

Research Article

Multiscale Feature Fusion for Skin Lesion Classification

Gang Wang ^{1,2}, Pu Yan ^{1,3}, Qingwei Tang^{1,2}, Lijuan Yang^{1,2} and Jie Chen^{1,2}

¹College of Electronic and Information Engineering, Anhui Jianzhu University, Hefei 230000, China

²Anhui International Joint Research Center for Ancient Architecture Intellisencing and Multi-Dimensional Modeling, Anhui Jianzhu University, Hefei 230000, China

³Anhui Provincial Key Laboratory of Intelligent Building and Building Energy Conservation, Anhui Jianzhu University, Hefei 230000, China

Correspondence should be addressed to Pu Yan; yp8188@ahjzu.edu.cn

Received 16 July 2022; Revised 13 December 2022; Accepted 23 December 2022; Published 5 January 2023

Academic Editor: Kwang Gi Kim

Copyright © 2023 Gang Wang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Skin cancer has a high mortality rate, and early detection can greatly reduce patient mortality. Convolutional neural network (CNN) has been widely applied in the field of computer-aided diagnosis. To improve the ability of convolutional neural networks to accurately classify skin lesions, we propose a multiscale feature fusion model for skin lesion classification. We use a two-stream network, which are a densely connected network (DenseNet-121) and improved visual geometry group network (VGG-16). In the feature fusion module, we construct multireceptive fields to obtain multiscale pathological information and use generalized mean pooling (GeM pooling) to reduce the spatial dimensionality of lesion features. Finally, we built and tested a system with the developed skin lesion classification model. The experiments were performed on the dataset ISIC2018, which can achieve a good classification performance with a test accuracy of 91.24% and macroaverages of 95%.

1. Introduction

The skin epidermis consists of three kinds of cells: basal, squamous, and melanocyte [1]. Pigmented skin diseases [2] are formed because of abnormalities in melanin and melanocytes. The deadliest skin disease is melanoma, a highly malignant tumor derived from melanocytes. Due to changes in skin color and inconspicuous early lesions, it is very difficult to classify them. However, it can be cured if detected and treated in time, so the accurate classification of pigmented skin diseases is particularly important for later treatment [3]. Clinically, it is a common method to use modern skin imaging technology and expert experience for treatment and diagnosis [4]. However, the subjective factors of experts are not fully authoritative, and imaging analysis is time-consuming and laborious, so it is difficult to automatically classify malignant skin diseases from dermoscopic images. With the development of computer image processing technology, computer-aided design (CAD) system [5] can realize automatic classification based on skin lesion images. How-

ever, pigmented skin diseases have high interclass similarity and intraclass variations, and the pathological characteristics of early and late stages are completely different. Benign and malignant diseases have little difference in early symptoms, but when the lesions are formed, the appearance is very different, for example, melanoma (malignant) and melanoma nevus (benign). Both diseases are tumors derived from melanocytes, and their clinical manifestations are the same, which makes classifying them difficult. Figure 1 shows the early symptoms and later manifestations of the seven skin lesions. The first column is the early stage of seven skin lesions, and the others are random symptoms of skin lesions. We can see that the color, boundary size, and abnormal appearance of lesions in the early stage are very similar, while the appearance of late stage is very different. Therefore, it is impossible to infer with the naked eye whether the early skin lesions will become malignant or benign. It makes it more difficult to classify them.

Accurate skin lesion classification remains a challenge due to four factors. Firstly, image classification technology



FIGURE 1: Seven classes of pigmented skin lesions.

based on CNN network requires a large number of training datasets. However, due to the scarcity and imbalance of medical datasets, it is difficult to collect a large number of skin lesion images. Thus, we use six data augmentation operations to expand and balance the ISIC2018 dataset. Secondly, local pathological information is an important judgment to improve the classification accuracy. To fully mine the local pathological information of skin lesion images, we introduce a residual structure to simplify the learning process and allow the network to learn the local and global features of skin lesion images. Thirdly, feature fusion is an important means to improve classification performance. Since the pathological features obtained from a single network are limited and feature correlation affects the classification performance, we use a two-stream network to eliminate the redundant information brought by the correlation between different feature sets. Finally, a single receptive field cannot obtain more comprehensive lesion information. Therefore, we use multireceptive

fields (small convolution kernel and large convolution kernel) to complement each other and make the network get more pathological regions. Overall, the contributions of this paper are as follows:

- (i) We fuse the residual structure for the traditional VGG-16 network model to reduce the risk of gradient disappearance in the network
- (ii) We use a two-stream network model (DenseNet-121 and VGG-16) for feature fusion, which can combine the advantages of a single network. And then, we use the multireceptive field module to obtain multiscale pathological information
- (iii) A skin lesion classification system is designed, which provides accurate diagnostic information for experts or patients

2. Related Works

2.1. DCNN-Based Skin Lesion Classification. To apply deep convolutional neural network (DCNN) to skin lesion classification [6, 7], experts and scholars have innovated many classical network structures. He et al. [8] introduced a deep residual learning framework to solve the degradation problem, making it possible to build very deep networks, such as AlexNet [9], VGGNet, GoogLeNet [10], and Inception [11]. However, deep network structures generate a large number of parameters, resulting in model redundancy. DenseNet [12] used a dense connection layer, and each layer can obtain the connected feature map of the previous layer. Model redundancy is reduced by feature reuse at each level of the network. The representative VGG-16 network [13] explored the relationship between the depth of a convolutional neural network and its performance by repeatedly stacking 3×3 convolution kernels and 2×2 max pooling layers. The convolution concatenation in VGG-16 network has fewer parameters than using a larger convolution kernel alone and has more nonlinear transformations than a single convolution layer. VGG-16 obtained more image features with a simple network structure, smaller convolutional kernels, and pooling layers, while avoiding excessive computational effort and overly complex structures.

2.2. Multireceptive Field-Based Skin Lesion Classification. High-level features are used to measure semantic similarity, and low-level features (edges and contours of pathological regions) can reflect image content. How to efficiently integrate the two features is the key to improving the classification model [14]. Szegedy et al. [15] combined the CNN features of different high and low layers applied to skin lesion classification. The different convolution layers would learn different weights according to different receptive fields, which can make the network explore more comprehensive pathological regions. Inception module of GoogLeNet [16] used multiple convolution layers with different kernels which are sampled at the same center to construct its receptive field module. Multireceptive fields are used to focus on different spatial positions of the object and its adjacent background, which is helpful to obtain high-quality features and enhance the distinguishability of features.

2.3. Recent Deep Learning-Based Skin Lesion Classification. Recent skin lesion classification studies have achieved recent performance [17, 18]. Gessert et al. [19] pretrained three neural network models and validated multiple balancing methods. They used metalearning methods for prediction and obtained 85.1% classification accuracy on ISIC2018 dataset. Shahin et al. [20] used an ensemble method to classify seven skin lesions by combining two network models, ResNet-50 and Inception-v3, and verified that the classification accuracy is as high as 89.9%. Amirreza et al. [21] studied the effect of image size on skin lesion classification based on pretrained CNN and transfer learning. On the ISIC2018 classification challenge testing set, the author's multiscale and multinet method yielded a balanced multiclass accuracy of 86.2%. Al-Masni et al. [22] segmented the path-

ological regions and then applied multiple convolutional network classifiers for lesion classification. The authors chose ResNet-50 by testing a number of established neural networks, which demonstrated its excellent performance with a classification accuracy of 89.28%. Zillur et al. [23] proposed a weighted average ensemble learning-based model to classify skin lesions, using five deep neural network models (ResNeXt, SeResNeXt, ResNet, Xception, and DenseNet) to find the best combination in the ensemble. Finally, an average classification accuracy of 88% was obtained. Abayomi-A et al. [24] created synthetic melanoma images by oversampling data in a nonlinear lower-dimensional embedding manifold. The augmented images were used to train the Squeeze Net deep learning model. Nawaz et al. [25] proposed a UNET model based on DenseNet77. The authors introduced the DenseNet77 network at the encoder unit of the UNET method to compute a more representative set of image features. The key points of the computation are subsequently segmented by the decoder of the UNET model.

In summary, a single network needs to be designed with a deeper network structure to learn as many pathological features as possible, but widening the depth and width of the network does not improve the classification performance of the model [26, 27]. Therefore, we employ a two-stream network to compensate for the shortcomings of a single network. Meanwhile, VGG-16 in our two-stream network fuses the residual structure, broadening the network depth while also focusing on more pathological regions. Thus, we decided to use DenseNet and VGG-16 to design our two-stream network model to obtain finer-grained pathological features and improve the classification accuracy of skin lesions.

