Journal of Machine Learning and Information Security
https://www.sciltp.com/journals/jmlis

Article

A Multi-View Ensemble-Based Weakly Supervised Model for Skin Lesion Images Diagnosis in Dermoscopic Images

Qi Han, Hong Zhang *, Tengfei Weng, Yuan Tian, Zhong Li, Yadong Lan, Yuhui Lin, Shaoshun Yi, Yutong Wu and Yangjun Pei

School of Computer Science and Engineering, Chongqing University of Science and Technology, Chongqing 401331, China
* Correspondence: zhzhang2024@163.com

How To Cite: Han, Q.; Zhang, H.; Weng, T.; et al. A Multi-View Ensemble-Based Weakly Supervised Model for Skin Lesion Images Diagnosis in Dermoscopic Images. *Journal of Machine Learning and Information Security* **2026**, *2*(2), 13. https://doi.org/10.53941/jmlis.2026.100013

Received: 15 October 2025
Revised: 30 March 2026
Accepted: 7 April 2026
Published: 26 June 2026

Abstract: Skin cancer has become one of the most common causes of death, and accurate diagnosis of skin lesions is essential for early detection of melanoma. Beyond traditional approaches, computer-aided diagnosis is increasingly applied to cancer detection. A key benefit is that it removes the potential for human error. However, existing methods are unable to achieve quite high accuracy due to noise (such as hair, ink dots, scales, etc.) and the small inter-class and large intra-class differences in skin images. Therefore, a multi-view ensemble-based weakly supervised model for skin lesion image diagnosis in dermoscopy is proposed. In this method, a weakly supervised multi-view(WSM) module is proposed to deal with noise in data images. We propose a multi-scale feature fusion (MFF) module to address the challenge of small inter-class variance and large intra-class variance in skin disease images. The model can integrate features of multiple basic models well, capture information of different scales, and explore and utilize the advantages of different features. We conducted a series of experiments in the open dataset HAM10000. The experimental results show that the performance of the proposed model is superior to that of other models, with an accuracy of 95.90%. In conclusion, our model solves the problems of noise and small inter-class and large intra-class differences in skin images well, and achieves advanced performance in skin disease image recognition.

Keywords: deep learning; intelligent assisted diagnosis; skin cancer; ensemble learning

1. Introduction

Skin disease ranks among the most prevalent and clinically serious conditions, which can develop into serious skin cancer if not properly treated. In the United States, another 5 million patients suffer from skin cancer each year [1]. In the clinic, skin cancer is generally divided into two categories: melanoma and non-melanoma. Melanoma is usually more aggressive and metastatic and is therefore considered a more serious type of skin cancer [2]. The major diagnostic categories within pigmented lesions are illustrated in Figure 1, which includes actinic keratoses and intraepithelial carcinoma/Bowen's disease (akiec), basal cell carcinoma (bcc), benign keratosis-like lesions (solar lentigines/seborrheic keratoses and lichen-planus-like keratoses, bkl), dermatofibroma (df), melanoma (mel), melanocytic nevi (nv), and vascular lesions (angiomas, angiokeratomas, pyogenic granulomas and hemorrhage, vasc). While non-melanoma skin cancers, including basal cell carcinoma (BCC) and squamous cell carcinoma (SCC), typically exhibit lower aggressiveness and better prognosis, melanoma is considerably more dangerous, contributing to over 75% of deaths related to skin cancer [3]. Early detection of skin cancer can help to detect and diagnose skin cancer as early as possible, so that timely treatment can be carried out. In the early stages, the success rate of skin cancer treatment is higher, and the survival rate of patients is relatively high.

The diversity of skin textures and lesions makes skin cancer screening particularly challenging. Dermatologists use methods such as ABCD rule [4] and dermoscopy to diagnose skin lesions. Melanoma has similar characteristics to other types of skin diseases and can be difficult to detect even by experienced dermatologists, so accurate diagnosis

of lesions is essential. However, for experienced dermatologists, manually identifying skin lesions via dermoscope is also a time-consuming and error-prone process. To overcome this challenge, the researchers proposed using computer-aided diagnosis (CAD) methods to avoid these problems and save time [5]. The diagnostic procedure of deep learning and convolutional neural network (CNN) features can effectively and automatically detect dermoscopic images of skin injury at an early stage [6,7]. Deep learning has achieved remarkable performance in skin disease image recognition. ResGANet [8] is a modular attention mechanism that can capture the correlation characteristics of medical images across two independent dimensions. The work in [9] introduces an octave convolutional capsule network of concern (AOC Caps) for medical image classification. This network incorporates an AOC module that jointly processes and merges high- and low-frequency image features, with an automatic weighting mechanism for salient components. In [10], a pre-processed image pipeline is developed, which removes hair from the images, enhances the data set, and resizes the images to meet the requirements of each model. By designing a custom 26-layer CNN architecture for skin lesion segmentation, the authors of [11] achieve a reduction in dermatologists' identification time without compromising detection accuracy.

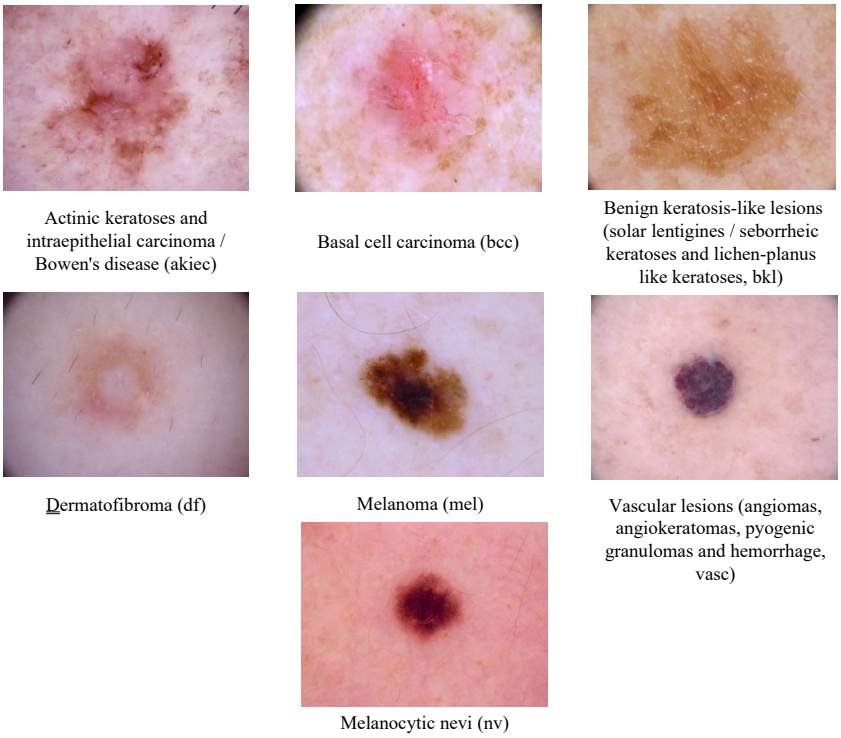


Figure 1. Examples spanning the full range of important diagnostic categories in pigmented lesions.

Recently, several high-impact studies have advanced the field of dermatological image analysis. For instance, Priyeshkumar et al. [12] proposed a deep learning approach for skin cancer diagnosis on the HAM10000 dataset, achieving an average accuracy of 98.19% and an average F1-score of 98.15%. Shetty et al. [13] introduced a customized CNN model, DeepSkinNet, for skin cancer detection, reporting an accuracy of 97.35%, precision of 98%, recall of 97%, and F1-score of 97%. More recently, panDerm [14] presented a multimodal vision foundation model trained on large-scale clinical data, which attained an average accuracy of 92.6% and F1-score of 93.9% on HAM10000. Beyond CNN-based and foundation model approaches, other directions have also emerged: a knowledge graph-enhanced framework [15] integrates clinical guidelines with gradient-based neural systems for melanoma diagnosis; SkinGPT-4 [16] leverages a multimodal large language model for interactive dermatology diagnosis; and GloW-VSNet [17] introduces a scribble-based weakly supervised method for vitiligo lesion segmentation. These works represent distinct paradigms, including pure CNN-based classification, multimodal large language models, and large-scale pre-trained foundation models. In contrast, our work focuses on a weakly supervised preprocessing pipeline and a multi-view ensemble learning framework with a dedicated multi-ensemble loss. By leveraging multiple views and a carefully designed loss function, our method achieves competitive average accuracy (98.81%) while maintaining a lightweight and interpretable architecture, offering a practical solution for computer-aided skin cancer diagnosis.

In recent years, the accurate recognition of skin disease images has brought some challenges due to various factors. Such as artifacts, air bubbles, hair, dark borders, measurement markers, uneven color illumination, make the task of identifying lesions more complex [18]. In addition, the small inter-class and large intra-class differences in

skin image recognition are also an important reason for the low accuracy of skin image recognition [9].

In order to solve the above problems, a new skin disease image identification network is proposed in this paper. The network has carried out weakly supervised preprocessing for the noise in the skin lesion images, which not only removes the hair and scale noise on some skin images, but also makes the lesion area more obvious to a certain extent. In addition, to solve the problem of small inter-class difference and large intra-class difference in skin disease images, the idea based on ensemble learning is adopted, which greatly improves the classification accuracy.

The primary contributions of this study can be stated as follows:

- A multi-view ensemble-based weakly supervised model for skin lesion images diagnosis in dermoscopic images is proposed.
- A weakly supervised multi-view(WSM) module is proposed. It does not require additional training and can automatically remove noise (such as hair, scales, etc.) from images of skin lesions. The WSM module improves the learning effect by integrating information from different views to complement and cooperate with each other.
- A multi-scale feature fusion(MFF) module is proposed, which combines low-level features (high resolution, containing location and detail information) with high-level features (with strong semantic information) to improve accuracy and robustness.
- Multi-ensemble(ME) loss function is proposed. The ME loss can promote the learning of richer and more abstract feature representations. Therefore, the performance of the model is improved.

