Received August 8, 2020, accepted August 20, 2020, date of publication August 24, 2020, date of current version September 4, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3019210

Classification for Dermoscopy Images Using Convolutional Neural Networks Based on the Ensemble of Individual Advantage and Group Decision

AN GONG[1], XINJIE YAO[1], AND WEI LIN[2,3]

[1]College of Computer Science and Technology, China University of Petroleum (East China), Qingdao 266580, China
[2]School of Engineering Science, University of Chinese Academy of Sciences, Beijing 100049, China
[3]Department of Earth and Planetary Science, University of California at Berkeley, Berkeley, CA 94720, USA

Corresponding author: Xinjie Yao (z18070013@s.upc.edu.cn)

This work was supported in part by the National Science and Technology Major Project of China under Grant 2017ZX05013-001, in part by the Major Scientific and Technological Projects of CNPC under Grant ZD2019-183-004, and in part by the Fundamental Research Funds for the Central Universities under Grant 20CX05019A.

ABSTRACT Skin cancer is a common and deadly cancer. Dermoscopy is an effective tool for the observation of abnormal skin pigmentation. However, dermoscopy images are extremely complex and present great challenges for diagnosis. Therefore, we proposed a classification method based on the ensemble of individual advantage and group decision in dermoscopy images, including the ensemble strategy of group decision, the ensemble strategy of maximizing individual advantage, and the ensemble strategy of block-integrated voting. We used generative adversarial networks (GANs) to create a balanced sample space to better train convolutional neural networks (CNNs). Through transfer learning, the pre-training CNNs were used for fine-tuning, then the effects of different CNNs on the classification of different categories of dermoscopy images were compared, and the CNNs with better classification effect were selected for the ensemble of different strategies. This study is based on the ISIC 2018 dataset and ISIC 2019 dataset. Compared with the different individual CNNs and the frameworks, the proposed ensemble strategies achieve a better improvement in the evaluation criteria.

INDEX TERMS Dermoscopy images, ensemble strategies, convolutional neural networks, transfer learning.

I. INTRODUCTION

Skin cancer is one of the most common cancers in humans and is easily confused with other skin diseases. Skin cancer is particularly common in the United States [1]–[3]. Nearly 5 million adults are treated for skin cancer annually, with average treatment costs of $8.1 billion each year [1]. In the United States, 95,830 cases of melanoma are newly diagnosed in 2019 [2]. Skin cancer is a serious problem for the world. The skin cancer rates is higher as compared to other cancers [4]. In Australia, more than 14,000 new cases of melanoma are reported yearly, leading to nearly 2,000 deaths [5]. In Europe, over 100,000 new melanoma cases and 22,000 melanoma related deaths are reported annually [6]. Malignant melanoma is the deadliest form of skin cancer, which accounting for 79% of skin cancer deaths [7]–[9]. Early diagnosis is of great importance for treating skin cancer as it can be cured better at early stages [7]–[10]. Considerable attention has been paid to dermoscopy research in the field of dermatological research. Cascinelli *et al.* [11] first applied dermoscopy technique to the clinical diagnosis of malignant melanoma. Dermoscopy is currently one of the most effective tools to assist dermatologists in their diagnosis and is gradually being introduced into clinical diagnosis [12], [13], but it is highly dependent on the clinical experience of dermatologists and the dermoscopy images themselves are extremely complex, as shown in Fig. 1.

Such as intra-class variation, inter-class similarity, and blurring of the boundary of skin lesions have a great influence on the diagnosis. Therefore, computer aided diagnosis (CAD) has gradually become the focus of research. Early dermoscopy image classification methods

The associate editor coordinating the review of this manuscript and approving it for publication was Jiju Poovvancheri.

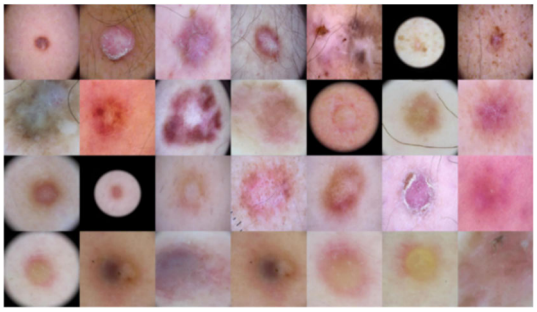

FIGURE 1. Examples of dermoscopy images.

usually focused on the manual extraction of features [14]–[19]. Kusumoputro and Ariyanto [20] extracted shape and color features from dermoscopy images and used artificial neural network to classify malignant melanomas. Schaefer *et al.* used an automatic border detection approach [21] and assembled the extracted features for melanoma recognition [22].

The problem faces a number of challenges due to interclass similarities within skin cancer and large intra-class variations in background, color, and illumination. In addition, the method of extracting features manually is complex and requires a lot of manpower. The extraction process is unstable and the generalization ability is limited. With the development of deep learning, convolutional neural networks (CNNs) have an outstanding performance in image processing and powerful generalization ability, including but not limited to segmentation, classification and detection. CNNs are able to learn multilevel features from original data, and the extracted features are more high-level and more robust. Many researchers established a system that combined recent developments in deep learning and machine learning for skin lesion segmentation and classification [23]–[28]. Schaefer *et al.* [23] segmented the area of the lesionusing an approach based on thresholding, region growingand region merging, and the extracted features are analysed in apattern classification stage. Demyanov *et al.* [24] employed a convolutional neural network with 8 layers to solve the problem of dermoscopy pattern classification. Kawahara *et al.* [25] gained consistent additional improvements toaccuracy using a per image normalization, a fully convolu-tional network to extract multi-scale features, and by poolingover an augmented feature space. In addition, ensembles and attention mechanisms were also considered in the image classification and segmentation of skin lesion [29]–[31]. Gessert *et al.* [29] combined the crops both by simple averaging and a meta learning strategy to improve the accuracy of model classification. Wei *et al.* [31] proposed a novel Attention Based DenseUnet network with adversarial training for skin lesion segmentation. Although some research has been done on the computer-aided diagnosis, the problem of insufficient medical samples and uneven distribution still exists. To alleviate the above problems, researchers have used transfer learning and generative adversarial networks (GANs), which have achieved excellent results in CAD [32]–[38]. Ghazi *et al.* [33] used transfer learning to fine-tune the three powerful and popular deep learningarchitectures, namely GoogLeNet, AlexNet, and VGGNet. Through transfer learning, CNNs can converge faster. Burlina *et al.* [37] used GANs to generate fundus images of retinal diseases and used deep learning for discriminative tasks in ophthalmology. However, few dermoscopy studies have been proposed to use transfer learning and GANs. In addition, the loss function can help the CNNs to learn effective information more accurately and pay more attention to special samples. Yang *et al.* [39] presented to penalize the loss function with danger samples to enable the CNN to pay more attention to danger samples. Different from the extensively studied lesion classification, dermoscopy feature extraction is a new task in the area. Traditional ensemble strategies limit the advantages of each CNN, and each CNN has different recognition capabilities for different categories of images. Most research focuses on parameter tuning and CNN structure, and there is still much room for improvement on multiple evaluation criteria.

Based on the above analysis, we proposed a classification method based on the ensemble of individual advantage and group decision to automatically classify the dermoscopy image. The main contributions of our work can be summarized as follows:

1) We propose a classification method based on the ensemble of individual advantage and group decision, including the ensemble strategy of group decision, the ensemble strategy of maximizing individual advantage, and the ensemble strategy of block-integrated voting. The method can solve the problem of a individual convolutional neural network (CNN) in solving multiple problems. Different from the traditional voting method, our method only performs binary classification voting on one dermoscopy image at a time. Our method is more robust and stable than the different individual CNNs and the traditional ensemble strategies. The basic principle of the method we propose is simply to use the idea of divide and conquer. The complex classification problem is divided into several simple binary classification problems, and finally the dermoscopy image classification is completed through different strategies.
2) We train multiple CNNs based on transfer learning. Through transfer learning, the pre-training CNNs are used for fine-tuning, then the effects of different CNNs on the classification of different categories of dermoscopy images are compared, and the CNNs with better classification effect are selected for the ensemble of different strategies. The experiment finds that transfer learning has excellent performance on the dermoscopy image dataset.
3) We use the method of GANs to create balanced sample space and combine rotation, scaling, and cropping for data augmentation to improve the training effect of CNNs.

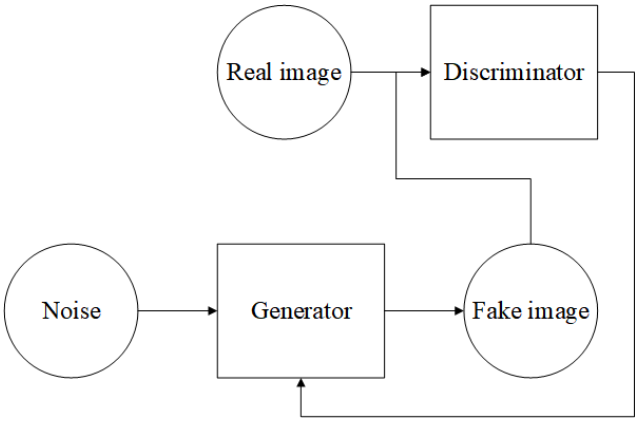

FIGURE 2. The flow chart of the GANs.

II. MATERIALS AND METHODS

A. DATA AUGMENTATION

We collected 10,015 dermoscopy images from the International Skin Imaging Collaboration (ISIC) 2018 dataset. From the 10,015 images, we selected all images as our experimental dataset, which include 7 skin diseases: actinic keratosis (AKIEC), basal cell carcinoma (BCC), benign keratosis (BKL), dermatofibroma (DF), melanoma (MEL), melanocytic nevus (NV) and vascular lesion (VASC). The outstanding feature extraction capabilities of CNNs require large and balanced dataset. The problem of insufficient medical dataset and uneven distribution poses an important challenge, which will reduce the training effect of CNNs and easily lead to over-fitting. In view of the uneven distribution of the dataset, we divide the training set into a balanced sample space. By using GANs to generate some fake sample images to expand the categories with fewer samples. The flow chart of the GANs is shown in Fig. 2. Noise is used as the input of the generator. The generator attempts to generate a fake image that is sufficiently similar to the real image, making it difficult for the discriminator to correctly distinguish the authenticity of the image.

If the discriminator can effectively distinguish the authenticity of the image, then the effect of the generator needs to be improved, and the parameters need to be adjusted. The training of the generator and discriminator can be defined as:

$$\begin{aligned} &\max_D V(D,G) \\ &= E_{x\sim P_{data}(x)}[log(D(x))] + E_{z\sim P_z(z)}[log(1-D(G(z)))] \end{aligned} \tag{1}$$

$$\begin{aligned} &\max_G E_{z\sim P_z(z)}[log(D(G(z)))] \\ &\Longleftrightarrow \min_G E_{z\sim P_z(z)}[log(1-D(G(z)))] \end{aligned} \tag{2}$$

$$\begin{aligned} &\min_G V(D,G) \\ &= E_{z\sim P_z(z)}[log(1-D(G(z)))] \end{aligned} \tag{3}$$

$$\begin{aligned} &\min_G \max_D V(D,G) \\ &= E_{x\sim P_{data}(x)}[log(D(x))] + E_{z\sim P_z(z)}[log(1-D(G(z)))] \end{aligned} \tag{4}$$

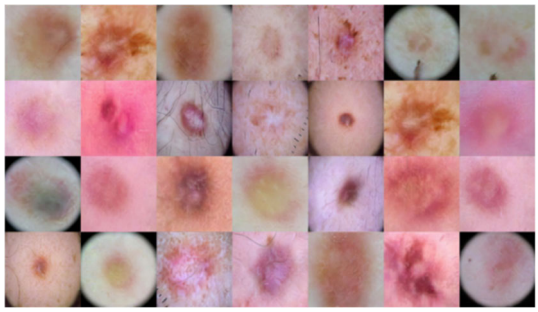

FIGURE 3. Examples of fake dermoscopy images.

