et al. Skin cancer classification with deep learning: a systematic review. *Front Oncol* 2022; 12: 893972.
6. Goceri E. Convolutional neural network based desktop applications to classify dermatological diseases. In: 2020 IEEE 4th International Conference on Image Processing, Applications and Systems (IPAS), pp.138–143.
7. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017; 542: 115–118.
8. Duman E and Tolan Z. Ensemble the recent architectures of deep convolutional networks for skin diseases diagnosis. *Int J Imaging Syst Technol* 2023; 33: 1293–1305.
9. Yu L, Chen H, Dou Q, et al. Automated melanoma recognition in dermoscopy images via very deep residual networks. *IEEE Trans Med Imag* 2017; 36: 994–1004.
10. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Adv Neural Inf Process Syst* 2017; 30.
11. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16 (16 words: transformers for image recognition at scale. Epub ahead of print 3 June 2021. DOI: 10.48550/arXiv.2010.11929.
12. Codella NCF, Gutman D, Celebi ME, et al. Skin lesion analysis toward melanoma detection: a challenge at the 2017 International symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In: 2018 *IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. 2018, pp. 168–172.
13. Wang T, Lan J, Han Z, et al. O-Net: a novel framework with deep fusion of CNN and transformer for simultaneous segmentation and classification. *Front Neurosci* 2022; 16: 876065.
14. Mehta S and Rastegari M. MobileViT: light-weight, general-purpose, and mobile-friendly vision transformer. Epub ahead of print 4 March 2022. DOI: 10.48550/arXiv.2110.02178.
15. Carcagni P, Leo M, Cuna A, et al. Classification of skin lesions by combining multilevel learnings in a DenseNet architecture. In: *Image Analysis and Processing-ICIAP 2019: 20th International Conference, Trento, Italy, September 9–13, 2019, Proceedings, Part I* 20, 2019, pp.335–344: Springer.
16. Jain S, Singhanian U, Tripathy B, et al. Deep learning-based transfer learning for classification of skin cancer. *Sens* 2021; 21: 8142.
17. Hu J, Qi Y and Wang J. Skin disease classification using Mobilenet-RseSK network. *J Phys Conf Ser* 2022; 2405: 012017.
18. Muhaba KA, Dese K, Aga TM, et al. Automatic skin disease diagnosis using deep learning from clinical image and patient information. *Skin Health Dis* 2022; 2: e81.
19. Xin C, Liu Z, Zhao K, et al. An improved transformer network for skin cancer classification. *Comput Biol Med* 2022; 149: 105939.
20. He X, Tan E-L, Bi H, et al. Fully transformer network for skin lesion analysis. *Med Image Anal* 2022; 77: 102357.
21. Sarker MMK, Moreno-García CF, Ren J, et al. TransSLC: Skin lesion classification in dermatoscopic images using transformers. In: *Annual Conference on Medical Image Understanding and Analysis*. Springer, 2022, pp. 651–660.
22. Nakai K, Chen Y-W and Han X-H. Enhanced deep bottleneck transformer model for skin lesion classification. *Biomed Signal Process Control* 2022; 78: 103997.
23. Aladhadh S, Alsanea M, Aloraini M, et al. An effective skin cancer classification mechanism via medical vision transformer. *Sens* 2022; 22: 4008.
24. Goceri E. Classification of skin cancer using adjustable and fully convolutional capsule layers. *Biomed Signal Process Control* 2023; 85: 104949.
25. Anand V, Gupta S, Koundal D, et al. Fusion of U-Net and CNN model for segmentation and classification of skin lesion from dermoscopy images. *Expert Syst Appl* 2023; 213: 119230.
26. Ayas S. Multiclass skin lesion classification in dermoscopic images using swin transformer model. *Neural Computing and Applications* 2023; 35: 6713–6722.
27. Cai G, Zhu Y, Wu Y, et al. A multimodal transformer to fuse images and metadata for skin disease classification. *Visual Comput* 2023; 39: 2781–2793.
28. Mukadam SB and Patil HY. Skin cancer classification framework using enhanced super resolution generative adversarial network and custom convolutional neural network. *Appl Sci* 2023; 13: 1210.
29. Sandler M, Howard A, Zhu M, et al. MobileNetV2: inverted residuals and linear bottlenecks. Epub ahead of print 21 March 2019. DOI: 10.48550/arXiv.1801.04381.
30. Codella N, Rotemberg V, Tschandl P, et al. Skin lesion analysis toward melanoma detection 2018: a challenge hosted by the International Skin Imaging Collaboration (ISIC). Epub ahead of print 29 March 2019. DOI: 10.48550/arXiv.1902.03368.