3. Materials and Methods

3.1. Dataset. ISIC means International Skin Imaging Collaboration. It is the largest and public dataset of skin images and for medical image classification. In our model, the published ISIC2018 Task 3 dermoscopic image dataset [28] is used (also called the Human Against Machine (HAM) 10000 dataset) [29], which consists of 10,015 dermoscopic images, and each of which is a 600×450 three-channel RGB image. The dataset includes seven skin lesions, namely, actinic keratosis and intraepithelial carcinoma (Akiec), basal cell carcinoma (Bcc), benign lesions of the keratosis (Bkl), dermatofibroma (Df), melanoma (Mel), melanocytic nevus (Nv), and vascular disease (Vasc). The seven skin lesion types are shown in Figure 1. As can be seen from Figure 1, the high interclass similarity makes it difficult to classify lesions by the naked eye. Moreover, many images include hairs that are present on the skin, which significantly reduces the classification ability of the model. Although the ISIC2018 dataset provides 10,015 images as the training set and ground truth category labels with annotations, it is far from the amount of data needed for our model in the training and evaluation process. Thus, the dataset still needs to be manually divided. The component distribution of each lesion in the dataset is shown in Figure 2. The maximum and minimum numbers of images for different classes are 6705 and 115, respectively. Therefore, we perform a

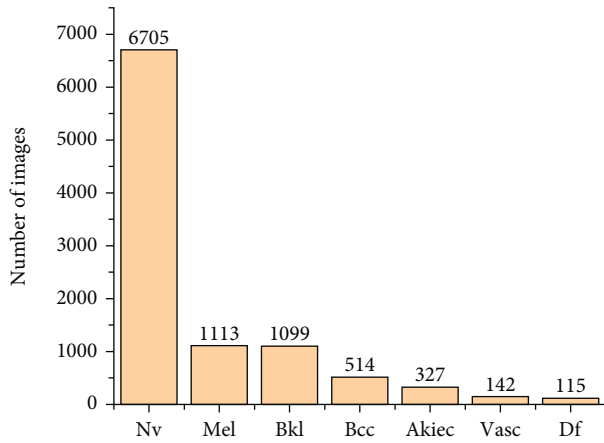


FIGURE 2: Distribution of lesion types in the dataset.

dataset preprocessing operation before the images are sent to the network.

3.2. Skin Lesion Multiclassification Model. To better improve the classification accuracy of skin lesions, we design a multiscale feature fusion model. Our model structure is based on DenseNet-121 network and improved VGG-16 network. We concatenate the output features of the two networks and use the multireceptive field to obtain multigranularity and multiscale global features. The feature fusion module is to enhance the ability to distinguish the pathological regions and background regions.

Specifically, our network framework consists of preprocessing dataset, two-stream network, feature fusion module, and multiclassification. In the preprocessing step, we dehair the skin lesion images and use six data augmentation operations to improve the generalization ability of the network. To obtain and exploit more pathological features, we use two networks in parallel in the two-stream network. The two-stream networks are DenseNet-121 and improved VGG-16, respectively. We add a residual structure to the original VGG-16 to deepen the depth of the network without increasing the parameters. We use a VGG-16 network with residual structure to remove redundant information from correlations between different feature sets and further obtain contextual correlations in pathological regions. In the feature fusion module, we obtain multiscale regions by adding multireceptive field. Moreover, we abandon the traditional pooling operation and use the GeM pooling operation to balance the compression of the features, which can improve the classification accuracy. We use the softmax function as a classifier for multiclassification. Overall, the framework of our model is shown in Figure 3.

The steps of our two-stream network for skin lesion classification are as follows:

- (1) The skin lesion images are resized to $224 \times 224 \times 3$ and divided into training set, test set, and validation set in the ratio of 7:2:1
- (2) The hair removal algorithm and six data enhancement operations are performed on the training set to reduce

noise interference and balance the number of seven types of skin lesion images after preprocessing

- (3) The improved two-stream network model based on multireceptive fields is built to obtain more pathological features to improve the model classification performance
- (4) The seven classifications of skin lesion images are achieved using softmax classifier

3.2.1. Data Preprocessing. In this subsection, we describe the preprocessing of the dataset in detail, including resizing, hair removal, and data augmentation, as can be seen in Figure 4.

- (i) **Resize:** the ISIC2018 dataset contains high-resolution images. The resolution of all skin lesion images is 600×450 pixels, which requires high computational cost if used directly for training. Therefore, we resize all images from 600×450 to 224×224 as required by the model. Based on the literature [30], we select 70% of the data as the basic training set, 20% of the data as the basic test set, and the remainder of the data as the basic verification set, and the number of each lesion in the training set, test set, and verification sets is divided into a ratio of 7:2:1. The sample distribution after dataset division is shown in Table 1
- (ii) **Hair removal [31]:** the ISIC2018 datasets are often characterized by hair-like regions within the skin lesions, which would interfere with the model's extraction of pathological features. Thus, we dehair all images to reduce the hair interfere. We convert the original image to a grayscale image, detect the hair contour using the black-hat operation, and create the mask. We apply an image inpainting technique based on the fast marching method (FMM) [32] to remove the mask containing only the hair from the original image and repair the void pixels. Hair removal operation can remove hair in pathological regions very well. Although some pixel information is lost, the removed pixels are mostly hair pixels and do not have a large impact on the feature extraction of the model
- (iii) **Data augmentation:** although the ISIC2018 dataset contains 10015 images, it cannot meet the number of large-scale data required for deep learning methods. Thus, we perform six data augmentation operations including randomly rotating, horizontal and vertical shifting, random zooming, random transforming, flipping, and resizing on hair removal training samples. We expanded the samples of the other six skin lesions in the training set except for melanocytic nevus (Nv) samples to enlarge the sample data. Thus, the other number of samples is approximately equal to the number of melanocytic nevus (Nv) samples, which avoid the overfitting

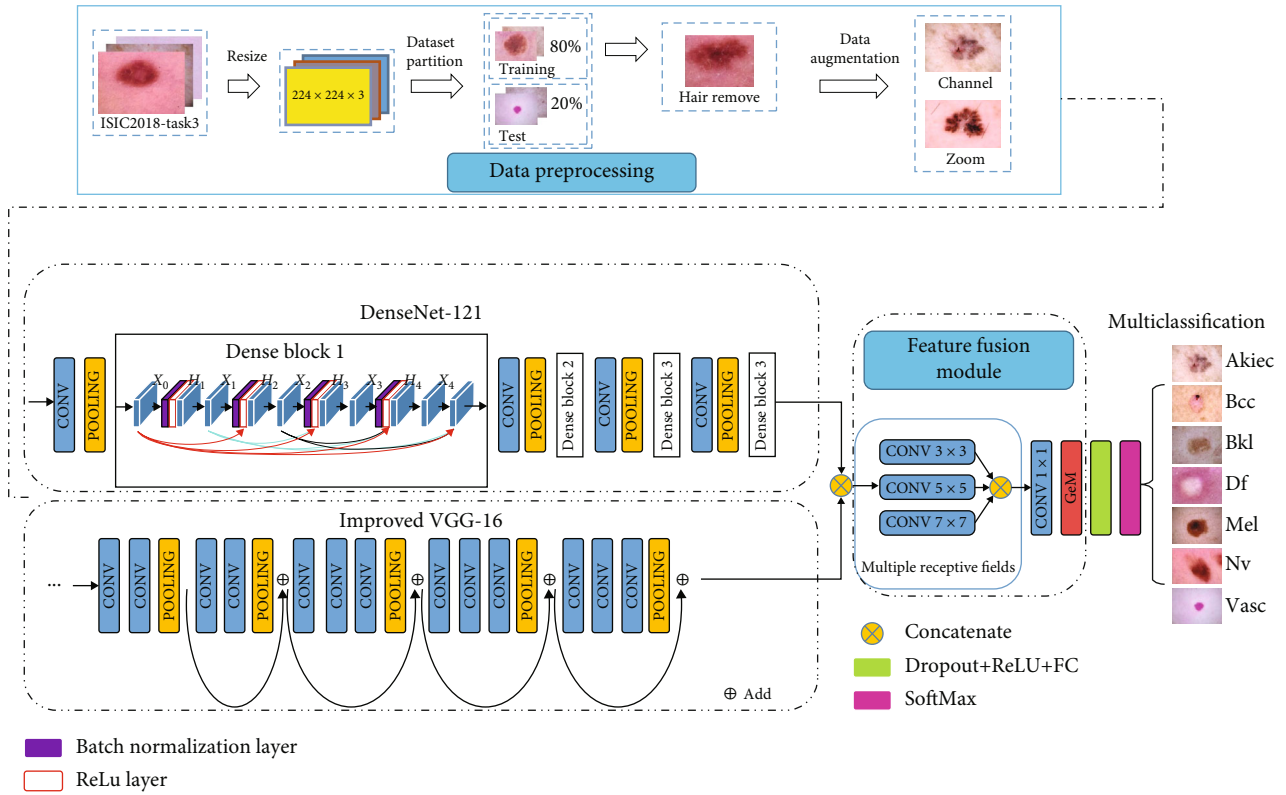


FIGURE 3: The framework of our model. Our model consists of four stages: data preprocessing, two-stream network, feature fusion module, and multiclassification.