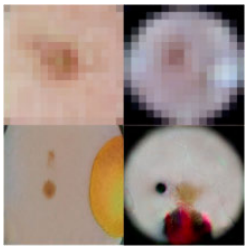

FIGURE 4. Examples of deformed or blurred dermoscopy images.

where D denotes discriminator, G denotes generator, x denotes real data, and z denotes the noise. The most important thing in GANs is to optimize the objective function. The purpose of the discriminator is to make that the probability of $D(G(z))$ is true as small as possible, that is, $1-D(G(z))$ is as large as possible. At the same time, the probability of $D(x)$ is true as large as possible. The purpose of the generator is to make that the probability of $D(G(z))$ is true as large as possible, that is, $1-D(G(z))$ is as small as possible. Because the samples generated by GANs are fake, we need to pick out more realistic or useful images for our training, as shown in Fig. 3.

After constructing the balanced sample space, we use GANs, rotation, scaling, and cropping methods to augment the dataset. The processed dataset is larger and more balanced than before. Compared with ISIC 2018 dataset, ISIC 2019 dataset added squamous cell carcinoma (SCC) with a total of 25,331 dermoscopy images. For the ISIC 2019 dataset, we also use the above processing method. GANs may generate deformed or blurred dermoscopy images, as shown in Fig. 4.

We need to pick as clear and undistorted images as possible. Because these images are relatively good for us to train CNNs. The reason for using the dataset for two consecutive years is to better verify the generalization of our proposed method and the feasibility of transfer learning on dermoscopy images.

B. IMAGE SIZE

The size and integrity of the image have an important effect on the feature extraction of CNNs. Relatively complete dermoscopy images are of great significance to CNNs. Most cropping causes skin lesions to deform, but shape contour information is an important basis for extracting features.

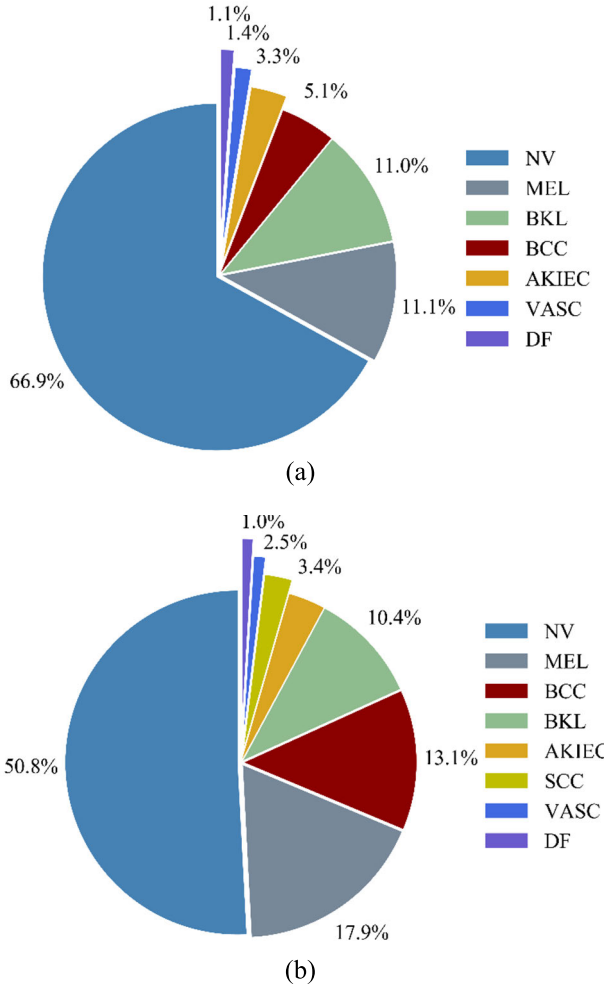

FIGURE 5. Dermoscopy image distribution of different categories. (a) is ISIC 2018 dataset, and (b) is ISIC 2019 dataset.

When images of different sizes are used as training set, the training time is also different. Therefore, we consider the time cost and the effect of feature extraction, and use 600×600 as the image size.

C. WEIGHTED LOSS FUNCTION

Fig. 5 shows the uneven distribution of dermoscopy images. Insufficient medical samples make it difficult to extract key features, and uneven distribution of samples easily leads to overfitting. To solve this problem, we use transfer learning and construct a weighted loss function.

We use CNNs based on pre-trained models on ImageNet. With the help of transfer learning, we fine-tune the pre-trained model and modify the last fully connected layer for training. Because we use ISIC 2018 dataset and ISIC 2019 dataset, we first use the pre-trained model based on ImageNet to train CNNs based on the ISIC 2018 dataset. Then we use the pre-trained model based on ISIC 2018 dataset to train CNNs based on ISIC 2019 dataset. Although we constructed a relatively balanced sample space, the generated samples are limited in terms of feature extraction. Considering the uneven distribution of dermoscopy image dataset, this experiment sets the weight coefficient in the cross entropy loss function to multiply the larger sample by the smaller weight and the fewer sample by the larger weight, so as to alleviate the problem of uneven distribution in the dataset. Hence, the weighted cross entropy loss function is defined as:

$$\text{loss} = -\frac{1}{m}\left[\sum_{j=1}^{m}\sum_{i=1}^{n} w_i p_{ji} log\left(q_{ji}\right)\right] \tag{5}$$

where $w_i = \mathrm{N}/\mathrm{N_i}$ denotes weight of loss function, N denotes total number of samples, and $\mathrm{N_i}$ denotes the number of samples for class i, p_{ji} denotes the desired output, q_{ji} denotes the actual output, m denotes the number of samples in batches, n denotes the number of categories of dermoscopy images.

D. IMAGE NORMALIZATION

This study normalized the dermoscopy images to reduce the interference of uneven light in dermoscopy images. At present, image normalization methods often use the image pixel value subtracts the average pixel value of the entire dataset. However, the illumination and observation angles of different dermoscopy images in the dataset are different. There are certain limitations in subtracting the uniform average value of the whole dataset, so the standardization effect of a individual image is not ideal. In view of the above problems, this experiment normalizes each dermoscopy image by subtracting the channel average intensity value calculated for a individual image from the image. By giving a dermoscopy image, the normalized image is calculated and defined as:

$$\begin{cases} \mathrm{X_{norm,R}} = \mathrm{X_R} - \mathrm{u}\left(\mathrm{X_R}\right) \\ \mathrm{X_{norm,G}} = \mathrm{X_G} - \mathrm{u}\left(\mathrm{X_G}\right) \\ \mathrm{X_{norm,B}} = \mathrm{X_B} - \mathrm{u}\left(\mathrm{X_B}\right) \end{cases} \tag{6}$$

where $\mathrm{u}\left(\mathrm{X_R}\right)$, $\mathrm{u}\left(\mathrm{X_G}\right)$, and $\mathrm{u}\left(\mathrm{X_B}\right)$ denote the average pixel values of the 3 color channels respectively.

E. THE ENSEMBLE STRATEGY OF GROUP DECISION

In order to take advantage of group decision, we proposed the ensemble strategy of group decision, the flow chart of which is presented in Fig. 6.

Except for the large sub-modules containing m CNNs, each of the small sub-modules only performs binary classification for each corresponding dermoscopy category. Among them, each small sub-module contains n CNNs. The result of each small sub-module after voting is a minimum of 0 and a maximum of n. The maximum value of the voting result of the small sub-modules is compared with the threshold (If the small sub-module of the maximum value of the voting result is not unique, the small sub-module with the highest probability is selected). If the threshold is satisfied, the category of dermoscopy image corresponding to the maximum value of the voting result is output as a prediction result. Otherwise, voting is performed by a large sub-module including m CNNs, and the category of dermoscopy image corresponding to the maximum value of the voting result is output as a

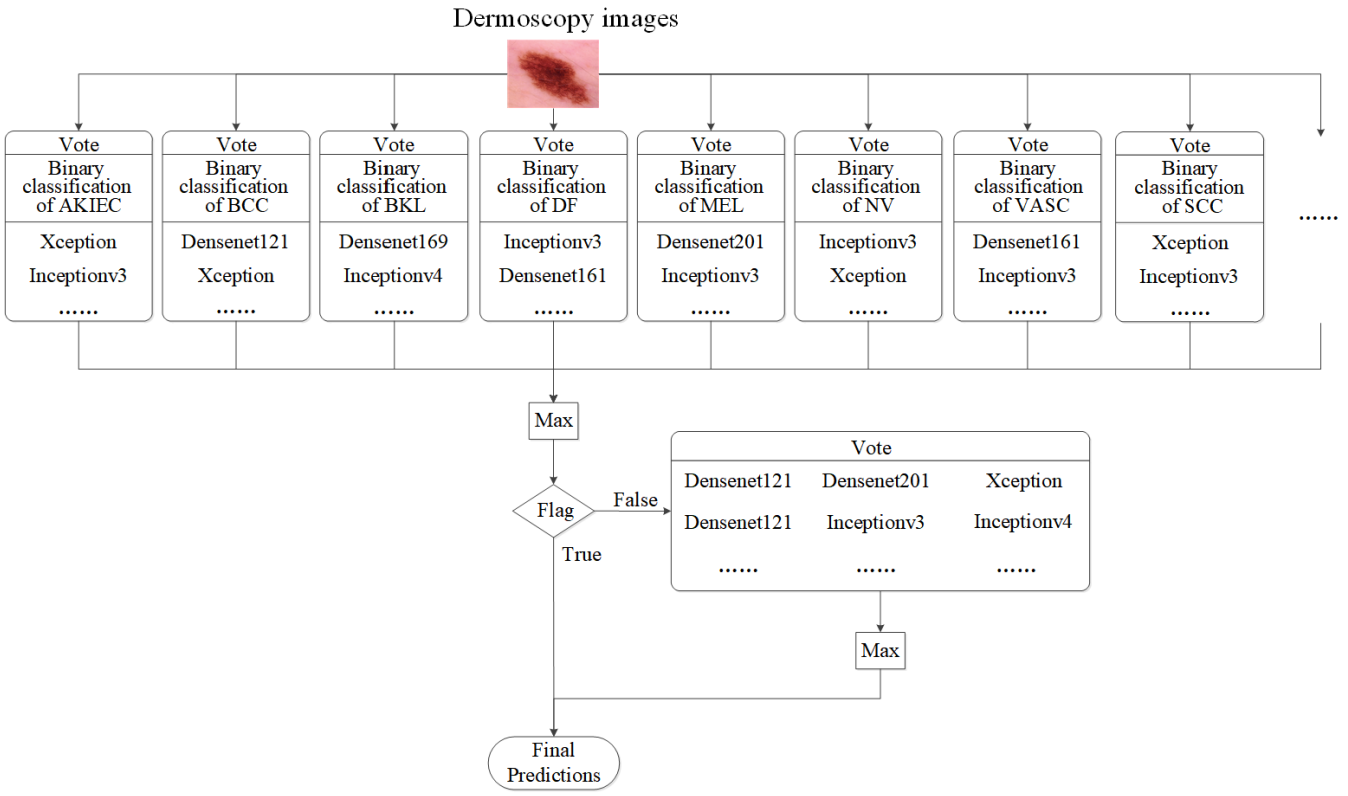

FIGURE 6. The flow chart of the ensemble strategy of group decision.

prediction result (If the category of dermoscopy image corresponding to the maximum value of the voting result is not unique, the category of dermoscopy image with the highest probability is selected). The threshold is 2. We select 9 CNNs as a group. Our large sub-module contains 9 CNNs, and each of the small sub-modules contains 3 CNNs. The CNNs in the small sub-modules are selected based on the classification effect of the dermoscopy images. The first three CNNs with better classification effect of the corresponding dermoscopy image category in the group were put into the corresponding small sub-module. The pseudocode of the ensemble strategy of group decision is shown in Algorithm 1.