31. Howard A, Sandler M, Chu G, et al. Searching for MobileNetV3. Epub ahead of print 20 November 2019. DOI: 10.48550/arXiv.1905.02244.
32. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. Epub ahead of print 10 December 2015. DOI: 10.48550/arXiv.1512.03385.
33. Howard AG, Zhu M, Chen B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications. Epub ahead of print 16 April 2017. DOI: 10.48550/arXiv.1704.04861.
34. Glorot X, Bordes A and Bengio Y. Deep sparse rectifier neural networks. In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. JMLR Workshop and Conference Proceedings*, pp. 315–323.
35. Dubey SR, Singh S and Chaudhuri B. A comprehensive survey and performance analysis of activation functions in deep learning. 2021.
36. He X, Wang Y, Zhao S, et al. Deep metric attention learning for skin lesion classification in dermoscopy images. *Complex Intell Syst* 2022; 8: 1487–1504.
37. Wan Y, Cheng Y and Shao M. MSLANet: multi-scale long attention network for skin lesion classification. *Appl Intell*.

- Epub ahead of print 29 September 2022. DOI: 10.1007/s10489-022-03320-x.
38. He K, Gan C, Li Z, et al. Transformers in medical image analysis. *Intell Med* 2023; 3: 59–78.
 39. Goceri E. Medical image data augmentation: techniques, comparisons and interpretations. *Artif Intell Rev*. Epub ahead of print 20 March 2023. DOI: 10.1007/s10462-023-10453-z.
 40. Goceri E. Comparison of the impacts of dermoscopy image augmentation methods on skin cancer classification and a new augmentation method with wavelet packets. *Int J Imaging Syst Technol* 2023; 33: 1727–1744.
 41. Goceri E. Image augmentation for deep learning based lesion classification from skin images. In: 2020 IEEE 4th International Conference on Image Processing, Applications and Systems (IPAS), pp.144–148.
 42. Goceri E. Intensity normalization in brain MR images using spatially varying distribution matching. In: 11th International Conference on computer graphics, visualization, computer vision and image processing (CGVCVIP 2017), 2017, pp.300–304.
 43. Goceri E. Fully automated and adaptive intensity normalization using statistical features for brain MR images. *Celal Bayar Üniversitesi Fen Bilimleri Dergisi* 2018; 14: 125–134.
 44. Goceri E. Evaluation of denoising techniques to remove speckle and Gaussian noise from dermoscopy images. *Comput Biol Med* 2023; 152: 106474.
 45. Loshchilov I and Hutter F. Fixing Weight Decay Regularization in Adam. 2018.
 46. Dozat T. Incorporating nesterov momentum into adam. 2016.
 47. Lin T-Y, Goyal P, Girshick R, et al. Focal loss for dense object detection. In: *Proceedings of the IEEE international conference on computer vision* 2017, pp.2980–2988.
 48. Leng Z, Tan M, Liu C, et al. Polyloss: a polynomial expansion perspective of classification loss functions. arXiv preprint arXiv:220412511. 2022.
 49. Hendrycks D and Gimpel K. Gaussian error linear units (gelus). arXiv preprint arXiv:160608415. 2016.
 50. Clevert D-A, Unterthiner T and Hochreiter S. Fast and accurate deep network learning by exponential linear units (ELUs). arXiv preprint arXiv:151107289. 2015.
 51. Maas AL, Hannun AY and Ng AY. Rectifier nonlinearities improve neural network acoustic models. In: *Proc. icml*. Atlanta, GA, 2013, p. 3.
 52. He K, Zhang X, Ren S, et al. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1026–1034.
 53. Lin M, Chen Q and Yan S. *Network in network*. arXiv preprint arXiv:1312.4400. 2013.
 54. Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.2921–2929. Las Vegas, NV, USA: IEEE.
 55. Xie S, Girshick R, Dollár P, et al. Aggregated residual transformations for deep neural networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.5987–5995. Honolulu, HI: IEEE.
 56. Ma N, Zhang X, Zheng H-T, et al. ShuffleNet V2: practical guidelines for efficient CNN architecture design. pp. 116–131.
 57. Tan M, Chen B, Pang R, et al. MnasNet: platform-aware neural architecture search for mobile. pp. 2820–2828.
 58. Vasu PKA, Gabriel J, Zhu J, et al. MobileOne: an improved one millisecond mobile backbone. Epub ahead of print 28 March 2023. DOI: 10.48550/arXiv.2206.04040.
 59. Woo S, Debnath S, Hu R, et al. ConvNeXt V2: co-designing and scaling ConvNets with masked autoencoders. Epub ahead of print 2 January 2023. DOI: 10.48550/arXiv.2301.00808.