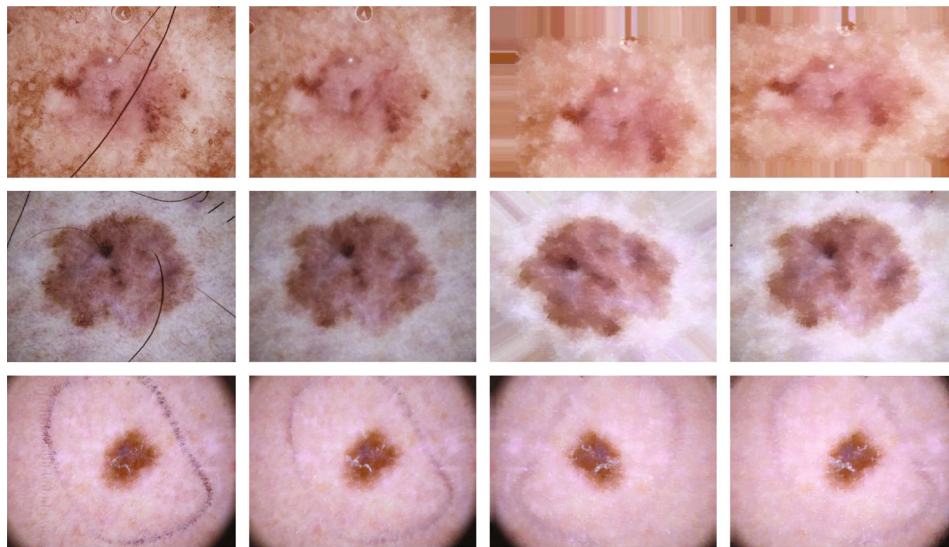


FIGURE 4: Examples of three groups of skin lesions by preprocessing. The first column shows the three sets of original images. The second column shows the corresponding dehaired images, and the third and fourth columns show the augmented images. We can observe that the hair in the pathological region is well removed.

caused by too few samples. The sample distribution of the training set after data augmentation is shown in Table 1

3.2.2. *Our Two-Stream Network Structure.* We chose a deep network DenseNet-121 and a shallow network VGG-16 to

construct our two-stream network. To increase the network depth and exploit more pathological features without adding extra parameters, we improve the VGG-16 by adding the residual structure. We use two networks to better distinguish pathological regions from background regions and improve the classification accuracy of skin lesions.

TABLE 1: The sample distribution after dataset division and the training set after data augmentation.

Category	Basic training set	Augmentation training set	Basic test set	Basic verification set
Nv	4697	4697	1339	669
Mel	780	5439	222	111
Bkl	771	5383	219	109
Bcc	361	5040	102	51
Akiec	230	4251	65	32
Vasc	100	5600	28	14
Df	81	4536	23	11
Total	7020	34946	1998	997

DenseNet-121 is a tightly connected CNN with well resistance to overfitting. The complexity of the network increases with depth (combination of more nonlinear functions), but DenseNet-121 has a direct connection between any two layers. Therefore, it can make full use of previous features, and it is easier to obtain richer pathological information. Compared with the sparsity of VGG-16, the DenseNet-121 is more compact, and the feature generated by the DenseNet-121 is more powerful, which makes up for the deficiency of VGG-16. Moreover, the compact structure can effectively reduce the gradient disappearance and improve the efficiency of feature utilization and enable the model to pay attention to the pathological information of a larger region. Table 2 is the network structure and output size of DenseNet-121.

Since the densely connected layer repeatedly utilizes features from the previous layer, the correlation between features affects the model classification performance. To compensate for the redundancy caused by feature reuse and enhance the transfer of pathological information between layers in the VGG block, we use VGG-16 fusion residual network for feature fusion to eliminate redundancy. The original VGG-16 network structure is shown in Table 3 and mainly includes 5 VGG blocks. Each VGG block contains 2 or 3 convolution layers and a max pooling layer. It can be observed that the VGG-16 network structure is simple without redundant layers to interfere with the model classification performance. For skin lesion images, we must explore fine-grained local features to ensure accurate classification, especially contour descriptors with lesion tendency. We add residual structures before and after each block of the VGG-16 network to extract multiscale features. The residual structure is divided into two parts: identity mapping and residual mapping. Identity mapping occurs when the output features of the previous layer are directly input to the next layer, and we use a 1×1 convolutional layer to match the number of feature channels. Residual mapping occurs when the input features of the previous layer are input to the next layer through the superposition of nonlinear changes. We use a unit addition operation on the residual structure output features to obtain more pathological features, which does not add additional trainable parameters. Moreover, we use a

TABLE 2: Network structure and output size of DenseNet-121.

Layers	Output size	DenseNet-121
Convolution	$112 \times 112 \times 64$	
Pooling	$56 \times 56 \times 64$	
Dense block (1)	$56 \times 56 \times 256$	$(1 \times 1 \text{ conv}) \times 6,$ $(3 \times 3 \text{ conv}) \times 6$
Transition layer (1)	$56 \times 56 \times 128$ $28 \times 28 \times 128$	
Dense block (2)	$28 \times 28 \times 512$	$(1 \times 1 \text{ conv}) \times 12,$ $(3 \times 3 \text{ conv}) \times 12$
Transition layer (2)	$28 \times 28 \times 256$ $14 \times 14 \times 256$	
Dense block (3)	$14 \times 14 \times 1024$	$(1 \times 1 \text{ conv}) \times 24,$ $(3 \times 3 \text{ conv}) \times 24$
Transition layer (3)	$14 \times 14 \times 512$ $7 \times 7 \times 512$	
Dense block (4)	$7 \times 7 \times 768$	$(1 \times 1 \text{ conv}) \times 16,$ $(3 \times 3 \text{ conv}) \times 16$

TABLE 3: Network structure and output size of VGG-16.

Layers	Output size	VGG-16
VGG block (1)	$224 \times 224 \times 64$	$(3 \times 3 \text{ conv}) \times 2$
Pooling	$112 \times 112 \times 128$	
VGG block (2)	$112 \times 112 \times 128$	$(3 \times 3 \text{ conv}) \times 2$
Pooling	$56 \times 56 \times 256$	
VGG block (3)	$56 \times 56 \times 256$	$(3 \times 3 \text{ conv}) \times 3$
Pooling	$28 \times 28 \times 512$	
VGG block (4)	$28 \times 28 \times 512$ $28 \times 28 \times 1024$ $28 \times 28 \times 1536$	$(3 \times 3 \text{ conv}) \times 3$
Pooling	$14 \times 14 \times 768$	
VGG block (5)	$14 \times 14 \times 768$ $14 \times 14 \times 512$ $14 \times 14 \times 768$	$(3 \times 3 \text{ conv}) \times 3$
Pooling	$7 \times 7 \times 768$	

residual structure to obtain the feature information of the upper layer input to reduce the feature loss between different blocks. The fuse of the residual structure [8] simplifies the training of the network while preserving the information integrity. The residual structure increases the ability of gradient cross-layer propagation and further improves the classification accuracy of our model. The model structure of the VGG-16 fusion residual network is shown in Figure 5.