F. THE ENSEMBLE STRATEGY OF MAXIMIZING INDIVIDUAL ADVANTAGE

In order to further reduce the computational cost and better exert the effect of different CNNs on the classification of different categories of dermoscopy images, we used more recent CNNs. Based on these CNNs, the ensemble strategy of maximizing individual advantage is proposed. The flow chart is presented in Fig. 7.

The ensemble strategy of maximizing individual advantage consists of small sub-modules, each of which only performs binary classification for each corresponding dermoscopy category. Among them, each sub-module contains n CNNs. The result of each sub-module after voting is a minimum of 0 and a maximum of n, and the category of dermoscopy image corresponding to the maximum value of the voting results of the small sub-modules is output as a prediction result (If the sub-module of the maximum value of the voting result is not unique, the sub-module with the highest probability is selected). We select 9 CNNs as a group. Each sub-module of our experiment contains 3 CNNs. The CNNs in the sub-modules are selected based on the classification effect of the dermoscopy images. The first three CNNs with better classification effect of the corresponding dermoscopy image category in the group were put into the corresponding sub-module. The pseudocode is shown in Algorithm 2.

G. THE ENSEMBLE STRATEGY OF BLOCK-INTEGRATED VOTING

In order to better play the role of blocks for different categories of dermoscopy image classification, we proposed the ensemble strategy of block-integrated voting based on multiple CNNs. The flow chart is presented in Fig. 8.

The ensemble strategy of block-integrated voting is to integrate m different CNNs into one block (The CNNs in each block are preferably not combined repeatedly with each other). Each block performs binary classification on different categories of dermoscopy images, and the category of dermoscopy image corresponding to the maximum value of the voting result is added as a prediction result to the next block (If the category of dermoscopy image corresponding to the maximum value of the voting result is not unique, the category of dermoscopy image with the highest probability is selected). After a number of blocks are accumulated,

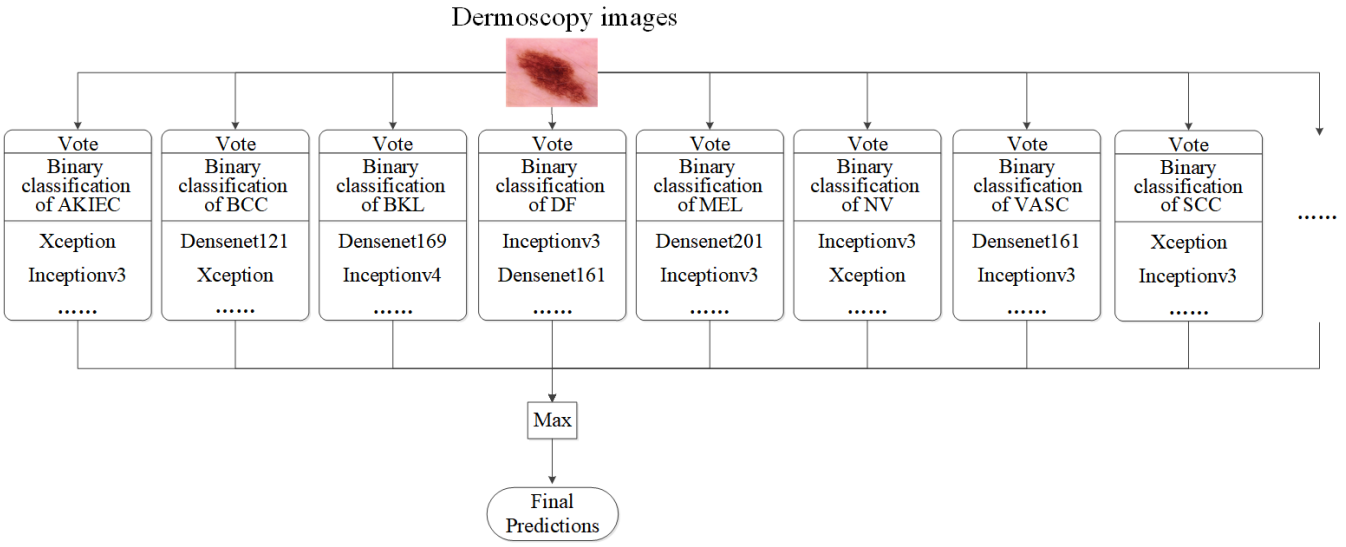

FIGURE 7. The flow chart of the ensemble strategy of maximizing individual advantage.

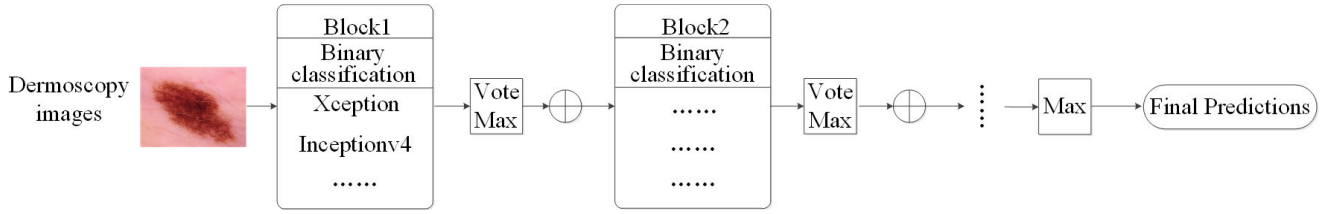

FIGURE 8. The flow chart of the ensemble strategy of block-integrated voting.

the category of dermoscopy image corresponding to the maximum value of the cumulative voting result is selected as the predicted result output (If the category of dermoscopy image corresponding to the maximum value of the voting result is not unique, the category of dermoscopy image with the highest probability is selected). We select 9 CNNs as a group. 3 different CNNs were selected for each block in our experiment, and accumulated through 84 blocks. CNNs in each block are selected by combining formulas, which is defined as:

$$C_n^m = \frac{n!}{(n-m)!m!} \tag{7}$$

where n and m denote the total number of CNNs per group and the total number of CNNs in each block. The pseudocode is shown in Algorithm 3.

III. EXPERIMENT RESULTS AND ANALYSIS

Experimental images are from ISIC 2018 and ISIC 2019 Skin Lesion Analysis Towards Melanoma Detection Challenge dataset [40]–[42]. For the optimizer, Adam was used as the optimizer. Experiments were implemented with PyTorch framework on a computer equipped with four NVIDIA Tesla P100 GPUs with 16GB of memory. The CNNs performed 100 epoch of training. In the experiment, the batch size was 64, the learning rate was 0.0001, and the learning rate decay was 0.00001. In order to ensure the stability and fidelity of the experiment, We used ten-fold cross-validation technique on the training set to train the model and independent test to evaluate the performance. The independent dataset was re-used to evaluate the performance accuracy to avoid any systematic bias in the cross-validation set. Our dataset is from ISIC, and the dermoscopy images are classified by our proposed method. Take the dermoscopy image as input, and then output the category to which the dermoscopy image belongs. Since the ensemble strategy of group decision and the ensemble strategy of maximizing individual advantage contain multiple small sub-modules, each small sub-module performs binary classification for the corresponding dermoscopy image category. Therefore, we need to select CNNs that have a better classification effect in the corresponding dermoscopy image category. For example, the first small sub-module is for AKIEC, and we need to select CNNs with better criteria such as accuracy in the binary classification of AKIEC. The ensemble strategy of block-integrated voting accumulates the results of multiple blocks and then makes a comprehensive decision. Therefore, the ensemble strategy of block-integrated voting pays more attention to the overall classification effect of CNNs. We need to select CNNs that have good performance in the average value of each criterion. Different CNNs have different classification effects, so we select different CNNs as much as possible based on the classification effects.

Algorithm 1 The Ensemble Strategy of Group Decision

```
for number of categories of dermoscopy images do
  for each small sub-module contains the number of
    CNNs do
    if judge_kj (x) = k then
      count[k]=count[k]+1;
//k represents a category of dermoscopy image, and
judge_kj (x)
//performs binary classification on the input image x.
for i=0;i<count.length;i++ do
  if count[i]>max then
    max=count[i];
    max_category=i;
//max refers to the largest number of votes in the small sub-
//modules.
// max_category refers to the category corresponding to
max.
if max>flag then
  return max_category;
else
  for number of categories of dermoscopy images do
    for large sub-module contains the number of CNNs
      do
    if judge_klarge (x) =k then
      count_large[k]=count_large[k]+1;
  for i=0;i<count_large.length;i++ do
    if count_large[i]>max_large then
      max_large=count_large[i];
      max_category=i;
return max_category;
```

Algorithm 2 The Ensemble Strategy of Maximizing Individual Advantage

```
for number of categories of dermoscopy images do
  for each small sub-module contains the number of
    CNNs do
    if judge_kj (x) = k then
      count[k]=count[k]+1;
//k represents a category of dermoscopy image, and
judge_kj (x)
//performs binary classification on the input image x.
for i=0;i<count.length;i++ do
  if count[i]>max then
    max=count[i];
    max_category=i;
//max refers to the largest number of votes in the small sub-
//modules.
//max_category refers to the category corresponding to
max.
  return max_category;
```

A. EVALUATION CRITERIA

In order to comprehensively evaluate the classification performance of the model, accuracy (ACC), sensitivity (SE), specificity (SP), F_1 and area under ROC curve (AUC) are used as evaluation criteria. In the experiment, the CNN with better classification of dermoscopy images was selected based on the ACC and AUC as the main evaluation criteria. SE and SP are also important criteria in medical diagnosis. SE is also called true positive rate or recall. The higher the value, the lower the probability of missed diagnosis. The SP is also called the true negative rate, and the higher the value, the higher the probability of diagnosis. The criteria are defined as:

$$\mathrm{ACC} = \frac{N_{tp} + N_{tn}}{N_{tp} + N_{fp} + N_{fn} + N_{tn}}, \tag{8}$$

$$\mathrm{SE} = \frac{N_{tp}}{N_{tp} + N_{fn}}, \tag{9}$$

$$\mathrm{SP} = \frac{N_{tn}}{N_{tn} + N_{fp}}, \tag{10}$$

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R}, \tag{11}$$

$$\mathrm{AUC} = \int_0^1 t_{pr}\left(f_{pr}\right) df_{pr} \tag{12}$$

where N_{tp}, N_{tn}, N_{fp}, N_{fn}, P, R, t_{pr} and f_{pr} denote the number of true positive, true negative, false positive, false negative, precision, recall, true positive rate and false positive rate, respectively, and they are all defined on the image level. A lesion image is considered as a true positive if its prediction is lesion; otherwise it is regarded as a false negative. A non-lesion image is considered as a true negative if its prediction is non-lesion; otherwise it is regarded as a false positive.