Motivated by the pyramid structure [19] and the attention module [20,21], this work presents a self-interactive attention module.

2. Related Work

2.1. Ensemble Learning for Medical Image Analysis

Ensemble learning is widely used in medical imaging, where medical images often have complex structures and noise [22]. By combining the prediction results of multiple models, ensemble learning can improve overall performance, ultimately leading to improved accuracy and robustness in medical diagnostic tasks. Li et al. [23] presented a semi-supervised mechanism model based on ensemble learning, which introduces a novel integrated mechanism known as Alternative Adaptive Boosting (Al-Adaboost) that combines the decisions of two hierarchical models. Thomas [24] proposed a scalable and intuitive framework for uncertainty quantification in medical image segmentation, which generates measurement values approximating classification probabilities. Yifei et al. [25] proposed a multi-view ensemble learning method based on a voting mechanism. The proposed method fuses three sources of diagnostic evidence: ultrasound images of thyroid nodules, medical features derived from U-Net outputs, and features selected via mRMR from statistical and texture attributes. Zhao et al. [19] proposed an effective self-ensembling learning framework to enhance feature representation for discriminating rare cases through feature distillation. Ensemble learning models can indeed improve the performance of computer-aided diagnosis, but the above models did not individually design loss functions for each ensemble submodel, which limits the performance of the ensemble model. This paper utilizes ensemble models for skin disease recognition and separately designs loss functions for each integrated submodel during training to improve diagnostic accuracy and stability, which is of great significance for clinical diagnosis and medical research.

2.2. Noise Preprocessing Method for Skin Diseases Image

The noise in skin disease images may cause the details of the images to become blurred or distorted, thus affecting the accurate diagnosis of lesions by doctors. Through noise pre-processing methods, the noise can be removed or reduced, making the images clearer and more realistic, which helps doctors accurately diagnose the lesions. Joseph [20] studied the impact of image preprocessing on the performance of skin lesion saliency segmentation methods. The CHC-Otsu algorithm was achieved by using the collaborative implementation of color histogram clustering and Otsu's threshold. Hadi [26] proposed a new preprocessing technique, which compares the performance of state-of-the-art CNN classifiers using two datasets containing original images and RoI (Region of Interest) extracted images. Piotr [27] proposed a novel and effective image denoising method, which utilizes noise obtained from BM3D filtering and clean scans for training deep learning models. Moreover, this method serves as a preprocessing step that enhances disease classification performance and extends to other image analysis tasks. Guy et al. [28] used edge detectors based on the Sobel and Scharr operators to generate images with detected boundaries between the original X-ray elements. The generated images were used to train a shallow CNN to classify images as either clear or lacking in quality. In this study, a new weakly supervised preprocessing method is proposed

to address some noise in dermatological images, such as hair, light, and scale markers. This method relies on morphological dilation and erosion to eliminate most of the noise, enabling accurate analysis and diagnosis of skin diseases by the model.

2.3. Feature Fusion Network

Medical images typically contain various types of features, such as morphological features, texture features, and color features. These features can provide crucial information about diseases, lesions, or organ structures. By integrating multiple features and merging features at different scales, it is possible to capture both detailed information and the overall structure of the image, which helps comprehensively assess the features of the lesion area. Wenyu et al. [29] proposed a precise skin lesion segmentation deep network based on a U-shaped structure. Specifically, the network incorporates two lightweight attention modules: the Adaptive Channel Context-Aware Pyramid Attention (ACCAPA) module and the Global Feature Fusion (GFF) module. Yilan [30] designed a dual-branch HMT module for image modality fusion, and the proposed TFormer achieved state-of-the-art performance on benchmark datasets. To perform skin lesion classification by integrating features from dermatoscopic and clinical images, Yiguang et al. [31] introduced a multi-scale fully shared fusion network (MFF-Net). Shao et al. [32] proposed a multi-scale feature fusion network (MSF-Net) based on the Comprehensive Attention Convolutional Neural Network (CA-Net). They introduced a spatial attention mechanism into the convolutional blocks through residual connections, which focuses on key regions. Inspired by [33], this paper proposes an improved multi-scale feature fusion (MFF) module based on FPN, channel attention, and spatial attention. MFF allows for more accurate capture of target details and structural information, enhancing the recognition stability of recognition.

3. Method

3.1. Overview of the Model Architecture

Figure 2 illustrates the overall architecture of our method, which integrates three components: a weakly supervised preprocessing module, a multi-view ensemble module, and a multi-scale feature fusion module. The preprocessing module removes noise (e.g., hair, background clutter) from input images so that the network can focus on the lesion area. The multi-view ensemble module combines three distinct backbone networks (VGG, ResNet, ConvNeXt) to extract complementary features. The multi-scale feature fusion module then aggregates these features at multiple scales to enhance the representation of skin lesion regions. Adaptive pooling, two fully connected layers, and softmax form the final classifier, which outputs the prediction.

3.2. Weakly Supervised Multi-View (WSM) Module

In Figure 2, the proposed model has three different inputs. The input of view 1 is the original image, the input of view 2 is the image after segmentation of the lesion, and the input of view 3 is the image after removing the noise in the original image. The WSM module improves the learning effect by integrating information from different views to complement and cooperate with each other. In addition, the WSM module helps the model discover deeper and more discriminative feature representations to improve learning efficiency and performance.

3.2.1. Weakly Supervised ROI algorithm

Skin disease images often contain occlusions and background interference that can distract the model. To mitigate this, we extract the region of interest (ROI) using a weakly supervised approach based on GrabCut [34]. GrabCut is an interactive segmentation algorithm that models foreground and background color distributions via Gaussian mixture models and iteratively refines the segmentation through graph cuts. In our implementation, the initialization is performed automatically: we first apply a simple thresholding step on the grayscale image to obtain a coarse binary mask, then use the bounding box of the largest connected component as the initial ROI. No manually labeled data or additional training is required—the segmentation relies solely on the intrinsic color and spatial properties of the image. This distinguishes our method from fully supervised segmentation models (e.g., U-Net, FCN) that demand large annotated datasets, and from purely unsupervised methods that lack any guiding prior. Hence, we term it "weakly supervised," as it leverages domain-agnostic heuristics and image statistics to obtain the ROI without extra supervision.

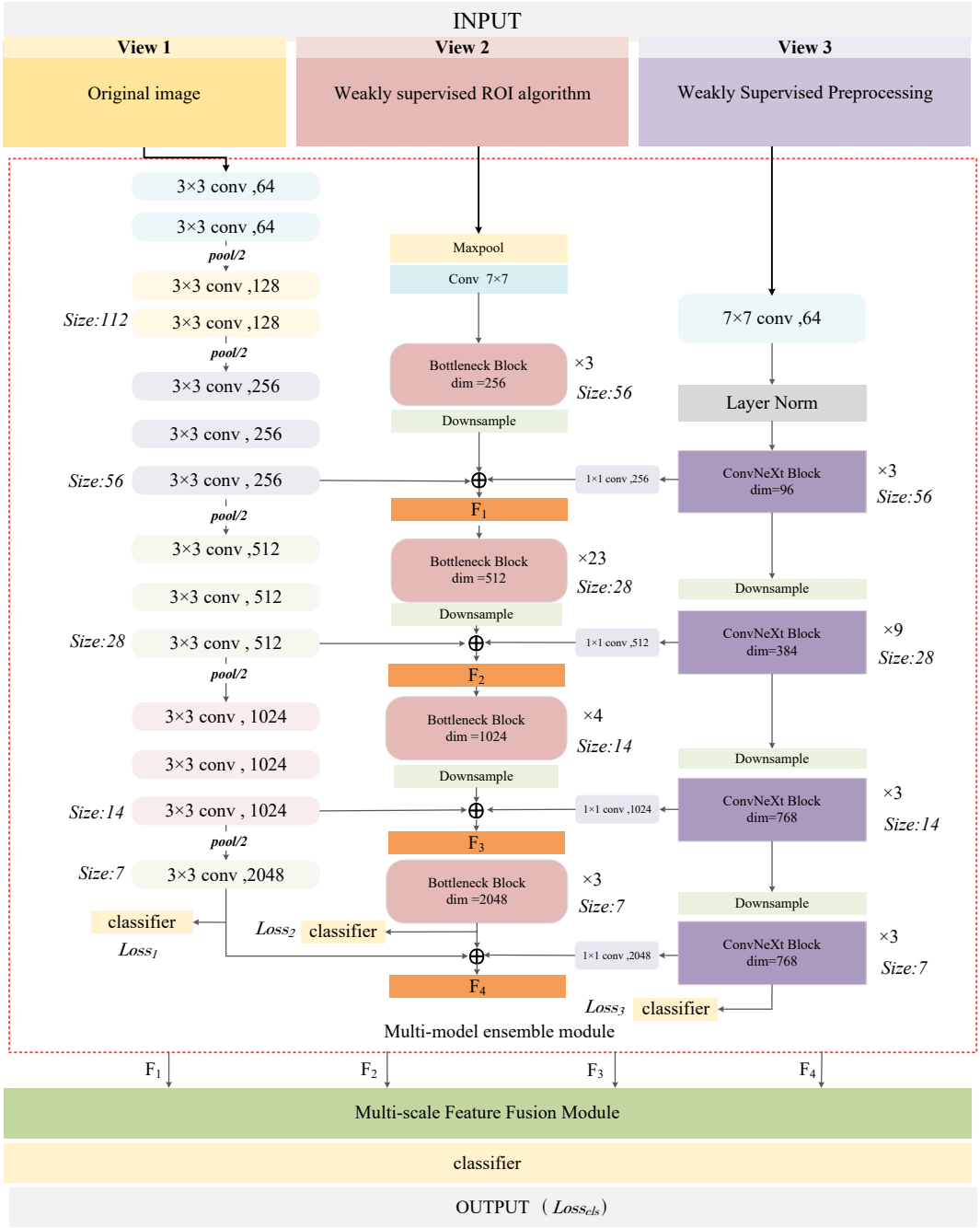


Figure 2. Illustration of the proposed method. The structure mainly includes three parts: weakly supervised preprocessing, a multi-view ensemble module, and a multi-scale feature fusion module.