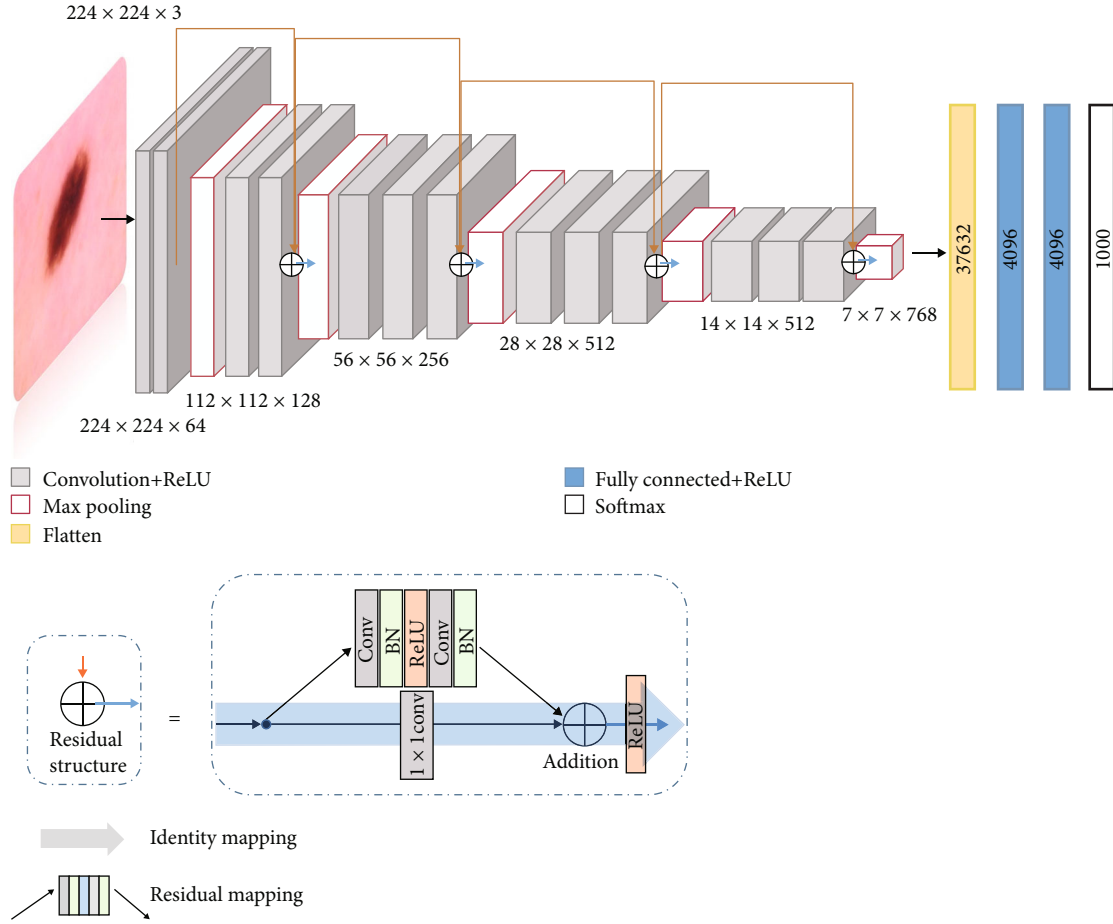


FIGURE 5: VGG-16 fusion residual structure network model.

Finally, we remove the FC layer from both DenseNet-121 and improved VGG-16. This keeps the consistent dimensions of the output feature maps, so that the feature fusion module can be performed on the output feature maps.

3.2.3. Feature Fusion Module. Due to the uneven distribution of pathological regions in the skin lesion images with different sizes and poor continuity, we determine the distribution of the extracted features by the size of the convolution kernel. The use of smaller convolutional kernels is biased towards extracting more local feature information, while the use of larger convolutional kernels is biased towards extracting more global pathology image features. However, to ensure that the output feature mapping is large enough, if we only use smaller convolutional kernels, a deeper network is often required, which will easily lead to network overfitting. The use of larger convolutional kernels will ignore local information, and the stacking of larger convolutional kernels will increase the computational effort and lead to a decrease in model efficiency.

To obtain information on a larger range of skin lesions, we have designed the feature fusion module to obtain the multireceptive field of skin lesions. As shown in the feature fusion module in Figure 3, the multireceptive fields are composed of multiple convolution layers with different kernel sizes, including 3×3 convolutional layers, 5×5 convolutional

layers, and 7×7 convolutional layers. The role of multireceptive field is to cover a larger region of skin lesions for obtaining more pathological regions. Different convolutional layers will learn different weights according to different receptive fields, and the smaller convolution kernel and the larger convolution kernel complement each other. They explore a more comprehensive pathological region, which helps to improve the overall accuracy of the model. We concatenate the feature maps of all convolutions with different receptive fields and enter them into the 1×1 convolutional layer+ReLU layer to perform channel-integrated and non-linear processing.

To weigh the pathological regions obtained by the features, we use a GeM pooling operation [33, 34] in the feature fusion module. We take a $1 \times 1 \times 1536$ feature vector X as the input and a vector H as the output of the pooling process. In the case of using max pooling, this vector $H_n^{(m)}$ is given by

$$H^{(m)} = [H_1^{(m)} \dots H_n^{(m)} \dots H_K^{(m)}]^T, \quad H_n^{(m)} = \max_{x \in X_k} x, \quad (1)$$

where K is the number of channels of the feature map. Let X_k be the set for feature map $n \in (1, K)$. The network output

consists of K such feature maps. The m refers to the max pooling operation and the x is all the features obtained in X_k .

And in the case of using average pooling, this vector $H_n^{(a)}$ is given by

$$H^{(a)} = \left[H_1^{(a)} \dots H_n^{(a)} \dots H_K^{(a)} \right]^T, \quad H_n^{(a)} = \frac{1}{|X_k|} \sum_{x \in X_k} x. \quad (2)$$

Instead, we exploit the GeM pooling, and this vector $H_n^{(g)}$ is given by

$$H^{(g)} = \left[H_1^{(g)} \dots H_n^{(g)} \dots H_K^{(g)} \right]^T, \quad H_n^{(g)} = \left(\frac{1}{|X_k|} \sum_{x \in X_k} x^{P_k} \right)^{1/P_k}. \quad (3)$$

P_k is a hyperparameter that adjusts the weights of the two pooling operations. Max pooling and average pooling are special cases of GeM pooling. When $P_k \rightarrow \infty$, it is max pooling and average pooling for $P_k \rightarrow 1$. The $H^{(g)}$ ultimately consists of the values of each feature map after GeM pooling, and its dimensionality is equal to K . We have trained and tested different P_k values and selected appropriate parameters to get better classification results in the following comparative experiments.

In the multiclassification module, we use the softmax classifier to achieve multiclassification of skin lesions.

4. Experiments and Results

In this section, we first introduce our experimental environment and the evaluation metrics. Then, we present the results of our ablation experiments to evaluate the performance of the network model. Next, we compare the performance of our model with other models. Finally, we introduce the skin lesion classification system, which can be used to assist doctors in diagnosing.

We use the deep learning framework TensorFlow [35] to build the network model. We use Python 3.8 as the programming language on the Windows 10 operating system with an Intel i7-10700F CPU and NVIDIA GeForce GTX2060 GPU.

4.1. Evaluation Metrics. The evaluation metrics include precision, recall, $F1$ -score, and accuracy. Precision in Equation (4) is used to measure the prediction accuracy of the classifier for a certain category. Recall in Equation (5) is used to measure the coverage of the classifier's prediction results for a certain category. $F1$ -score in Equation (6) is used to measure the accuracy and coverage of the classifier's prediction for a certain category. Accuracy in Equation (7) is the average accuracy that is used to calculate the overall, where TP, TN, FP, and FN represent the true positive, true negative, false positive, and false negative, respectively. We also use the receiver operating characteristic (ROC) curve and area under the curve (AUC) as the evaluation criteria of model performance [36]. The ROC and AUC can well describe the classification performance of the classifier for samples with uneven distribution. To extend our metrics to

multiclassification, the macro- and microaverages are also calculated [23]. The macro- and microaverages mean that when calculating multiclass indicators, different weight methods are used to assign all samples.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad (4)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (5)$$

$$F1\text{-score} = \frac{2 \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \quad (6)$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \quad (7)$$

4.2. Ablation Experiment. We conducted three sets of ablation experiments in Section 4.2, and the training set is included by hair removal and data augmentation. The distribution of the training and test sets listed is in Table 1.

4.2.1. The Performance of a Two-Stream Network. We compare our two-stream network model with the VGG-16, improved VGG-16, and DenseNet-121, respectively. The experimental results are shown in Table 4. Compared with VGG-16, our improved VGG-16 has better classification performance, with precision, recall, $F1$ -score, and accuracy increase of 11.42%, 10.43%, 11%, and 15.12%, respectively. It can be seen from Table 3 that the main layers in the VGG-16 network are only the convolutional layer and max pooling layer, which causes the VGG-16 network to ignore the correlation between different blocks. And the VGG module has poor ability to aggregate feature information due to its simple stacking. The residual structure strengthens the correlation between blocks. It well solves the gradient degradation due to feature redundancy as the depth of the network increases.

Compared with the classical networks, our two-stream network has better classification performance; the precision, recall, $F1$ -score, and accuracy are 83.53%, 95.04%, 88.91%, and 91.24%, respectively. Our two-stream network is a sparse-compact network structure that can complement the strengths of both networks, where the DenseNet-121 obtains more detailed pathological features due to its compact network structure and the improved VGG-16 can well capture the local features of pathological regions due to the added residual structure. This shows that our two-stream network can effectively obtain more fine-grained information in the skin lesion images, making the feature representation more robust and improving the classification accuracy of the model.