B. COMPARISON OF CLASSIFICATION EFFECTS UNDER DIFFERENT IMAGE SIZES

CNNs usually use fixed size and square images as input, and different size images can get different feature extraction effect. All images are cropped to the desired size before entering the CNN, but most cropping causes skin lesions to deform, and shape contour information is an important basis for discriminating skin cell damage categories. However, if the input image is large, it will cause the training cost of CNNs to be high. Therefore, we tried a variety of image sizes to the optimal image size, and the experimental results based on ISIC 2019 dataset are shown in Table 1.

Because our proposed method requires multiple CNNs, we need to weigh the classification effect and training cost. As the size increases, the classification effect is better. However, the change in classification effect starting from the size of 600 × 600 is not so obvious. By comparing the experimental results of multiple image sizes as input, we find that images with an image dimension of 600 × 600 suit us better as input. Experiments show that the relatively complete dermoscopy image information is beneficial to feature extraction, and the CNN can be performed better to some extent. The comparison of the size of 600 × 600 and the

TABLE 1. Comparison of different image sizes classification effects based on ISIC 2019 dataset.

Model	Accuracy of different categories								Mean Value				
	AKIEC	BCC	BKL	DF	MEL	NV	VASC	SCC	ACC	AUC	F_1	SE	SP
DenseNet121-800×800	0.909	0.912	0.890	0.927	0.849	0.847	0.919	0.931	0.898	0.872	0.405	0.399	0.957
DenseNet121-600×600	**0.908**	**0.909**	**0.889**	**0.925**	**0.848**	**0.844**	**0.916**	**0.928**	**0.896**	**0.871**	**0.402**	**0.395**	**0.957**
DenseNet121-300×300	0.880	0.863	0.858	0.924	0.768	0.760	0.898	0.922	0.859	0.788	0.310	0.311	0.939
DenseNet121-150×150	0.843	0.799	0.821	0.920	0.667	0.684	0.888	0.889	0.814	0.600	0.165	0.167	0.910
InceptionV3-800×800	0.915	0.909	0.899	0.928	0.857	0.855	0.921	0.932	0.902	0.873	0.418	0.411	0.962
InceptionV3-600×600	**0.910**	**0.907**	**0.897**	**0.926**	**0.853**	**0.854**	**0.918**	**0.929**	**0.899**	**0.869**	**0.414**	**0.407**	**0.959**
InceptionV3-300×300	0.891	0.865	0.848	0.924	0.768	0.757	0.905	0.926	0.861	0.774	0.310	0.289	0.941
InceptionV3-150×150	0.882	0.805	0.806	0.921	0.651	0.615	0.890	0.920	0.811	0.587	0.175	0.137	0.921
ResNet101-800×800	0.909	0.907	0.892	0.928	0.852	0.837	0.918	0.928	0.896	0.857	0.397	0.382	0.955
ResNet101-600×600	**0.906**	**0.904**	**0.888**	**0.924**	**0.840**	**0.833**	**0.917**	**0.928**	**0.893**	**0.856**	**0.391**	**0.376**	**0.954**
ResNet101-300×300	0.888	0.844	0.800	0.922	0.752	0.702	0.895	0.924	0.841	0.759	0.272	0.263	0.932
ResNet101-150×150	0.845	0.765	0.747	0.919	0.738	0.637	0.886	0.917	0.807	0.614	0.163	0.160	0.908
VGG13BN-800×800	0.907	0.902	0.889	0.928	0.841	0.838	0.914	0.927	0.893	0.844	0.382	0.367	0.954
VGG13BN-600×600	**0.904**	**0.896**	**0.887**	**0.924**	**0.835**	**0.835**	**0.913**	**0.927**	**0.890**	**0.841**	**0.379**	**0.364**	**0.952**
VGG13BN-300×300	0.896	0.873	0.847	0.922	0.771	0.770	0.904	0.924	0.863	0.735	0.300	0.278	0.935
VGG13BN-150×150	0.891	0.805	0.764	0.921	0.730	0.631	0.898	0.903	0.818	0.595	0.176	0.182	0.902
Xception-800×800	0.913	0.918	0.897	0.926	0.862	0.865	0.921	0.929	0.904	0.865	0.426	0.418	0.963
Xception-600×600	**0.912**	**0.913**	**0.894**	**0.926**	**0.859**	**0.863**	**0.919**	**0.929**	**0.902**	**0.859**	**0.419**	**0.411**	**0.960**
Xception-300×300	0.894	0.870	0.862	0.925	0.803	0.805	0.902	0.925	0.873	0.773	0.333	0.324	0.942
Xception-150×150	0.888	0.793	0.817	0.920	0.693	0.644	0.898	0.897	0.819	0.601	0.152	0.132	0.929

size of 150 × 150 is the most obvious. InceptionV3 improves 0.088 and 0.282 in ACC and AUC, Xception improves 0.267 and 0.279 in F_1 and SE, and VGG13BN improves 0.050 in SP.

C. COMPARISON OF CLASSIFICATION EFFECTS BASED ON ISIC 2018 DATASET

In the experiment, the effects of different CNNs and strategies on the classification of different categories of dermoscopy images were compared. The experimental results based on ISIC 2018 dataset are shown in Table 2.

By comparing the effects of 43 different CNNs on the classification of 7 different categories of dermoscopy images, it is found that the best average accuracy of the CNN is not necessarily the best in the classification of different categories of dermoscopy images. InceptionV4 performs best on average accuracy, but DenseNet201 is 0.002 higher than InceptionV4 on DF. InceptionResNetV2, DenseNet121 and SENet154 perform better than InceptionV4 on BKL, and InceptionResNetV2, InceptionV3, DenseNet161, SE_ResNeXt101_32×4d and VGG11BN perform more outstanding on AKIEC. According to Table 2, the ACC and AUC of classification of different dermoscopy image categories are used as the main evaluation criteria, and the CNNs with better classification of dermoscopy images are selected. Based on the consideration of computational cost, two sets of CNNs are selected in the experiment. The first group includes DenseNet121, DenseNet161, DenseNet169, DenseNet201, InceptionV3, InceptionV4, InceptionResNetV2, SE_ResNeXt50_32 × 4d and Xception. The second group includes AlexNet, ResNet18, ResNet50, ResNet101, ResNet152, SENet154, SqueezeNet1_0, VGG13BN and VGG16BN. IA indicates the ensemble strategy of maximizing individual advantage. GD represents the ensemble strategy of group decision. VWB denotes the ensemble strategy of block-integrated voting. AVR denotes the ensemble strategy of averaging. VT represents the ensemble strategy of voting. AVR and VT are common traditional ensemble strategies. According to Table 2, the improvement of VWB on each criterion is more obvious, but the improvement of GD on AUC is better.

D. COMPARISON OF CLASSIFICATION EFFECTS BASED ON ISIC 2019 DATASET

In order to better compare the performance of three different ensemble strategies, we also performed experiments on the ISIC 2019 dataset and compared it with the different individual CNNs and the traditional ensemble strategies. The experimental results based on ISIC 2019 dataset are shown in Table 3. In order to compare the impact of choosing different numbers of CNNs on our proposed method based on ISIC 2019 dataset, we took DenseNet169, DenseNet201, InceptionV3, InceptionV4, InceptionResNetV2 from the first group as the third group, and took ResNet152, SENet154, SqueezeNet1_0, VGG13BN, VGG16BN from the second group as the fourth group. In the two groups, our large sub-module contains 5 CNNs, and each of the small sub-modules contains 3 CNNs. 3 different CNNs were selected for each block in our experiment, and accumulated through 10 blocks. The method of selecting CNNs is explained in Section II. Because the ISIC 2019 dataset is more comprehensive, we performed more experiments.

The experimental results show that the ensemble performance based on the first group is better than the second group,

TABLE 2. Comparison of classification effects based on ISIC 2018 dataset.