3.2.2. Weakly Supervised Preprocessing

There are some noise problems in some skin lesion images, such as boundary blurring and hair artifacts. This is a feature that we do not need and do not want the network to learn. To tackle the issues described above, we present a weakly supervised preprocessing method (Figure 3). The method is a two-branch structure. The first branch [35] includes the original image, grayscale conversion, black-hat morphological operation for highlighting hair-like structures, mask generation, and image restoration using the obtained mask. The second branch consists of dilating and then eroding the original image [36], where the dilation kernel size is 3×3, and the erosion kernel size is 5×5. Summing the outputs of the two branches yields the cleaned image. This process requires no additional training, hence the term "weakly supervised." After preprocessing, each input image $I \in R^{(w \times h \times d)}$ yields three cleaned views of size $3 \times 224 \times 224$:

$$M_{\text{image}} = WSP(I). \tag{1}$$

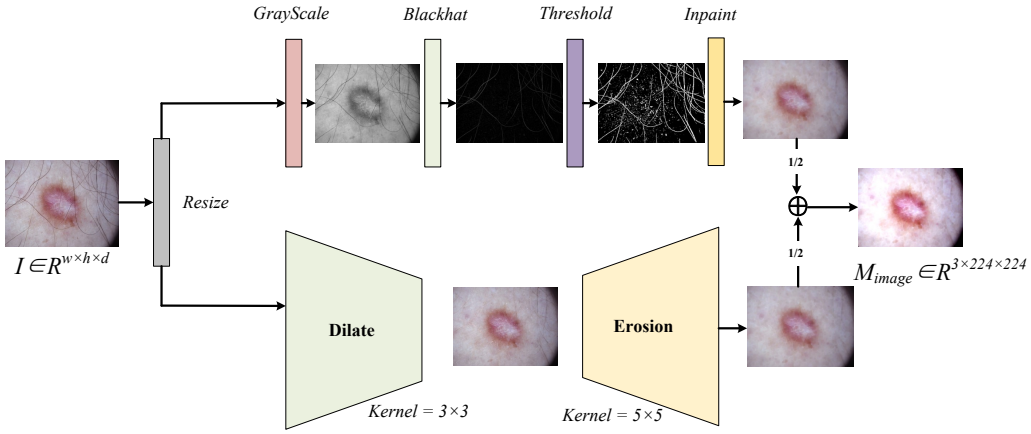


Figure 3. Illustration of the weakly supervised preprocessing module.

3.3. Multi-View Ensemble Module

Ensemble learning improves prediction performance and robustness by combining multiple base learners. Our multi-view ensemble module (Figure 1) consists of three parallel branches, each built on a different backbone: VGG [37], ResNet [38], and ConvNeXt [39]. Lateral connections allow information exchange among branches. The VGG branch uses 3 × 3 convolutions and pooling; the ResNet branch employs bottleneck blocks (Figure 4a) with 1 × 1 and 3 × 3 convolutions, downsampling, and skip connections; the ConvNeXt branch (Figure 4b) uses depthwise convolution [40], LayerNorm [41], GELU [42], 1×1 convolution, drop path, and skip connections.

Therefore, the outputs F_1, F_2, F_3, F_4 of the multi-view ensemble module can be expressed as follows.

$$\begin{aligned}
F_1 &= (vgg_{(256\times56\times56)} \oplus \text{resnet}_{(256\times56\times56)} \oplus \textbf{Cov}\,(\text{ ConvNext }_{96\times56\times56})_{(256\times56\times56)})/3,\\
F_2 &= (vgg_{(512\times28\times28)} \oplus \text{resnet}_{(512\times28\times28)} \oplus \textbf{Cov}\,(\text{ConvNext}_{384\times28\times28})_{(512\times28\times28)})/3,\\
F_3 &= (vgg_{(1024\times14\times14)} \oplus \text{resnet}_{(1024\times14\times14)} \oplus \textbf{Cov}\,(\text{ConvNext}_{768\times14\times14})_{(1024\times14\times14)})/3,\\
F_4 &= (vgg_{(2048\times7\times7)} \oplus \text{resnet}_{(2048\times7\times7)} \oplus \textbf{Cov}\,(\text{ConvNext}_{96\times7\times7})_{(2048\times7\times7)})/3,
\end{aligned} \tag{2}$$

where **Conv** stands for 1 × 1 convolution operations to change the number of channels, $\oplus$ stands for element-wise addition.

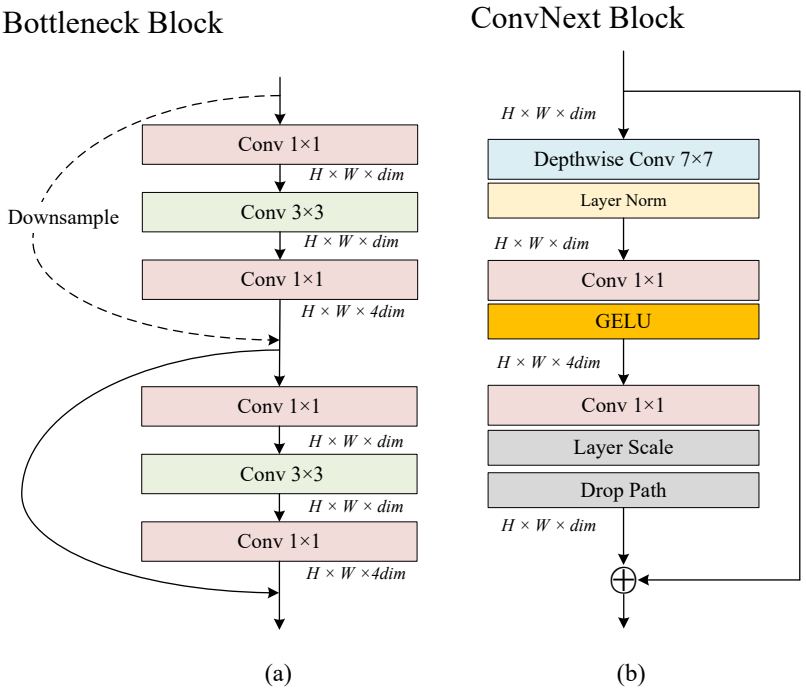


Figure 4. Illustration of (**a**) Bottleneck block and (**b**) ConvNext block.

3.4. Multi-Scale Feature Fusion Module

The multi-scale feature fusion (MFF) module, detailed in Algorithm 1, processes the four multi-view outputs $\{F_1, F_2, F_3, F_4\}$. First, a Feature Pyramid Network (FPN) [43] generates pyramid features $\{P_1, P_2, P_3, P_4\}$.

$$P_{\mathrm{n}} = \mathbf{FPN}(F_n)\{n = 1, 2, 3, 4\} \tag{3}$$

where **FPN** refers to the feature pyramid structure [43].

Algorithm 1 Multi-scale Feature Fusion (MFF)

Input: P_1, P_2, P_3, P_4
Output: C
for all $n = 1$ to 4 **do**
  $P_n \leftarrow FPN(F_n)$
end for
for all $n = 1$ to 4 **do**
  $S_n^c \leftarrow MaxPool\,(P_n) + AvgPool\,(P_n)$
  $A_n^{(s)} \leftarrow \sigma\left(f^{3\times3}\left(\boldsymbol{S}_n^c\right)\right)$
  $A_n^{(c)} = \sigma\left(W_2 \cdot ReLU\left(W_1 \cdot AvgPool\,(P_n)\right)\right)$
  if $l \neq 2, 4$ **then**
    $B_n \leftarrow P_n \otimes (A_n^c \oplus A_n^s)/2$
  else
    $S_n^{cw} \leftarrow MaxPool\,(\boldsymbol{P}_n) + AvgPool\,(P_n)$
    $S_n^{ch} \leftarrow MaxPool\,(P_n) + AvgPool\,(P_n)$
    $A_n^{(i)} \leftarrow \left(\sigma(f^{7\times7}(S_{cw})) + \sigma(f^{7\times7}(S_{cw}))\right)/2$
    $B_n \leftarrow P_n \otimes (A_n^c \oplus A_n^s) \oplus A_n^i)/3$
  end if
end for
$C \leftarrow Concatenate\{B_1, B_2, B_3, B_4\}$

Then, $\{P_1, P_2, P_3, P_4\}$ will enter the attention module to get $\{B_1, B_2, B_3, B_4\}$. As shown in Figure 5, the attention module can be divided into channel attention($A_n^{(c)}$) , spatial attention($A_n^{(s)}$) [44] and channel-spatial Interactive attention [45], which are obtained by the following formula.