 60. Liu Z, Lin Y, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows. Epub ahead of print 17 August 2021. DOI: 10.48550/arXiv.2103.14030.
 61. Yu W, Luo M, Zhou P, et al. MetaFormer is actually what you need for vision. Epub ahead of print 4 July 2022. DOI: 10.48550/arXiv.2111.11418.
 62. Li Y, Yuan G, Wen Y, et al. EfficientFormer: vision transformers at MobileNet speed. Epub ahead of print 10 October 2022. DOI: 10.48550/arXiv.2206.01191.
 63. Tu Z, Talebi H, Zhang H, et al. MaxViT: multi-axis vision transformer. Epub ahead of print 9 September 2022. DOI: 10.48550/arXiv.2204.01697.
 64. Hatamizadeh A, Heinrich G, Yin H, et al. FasterViT: fast vision transformers with hierarchical attention. Epub ahead of print 9 June 2023. DOI: 10.48550/arXiv.2306.06189.
 65. Zhuang J-X, Li W, Manivannan S, et al. *Skin lesion analysis towards melanoma detection using deep neural network ensemble*. 2018. Epub ahead of print 23 December 2018. DOI: 10.13140/RG.2.2.20668.26240.
 66. Li KM and Li EC. Skin lesion analysis towards melanoma detection via end-to-end deep learning of convolutional neural networks. Epub ahead of print 22 July 2018. DOI: 10.48550/arXiv.1807.08332.
 67. Amro MK, Singh B and Rizvi A. Skin lesion classification and segmentation for imbalanced classes using deep learning.
 68. Bissoto A, Perez F, Ribeiro V, et al. Deep-learning ensembles for skin-lesion segmentation, analysis, classification: RECOD Titans at ISIC challenge 2018. Epub ahead of print 25 August 2018. DOI: 10.48550/arXiv.1808.08480.
 69. Pan Y and Xia Y. Residual network based aggregation model for skin lesion classification. Epub ahead of print 24 July 2018. DOI: 10.48550/arXiv.1807.09150.
 70. Alam M, Mohammad MS, Hossain MAF, et al. S2C-DeLeNet: a parameter transfer based segmentation-classification integration for detecting skin cancer lesions from dermoscopic images. *Comput Biol Med* 2022; 150: 106148.
 71. Anand V, Gupta S, Altameem A, et al. An enhanced transfer learning based classification for diagnosis of skin cancer. *Diagnostics* 2022; 12: 1628.
 72. Popescu D, El-Khatib M and Ichim L. Skin lesion classification using collective intelligence of multiple neural networks. *Sens* 2022; 22: 4399.
 73. Tada M and Han X-H. Bottleneck transformer model with channel self-attention for skin lesion classification. In: 2023 18th International Conference on Machine Vision and Applications (MVA), 2023, pp.1–5.

74. Durães PF and Véstias MP. Smart embedded system for skin cancer classification. *Future Internet* 2023; 15: 52.
75. Li S, Li C, Liu Q, et al. An actinic keratosis auxiliary diagnosis method based on an enhanced MobileNet model. *Bioengineering* 2023; 10: 732.
76. Ringnér M. What is principal component analysis? *Nat Biotechnol* 2008; 26: 303–304.
77. van der Maaten L and Hinton G. Visualizing data using t-SNE. *J Mach Learn Res* 2008; 9: 2579–2605.
78. McInnes L, Healy J and Melville J. Umap: uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:180203426*.
79. Selvaraju RR, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. pp. 618–626.
80. Chattopadhyay A, Sarkar A, Howlader P, et al. Grad-CAM+: improved visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp.839–847.
81. Desai S and Ramaswamy HG. Ablation-CAM: visual explanations for deep convolutional network via gradient-free localization. pp. 983–991.
82. Wang H, Wang Z, Du M, et al. Score-CAM: score-weighted visual explanations for convolutional neural networks. pp. 24–25.
83. Muhammad MB and Yeasin M. Eigen-CAM: class activation map using principal components. In: 2020 International Joint Conference on Neural Networks (IJCNN), pp.1–7.
84. Raghu M, Unterthiner T, Kornblith S, et al. Do vision transformers see like convolutional neural networks? Epub ahead of print 3 March 2022. DOI: 10.48550/arXiv.2108.08810.
85. Wu H, Chen S, Chen G, et al. FAT-Net: feature adaptive transformers for automated skin lesion segmentation. *Med Image Anal* 2022; 76: 102327.