Model	Accuracy of different categories							Mean Value				
	AKIEC	BCC	BKL	DF	MEL	NV	VASC	ACC	AUC	F_1	SE	SP
AlexNet	0.912	0.911	0.886	0.922	0.885	0.861	0.927	0.901	0.867	0.398	0.380	0.954
BN-Inception	0.911	0.907	0.895	0.925	0.857	0.852	0.918	0.895	0.857	0.403	0.390	0.956
Caffe-ResNet101	0.906	0.904	0.890	0.925	0.843	0.849	0.914	0.890	0.848	0.387	0.376	0.952
DenseNet121	0.920	0.923	0.907	0.928	0.909	0.893	0.928	0.915	0.895	0.456	0.444	0.968
DenseNet161	0.921	0.923	0.903	0.925	0.905	0.893	0.928	0.914	0.898	0.446	0.434	0.966
DenseNet169	0.918	0.923	0.905	0.927	0.911	0.893	0.928	0.915	0.899	0.451	0.440	0.966
DenseNet201	0.920	0.923	0.901	0.929	0.911	0.891	0.928	0.915	0.901	0.457	0.448	0.966
DPN68b	0.910	0.901	0.891	0.925	0.847	0.841	0.914	0.890	0.840	0.396	0.401	0.952
DPN68	0.919	0.923	0.894	0.926	0.893	0.870	0.929	0.908	0.870	0.440	0.437	0.959
DPN92	0.899	0.893	0.876	0.922	0.828	0.816	0.904	0.877	0.808	0.355	0.367	0.945
DPN98	0.911	0.922	0.894	0.921	0.880	0.853	0.927	0.901	0.852	0.411	0.406	0.949
DPN107	0.896	0.884	0.860	0.913	0.814	0.803	0.896	0.867	0.790	0.318	0.330	0.940
DPN131	0.905	0.911	0.887	0.917	0.873	0.844	0.926	0.895	0.832	0.376	0.380	0.950
FbResNet152	0.905	0.900	0.877	0.923	0.841	0.836	0.914	0.885	0.853	0.367	0.348	0.950
InceptionResNetV2	0.921	0.925	0.907	0.927	0.905	0.893	0.929	0.915	0.894	0.457	0.438	0.966
InceptionV3	0.921	0.922	0.905	0.926	0.904	0.889	0.927	0.913	0.902	0.447	0.441	0.966
InceptionV4	0.919	0.925	0.905	0.927	0.913	0.899	0.929	0.917	0.896	0.454	0.441	0.968
NASNet-A_Large	0.908	0.900	0.886	0.926	0.842	0.840	0.916	0.888	0.825	0.385	0.352	0.950
NASNet-A_Mobile	0.908	0.900	0.886	0.926	0.842	0.840	0.916	0.888	0.825	0.385	0.352	0.950
PNASNet-5_Large	0.915	0.918	0.892	0.925	0.887	0.869	0.928	0.905	0.856	0.423	0.394	0.955
PolyNet	0.915	0.918	0.892	0.925	0.887	0.869	0.928	0.905	0.856	0.423	0.394	0.955
ResNet18	0.919	0.923	0.903	0.925	0.905	0.885	0.927	0.912	0.883	0.441	0.424	0.963
ResNet34	0.920	0.925	0.895	0.927	0.899	0.893	0.929	0.913	0.879	0.447	0.432	0.965
ResNet50	0.919	0.921	0.901	0.927	0.903	0.890	0.927	0.913	0.882	0.443	0.426	0.965
ResNet101	0.917	0.921	0.902	0.925	0.901	0.891	0.929	0.912	0.884	0.438	0.425	0.965
ResNet152	0.919	0.916	0.894	0.923	0.887	0.878	0.928	0.906	0.879	0.425	0.415	0.959
ResNeXt101_32x4d	0.917	0.921	0.900	0.926	0.903	0.887	0.928	0.912	0.880	0.441	0.426	0.963
ResNeXt101_64x4d	0.913	0.916	0.897	0.921	0.899	0.876	0.926	0.907	0.875	0.413	0.411	0.959
SE_ResNet152	0.915	0.921	0.893	0.923	0.893	0.879	0.925	0.907	0.871	0.412	0.397	0.961
SE_ResNeXt50_32x4d	0.915	0.921	0.905	0.926	0.907	0.890	0.928	0.913	0.880	0.443	0.438	0.966
SE_ResNeXt101_32x4d	0.921	0.925	0.899	0.927	0.906	0.889	0.928	0.914	0.888	0.452	0.444	0.965
SENet154	0.918	0.921	0.907	0.926	0.901	0.889	0.927	0.913	0.879	0.443	0.444	0.966
SqueezeNet1_0	0.900	0.892	0.868	0.918	0.852	0.820	0.923	0.882	0.820	0.287	0.248	0.937
SqueezeNet1_1	0.903	0.895	0.865	0.918	0.852	0.819	0.921	0.882	0.798	0.293	0.246	0.935
VGG11	0.909	0.912	0.887	0.924	0.895	0.870	0.928	0.904	0.860	0.407	0.376	0.955
VGG11BN	0.921	0.917	0.899	0.923	0.899	0.877	0.928	0.909	0.867	0.430	0.414	0.959
VGG13	0.914	0.913	0.894	0.925	0.892	0.862	0.927	0.904	0.852	0.416	0.388	0.955
VGG13BN	0.909	0.917	0.872	0.922	0.844	0.822	0.918	0.886	0.783	0.347	0.311	0.935
VGG16	0.897	0.892	0.873	0.922	0.831	0.821	0.909	0.878	0.825	0.329	0.304	0.945
VGG16BN	0.917	0.922	0.898	0.926	0.897	0.884	0.928	0.910	0.880	0.440	0.431	0.962
VGG19	0.909	0.911	0.879	0.919	0.867	0.845	0.926	0.894	0.838	0.356	0.313	0.948
VGG19BN	0.916	0.923	0.903	0.924	0.897	0.883	0.927	0.910	0.876	0.430	0.416	0.961
Xception	0.912	0.913	0.894	0.926	0.859	0.863	0.919	0.898	0.852	0.410	0.402	0.957
AVR-1	0.927	0.929	0.922	0.928	0.917	0.913	0.929	0.924	0.906	0.479	0.467	0.972
VT-1	0.913	0.915	0.877	0.921	0.857	0.871	0.926	0.897	0.913	0.357	0.283	0.979
VWB-1	**0.929**	**0.929**	**0.926**	**0.929**	**0.921**	**0.922**	**0.929**	**0.926**	**0.918**	**0.491**	**0.484**	**0.978**
GD-1	0.927	0.929	0.919	0.929	0.914	0.918	0.929	0.924	0.923	0.481	0.477	0.978
IA-1	0.927	0.929	0.919	0.929	0.914	0.918	0.929	0.924	0.917	0.481	0.477	0.978
AVR-2	0.917	0.919	0.912	0.918	0.907	0.904	0.919	0.914	0.897	0.475	0.472	0.963
VT-2	0.904	0.906	0.867	0.911	0.848	0.862	0.916	0.888	0.903	0.353	0.280	0.969
VWB-2	0.929	0.929	0.921	0.929	0.910	0.915	0.929	0.923	0.911	0.480	0.470	0.975
GD-2	0.913	0.919	0.883	0.921	0.899	0.881	0.926	0.906	0.913	0.389	0.326	0.974
IA-2	0.913	0.918	0.883	0.921	0.899	0.880	0.926	0.906	0.912	0.386	0.326	0.973

the third group is better than the fourth group. By comparing these groups, we find that choosing better CNNs can make our proposed method perform well in dermoscopy image classification. By comparing the first group with the third group, the second group and the fourth group, experiments show that selecting more and better CNNs can make our proposed method more robust and stable. The ensemble strategy of block-integrated voting is the most prominent in terms of ACC. VWB-1, VWB-2, VWB-3, VWB-4 have performed well in their respective groups, especially VWB-1 can reach 0.924. However, the ensemble strategy of group decision on the AUC evaluation criterion is significantly

TABLE 3. Comparison of classification effects based on ISIC 2019 dataset.

Model	Accuracy of different categories								Mean Value				
	AKIEC	BCC	BKL	DF	MEL	NV	VASC	SCC	ACC	AUC	F_1	SE	SP
AlexNet	0.905	0.898	0.876	0.923	0.825	0.817	0.909	0.928	0.885	0.836	0.351	0.318	0.950
BN-Inception	0.911	0.907	0.895	0.925	0.857	0.852	0.918	0.928	0.899	0.861	0.409	0.395	0.959
Caffe-ResNet101	0.906	0.904	0.890	0.925	0.843	0.849	0.914	0.927	0.895	0.857	0.392	0.387	0.956
DenseNet121	0.908	0.909	0.889	0.925	0.848	0.844	0.916	0.928	0.896	0.871	0.402	0.395	0.957
DenseNet161	0.904	0.905	0.892	0.926	0.848	0.843	0.918	0.927	0.895	0.863	0.401	0.397	0.956
DenseNet169	0.909	0.907	0.894	0.925	0.852	0.848	0.917	0.928	0.898	0.873	0.407	0.402	0.958
DenseNet201	0.908	0.908	0.891	0.926	0.855	0.852	0.918	0.928	0.898	0.874	0.413	0.412	0.958
DPN68b	0.910	0.901	0.891	0.925	0.847	0.841	0.914	0.927	0.895	0.846	0.400	0.408	0.955
DPN68	0.907	0.902	0.888	0.923	0.844	0.837	0.914	0.927	0.893	0.841	0.392	0.401	0.954
DPN92	0.899	0.893	0.876	0.922	0.828	0.816	0.904	0.926	0.883	0.818	0.362	0.379	0.949
DPN98	0.900	0.892	0.872	0.911	0.823	0.813	0.906	0.924	0.880	0.819	0.350	0.376	0.947
DPN107	0.896	0.884	0.860	0.913	0.814	0.803	0.896	0.916	0.873	0.804	0.319	0.348	0.943
DPN131	0.894	0.892	0.861	0.912	0.814	0.800	0.899	0.925	0.875	0.798	0.330	0.349	0.944
FbResNet152	0.905	0.900	0.877	0.923	0.841	0.836	0.914	0.927	0.890	0.860	0.376	0.364	0.954
InceptionResNetV2	0.910	0.903	0.893	0.926	0.860	0.855	0.917	0.928	0.899	0.868	0.409	0.398	0.958
InceptionV3	0.912	0.910	0.898	0.927	0.865	0.864	0.920	0.929	0.903	0.871	0.425	0.410	0.961
InceptionV4	0.910	0.907	0.897	0.926	0.853	0.854	0.918	0.929	0.899	0.869	0.414	0.407	0.959
NASNet-A_Large	0.902	0.897	0.870	0.924	0.813	0.807	0.909	0.926	0.881	0.814	0.356	0.348	0.946
NASNet-A_Mobile	0.908	0.900	0.886	0.926	0.842	0.840	0.916	0.927	0.893	0.833	0.392	0.363	0.954
PNASNet-5_Large	0.901	0.899	0.874	0.923	0.826	0.824	0.913	0.921	0.885	0.818	0.359	0.358	0.949
PolyNet	0.903	0.894	0.873	0.923	0.824	0.819	0.911	0.927	0.884	0.836	0.362	0.348	0.949
ResNet18	0.908	0.902	0.886	0.925	0.841	0.839	0.917	0.928	0.893	0.864	0.395	0.382	0.956
ResNet34	0.909	0.907	0.890	0.926	0.852	0.849	0.918	0.928	0.897	0.862	0.408	0.395	0.957
ResNet50	0.905	0.902	0.886	0.925	0.845	0.839	0.915	0.928	0.893	0.861	0.392	0.380	0.955
ResNet101	0.906	0.904	0.888	0.924	0.840	0.833	0.917	0.928	0.893	0.856	0.391	0.376	0.954
ResNet152	0.904	0.897	0.877	0.922	0.842	0.833	0.912	0.926	0.889	0.852	0.370	0.366	0.953
ResNeXt101_32x4d	0.907	0.901	0.887	0.924	0.848	0.844	0.914	0.927	0.894	0.852	0.389	0.382	0.956
ResNeXt101_64x4d	0.900	0.897	0.880	0.922	0.836	0.830	0.913	0.926	0.888	0.843	0.371	0.372	0.952
SE_ResNet152	0.906	0.899	0.886	0.926	0.842	0.833	0.918	0.926	0.892	0.856	0.396	0.393	0.954
SE_ResNeXt50_32x4d	0.915	0.908	0.897	0.927	0.860	0.859	0.920	0.928	0.902	0.858	0.422	0.413	0.959
SE_ResNeXt101_32x4d	0.909	0.909	0.891	0.926	0.852	0.852	0.915	0.927	0.898	0.853	0.403	0.396	0.956
SENet154	0.903	0.903	0.889	0.925	0.839	0.836	0.911	0.927	0.892	0.841	0.386	0.383	0.953
SqueezeNet1_0	0.897	0.859	0.853	0.920	0.804	0.772	0.907	0.926	0.867	0.762	0.268	0.218	0.943
SqueezeNet1_1	0.897	0.863	0.855	0.921	0.808	0.780	0.906	0.926	0.870	0.772	0.277	0.226	0.944
VGG11	0.904	0.897	0.875	0.923	0.833	0.824	0.912	0.927	0.887	0.827	0.364	0.341	0.950
VGG11BN	0.902	0.901	0.883	0.924	0.841	0.835	0.915	0.928	0.891	0.840	0.382	0.370	0.952
VGG13	0.902	0.893	0.872	0.923	0.833	0.824	0.912	0.927	0.886	0.829	0.360	0.340	0.949
VGG13BN	0.904	0.896	0.887	0.924	0.835	0.835	0.913	0.927	0.890	0.841	0.379	0.364	0.952
VGG16	0.897	0.892	0.873	0.922	0.831	0.821	0.909	0.927	0.884	0.832	0.340	0.317	0.949
VGG16BN	0.901	0.901	0.879	0.924	0.837	0.838	0.912	0.928	0.890	0.839	0.374	0.362	0.953
VGG19	0.902	0.887	0.872	0.920	0.820	0.810	0.909	0.927	0.881	0.815	0.326	0.297	0.948
VGG19BN	0.903	0.897	0.875	0.925	0.841	0.834	0.911	0.928	0.889	0.837	0.372	0.353	0.952
Xception	0.912	0.913	0.894	0.926	0.859	0.863	0.919	0.929	0.902	0.859	0.419	0.411	0.960
AVR-1	0.914	0.912	0.900	0.926	0.865	0.861	0.921	0.929	0.904	0.854	0.429	0.426	0.960
VT-1	0.910	0.898	0.882	0.926	0.844	0.836	0.917	0.928	0.893	0.841	0.383	0.320	0.974
VWB-1	**0.927**	**0.927**	**0.925**	**0.929**	**0.910**	**0.916**	**0.928**	**0.929**	**0.924**	**0.919**	**0.488**	**0.483**	**0.977**
GD-1	0.918	0.924	0.917	0.929	0.899	0.908	0.926	0.929	0.919	0.921	0.468	0.443	0.978
IA-1	0.918	0.924	0.917	0.929	0.899	0.908	0.926	0.929	0.919	0.916	0.467	0.443	0.978
AVR-3	0.912	0.911	0.893	0.922	0.862	0.854	0.913	0.927	0.899	0.776	0.371	0.346	0.955
VT-3	0.903	0.899	0.874	0.906	0.828	0.827	0.906	0.925	0.884	0.772	0.285	0.237	0.964
VWB-3	0.909	0.909	0.897	0.926	0.859	0.857	0.919	0.928	0.901	0.840	0.416	0.415	0.959
GD-3	0.910	0.906	0.889	0.926	0.856	0.848	0.919	0.928	0.898	0.856	0.405	0.376	0.966
IA-3	0.908	0.906	0.886	0.926	0.856	0.844	0.918	0.928	0.897	0.844	0.402	0.381	0.963
AVR-2	0.904	0.897	0.881	0.923	0.836	0.826	0.913	0.928	0.889	0.768	0.368	0.343	0.948
VT-2	0.895	0.878	0.856	0.921	0.819	0.797	0.906	0.925	0.875	0.764	0.283	0.236	0.957
VWB-2	0.908	0.903	0.887	0.923	0.842	0.842	0.914	0.928	0.893	0.806	0.374	0.335	0.958
GD-2	0.900	0.893	0.863	0.921	0.829	0.805	0.908	0.926	0.881	0.821	0.292	0.216	0.965
IA-2	0.900	0.891	0.862	0.921	0.827	0.805	0.908	0.926	0.880	0.788	0.292	0.220	0.964
AVR-4	0.904	0.897	0.881	0.921	0.836	0.826	0.907	0.925	0.887	0.766	0.366	0.341	0.942
VT-4	0.895	0.878	0.856	0.915	0.819	0.797	0.901	0.924	0.873	0.762	0.281	0.234	0.953
VWB-4	0.905	0.895	0.877	0.923	0.828	0.817	0.912	0.928	0.886	0.772	0.356	0.316	0.952
GD-4	0.899	0.897	0.863	0.921	0.822	0.806	0.908	0.926	0.880	0.801	0.296	0.222	0.964
IA-4	0.895	0.888	0.855	0.921	0.818	0.797	0.906	0.925	0.876	0.767	0.288	0.245	0.955