Figure 5a presents the expression of $A_n^{(s)}$ as follows.

$$\begin{aligned} \boldsymbol{S}_n^c &= \mathrm{MaxPool}\left(\boldsymbol{P}_n\right) + \mathrm{AvgPool}\left(\boldsymbol{P}_n\right) \\ \boldsymbol{A}_n^{(s)} &= \sigma\left(f^{3\times3}\left(\boldsymbol{S}_n^c\right)\right), \{n = 1, 2, 3, 4\} \end{aligned} \tag{4}$$

Here "+" indicates a feature map connection, σ is the sigmoid activation function, $AvgPool$ and $MaxPool$ correspond to average and max pooling, $f^{3\times3}$ represents a 3×3 convolution operation, respectively. As shown in Figure 5b, $A_n^{(c)}$ can be as follows.

$$\boldsymbol{A}_n^{(c)} = \sigma\left(\boldsymbol{W}_2 \cdot \mathrm{ReLU}\left(\boldsymbol{W}_1 \cdot \mathrm{AvgPool}\left(\boldsymbol{P}_n\right)\right)\right), \{n = 1, 2, 3, 4\} \tag{5}$$

The sigmoid and ReLU activation functions are denoted by σ and ReLU, respectively. Two 2D convolutional layers have weight matrices W_1 and W_2, and $\cdot$ denotes element multiplication. Figure 5c is the diagram of channel-spatial Interactive attention, which can be divided into two branches. One branch is designed to capture interactions along the channel C and spatial W dimensions, while the other focuses on interactions along the channel C and spatial H dimensions. Permute indicates the change of the dimensional feature [45]. $A^{(i)}$ can be as follows:

$$\begin{aligned} \boldsymbol{S}_n^{cw} &= \mathrm{MaxPool}\left(\boldsymbol{P}_n\right) + \mathrm{AvgPool}\left(\boldsymbol{P}_n\right), \\ \boldsymbol{S}_n^{ch} &= \mathrm{MaxPool}\left(\boldsymbol{P}_n\right) + \mathrm{AvgPool}\left(\boldsymbol{P}_n\right), \\ \boldsymbol{A}_n^{(i)} &= \left(\sigma(f^{7\times7}(\boldsymbol{S}_{cw})) + \sigma(f^{7\times7}(\boldsymbol{S}_{cw}))\right)/2, \{n = 2, 3\}, \end{aligned} \tag{6}$$

where "+" indicates a feature map connection, $f^{7\times7}$ represents a 7×7 convolution operation, respectively. Figure

5c presents $\boldsymbol{B}$ in the following form:

$$\begin{aligned} \boldsymbol{B}_n &= \boldsymbol{P}_n \otimes \left(\boldsymbol{A}_n^{(c)} \oplus \boldsymbol{A}_n^{(s)} \right), \{n = 1, 4\}, \\ \boldsymbol{B}_n &= \boldsymbol{P}_n \otimes \left(\boldsymbol{A}_n^{(c)} \oplus \boldsymbol{A}_n^{(s)} \oplus \boldsymbol{A}^{(i)} \right), \{n = 2, 3\}. \end{aligned} \tag{7}$$

Here, $\otimes$ denotes element-wise multiplication, while $\oplus$ indicates brodcadting addition. Finally, the output C of the MFF module can be obtained as follows:

$$\boldsymbol{C} = Concatenate\{\boldsymbol{B_1}, \boldsymbol{B_2}, \boldsymbol{B_3}, \boldsymbol{B_4}\}, \tag{8}$$

where $Concatenate$ represents the concatenation of channels.

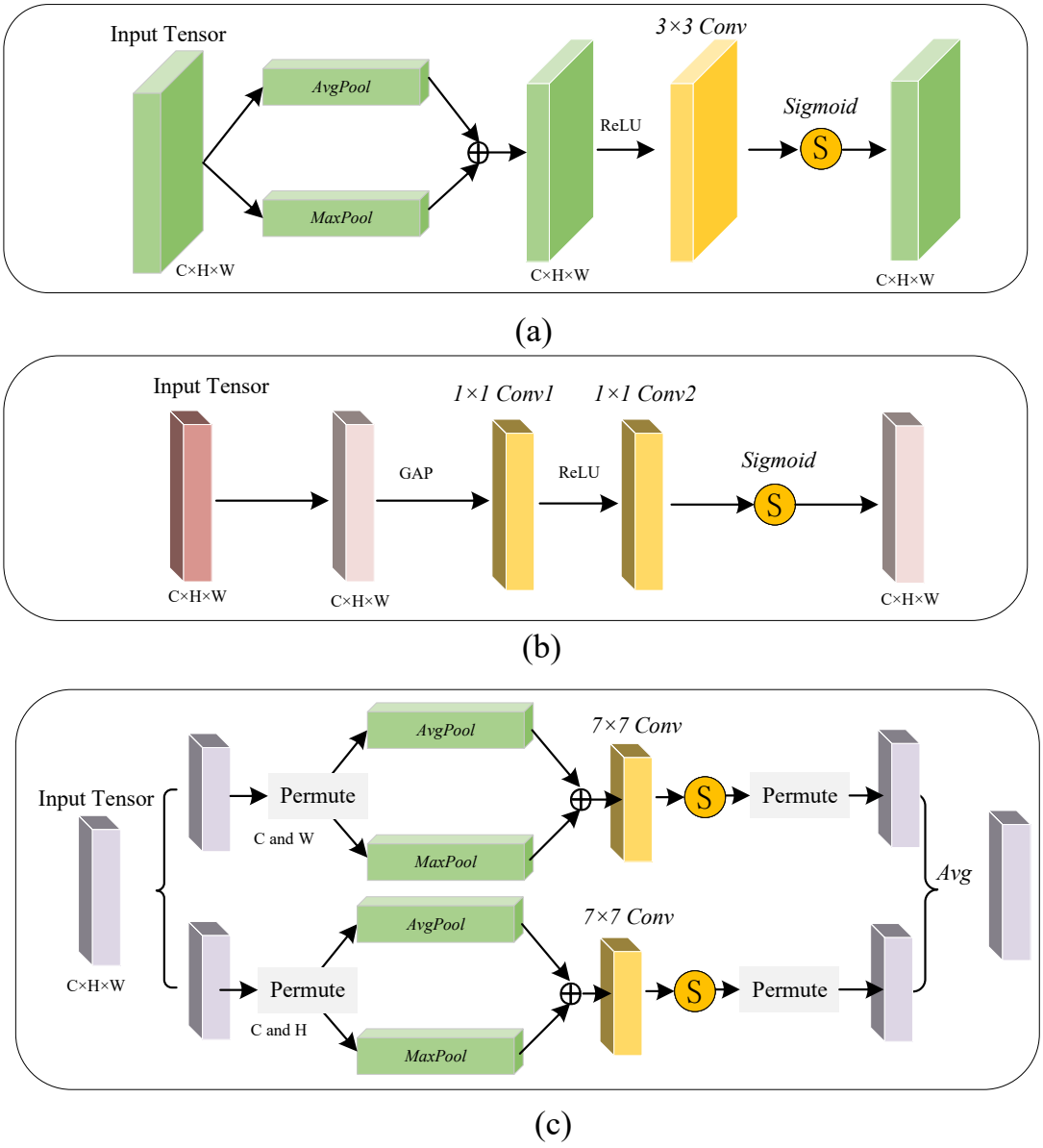


Figure 5. (**a**) Channel attention, (**b**) Spatial attention, and (**c**) Channel-spatial Interactive attention.

3.5. Multi-Ensemble (ME) Loss

In multi-view learning, there will be multiple tasks that need to be trained together and multiple tasks need to be optimized at the same time. This paper proposes ME loss (Algorithm 2), which combines the loss functions of multiple tasks together for joint optimization, and balances the influence of each task in the overall loss function by setting the weights of different tasks.

Algorithm 2 ME loss

Input: I: skin lesion images with RGB.

Require: $s = 0$: initialize iteration; θ:parameters after iteration; w_s: weight after the s-th iteration; b_s: bias after the s-th iteration; E : the number of iterations.

Require: $Loss_1$, $Loss_2$, $Loss_3$ and $Loss_{cls}$ represent the loss between the true value T and predicted value O_1, O_2, O_3, O_{cls}, respectively. T: ground-truth labels (one-hot encoded), each branch output (O_1, O_2, O_3, O_cls) corresponds to the logits of a specific module.

Require: $f(\theta, T)$: cross-entropy loss function with a set θ of all parameters, the loss between the predicted results (O) and (T).

Require: η: learning rate; Three hyper-parameters, namely λ, γ, and α, are introduced to respectively control the contribution of view 1, view 2, and view 3 to the classification.

Output: the ME loss of s-th iteration:$Loss_{me}(s)$.