4.2.2. The Performance of Multireceptive Field. To better illustrate the effectiveness of multireceptive field, we compare the improved VGG-16 and our model that both removed the multireceptive field. The results are shown in Table 5, where (-) stands for removing the multireceptive field module. Compared with the model without multireceptive field, our model has better classification performance, with precision, recall, $F1$ -score, and accuracy increase of 0.65%, 0.94%, 0.78%, and 1.07%, respectively. In the absence

TABLE 4: Our model performance compared to classical networks.

Methods	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)
VGG-16	66.43	75.09	70.50	72.48
Improved VGG-16	77.85	85.52	81.50	87.60
DenseNet-121	80.94	87.61	84.14	88.14
Our model	83.53	95.04	88.91	91.24

TABLE 5: Our model performance compared to multireceptive field.

Methods	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)
Improved VGG-16 (-)	76.47	83.23	79.71	86.32
Improved VGG-16	77.85	85.52	81.50	87.60
Our model (-)	82.88	94.10	88.13	90.17
Our model	83.53	95.04	88.91	91.24

of multireceptive field, the network is unable to distinguish the boundaries of pathological regions and obtain more local information. We can obtain a wider range of local information by using multireceptive field. Avoiding the loss of image detail texture by too small or too large receptive fields can effectively enhance the discriminability of pathological features and improve the accuracy of model classification.

4.2.3. The Performance of Different P_k . According to literature [33], we debug the P_k value in the GeM pooling operation ranging from 1 to ∞ . The accuracy of the model drops when P_k is 5, at which point we stop testing. The GeM pooling operation is a weighting of the max pooling and average pooling to highlight the advantages of both pooling operations. As can be seen from Table 6, when the parameter P_k is 4, the classification accuracy of our model is the highest, reaching 91.24. So we set the value of P_k to 4. The GeM pooling operation we used is better than the classification accuracy obtained by the max pooling ($P_k = 1$) and average pooling ($P_k = \infty$) operations by 3.09% and 2.21%, respectively.

4.3. Visualization Results. Grad-CAM [37] is used to display the visualization results of the output feature map, as shown in Figure 6. We visualized some of the random images of the training set for judgment. To test whether the proposed network model can focus on the location of the center of the pathology, we visualized some of the random images of the training set. Figure 6(a) shows the visualization results of the original VGG-16, and it is observed that the network only pays attention to two small parts of the pathological region. Figure 6(b) shows the visualization results of the improved VGG-16, and it is observed that the network focuses on larger and broader pathological regions. The improved VGG-16 enables to pay more attention to the semantically meaningful parts of the lesions and enhance the ability of the network to learn discriminative representations. The results show that combining the residual structure

TABLE 6: Our model accuracy compared to used different P_k .

Parameter value of P_k	1	2	3	4	5	∞
Accuracy	88.15	89.87	90.58	91.24	89.17	89.03

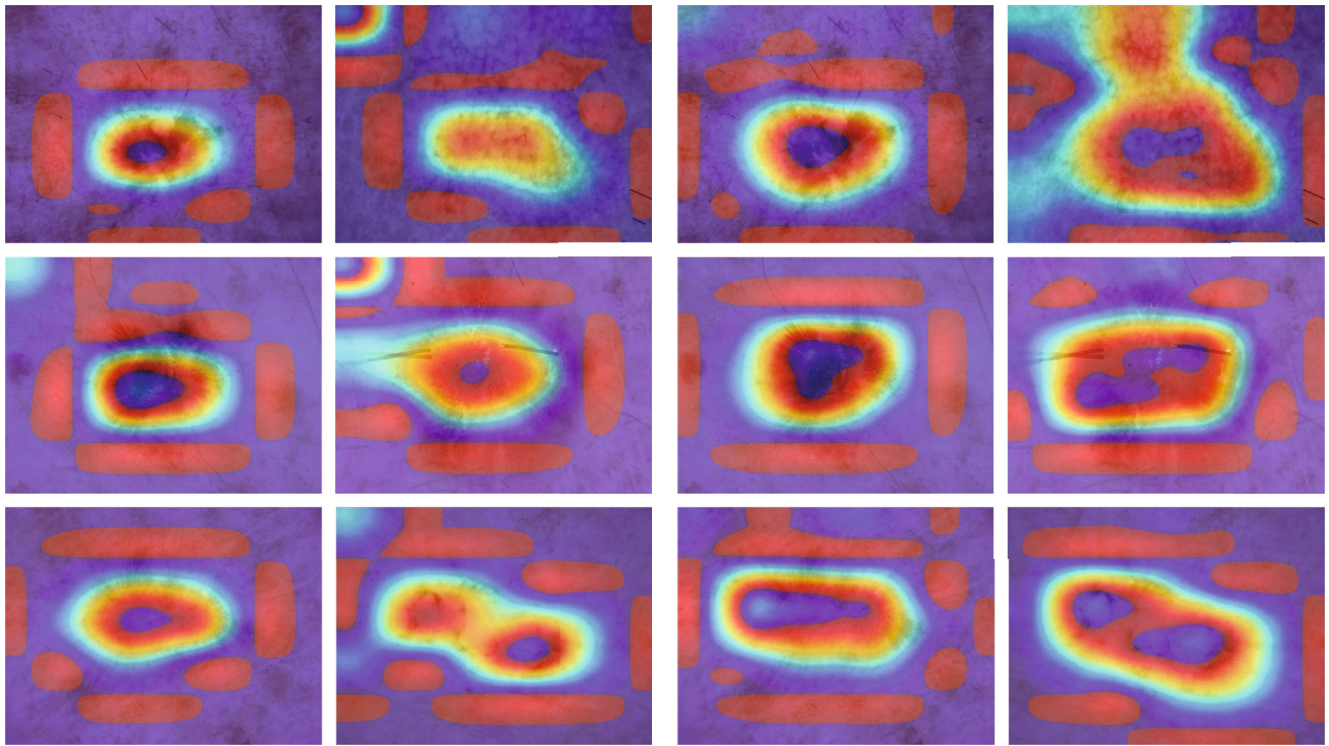
can make the model better detect the lesion region. It can be seen that the improved VGG-16 network is more accurate for the feature extraction of pathological regions.

4.4. Training and Testing Results of Our Model Performance

4.4.1. Model Loss. Figure 7 shows the loss between the training and validation processes of our model. The x -axis represents the epoch, and the y -axis represents the loss. We choose a batch size of 16 and a learning rate of 0.0001 initially, which decreases to 1/10 of the original learning rate after a period of training. It is observed that the model is run 28 epochs with GPU acceleration. In the trials for the training model, if the epoch does not change for five consecutive times, the training will be stopped. Considering that increasing the number of epochs does not increase the performance while extending the training period, it is deemed appropriate to limit the training process to 28 epochs. As can be seen from Figure 7, the training loss of our model reaches a very low value, which indicates that our model has been effectively trained.

4.4.2. Confusion Matrix. As a result of testing the model, the confusion matrix of seven different classes is shown in Figure 8. The x -axis represents the predicted skin lesion label (classes 0 to 6 represent Akiec, Bcc, Bkl, Df, Mel, Nv, and Vasc, respectively), and the y -axis represents the actual skin lesion label. The darker the color in the table, the greater the indicators of representative classification. It can be seen that there are a total of 222 Mel images in the testing set, and 115 of these images are misclassified as Nv, because the pathological region of the Mel image in the early stage is similar to Nv images. Therefore, it was misclassified the most, with an accuracy rate of only 47%. The classification accuracy of Vasc is 100% because of the significant difference in texture and color. It can be seen from the confusion matrix that our model has obtained a satisfactory result.

4.4.3. Area under the ROC Curve. To evaluate our classification model holistically across all classes, we also use macroaverages and microaverages. Because the ISIC2018 dataset has an unbalanced distribution. The macroaverage takes into account the distinction between classes, so we use this parameter to provide a good measure of the performance of our model on the studied dataset. The area under the ROC curve of our model is shown in Figure 9. AUCs of seven skin lesions are 1, 0.98, 0.92, 0.92, 0.94, 0.93, and 0.96, respectively. The microaverage of the ROC curve is 0.98, whereas its macroaverage is 0.95. This demonstrates that the model has excellent ROC curve scores for all classes. Therefore, our model has good classification performance in the multiclassification task of skin lesions.



(a) The original VGG-16

(b) The improved VGG-16

FIGURE 6: Visualization comparison between the original VGG-16 and the improved VGG-16.