Algorithm 3 The Ensemble Strategy of Block-Integrated Voting

```
for number of blocks do
  for number of categories of dermoscopy images do
    for each block contains the number of CNNs do
      if judge_kj (x) =k then
        count[k]=count[k]+1;
//k represents a category of dermoscopy image, and judge_kj (x)
//performs binary classification on the input image x.
  for i=0;i<count.length;i++ do
    if count[i]>max then
      max=count[i];
      max_category=i;
//max refers to the largest number of votes in the small sub-
//modules.
//max_category refers to the category corresponding to max.
  max_category_array[max_category]=max_category
    _array[max_category]+1;
//max_category_array is used to accumulate the output of each //block.
for i=0;i<max_category_array.length;i++ do
  if max_category_array[i]>max then
    max=max_category_array[i];
    max_category_final=i;
return max_category_final;
```

better than the ensemble strategy of block-integrated voting. GD-1 is 0.002 higher than VWB-1, GD-2 is 0.015 higher than VWB-2, GD-3 is 0.016 higher than VWB-3, and GD-4 is 0.029 higher than VWB-4. AUC can more accurately evaluate the classifiers for unbalanced data. The ensemble strategy of group decision performs best on AUC. Our proposed method is highly dependent on selected CNNs. If the selected CNNs are better, the ensemble strategy of group decision and the ensemble strategy of maximizing individual advantage perform similarly, but the ensemble strategy of maximizing individual advantage can reduce certain calculation costs. The ensemble strategy of group decision is not much different from the ensemble strategy of maximizing individual advantage in the first group, only 0.005 and 0.001 differ from AUC and F_1. In addition, when the selected CNNs perform well, our proposed method improves each evaluation criterion compared to the traditional ensemble strategies, such as the first group. However, if the selected CNNs perform poorly, the performances of the ensemble strategy of group decision and the ensemble strategy of maximizing individual advantage differ greatly and the proposed method may not be as effective as traditional ensemble strategies on some evaluation criteria, especially on SE.

By comparing Table 3, the evaluation criteria of the different individual CNNs are obviously not as good as our proposed method. Different strategies and individual CNNs have different classification effects, but the classification effect of the ensemble strategy is usually better than that of the individual CNN. DenseNet201 has performed well in these individual CNNs, but compared with the ensemble strategies, the classification effect still has a big gap. In general, the ensemble strategy of block-integrated voting has shown an advantage in various evaluation criteria, especially in terms of SE, VWB-1 is 0.071 higher than DenseNet201, 0.040 higher than GD-1 and IA-1, 0.057 higher than AVR-1, and 0.163 higher than VT-1. On the whole, the classification effect of the VWB-1 is better.

E. COMPARISON OF CLASSIFICATION EFFECTS OF DIFFERENT FRAMEWORKS

Over the years, numerous studies on lesion classification have been conducted. We compare different frameworks based on the ISIC 2019 dataset, as shown in Table 4.

Among the three strategies proposed by us, the comprehensive performance of VWB-1 is more outstanding. We can see that VWB-1, which we propose, has achieved good results in various evaluation criteria. A individual classifier has limitations on various evaluation criteria. The performance of VWB-1 is similar to that of [10], but VWB-1 has improved in all evaluation criteria, especially ACC, AUC and SE increased by 0.011, 0.018 and 0.012 respectively. Compared with [10], VWB-1 is mainly based on ensemble learning to combine multiple CNNs. Comparing VWB-1 and [44] shows that the performance in ACC is very similar, but our proposed method performs better on the overall classification effect. When the sample distribution is uneven, especially when the proportion of positive and negative samples is seriously unbalanced, it is difficult for ACC to evaluate the classifier effect. Compared with [45], VWB-1 is close in most evaluation criteria, but VWB-1 has obvious improvement in AUC and SP. On the whole, [46] performs better than VWB-1. However, VWB-1 is 0.002 higher in SP than [46]. In addition, the calculation cost of [46] is relatively high and the realization process is complex. Reference [47] is mainly based on Google InceptionV3 CNN architecture. Overall, VWB-1 is better than [47] in most evaluation criteria. Compared with [47], VWB-1 increases by 0.012 in AUC. By comparing different frameworks, it is found that combining multiple deep learning techniques can improve the classification effect to a certain extent.

F. TRANSFER LEARNING

We analyze the application of transfer learning in ISIC 2019 dataset by comparing the accuracy and loss of ResNet50 in three cases. The three cases we compared are in the absence of pre-training, in the case of ImageNet pre-training, and in the case of ISIC 2018 dataset pre-training. The loss function can be a good tool to reflect the gap between the model and the actual data. Through the loss function, we can better analyze and understand the subsequent optimization of the model. In addition, the different distribution of the dataset will cause our optimized loss function to be

TABLE 4. Comparison of classification effects of different frameworks.

Method	Mean Value				
	ACC	AUC	F_1	SE	SP
Deep learning, sparse coding, SVM[10]	0.913	0.901	0.483	0.471	0.974
Linear classifier CNN deep learning[25]	0.884	0.858	0.458	0.462	0.956
Feature-based method[43]	0.859	0.838	0.451	0.454	0.931
Two deep learning method[44]	0.921	0.892	0.477	0.466	0.966
13 models + hierarchical approach to select outliers[45]	0.923	0.905	0.491	0.487	0.968
Ensemble of multi-Res EfficientNets + SEN154[46]	0.925	0.921	0.502	0.492	0.975
Deep CNN[47]	0.917	0.907	0.481	0.485	0.971
GD-1(Ours)	0.919	0.921	0.468	0.443	0.978
IA-1(Ours)	0.919	0.916	0.467	0.443	0.978
VWB-1(Ours)	**0.924**	**0.919**	**0.488**	**0.483**	**0.977**

different from the real data loss function, so we need to choose the dataset carefully. We use GANs to create a relatively balanced sample space and use ten-fold cross-validation to ensure the stability and reliability of training. In general, the larger the loss function value, the smaller the classification probability of the classifier on the real label, and the worse the performance. ResNet50 is selected to compare the loss and accuracy in three cases on the validation set, which is presented in Fig. 9.

Through experiments, we find that within 100 epochs, the pre-trained ResNet50 performed well on accuracy and loss. In particular, ResNet50 pre-trained by ISIC 2018 dataset has higher starting accuracy, faster convergence speed, and higher approximation accuracy. After ISIC 2018 dataset pre-training, ResNet50 can converge after about 15 epochs, thereby achieving a more stable state. InceptionV3 and ResNet50 have similar performance, but InceptionV3 has higher accuracy and lower loss. Through transfer learning, we can share the learned model parameters with the new model in a certain way, thereby speeding up and optimizing the learning efficiency of the model.

IV. DISCUSSION

Skin cancer is the most common form of cancer. While amenable to early detection by direct inspection, visual similarity with benign lesions makes the task difficult. Dermoscopy images are introduced to better visualize key details in skin lesions to improve diagnostic accuracy. We tried 43 CNNs, including DenseNet, SE-ResNet and other CNNs. By comparing the effects of different CNNs on the classification of different categories of dermoscopy images, in most cases, the deeper and more recent CNN has better feature extraction ability and classification ability than the relatively shallow CNN. By selecting more complete images to ensure the integrity of the feature information, the data augmentation and transfer learning are used to effectively alleviate insufficient samples during training. We trained CNNs by using GANs to create a balanced sample space. Setting different loss weight for different categories of samples alleviates uneven data distribution to some extent. We conduct extensive experiments and prove the effectiveness of the ensemble strategies proposed by the classification method based on the ensemble of individual advantage and group decision, which satisfies the higher clinical demand to some extent. VWB-1 performs similarly to [10], but the method we proposed is higher than 0.024 on SE. Comparing VWB-1 and [43] shows that the performance on ACC is very similar, but our proposed method performs better on the overall classification effect. By comparing different frameworks, it is found that combining multiple deep learning techniques can improve the classification effect to a certain extent. In addition, we also considered the influence of age, gender, and lesion location on the classification, which is presented in Fig. 10.