By using the proposed model, T and θ_s can be obtained according to X_{in}.
for all $Loss_{me}(s) \neq 0$ and $0 < epoch < E$ **do**
  $O_1, O_2, O_3, O_{cls} = f(\theta_s, T)$
  $Loss_i = -\sum_{c=1}^{C} \tilde{y}_c \log\left(\frac{\exp(O_{i,c})}{\sum_j \exp(O_{i,j})}\right), i \in (1,2,3)$
  $Loss_1 \leftarrow O_1 - T$
  $Loss_2 \leftarrow O_2 - T$
  $Loss_3 \leftarrow O_3 - T$
  $Loss_{cls} = O_{cls} - T.$
  $Loss_{me}(s) \leftarrow Loss_{cls} + \lambda Loss_1 + \gamma Loss_2 + \alpha Loss_3$
  $\theta_s \leftarrow w_s, b_s$
  $\nabla_{\theta_s} \leftarrow \frac{\partial Loss_{me}}{\partial \theta_s}$
  $\eta \leftarrow 1 \times 10^{-3}$
  $\theta \leftarrow \theta - \eta * \nabla_{\theta_s}$
  $s \leftarrow s+1$
end for

We assume that $\boldsymbol{Z}$ is the output of the $\boldsymbol{C}$ input classifier.

$$\boldsymbol{Z}(\boldsymbol{C}) = Softmax(FC(Pooling(\boldsymbol{C}))), \tag{9}$$

where $Pooling$ is adaptive pooling, FC a fully connected layer, and $Softmax$ the softmax activation function. We then can define a classification loss function L_{cls} as:

$$Loss_{cls} = -\tilde{\mathbf{y}} \cdot \log\left(Z\left(\mathbf{C}\right)\right), \tag{10}$$

$$Loss_i = -\sum_{c=1}^{C} \tilde{y}_c \log\left(\frac{\exp\left(O_{i,c}\right)}{\sum_j \exp\left(O_{i,j}\right)}\right), i \in (1,2,3). \tag{11}$$

The ground-truth category label is encoded into a one-hot vector denoted by $\tilde{\mathbf{y}}$. The proposed ME loss combines the objectives of multiple ensemble modules($Loss_1$, $Loss_2$, $Loss_3$ in Figure 2) into a single loss function to simplify the problem-solving process. Compared with optimizing each objective function separately, the ME loss function can incorporate the relationship and trade-off between each objective into a unified framework, thereby reducing the complexity of the optimization problem. The ME loss formula is as follows:

$$Loss_{me} = Loss_{cls} + \lambda Loss_1 + \gamma Loss_2 + \alpha Loss_3. \tag{12}$$

We set $\lambda = 0.4$, $\gamma = 0.3$, and $\alpha = 0.1$ to weigh the contributions from the three views in the overall loss. These values were determined empirically based on preliminary validation experiments and were kept fixed across all experiments.

The ME loss function can promote the learning of richer and more abstract feature representations, which is conducive to improving the performance of the model on new data. Therefore, the generalization ability of the model is improved.

4. Experiments and Results

4.1. Dataset

Dermoscopic images in HAM10000 [46] were obtained from multiple populations and encompass different modalities of acquisition and storage. The dataset contains a total of 10,015 samples across seven diagnostic categories: actinic keratoses and intraepithelial carcinoma (AKIEC), basal cell carcinoma (BCC), benign keratosis-like lesions (BKL), dermatofibroma (DF), melanoma (MEL), melanocytic nevi (NV), and vascular lesions (VASC). To preserve the original class distribution in both training and evaluation, we performed a stratified random split, allocating 90% of the samples to the training set and the remaining 10% to the test set while maintaining the proportion of each class. The split was performed with a fixed random seed (seed = 42) to ensure reproducibility. During training, we monitored the model's performance on the test set after every epoch and selected the checkpoint that achieved the highest test accuracy as the final model. All reported results correspond to this single run with the fixed seed, and the deterministic settings applied throughout the experiments ensure that the obtained performance metrics are reliable and reproducible. During training, we monitored the model's performance on the test set after every epoch and selected the checkpoint that achieved the highest test accuracy as the final model; this protocol was applied uniformly to all compared methods. All experiments were conducted with deterministic settings to guarantee that the obtained results are reliable and reproducible.

4.2. Evaluation Index

In order to evaluate the classification effect of the model, we selected a widely used evaluation index. Performance evaluation across all models was conducted using accuracy, precision, and specificity, derived from the classification outcomes of true positive (TP), false positive (FP), false negative (FN), and true negative (TN). The details are as follows:

$$\begin{aligned} \text{Accuracy} &= \frac{TP+TN}{TP+TN+FP+FN} \\ \text{Precision} &= \frac{TP}{TP+FP} \times 100\% \\ \text{Recall} &= \frac{TP}{TP+FN} \times 100\% \\ F_1 &= 2 \cdot \frac{\text{Precision} \cdot \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}, \end{aligned} \tag{13}$$

for the multi-class classification task with seven categories, the above binary metrics are extended as follows. Overall Accuracy is computed as the micro-average, i.e., defined as the ratio of correctly classified samples to the total number of samples, as derived from the confusion matrix. For Precision, Recall (Sensitivity), F1-score, and Specificity, we first calculate the class-wise values for each of the seven classes using the confusion matrix, and then take the arithmetic mean over all classes – this corresponds to macro-averaging, which treats all classes equally and is particularly suitable for imbalanced datasets like HAM10000. The confusion matrix is presented in a row-normalized form to show the proportion of predictions for each true class. If the Area Under the ROC Curve (AUC) is reported, we adopt the one-versus-rest strategy and compute the macro-average AUC.

To provide a more clinically relevant evaluation, Table 1 presents the per-class precision, recall (sensitivity), specificity, and F1-score for each of the seven diagnostic categories. These metrics are critical for understanding the model's behavior in a clinical context. For high-risk categories such as melanoma (MEL), minimizing missed diagnoses is critical. A recall of 90.09% reflects the model's ability to correctly detect the majority of such cases. The precision of 86.96% for MEL suggests that when the model predicts melanoma, it is correct in nearly 87% of cases, which is clinically acceptable given the severe consequences of false negatives. For the most common benign class, melanocytic nevi (NV), the model achieves near-perfect recall (99.33%) and high precision (98.96%), ensuring that most benign lesions are correctly classified without unnecessary concern. The specificity values across all classes are consistently above 97%, indicating that the model rarely misclassifies other classes as the target class—a desirable property to avoid false alarms. For the rarest class, dermatofibroma (DF), the recall is lower (81.82%) due to the limited number of training samples (approximately 100 images in the full dataset). In clinical practice, such a model could still serve as a triage tool, flagging ambiguous DF cases for expert review while maintaining high specificity (99.95%) to avoid overdiagnosis.

Table 1. Performance breakdown by class on the HAM10000 test set.

Class	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1-Score (%)
VASC	99.90	100.00	92.86	100.00	96.30
NV	98.85	98.96	99.33	97.87	99.14
MEL	97.40	86.96	90.09	98.31	88.50
DF	99.75	94.74	81.82	99.95	87.81
BKL	97.40	89.57	86.30	98.76	87.90
BCC	99.10	91.18	91.18	99.53	91.18
AKIEC	99.30	89.23	89.23	99.64	89.23

4.3. Training Strategy

PyTorch serves as the implementation framework for the proposed network. During training, an NVIDIA RTX3090 GPU was used as the training device. Training was carried out for 50 epochs with a batch size of 24, using an SGD optimizer (momentum 0.9, weight decay 5×10^{-4}, initial learning rate 0.001) coupled with a cosine annealing learning rate scheduler. Reproducibility was ensured by fixing the random seed to 1 across all runs. Input images were resized to 224 × 224 and augmented with random horizontal flips, color jitter, random resized cropping, and Cutout. Validation was performed at the end of every epoch. the checkpoint with the highest validation accuracy was saved as the best model. No early stopping was applied—the full 50 epochs were always run. All ensemble models were pre-trained on ImageNet.

4.4. Cross-Validation Evaluation

To further evaluate the robustness of the proposed model and to eliminate any concern regarding potential test-set leakage, we conducted a 5-fold stratified cross-validation on the HAM10000 dataset. The dataset was stratified by class and randomly divided into five folds. In each fold, 80% of the data was used for training, and the remaining 20% served as the test set. The model was trained for the same number of epochs as in the original single-split experiment, and the test performance was recorded at the final epoch. All metrics were computed from the confusion matrix using TP, FP, FN, and TN counts per class, as defined in Equation (13).

Table 2 summarizes the test accuracy for each fold, along with the macro-average metrics across the five folds. The mean accuracy is 97.28% with a standard deviation of 0.77%, and the macro-average F1-score is 94.52% ± 0.60%. These results are highly consistent with the original single-split result (95.90% accuracy). The low standard deviation across folds confirms that the model's performance is stable and not dependent on a particular data partition. Detailed per-class metrics for each fold are provided in the supplementary material.

Table 2. 5-fold cross-validation results on HAM10000.

Fold	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
1	96.80	95.24	94.15	94.37
2	98.10	96.32	96.46	96.34
3	96.11	94.49	94.02	94.14
4	97.50	94.86	95.17	94.98
5	97.90	94.87	95.90	95.31
Mean	97.28 ± 0.77	95.16 ± 0.70	95.14 ± 0.92	95.03 ± 0.85

4.5. Results

4.5.1. Research on Weakly Supervised Multi-View Modules

As shown in Figure 6, the inputs of three different views are visualized. View 1 is the original image. By inputting the original image into the model, the complexity of the model can be increased, so that the model can better learn image features and distinguish different image categories. View 2 is the segmented image of the region of interest. Image segmentation can provide accurate location information of the target, so that the classification model can locate and identify the target more accurately. View 3 is the de-noised image. The de-noising operation can remove the subtle noise and chaotic information in the image, making the features of the image more prominent and clear.