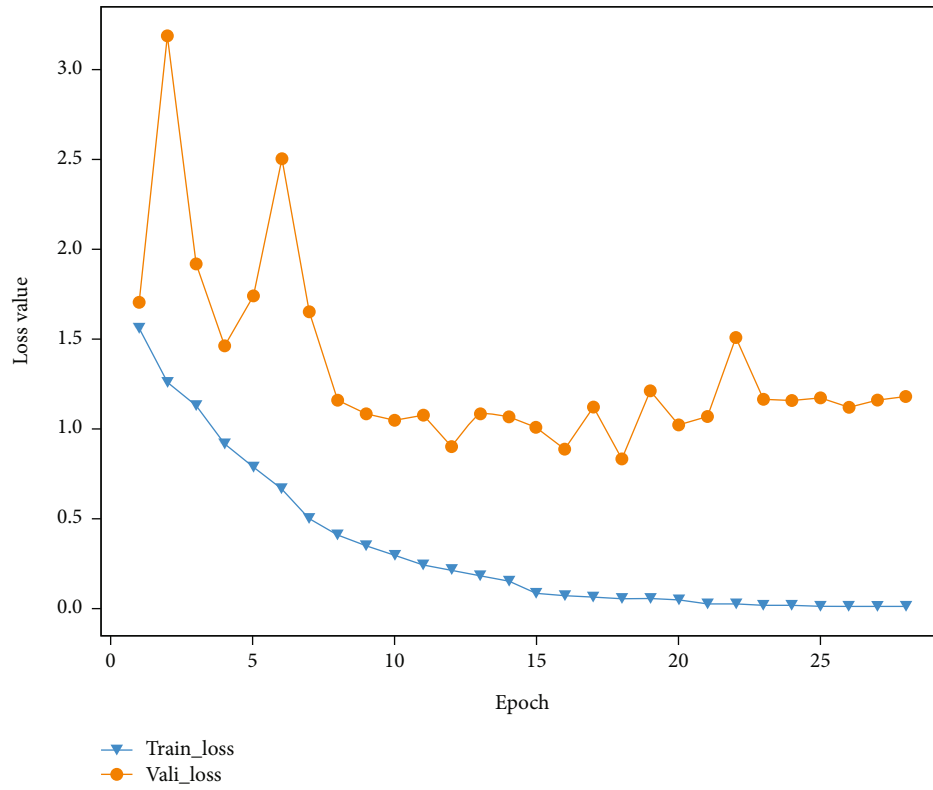


FIGURE 7: Line chart of training loss and validation loss.

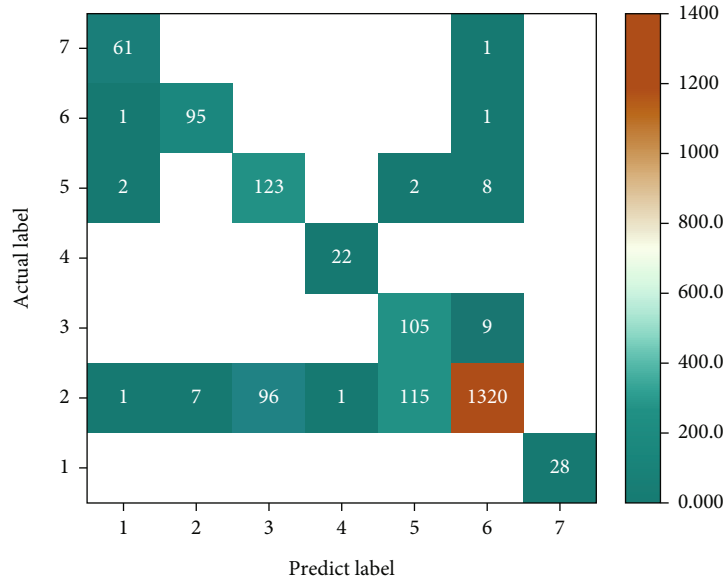


FIGURE 8: The confusion matrix of the model.

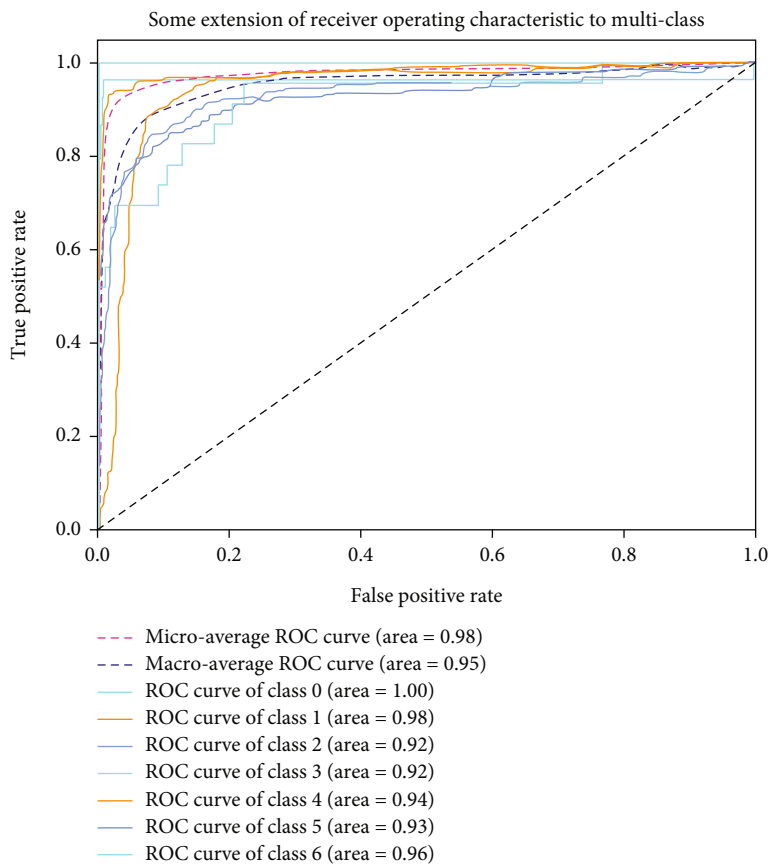


FIGURE 9: Area under the ROC curve. Note that classes 0 to 6 represent Akiec, Bcc, Bkl, Df, Mel, Nv, and Vasc, respectively.

4.5. Comparing to Other Existing Methods. To allow a fair comparison with previous works, all methods listed in Table 7 use the ISIC2018 dataset as the training set. With the exception of some Akiec images, the ISIC2018 dataset contains most of the HAM dataset images. Therefore, the

HAM dataset can also be used to compare. Literature [19–21] all used more than three network models for feature fusion, resulting in too many parameters and redundant features in the models. Literature [22] applied convolutional neural network classifiers (i.e., Inception-v3, ResNet-50,

TABLE 7: Comparison with other existing methods.

Method	Date	Precision (%)	Recall (%)	Para (M)	Accuracy (%)	Macroaverage (%)
DenseNet, SENet, ResNeXt [19]	2018	—	—	—	85.10	—
InceptionV3+ResNet-50 [20]	2018	86.2	79.60	—	89.90	—
MSM-CNN [21]	2020	91.30	—	—	86.20	98
ResNet-50 [22]	2020	—	81.00	23.54	89.28	—
ResNeXt, SeResNeXt, ResNet, Xception, DenseNet [23]	2021	87.00	94.00	—	88.00	93
RegNetY-3.2GF-Drop [38]	2021	—	—	15.30	86.40	97
Our method		83.53	95.04	19.70	91.24	98

Inception-ResNet-v2, and DenseNet-201) to segmented skin lesions for classification. Nevertheless, the authors fused the residual structure into Inception-v2 without better performance than the original ResNet-50 classification. Then, literature [23] proposed a weighted average ensemble learning-based model. They used five deep neural network models, namely, ResNeXt, SeResNeXt, ResNet, Xception, and DenseNet, as the base of the ensemble where the best weight combination was found by the grid search method. However, the authors did not consider the correlation between features. Literature [38] improved the data augmentation method and loss function and used the RegNetY-3.2GF-Drop model with medium complexity to achieve 86.4% classification accuracy. From the network improvement side, our two-stream network is a sparse and compact network structure. And the robustness of the final features can be improved by fusing the two networks. By adding the residual structures in VGG-16 and using multireceptive field, more pathological features are obtained for our model. Meanwhile, we fuse the output features of the two networks and appropriately remove the fully connected layer of the network, resulting in a significant reduction in the model parameters as well. We only use three different sets of convolution kernels to obtain multigranularity pathological features. In addition, our model obtains the complementary pathological region features as much as possible through the mechanism of multireceptive field. From the data processing side, our model simplifies the preprocessing process and further expands the data in terms of data augmentation. We have enough data to reduce the overfitting of the model. In summary, the classification accuracy of our model is 91.24%, and the macroaverage rate is 95%, which is better than the above methods.