Middle-aged and elderly people are the main group of skin cancer patients, especially middle-aged people suffer from BCC. Skin cancer is more common in men than women, except for NV. Different categories of skin cancer lesions vary, e.g., the main lesion location of DF is concentrated in Head or neck. The combination of computer-assisted diagnosis and physician experience may have great potential, which requires more clinical experience as a guide. When an individual CNN deals with a problem, it is easy to encounter a bottleneck in the model, and there is no guarantee that it will perform well when solving multiple problems. Therefore, it is common to use the ensemble of voting, the ensemble of averaging and other methods to fuse the excellent models to improve the generalization ability of the individual CNN, and to gather the advantages of each model to get an optimal solution. In the process of using the fusion CNNs, the problem of high computational cost is often encountered, because the excellent CNNs are generally deep CNNs, which are characterized by deeper levels and more parameters. Our proposed method includes three ensemble strategies, each of which has its advantages. We first performed experiments on ISIC 2018 dataset and found that CNNs with better classification results can make our proposed method more effective. Then, we continued the experiment on ISIC 2019 dataset, the purpose of which was to prove the feasibility of transfer learning on dermoscopy images. At the same time, it proves that our proposed method can improve the classification effect of dermoscopy images.

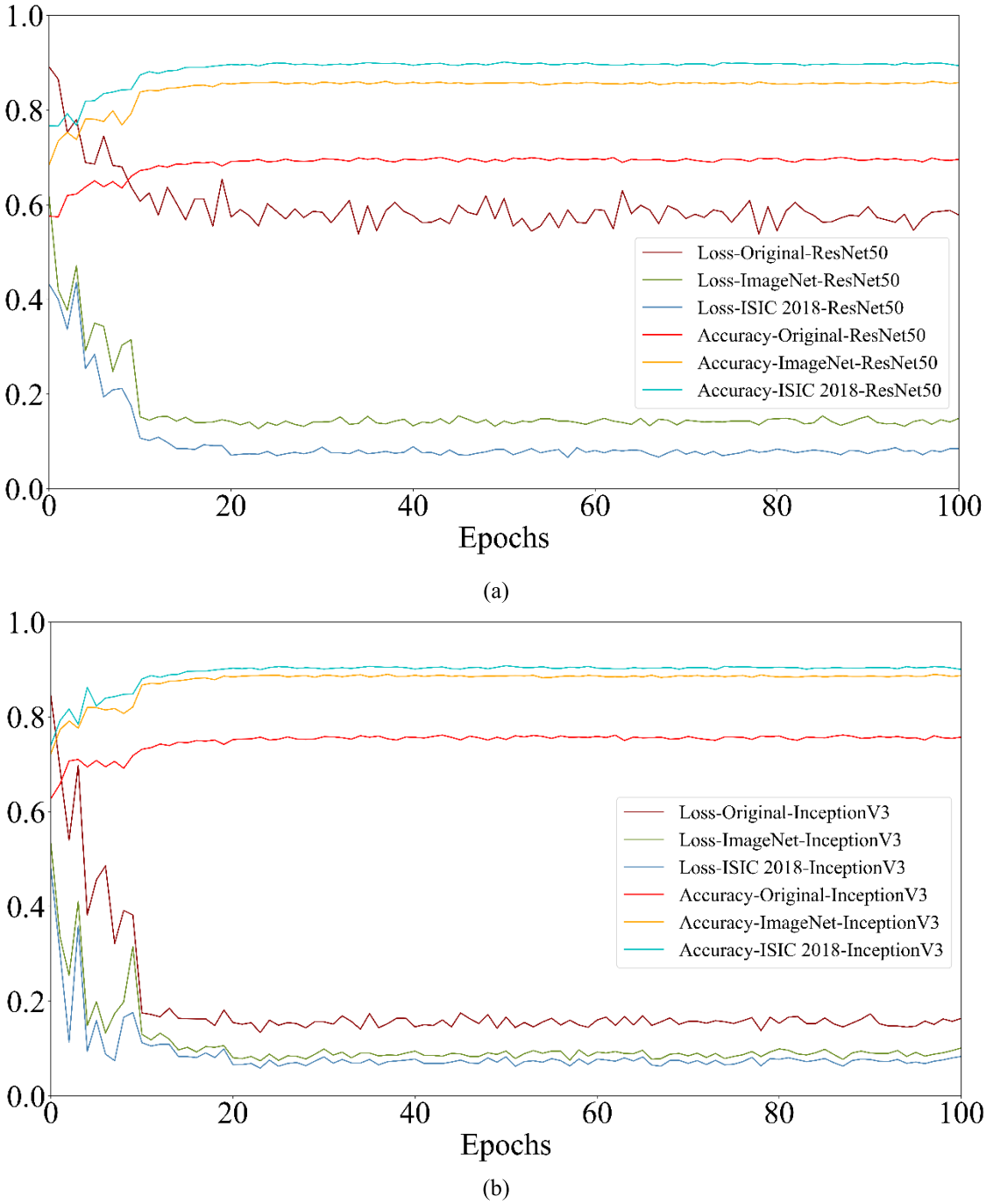

FIGURE 9. Loss and accuracy curves under different pre-trainings. (a) ResNet50. (b) InceptionV3.

The study deviated from a real-life clinical setting and was potentially affected by verification and selection bias. However, one goal of the classification method is to raise awareness among the people to check their moles regularly. It is not assumed that the higher accuracy in a diagnostic study with digital images substitutes the clinical expert, but that the application might be implemented in a screening process and brings the people to the expert dermatologist who will be able to precise the diagnosis.

V. CONCLUSION

In conclusion, A balanced and large dataset is crucial for CNNs. By creating balanced sample spaces through GANs and using transfer learning, CNNs are better trained. Different strategies and individual CNNs have different classification effects, but the classification effect of the ensemble strategy is usually better than that of the individual CNN. Based on the classification method of the ensemble of individual advantage and group decision, we proposed three kinds of ensemble

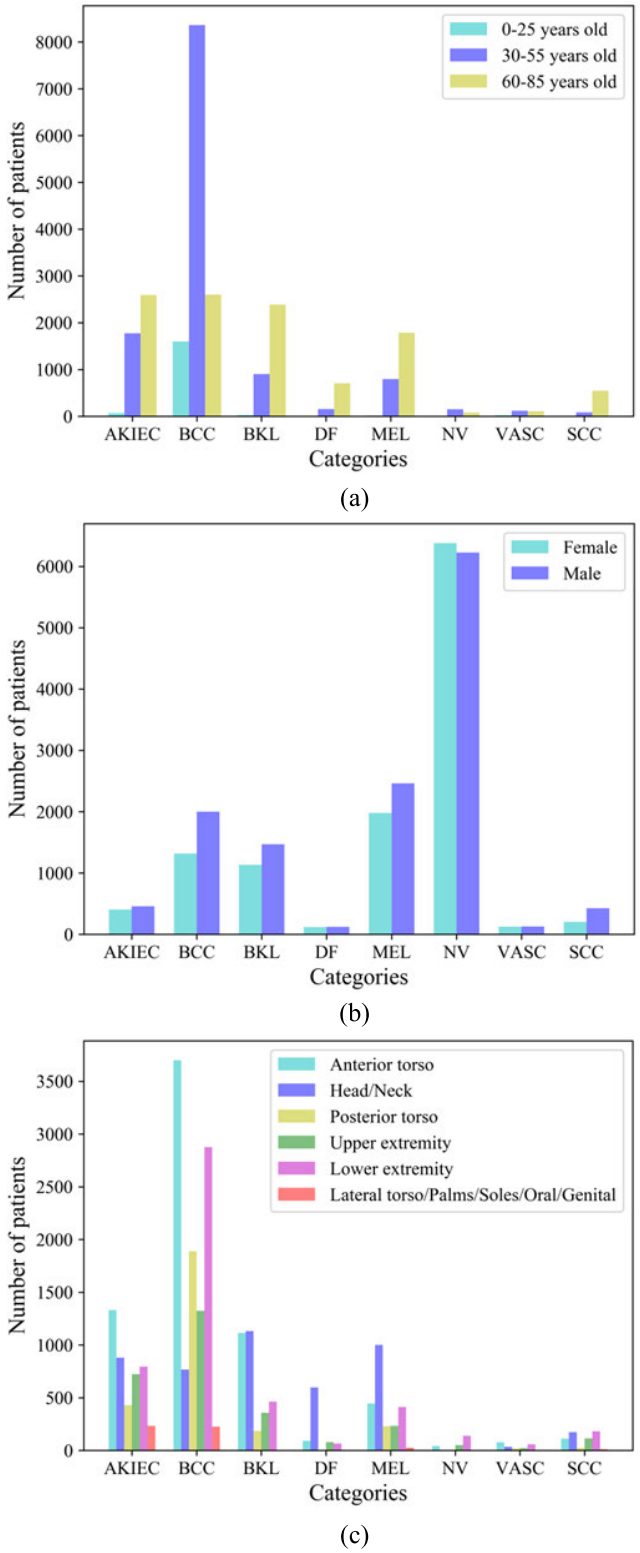

FIGURE 10. The distribution of patients under different categories of skin cancer. The influence of (a) age, (b) gender, and (c) lesion location.

strategies to enhance the classification effect of dermoscopy images. Compared with the different individual CNNs and traditional ensemble strategies, the method proposed in our study has a certain improvement in multiple evaluation criteria.

REFERENCES

[1] G. P. Guy, Jr., S. R. Machlin, D. U. Ekwueme, and K. R. Yabroff, "Prevalence and costs of skin cancer treatment in the U.S., 2002–2006 and 2007–2011," *Amer. J. Preventive Med.*, vol. 48, no. 2, pp. 183–187, Feb. 2015.

[2] American Cancer Society. (2019). *Cancer Facts & Figs*. [Online]. Available: https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-Figs/cancer-facts-Figs-2019.html

[3] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2019," *CA, Cancer J. Clin.*, vol. 69, no. 1, pp. 7–34, 2019.

[4] Skin Cancer Foundation. *Skin Cancer Facts & Statistics*. [Online]. Available: https://www.skincancer.org/skin-cancer-information/skin-cancer-facts

[5] *Australian Government, Melanoma of the Skin Statistics*. [Online]. Available: https://melanoma.canceraustralia.gov.au/statistics

[6] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA, Cancer J. Clin.*, vol. 68, no. 6, pp. 394–424, Nov. 2018.

[7] R. Kasmi and K. Mokrani, "Classification of malignant melanoma and benign skin lesions: Implementation of automatic ABCD rule," *IET Image Process.*, vol. 10, no. 6, pp. 448–455, Jun. 2016.