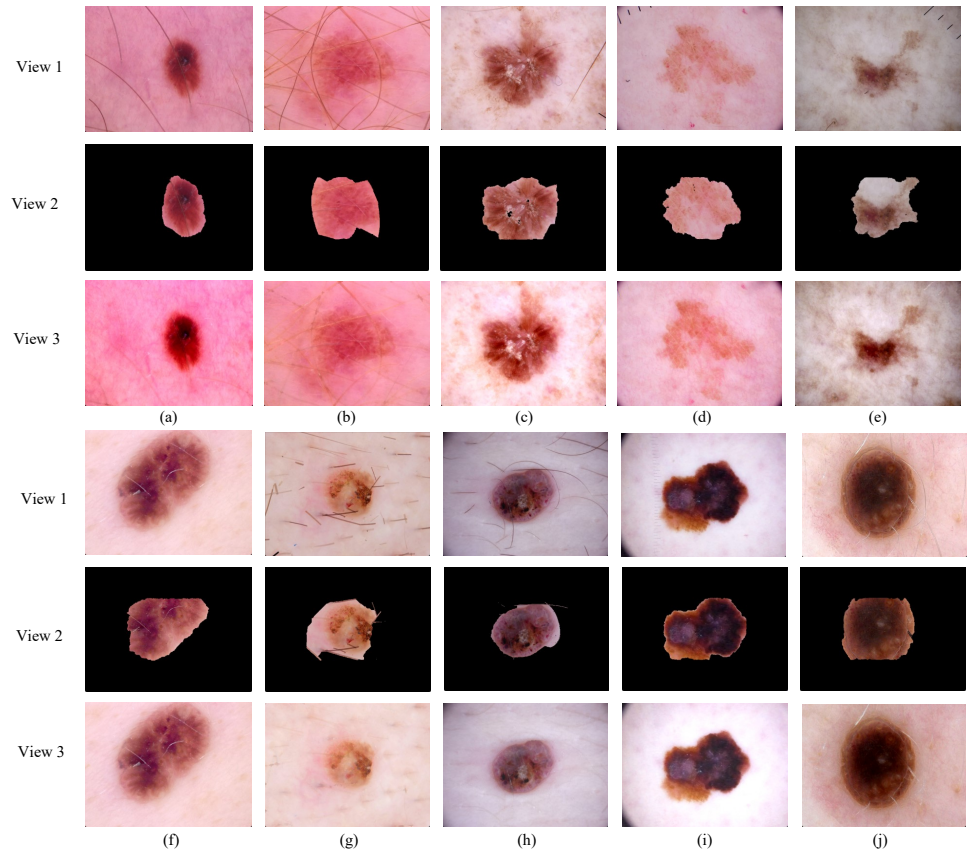


Figure 6. Visualization results for three different views.

Images from three different views can provide different information, and by feeding them into the classification model, the model can better learn and extract rich visual features. Each view may highlight different object parts or features. The integration of multi-view information yields richer and more precise feature representations, leading to improved classification performance.

4.5.2. Comparison with Traditional Models

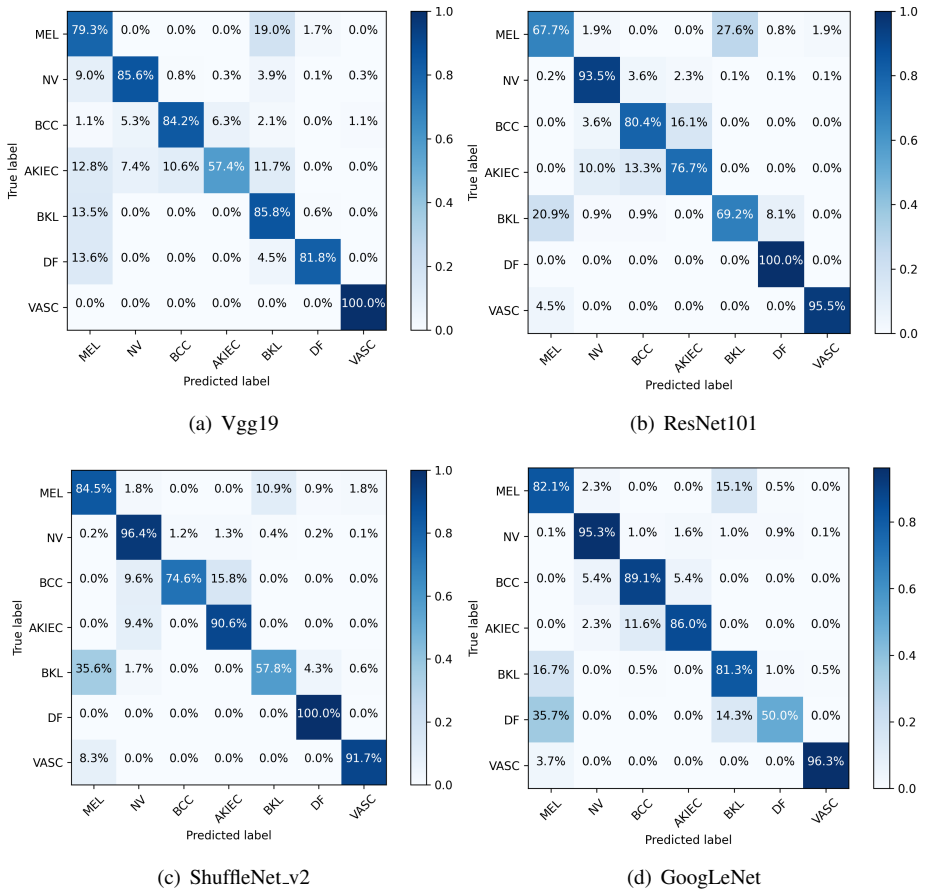
To validate the proposed method, we compare it against several widely used classification models, including Vgg19, ResNet101, ShuffleNet_v2, GoogLeNet, DenseNet121, Swin_transformer, and ConvNeXt. All models are trained under identical conditions to ensure a fair comparison.

The experiment results are shown in Table 3. From the accuracy point of view, the proposed model is 95.90%, while the accuracy of other models is about 90%, which indicates that our model is better than common models. From F1's point of view, the proposed model's 91.11% performance is far ahead of other models; such a result reflects a well-maintained balance between positive and negative instance identification. Accordingly, the proposed model outperforms the dominant traditional approaches on the HAM10000 dataset.

Table 3. The performance of the proposed method is compared with the traditional method on the HAM10000 dataset.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Vgg19 [47]	86.99	75.26	73.37	73.91
ResNet101 [38]	90.40	89.61	90.11	89.89
ShuffleNet_v2 [48]	88.59	77.17	69.57	71.31
GoogLeNet [49]	91.99	81.49	79.07	80.89
DensenNet121 [50]	93.74	87.37	83.64	85.36
ConvNeXt [39]	86.24	74.92	65.06	64.47
Swin_transformer [51]	89.74	81.34	78.57	79.12
Proposed	**95.90**	**93.30**	**90.16**	**91.44**

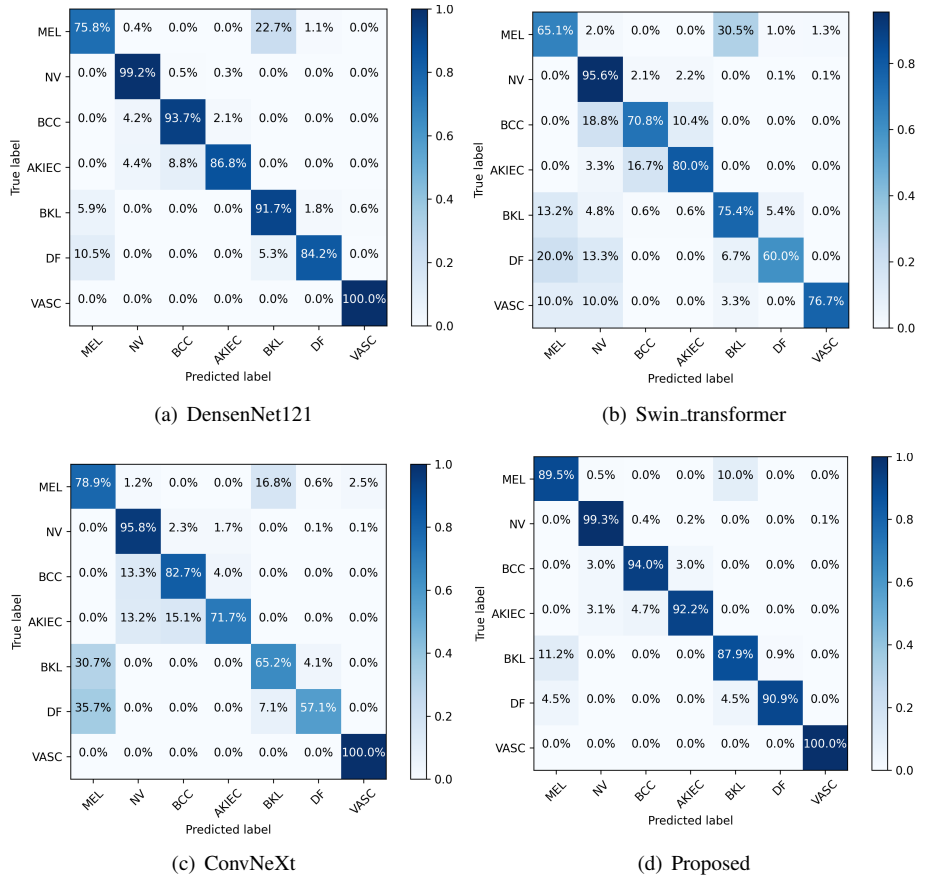
In Figures 7 and 8, where rows represent actual categories and columns denote predicted ones, the proposed method consistently outperforms all compared models, demonstrating its superior performance.



(a) Vgg19 (b) ResNet101

(c) ShuffleNet_v2 (d) GoogLeNet

Figure 7. The confusion matrix of vgg19,resnet101, shuffleNet_v2 and GoogLeNet.



(a) DensenNet121 (b) Swin_transformer

(c) ConvNeXt (d) Proposed

Figure 8. The confusion matrix of densenNet121, swin_transformer,convnext and proposed method.

4.5.3. Ablation Studies

To evaluate the contribution of individual components, we performed extensive ablation studies. In Table 4, we divide the proposed model into six key units: the weakly supervised preprocessing module (WSM), three independent branches (View 1, View 2, and View 3), the multi-scale feature fusion module (MFF), and the multi-ensemble loss (ME loss). All training optimization parameters and settings were fixed across experiments to ensure a fair comparison. The backbones of the three views are not shared; each is a distinct architecture (e.g., ResNet-101, VGG-19, ConvNeXt) pre-trained on ImageNet and fine-tuned independently. Their classification heads are also branch-specific. The MFF module operates on the features extracted by the branches, and the ME loss combines the cross-entropy losses of all branches with fixed weighting coefficients (as detailed in Section 3.5).