4.6. Skin Lesion Classification System Based on Two-Stream Model. We have designed a skin lesion classification system based on our two-stream model. The framework is built using the Python Flask [39]. Our web system has a simple interface for experts or patients to easily use. For classification, specific test images can be uploaded to the system (select images). Once the test images are classified, the results and time consumed will be displayed on the page.

The implementation process of the system is shown in Figure 10. Firstly, we input a skin lesion image, and we require the upload format to be RGB image. For the convenience of users, we provide two types of uploads: photo and

local upload. Users can take photos and upload images of any size. Secondly, the system uniformly adjusts the image to $224 \times 224 \times 3$. Thirdly, the image is sent to our two-stream network model to determine what kind of disease is being classified. Since the model uses the softmax classifier as the classification layer, this layer will output the probability of skin lesions of each category. The system will return the label corresponding to the maximum probability value and time consumed as the result to the user. Fourthly, the system determines whether it is below the threshold value. Here, to classify the malignant disease, melanoma, we set the threshold to 0.9. When the probability value of the largest category label is less than 0.9, the user is requested to reupload the images. Conversely, if the threshold is higher than 0.9, the result display is returned, showing the probability of the disease category at the highest threshold and the time required for analysis. And the classification results can be saved directly for easy viewing by users. The following is the system execution flow:

- (1) Input RGB image (you can upload locally or take photos)
- (2) Resize the image
- (3) Send to two-stream network model for classification
- (4) Judgment: if the threshold value is less than 0.9, return to the first step; otherwise, proceed to the fifth step
- (5) Return the results and save the analysis results
- (6) End

The test of the skin lesion classification system was carried out as follows. The skin lesion images uploaded by the user are first classified according to our two-stream model. Then, the accuracy is confirmed by comparing the dataset stored on the server, resulting in the lesion type with the highest probability and consumed time. We engineered tests by using uploaded data on different seven types of skin lesions. According to the test results in Table 8, the model correctly classified 548 of the 600 images. The classification system was able to accomplish a classification speed of less than 1 second. We tested from our own collection and operation of the skin dataset, and these results show the accurate classification performance of our system. We emphasize that

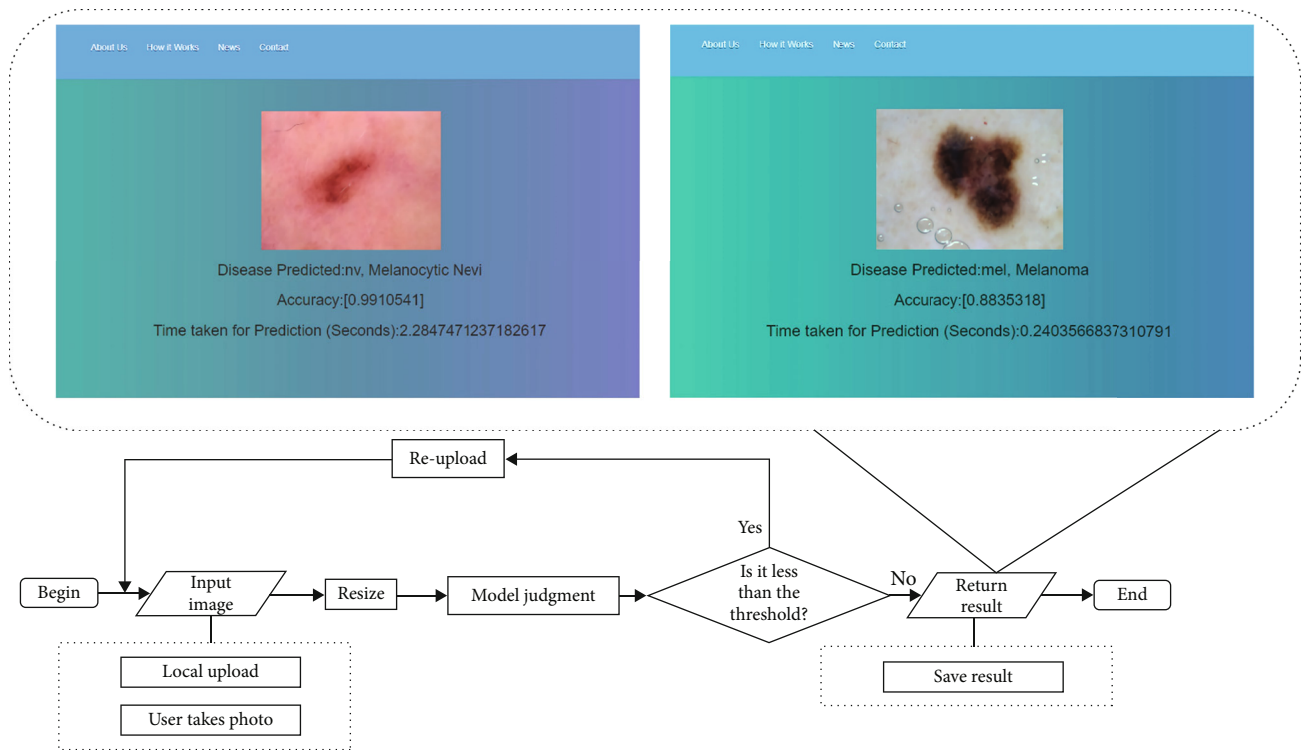


FIGURE 10: Skin lesion classification system.

TABLE 8: Skin lesion classification system results.

Groups	Number of correct predictions	Number of wrong predictions	Success rate (%)	Most confused lesion
1	93	7	93	Df
2	89	11	89	Df, Bkl
3	91	9	91	Nv, Df
4	93	7	93	Df
5	89	11	89	Df, Bkl
6	93	7	93	Df
	548	52	91	

our model can provide accurate diagnostic information for experts or patients.

From Table 8, we find that the test success rate of the model is 91%, which is roughly in line with the training accuracy of the model. The most misclassified lesion type observed in the test was Df, and it was also noted that Bkl and Nv were confused. The possible reason is that the color of the lesion region of Df and Bkl is similar to the skin color, making it difficult to classify correctly. As Nv is a common benign nevus with different color and shapes of growths, it is the most numerous in the dataset and the most easily confused with other skin lesions. In the ISIC2018 dataset, the pathological symptoms of Nv are similar to other categories, and Nv itself is asymmetric, irregular, and rough. Therefore, the classification success rate of Nv is also low. In addition, the classification success rate was lower in the presence of

hair and darker skin color around the lesion, so contrast enhancement of the image and hair removal would be beneficial. In future work, increasing the number of taken dermatological images in the test will also provide more generalized results.

Effective classification of skin lesions can allow patients to go to the hospital in time to improve the possibility of survival. In summary, our model has good performance on the skin lesion multiclassification task.

5. Discussion

The main contribution of this work is designing a two-stream network for multiscale feature fusion for skin lesion classification, which achieves good performance on the highly unbalanced seven-class dataset ISIC2018. The performance of our model in precision, recall, accuracy, and macroaverage is 83.53%, 95.04%, 91.24%, and 98%, respectively. We use the improved VGG-16 structure by using residual structures. Residual structures are fused before and after each layer. Thus, the features of the previous layer can be transported to the next layer through identity mapping and residual mapping. Another network model DenseNet-121 enhances the propagation of features by designing densely connected dense blocks, reduces the number of parameters of the network, and alleviates the gradient dispersion problem caused by the overly deep network model, thus improving the classification rate of the deep neural network. In the future, we will use different data augmentation methods to train and combine more models to

better achieve the classification performance to meet the medical diagnosis needs. The two-stream network we chose is the DenseNet-121 and improved VGG-16. The model is not a lightweight network, and we will use lightweight network to design the experiment later, which is a limitation of our method. And we only did the experiments on ISIC2018; later, we will choose more ISIC datasets and experiments using real data in the clinic.

6. Conclusions

Malignant skin lesions have a high mortality rate and have high interclass similarity and intraclass variations. Therefore, a reliable classification system would be of great help to clinicians in the early detection of malignant skin lesions. In this paper, we propose a multiscale feature fusion model for skin lesion classification. We use DenseNet-121 and an improved VGG-16 network as our two-stream network to complement the advantages of a single network. Noteworthy, we fuse the residual structure of the original VGG-16 model to optimize the model without adding parameters. Then, we exploit the feature fusion module to obtain multiscale pathological information. In summary, our model achieves 91.24% test accuracy and 95% macroaverages on the ISIC2018 dataset. Finally, we design a skin lesion classification system in our two-stream network to help physicians effectively classify a patient's early-stage skin lesions.

Therefore, improving the ability to automatically classify based on skin lesion images is necessary to help physicians classify skin lesions and assist in early medical diagnosis. In the future, we will try to design a complete auxiliary diagnosis system based on our model.