86. Matsunaga K, Hamada A, Minagawa A, et al. Image classification of melanoma, nevus and seborrheic keratosis by deep neural network ensemble. Epub ahead of print 8 March 2017. DOI: 10.48550/arXiv.1703.03108.
87. Díaz IG. Incorporating the knowledge of dermatologists to convolutional neural networks for the diagnosis of skin lesions. Epub ahead of print 2 June 2017. DOI: 10.48550/arXiv.1703.01976.
88. Menegola A, Tavares J, Fornaciali M, et al. RECOD titans at ISIC challenge 2017. Epub ahead of print 14 March 2017. DOI: 10.48550/arXiv.1703.04819.
89. Bi L, Kim J, Ahn E, et al. Automatic skin lesion analysis using large-scale dermoscopy images and deep residual networks. Epub ahead of print 16 March 2017. DOI: 10.48550/arXiv.1703.04197.
90. Yang X, Zeng Z, Yeo SY, et al. A novel multi-task deep learning model for skin lesion segmentation and classification. Epub ahead of print 2 March 2017. DOI: 10.48550/arXiv.1703.01025.
91. Salian SR and Sawarkar SD. Melanoma skin lesion classification using improved efficientnetb3. *Jordanian J Comp Inform Technol* 2022; 8(1).
92. Wu Y, Lariba AC, Chen H, et al. Skin lesion classification based on deep convolutional neural network. In: 2022 IEEE 4th International Conference on Power, Intelligent Computing and Systems (ICPICS), 2022, pp.376–380.
93. Golnoori F, Boroujeni FZ and Monadjemi A. Metaheuristic algorithm based hyper-parameters optimization for skin lesion classification. *Multimed Tools Appl* 2023; 82: 25677–25709.
94. Kim C, Jang M, Han Y, et al. Skin lesion classification using hybrid convolutional neural network with edge, color, and texture information. *Appl Sci* 13. Epub ahead of print 2023. DOI: 10.3390/app13095497
95. Zhu M, Zhang L, Wang L, et al. Robust co-teaching learning with consistency-based noisy label correction for medical image classification. *Int J Comput Ass Rad* 2023; 18: 675–683.
96. Tschandl P, Rosendahl C and Kittler H. The HAM10000 dataset: a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci Data* 5. Epub ahead of print 14 August 2018. DOI: 10.1038/sdata.2018.161.
97. Alwakid G, Gouda W, Humayun M, et al. Melanoma detection using deep learning-based classifications. *Healthcare* 2022; 10: 2481.
98. Ali K, Shaikh ZA, Khan AA, et al. Multiclass skin cancer classification using EfficientNets—a first step towards preventing skin cancer. *Neurosci Inform* 2022; 2: 100034.
99. Chaturvedi SS, Gupta K and Prasad PS. Skin lesion analyser: an efficient seven-way multi-class skin cancer classification using MobileNet. In: Hassanien AE, Bhatnagar R and Darwish A (eds) *Advanced machine learning technologies and applications*. Singapore: Springer, 2021, pp.165–176.
100. Peng Z, Huang W, Gu S, et al. Conformer: local features coupling global representations for visual recognition. Epub ahead of print 9 May 2021. DOI: 10.48550/arXiv.2105.03889.
101. Li B, Liu Y and Wang X. Gradient harmonized single-stage detector. Epub ahead of print 13 November 2018. DOI: 10.48550/arXiv.1811.05181.
102. Islam K. Recent advances in vision transformer: a survey and outlook of recent work. Epub ahead of print 22 August 2022. DOI: 10.48550/arXiv.2203.01536.
103. Patel S, Wang JV, Motaparthy K, et al. Artificial intelligence in dermatology for the clinician. *Clin Dermatol* 2021; 39: 667–672.
104. Chlap P, Min H, Vandenberg N, et al. A review of medical image data augmentation techniques for deep learning applications. *J Med Imag Radiat On* 2021; 65: 545–563.
105. Mikolajczyk A and Grochowski M. Data augmentation for improving deep learning in image classification problem. In: 2018 International Interdisciplinary PhD Workshop (IIPhDW), pp.117–122.
106. Caffery LJ, Clunie D, Curiel-Lewandrowski C, et al. Transforming dermatologic imaging for the digital era: meta-data and standards. *J Digit Imaging* 2018; 31: 568–577.
107. Kassem MA, Hosny KM, Damaševičius R, et al. Machine learning and deep learning methods for skin lesion classification and diagnosis: a systematic review. *Diagnostics* 2021; 11: 1390.
108. Goceri E. Analysis of capsule networks for image classification. In: International conference on computer graphics, visualization, computer vision and image processing, 2021.