Table 4. Performance contribution of the proposed model's submodules. The components are defined as: Preprocessing (WSM: Weakly Supervised Multi-view) enhances feature extraction; View 1, View 2, and View 3 are three independent branches with distinct backbones and branch-specific heads; Multi-scale feature fusion (MFF) serves to integrate information from multiple scales. ME loss (Multi-ensemble Loss) combines branch losses via fixed weights. All backbones are ImageNet-pretrained and updated independently; the final prediction is obtained by averaging branch logits.

No.	Model						Performance	
	WSM	View 1	View 2	View 3	MFF	ME Loss	Accuracy (%)	F1 (%)
1	✓	×	✓	✓	✓	✓	94.94	88.80
2	✓	✓	×	✓	✓	✓	89.39	74.91
3	✓	✓	✓	×	✓	✓	95.79	91.36
4	✓	✓	✓	✓	×	✓	95.24	89.31
5	✓	✓	✓	✓	✓	×	87.89	69.79
6	×	✓	✓	✓	✓	✓	94.69	90.36
7	✓	✓	✓	✓	✓	✓	95.90	91.44

The ME loss plays a critical role because it provides direct supervision to each view branch. Without it (row 5), the model is trained only with a loss on the final fused output, which offers weaker gradient signals to the individual branches, especially their early layers. This explains the substantial performance drop when ME loss is removed, and highlights that joint optimization of all branches is essential for learning discriminative features.

From the accuracy and F1-score analysis in Table 4, we observe that View 3 contributes the least to the proposed model, as removing it causes the smallest performance drop. Conversely, the ME loss contributes the most, since its removal leads to the largest degradation in performance.

4.5.4. Quantitative Results on the HAM10000 Dataset

We compared the proposed approach with other studies on the HAM10000 dataset. To ensure a fair comparison, all methods were trained on the same dataset described above. The results, presented in Table 5, show that our model significantly outperforms the others, confirming its advantage in lesion image recognition. Notably, unlike the competing approaches, our method is built upon ensemble learning. Ensemble learning takes into account the opinions of multiple models and is robust to noise and abnormal data. The proposed model is more reliable in the face of complex real data.

Table 5. The proposed method is evaluated against state-of-the-art techniques on the HAM10000 dataset.

Method	# of Images	Method	Precision (%)	Recall (%)	F1 (%)	Accuracy (%)
Karar Ali et al. [52]	10015	Transfer learning	86.00	86.00	85.00	87.90
Alhudhaif et al. [53]	10015	SAB-CNN	90.14	89.70	89.29	95.41
Muhammad et al. [54]	10015	Transfer learning	86.00	86.01	85.83	89.26
Mohamed et al. [55]	10015	IARO	90.02	89.65	89.52	89.65
Alani et al. [56]	10015	CNN	84.00	82.00	81.00	95.00
Karar et al. [21]	10015	EfficientNet	86.20	86.31	85.29	87.90
Proposed	10015	Ensemble learning	**93.30**	**90.16**	**91.44**	**95.90**

To better evaluate the performance of our proposed method, we benchmark it against several recent deep learning approaches reported in the literature. Table 5 presents a detailed per-class comparison based on the confusion matrix. In addition, Table 6 presents a performance comparison between our model and three recent state-of-the-art methods (Mg-EDCF [12], DeepSkinNet [13], and PanDerm [14]) in terms of average accuracy and F1-score. All results are obtained on the HAM10000 dataset.

Table 6. Detailed results (per-class and average) on HAM10000, contrasted with current state-of-the-art models.

Method	Metric	VASC	NV	MEL	DF	BKL	BCC	AKIEC	Average
MG-EDCF [12]	Accuracy	98.33	**99.00**	**98.00**	97.33	**99.00**	98.67	97.00	98.19
	F1-score	98.28	98.97	**97.95**	97.20	**98.95**	**98.69**	97.03	**98.15**
DeepSkinNet [13]	Accuracy	99.65	98.58	94.27	98.62	96.54	97.32	96.50	97.35
	F1-score	**100.00**	98.34	94.64	**98.37**	96.15	98.13	**97.32**	97.16
PanDerm [14]	Accuracy	-	-	-	-	-	-	-	92.60
	F1-score	-	-	-	-	-	-	-	93.90
Ours	Accuracy	**99.90**	98.85	97.40	**99.75**	97.40	**99.10**	**99.30**	**98.81**
	F1-score	96.30	**99.14**	88.50	87.81	87.90	91.18	89.23	91.44

With an average accuracy of 98.81%, our method outperforms Mg-EDCF (98.19%) and DeepSkinNet (97.35%), as shown in Table 6. Although the average F1-score of our model (91.44%) is lower than that of Mg-EDCF and DeepSkinNet, this is primarily because our per-class F1-scores (reported in Table 5) are computed from the confusion matrix using standard definitions, whereas the compared methods may employ different averaging schemes or evaluation protocols. Figure 9 presents the ROC curves of the proposed method for each disease category, where the x-axis is the false positive rate and the y-axis is the true positive rate. The curves consistently approach the top-left corner, demonstrating that our model achieves high sensitivity and specificity across all classes.

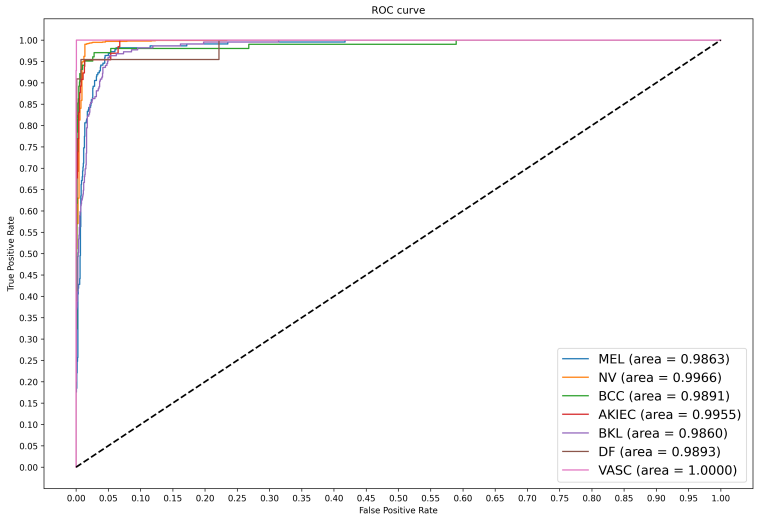


Figure 9. The ROC curve of the proposed method.

4.6. Discussion

Recent advances in multi-class classification have been driven by the growing depth of research in medical imaging and deep learning techniques. However, there are still some challenges in the field of skin lesion image recognition, such as hair and other artifacts in the images that distract the model. In addition to this, the small inter-class and large intra-class differences in skin image recognition are also a very thorny issue. Considering the morphological characteristics of skin disease images, a weakly supervised image preprocessing method is proposed to deal with the noise in skin disease images, which can clean the hair noise and make the lesion area more obvious. The effectiveness of individual components within the proposed models was evaluated through ablation experiments, which were designed to address the following problems. Aiming at the problem of the small inter-class and large intra-class differences, a multi-view and multi-scale ensemble learning model was proposed. The model can integrate features of multiple basic models well, capture information of different scales, and explore and utilize the advantages of different features. By reducing inter-class similarity and minimizing intra-class variability, the proposed method yields superior predictive performance, robustness, and generalization. In addition, the proposed model is a multi-scale model. By synthesizing information at different scales, the multi-scale model can better distinguish the differences between different categories, thus improving classification performance, especially for data sets with complex intra-class distributions.

Despite its strengths, the proposed model still exhibits certain shortcomings. For example, the proposed model requires large-scale data for training, especially for skin diseases, the demand for data is greater. This can be a challenge in some areas or for specific tasks, as getting high-quality, diverse data is not always easy.

5. Conclusions

In this paper, we proposed a multi-view ensemble-based weakly supervised model for skin lesion diagnosis in dermoscopic images. A weakly supervised preprocessing method is introduced to reduce noise in the input images. To address the challenges of small inter-class and large intra-class variations in skin disease images, we adopted ensemble learning and designed a multi-scale feature fusion (MFF) module. Experiments conducted on the public HAM10000 dataset demonstrate that: (1) the proposed model achieves competitive performance compared with several existing methods; (2) each component of the model contributes positively to the overall performance; and (3) the model attains an accuracy of 95.90% on this benchmark dataset. Overall, the proposed method shows promising results on the HAM10000 dataset for skin disease image recognition.

In future work, we plan to reduce the model's dependence on large-scale training data and explore model lightweighting techniques to improve its applicability in resource-constrained scenarios.

Author Contributions

Q.H.: conceptualization, methodology, supervision, writing—review and editing; H.Z.: methodology, software, validation, writing—original draft preparation; T.W.: data curation, software, visualization; Y.T.: investigation, validation, writing—review, and editing; Z.L.: formal analysis, resources, data curation; Y.L. (Yadong Lan): investigation, validation, resources; Y.L. (Yuhui Lin): software, methodology, experiments; S.Y.: supervision, project administration, funding acquisition; Y.W.: conceptualization, methodology, supervision, and editing; Y.P.: supervision, writing—review, and editing. All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by Technological Research Program of Chongqing Municipal Education Commission (KJQN202301517, HZ2021015, KJZD-K202100104, KJQN202301543).