[8] A.-R. A. Ali and T. M. Deserno, "A systematic review of automated melanoma detection in dermatoscopic images and its ground truth data," *Proc. SPIE*, vol. 8318, Feb. 2012, Art. no. 83181I.

[9] M. E. Celebi, H. A. Kingravi, B. Uddin, H. Iyatomi, Y. A. Aslandogan, W. V. Stoecker, and R. H. Moss, "A methodological approach to the classification of dermoscopy images," *Comput. Med. Imag. Graph.*, vol. 31, no. 6, pp. 362–373, Sep. 2007.

[10] N. Codella, J. Cai, M. Abedini, R. Garnavi, A. Halpern, and J. R. Smith, "Deep learning, sparse coding, and SVM for melanoma recognition in dermoscopy images," in *Proc. Int. Workshop Mach. Learn. Med. Imag.* Cham, Switzerland: Springer, 2015, pp. 118–126.

[11] N. Cascinelli, M. Ferrario, T. Tonelli, and E. Leo, "A possible new tool for clinical diagnosis of melanoma: The computer," *J. Amer. Acad. Dermatol.*, vol. 16, no. 2, pp. 267–361, 1987.

[12] A. J. Sober and J. M. Burstein, "Computerized digital image analysis: An aid for melanoma diagnosis: Preliminary investigations and brief review," *J. Dermatol.*, vol. 21, no. 11, pp. 885–890, 1994.

[13] C. Sinz *et al.*, "Accuracy of dermatoscopy for the diagnosis of nonpigmented cancers of the skin," *J. Amer. Acad. Dermatol.*, vol. 77, no. 6, pp. 1100–1109, 2017.

[14] N. K. Mishra and M. E. Celebi, "An overview of melanoma detection in dermoscopy images using image processing and machine learning," 2016, *arXiv:1601.07843*. [Online]. Available: http://arxiv.org/abs/1601.07843

[15] L. Ballerini, R. B. Fisher, B. Aldridge, and J. Rees, "A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions," in *Color Medical Image Analysis*. Dordrecht, The Netherlands: Springer, 2013, pp. 63–86.

[16] T. Tommasi, T. E. La, and B. Caputo, "Melanoma recognition using representative and discriminative kernel classifiers," in *Proc. Int. Workshop Comput. Vis. Approaches Med. Image Anal.* Berlin, Germany: Springer, 2006, pp. 1–12.

[17] R. J. Stanley, W. V. Stoecker, and R. H. Moss, "A relative color approach to color discrimination for malignant melanoma detection in dermoscopy images," *Skin Res. Technol.*, vol. 13, no. 1, pp. 62–72, Feb. 2007.

[18] Y. Cheng, R. Swamisai, S. E. Umbaugh, R. H. Moss, W. V. Stoecker, S. Teegala, and S. K. Srinivasan, "Skin lesion classification using relative color features," *Skin Res. Technol.*, vol. 14, no. 1, pp. 53–64, 2008.

[19] H. Ganster, P. Pinz, R. Rohrer, E. Wildling, M. Binder, and H. Kittler, "Automated melanoma recognition," *IEEE Trans. Med. Imag.*, vol. 20, no. 3, pp. 233–239, Mar. 2001.

[20] B. Kusumoputro and A. Ariyanto, "Neural network diagnosis of malignant skin cancers using principal component analysis as a preprocessor," in *Proc. IEEE Int. Joint Conf. Neural Netw., IEEE World Congr. Comput. Intell.*, vol. 1, May 1998, pp. 310–315.

[21] M. E. Celebi, H. Iyatomi, G. Schaefer, and W. V. Stoecker, "Lesion border detection in dermoscopy images," *Comput. Med. Imag. Graph.*, vol. 33, no. 2, pp. 148–153, Mar. 2009.

[22] G. Schaefer, B. Krawczyk, M. E. Celebi, and H. Iyatomi, "An ensemble classification approach for melanoma diagnosis," *Memetic Comput.*, vol. 6, no. 4, pp. 233–240, Dec. 2014.

[23] G. Schaefer, B. Krawczyk, M. E. Celebi, and H. Iyatomi, "Melanoma classification using dermoscopy imaging and ensemble learning," in *Proc. 2nd IAPR Asian Conf. Pattern Recognit.*, Nov. 2013, pp. 386–390.

[24] S. Demyanov, R. Chakravorty, M. Abedini, A. Halpern, and R. Garnavi, "Classification of dermoscopy patterns using deep convolutional neural networks," in *Proc. IEEE 13th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2016, pp. 364–368.

[25] J. Kawahara, A. BenTaieb, and G. Hamarneh, "Deep features to classify skin lesions," in *Proc. IEEE 13th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2016, pp. 1397–1400.

[26] J. Xu, X. Luo, G. Wang, H. Gilmore, and A. Madabhushi, "A deep convolutional neural network for segmenting and classifying epithelial and stromal regions in histopathological images," *Neurocomputing*, vol. 191, pp. 214–223, May 2016.

[27] J. Yang, F. Xie, H. Fan, Z. Jiang, and J. Liu, "Classification for dermoscopy images using convolutional neural networks based on region average pooling," *IEEE Access*, vol. 6, pp. 65130–65138, 2018.

[28] K. Møllersen, M. Zortea, T. R. Schopf, H. Kirchesch, and F. Godtliebsen, "Comparison of computer systems and ranking criteria for automatic melanoma detection in dermoscopic images," *PLoS ONE*, vol. 12, no. 12, Dec. 2017, Art. no. e0190112.

[29] N. Gessert, T. Sentker, F. Madesta, R. Schmitz, H. Kniep, I. Baltruschat, R. Werner, and A. Schlaefer, "Skin lesion diagnosis using ensembles, unscaled multi-crop evaluation and loss weighting," 2018, *arXiv:1808.01694*. [Online]. Available: http://arxiv.org/abs/1808.01694

[30] K. Matsunaga, A. Hamada, A. Minagawa, and H. Koga, "Image classification of melanoma, nevus and seborrheic keratosis by deep neural network ensemble," 2017, *arXiv:1703.03108*. [Online]. Available: http://arxiv.org/abs/1703.03108

[31] Z. Wei, H. Song, L. Chen, Q. Li, and G. Han, "Attention-based DenseUnet network with adversarial training for skin lesion segmentation," *IEEE Access*, vol. 7, pp. 136616–136629, 2019.

[32] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[33] M. M. Ghazi, B. Yanikoglu, and E. Aptoula, "Plant identification using deep neural networks via optimization of transfer learning parameters," *Neurocomputing*, vol. 235, pp. 228–235, Apr. 2017.

[34] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[36] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4401–4410.

[37] P. M. Burlina, N. Joshi, K. D. Pacheco, T. Y. A. Liu, and N. M. Bressler, "Assessment of deep generative models for high-resolution synthetic retinal image generation of age-related macular degeneration," *JAMA Ophthalmol.*, vol. 137, no. 3, pp. 258–264, 2019.

[38] K. M. Hosny, M. A. Kassem, and M. M. Foaud, "Classification of skin lesions using transfer learning and augmentation with alex-net," *PLoS ONE*, vol. 14, no. 5, May 2019, Art. no. e0217293.

[39] J. Yang, J. Zhao, L. Lu, T. Pan, and S. Jubair, "A new improved learning algorithm for convolutional neural networks," *Processes*, vol. 8, no. 3, p. 295, Mar. 2020.

[40] M. Combalia, N. C. F. Codella, V. Rotemberg, B. Helba, V. Vilaplana, O. Reiter, C. Carrera, A. Barreiro, A. C. Halpern, S. Puig, and J. Malvehy, "BCN20000: Dermoscopic lesions in the wild," 2019, *arXiv:1908.02288*. [Online]. Available: http://arxiv.org/abs/1908.02288

[41] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, H. Kittler, and A. Halpern, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC)," 2019, *arXiv:1902.03368*. [Online]. Available: http://arxiv.org/abs/1902.03368

[42] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci. Data*, vol. 5, no. 1, Dec. 2018, Art. no. 180161.

[43] S.-R.-S. Jianu, L. Ichim, D. Popescu, and O. Chenaru, "Advanced processing techniques for detection and classification of skin lesions," in *Proc. 22nd Int. Conf. Syst. Theory, Control Comput. (ICSTCC)*, Oct. 2018, pp. 498–503.

[44] Y. Li and L. Shen, "Skin lesion analysis towards melanoma detection using deep learning network," *Sensors*, vol. 18, no. 2, p. 556, Feb. 2018.

[45] A. G. C. Pacheco, A.-R. Ali, and T. Trappenberg, "Skin cancer detection based on deep learning and entropy to detect outlier samples," 2019, *arXiv:1909.04525*. [Online]. Available: http://arxiv.org/abs/1909.04525

[46] N. Gessert, M. Nielsen, M. Shaikh, R. Werner, and A. Schlaefer, "Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data," *MethodsX*, vol. 7, Mar. 2020, Art. no. 100864.

[47] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017.

AN GONG received the B.S. degree in computer and applications and the M.S. degree in oil and gas field development project from the China University of Petroleum (East China), Qingdao, China, in 1993 and 2001, respectively. He is currently an Associate Professor with the College of Computer Science and Technology, China University of Petroleum (East China). His research interest includes big data intelligent processing.

XINJIE YAO received the B.S. degree in information security from North China Electric Power University, Baoding, China, in 2018. He is currently pursuing the M.S. degree in computer technology under the supervision of Prof. A. Gong with the China University of Petroleum (East China), Qingdao, China. He holds ten software. His research interest includes computer vision. he was a Meritorious Winner from the Mathematical Contest in Modeling (MCM) in 2016.

WEI LIN received the B.S. degree in petroleum engineering from the China University of Geosciences, Beijing, China, in 2015. He is currently pursuing the Ph.D. degree in fluid mechanics under the supervision of Prof. X. Li and Z. Yang with the University of Chinese Academy of Sciences, Beijing. From 2018 to 2019, he visited with the Department of Earth and Planetary Science, University of California at Berkeley, Berkeley, under the supervision of Prof. M. Manga, for a period of 12 months. Since 2019, he has been a Bentham Ambassador with Bentham Science Publishers Ltd., Sharjah. His collaborative study is a combination of developing models and percolation theory for tight rocks and volcanic rocks and performing and analyzing x-ray tomography images of rocks using advanced light source (ALS) with the Lawrence Berkeley National Laboratory. He has authored more than 33 articles. He holds four patents and one software. His research interests include physical and chemical transport in porous materials, modeling and characterization of porous media, multiscale matter, and tertiary oil recovery, including oilfield chemistry and microbial enhanced oil recovery. He was the SPE President of the Research Institute of Petroleum Langfang SPE Student Chapter. He received awards and honors, include the CAS President Award, Chinese Academy of Sciences (CAS), the Chinese Government Scholarship (China Scholarship Council), and the National Scholarship for Doctoral Students, Ministry of Education, China. He serves as a Reviewer for *Marine and Petroleum Geology*, *Journal of Petroleum Science and Engineering*, IEEE Access, *Energies*, *RSC Advances*, *Journal of Dispersion Science and Technology*, *Heliyon*, and the *International Journal of Mechanics Research*.