Data Availability

The datasets generated or analysed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

Commons Attribution License, which permits use, distribution, and reproduction in any medium, provided in the original work is properly cited.

Acknowledgments

This work was supported by the University Synergy Innovation Program of Anhui Province under Grant GXXT-2019-007; by the National Natural Science Foundation of China under Grant 61901006, Grant 62105002, and Grant 62001004; by the Anhui Provincial Natural Science Foundation under Grant 1908085QF281 and Grant 2008085MF182; by the Anhui Provincial DOHURD Science Foundation under Grant 2020-YF22; by the Educational Commission of Anhui Province of China under Grant KJ2020A0471; and by the Key Research Project of Natural Science in Anhui Province under Grant 2022AH050249.

References

- [1] R. Kasmi and K. Mokrani, "Classification of malignant melanoma and benign skin lesions: implementation of automatic ABCD rule," *IET Image Processing*, vol. 10, no. 6, pp. 448–455, 2016.
- [2] R. J. Hay, N. E. Johns, H. C. Williams et al., "The global burden of skin disease in 2010: an analysis of the prevalence and impact of skin conditions," *Journal of Investigative Dermatology*, vol. 134, no. 6, pp. 1527–1534, 2014.
- [3] R. Z. Conic, C. I. Cabrera, A. A. Khorana, and B. R. Gastman, "Determination of the impact of melanoma surgical timing on survival using the national cancer database," *Journal of the American Academy of Dermatology*, vol. 78, pp. 40–46, 2018.
- [4] R. P. Braun, H. S. Rabinovitz, M. Oliviero, A. W. Kopf, and J. H. Saurat, "Dermoscopy of pigmented skin lesions," *Journal of the American Academy of Dermatology*, vol. 52, no. 1, pp. 109–121, 2005.
- [5] I. Bakkouri and K. Afde, "Correction to: Computer-aided diagnosis (CAD) system based on multi-layer feature fusion network for skin lesion recognition in dermoscopy images," *Multimed Tools and Applications*, vol. 79, no. 29–30, pp. 20519–20519, 2020.
- [6] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2921–2929, Las Vegas, NV, USA, 2016.
- [7] K. M. Hosny and M. A. Kassem, "Refined residual deep convolutional network for skin lesion classification," *Journal of Digital Imaging*, vol. 35, no. 2, pp. 258–280, 2022.
- [8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, USA, 2016.
- [9] A. Krizhevsky, I. Sutskever, and G. Hinton, "Image Net classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.
- [10] C. Szegedy, W. Liu, Y. Jia et al., "Going deeper with convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, Boston, MA, USA, 2015.
- [11] S. P. Xavier, R. Edgar, and D. S. Angel, "Dense extreme inception network: towards a robust CNN model for edge detection," in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1912–1921, Snowmass, CO, USA, 2020.
- [12] W. Shuihua and Z. Yu-Dong, "DenseNet-201-based deep neural network with composite learning factor and precomputation for multiple sclerosis classification," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 16, no. 2s, pp. 1–19, 2020.
- [13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *In Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*, San Diego, CA, 2015, abs/1409.1556.
- [14] M. A. Khan, M. I. Sharif, M. Raza, A. Anjum, T. Saba, and S. A. Shad, "Skin lesion segmentation and classification: a unified framework of deep neural network features fusion and selection," *Expert Systems*, vol. 39, no. 7, article e12497, 2022.
- [15] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, Las Vegas, NV, USA, 2016.

- [16] G. Huang, Z. Liu, D. M. L. Van, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, 2017.
- [17] R. N. Swetha, V. Shrivastava, and K. Parvathi, "Multiclass skin lesion classification using image augmentation technique and transfer learning models," *International Journal of Intelligent Unmanned Systems*, 2021.
- [18] A. Soujanya and N. Nandhagopal, "Automated skin lesion diagnosis and classification using learning algorithms," *Intelligent Automation and Soft Computing*, vol. 35, no. 1, pp. 675–687, 2023.
- [19] N. Gessert, T. Sentker, F. Madesta et al., "Skin lesion diagnosis using ensembles, unscaled multi-crop evaluation and loss weighting," *Computer Vision and Pattern Recognition*, 2018, <http://arxiv.org/abs/1808.01694>.
- [20] A. H. Shahin, A. Kamal, and M. A. Elattar, "Deep ensemble learning for skin lesion classification from dermoscopic images," in *In: IEEE 9th Cairo international biomedical engineering conference*, pp. 150–153, Cairo, Egypt, December 2018.
- [21] A. Mahbod, G. Schaefer, C. Wang, G. Dorffner, R. Ecker, and I. Ellinger, "Transfer learning using a multi-scale and multi-network ensemble for skin lesion classification," *Computer Methods and Programs in Biomedicine*, vol. 193, p. 105475, 2020.
- [22] M. A. Al-Masni, D. H. Kim, and T. S. Kim, "Multiple skin lesions diagnostics via integrated deep convolutional networks for segmentation and classification," *Computer Methods and Programs in Biomedicine*, vol. 190, p. 105351, 2020.
- [23] R. Zillur, H. Sabir, I. Rabiul, M. M. Hasan, and R. A. Hridhee, "An approach for multiclass skin lesion classification based on ensemble learning," *Informatics in Medicine Unlocked*, vol. 25, article 100659, 2021.
- [24] O. O. Abayomi-Alli, R. Damasevicius, S. Misra, R. Maskeliunas, and A. Abayomi-Alli, "Malignant skin melanoma detection using image augmentation by oversampling in nonlinear lower-dimensional embedding manifold," *Turkish Journal of Electrical Engineering and Computer Sciences.*, vol. 29, no. 8, pp. 2600–2614, 2021.
- [25] M. Nawaz, T. Nazir, M. Masood et al., "Melanoma segmentation: a framework of improved DenseNet77 and UNET convolutional neural network," *International Journal of Imaging Systems and Technology*, vol. 32, no. 6, pp. 2137–2153, 2022.
- [26] S. Kadry, D. Taniar, R. Damaševičius, V. Rajinikanth, and I. A. Lawal, "Extraction of abnormal skin lesion from dermoscopy image using VGG-SegNet," in *2021 Seventh International conference on Bio Signals, Images, and Instrumentation (ICBSII)*, pp. 1–5, Chennai, India, March 2021.
- [27] V. Rajinikanth, S. Kadry, R. Damaševičius, D. Sankaran, M. A. Mohammed, and S. Chander, "Skin melanoma segmentation using VGG-UNet with Adam/SGD optimizer: a study," in *Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT)*, pp. 982–986, Kannur, India, August 2022.
- [28] R. C. Hardie, R. Ali, M. S. De Silva, and T. M. Kebede, "Skin lesion segmentation and classification for ISIC 2018 using traditional classifiers with hand-crafted features," 2018, <http://arxiv.org/abs/1807.07001>.
- [29] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, no. 1, p. 180161, 2018.
- [30] M. S. Ali, M. S. Miah, J. Haque, M. M. Rahman, and M. K. Islam, "An enhanced technique of skin cancer classification using deep convolutional neural network with transfer learning models," *Machine Learning with Applications*, vol. 5, article 100036, 2021.
- [31] P. Yan, G. Wang, J. Chen, Q. Tang, and H. Xu, "Skin lesion classification based on the VGG-16 fusion residual structure," *International Journal of Imaging Systems and Technology*, pp. 1–16, 2022.
- [32] T. Alexandru, "An image inpainting technique based on the Fast Marching method," *Journal of Graphics Tools*, vol. 9, no. 1, pp. 23–34, 2004.
- [33] R. Filip, T. Giorgos, and C. Ondřej, "Fine-tuning CNN image retrieval with no human annotation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1655–1668, 2019.
- [34] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *British Machine Vision Conference, BMVC 2009*, London, UK, 2009.
- [35] L. Rampasek and A. Goldenberg, "TensorFlow: Biology's Gateway to Deep Learning?," *Cell Systems*, vol. 2, no. 1, pp. 12–14, 2016.
- [36] S. Ali, S. Miah, J. Haque, M. Rahman, and K. Islam, "An enhanced technique of skin cancer classification using deep convolutional neural network with transfer learning models," *Machine Learning with Applications*, vol. 5, pp. 2666–8270, 2021.
- [37] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, 2020.
- [38] P. Yao, S. Shen, M. Xu et al., "Single model deep learning on imbalanced small datasets for skin lesion classification," *IEEE Transactions on Medical Imaging*, vol. 41, pp. 1242–1254, 2021.
- [39] P. Tschandl, N. Codella, B. N. Akay et al., "Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study," *The Lancet Oncology*, vol. 20, no. 7, pp. 938–947, 2019.