Institutional Review Board Statement

Not applicable. This study used a publicly available, de-identified benchmark dataset (HAM10000) and did not involve any new data collection from human subjects or animals. Therefore, institutional review board approval was not required.

Informed Consent Statement

Not applicable. This study used a publicly available, de-identified benchmark dataset (HAM10000) and did not involve direct interaction with human subjects or collection of new patient data.

Data Availability Statement

Not applicable.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

During the preparation of this work, the authors used DeepSeek to polish the repetitive content. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

1. Siegel, R.L.; Kratzer, T.B.; Giaquinto, A.N.; et al. Cancer Statistics, 2025. *CA Cancer J. Clin.* **2025**, *75*, 10.
2. Mangione, C.M.; Barry, M.J.; Nicholson, W.K.; et al. Screening for Skin Cancer: US Preventive Services Task Force Recommendation Statement. *JAMA* **2023**, *329*, 1290–1295.

3. Hernández-Pérez, C.; Podlipnik, S.; Ficapal, J.; et al. Comparative analysis and interpretability of survival models for melanoma prognosis. *Comput. Biol. Med.* **2025**, *190*, 110027.
4. Jensen, J.D.; Elewski, B.E. The ABCDEF Rule: Combining the "ABCDE Rule" and the "Ugly Duckling Sign" in an Effort to Improve Patient Self-Screening Examinations. *J. Clin. Aesthetic Dermatol.* **2015**, *8*, 15.
5. Ravikumar, G.; Satpathy, S.K. Innovative Computer-Aided Techniques for Early Detection of Melanoma using Dermoscopic Image Analysis. In Proceedings of the 2025 International Conference on Electronics and Renewable Systems (ICEARS), Tuticorin, India, 11–13 February 2025; pp. 1565–1571.
6. Bhatt, H.; Shah, V.; Shah, K.; et al. State-of-the-Art Machine Learning Techniques for Melanoma Skin Cancer Detection and Classification: A Comprehensive Review. *Intell. Med.* **2023**, *3*, 180–190.
7. Ye, Z.; Zhang, D.; Zhao, Y.; et al. Deep Learning Algorithms for Melanoma Detection Using Dermoscopic Images: A Systematic Review and Meta-Analysis. *Artif. Intell. Med.* **2024**, *155*, 102934.
8. Gharawi, A.; Alahmadi, M.D.; Ramaswamy, L. Self-Supervised Skin Lesion Segmentation: An Annotation-Free Approach. *Mathematics* **2023**, *11*, 3805.
9. Wang, L.; Zhang, L.; Shu, X.; et al. Intra-Class Consistency and Inter-Class Discrimination Feature Learning for Automatic Skin Lesion Classification. *Med. Image Anal.* **2023**, *85*, 102746.
10. Cheng, J.; Tian, S.; Yu, L.; et al. ResGANet: Residual Group Attention Network for Medical Image Classification and Segmentation. *Med. Image Anal.* **2022**, *76*, 102313.
11. Zhang, H.; Li, Z.; Zhao, H.; et al. Attentive Octave Convolutional Capsule Network for Medical Image Classification. *Appl. Sci.* **2022**, *12*, 2634.
12. Selvan, V. Transforming Skin Cancer Diagnosis: A Deep Learning Approach with the Ham10000 Dataset. *Cancer Investig.* **2024**, *42*, 801–814.
13. Abhiram, A.; Anzar, S.; Panthakkan, A. DeepSkinNet: A Deep Learning Model for Skin Cancer Detection. In Proceedings of the 5th International Conference on Signal Processing and Information Security (ICSPIS), Dubai, United Arab Emirates, 7–8 December 2022; pp. 97–102.
14. Yan, S.; Yu, Z.; Primiero, C.; et al. A Multimodal Vision Foundation Model for Clinical Dermatology. *Nat. Med.* **2025**, *31*, 2691–2702.
15. Wang, Y.; Yu, T.; Cai, J.; et al. Integrating Clinical Knowledge Graphs and Gradient-Based Neural Systems for Enhanced Melanoma Diagnosis via the Seven-Point Checklist. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *37*, 37–51.
16. Zhou, J.; He, X.; Sun, L.; et al. Pre-Trained Multimodal Large Language Model Enhances Dermatological Diagnosis Using SkinGPT-4. *Nat. Commun.* **2024**, *15*, 5649.
17. Wang, Y.; Zheng, Y.; Yue, C.; et al. GloW-VSNet: A Scribble-Based Weakly Supervised Framework for Global-View Vitiligo Lesion Segmentation. *Med. Image Anal.* **2025**, *109*, 103920.
18. Wang, H.; Ahn, E.; Bi, L.; et al. Self-Supervised Multi-Modality Learning for Multi-Label Skin Lesion Classification. *Comput. Methods Programs Biomed.* **2025**, *265*, 108729.
19. Zhao, R.; Chen, X.; Chen, Z.; et al. Diagnosing Glaucoma on Imbalanced Data with Self-Ensemble Dual-Curriculum Learning. *Med. Image Anal.* **2022**, *75*, 102295.
20. Akram, T.; Khan, M.A.; Sharif, M.; et al. Skin Lesion Segmentation and Recognition Using Multichannel Saliency Estimation and M-SVM on Selected Serially Fused Features. *J. Ambient. Intell. Humaniz. Comput.* **2024**, *15*, 1083–1102.
21. Rahman, M.M.; Al Mahim, H.; Jeba, J.I.; et al. Deep Learning for Skin Cancer Detection: Multi-Class Lesion Classification Using CNN Architecture. In Proceedings of the 2025 2nd International Conference on Next-Generation Computing, IoT and Machine Learning (NCIM), Gazipur, Bangladesh, 27–28 June 2025.
22. Rana, M.; Bhushan, M. Machine Learning and Deep Learning Approach for Medical Image Analysis: Diagnosis to Detection. *Multimed. Tools Appl.* **2023**, *82*, 26731–26769.
23. Li, J.; Shi, H.; Chen, W.; et al. Semi-Supervised Detection Model Based on Adaptive Ensemble Learning for Medical Images. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *36*, 237–248.
24. Buddenkotte, T.; Sanchez, L.E.; Crispin-Ortuzar, M.; et al. Calibrating Ensembles for Scalable Uncertainty Quantification in Deep Learning-Based Medical Image Segmentation. *Comput. Biol. Med.* **2023**, *163*, 107096.
25. Chen, Y.; Li, D.; Zhang, X.; et al. Computer Aided Diagnosis of Thyroid Nodules Based on the Devised Small-Datasets Multi-View Ensemble Learning. *Med. Image Anal.* **2021**, *67*, 101819.
26. Zanddizari, H.; Nguyen, N.; Zeinali, B.; et al. A New Preprocessing Approach to Improve the Performance of CNN-Based Skin Lesion Classification. *Med. Biol. Eng. Comput.* **2021**, *59*, 1123–1131.
27. Bogacki, P.; Dziech, A. Effective Deep Learning Approach to Denoise Optical Coherence Tomography Images Using BM3D-Based Preprocessing of the Training Data Including Both Healthy and Pathological Cases. *IEEE Access* **2023**, *11*, 65395–65406.
28. Caseneuve, G.; Valova, I.; LeBlanc, N.; et al. Chest X-Ray Image Preprocessing for Disease Classification. *Procedia Comput. Sci.* **2021**, *192*, 658–665.
29. Zhang, W.; Lu, F.; Zhao, W.; et al. ACCPG-Net: A Skin Lesion Segmentation Network with Adaptive Channel-Context-Aware Pyramid Attention and Global Feature Fusion. *Comput. Biol. Med.* **2023**, *154*, 106580.

30. Zhang, Y.; Xie, F.; Chen, J. Tformer: A Throughout Fusion Transformer for Multi-Modal Skin Lesion Diagnosis. *Comput. Biol. Med.* **2023**, *157*, 106712.
31. Yang, Y.; Xie, F.; Zhang, H.; et al. Skin Lesion Classification Based on Two-Modal Images Using a Multi-Scale Fully-Shared Fusion Network. *Comput. Methods Programs Biomed.* **2023**, *229*, 107315.
32. Shao, D.; Ren, L.; Ma, L. MSF-Net: A Lightweight Multi-Scale Feature Fusion Network for Skin Lesion Segmentation. *Biomedicines* **2023**, *11*, 1733.
33. Ding, Y.; Ma, Z.; Wen, S.; et al. AP-CNN: Weakly Supervised Attention Pyramid Convolutional Neural Network for Fine-Grained Visual Classification. *IEEE Trans. Image Process.* **2021**, *30*, 2826–2836.
34. Novanti, A.I.; Harjoko, A. Modification of C-Grabcut for Segmentation and Classification of Coffee Leaf Diseases in Complex Backgrounds. *Int. J. Adv. Comput. Sci. Appl.* **2025**, *16*.
35. He, Y.; Wang, A.; Li, S.; et al. Nonfinite-Modality Data Augmentation for Brain Image Registration. *Comput. Biol. Med.* **2022**, *147*, 105780.
36. Maksimovic, V.; Jaksic, B.; Milosevic, M.; et al. Comparative Analysis of Edge Detection Operators Using a Threshold Estimation Approach on Medical Noisy Images with Different Complexities. *Sensors* **2024**, *25*, 87.
37. Liu, S.; Deng, W. Very Deep Convolutional Neural Network Based Image Classification Using Small Training Sample Size. In Proceedings of the 3rd IAPR Asian Conference on Pattern Recognition (ACPR), Kuala Lumpur, Malaysia, 3–6 November 2015; pp. 730–734.
38. He, K.; Zhang, X.; Ren, S.; et al. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
39. Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A ConvNet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 11976–11986.
40. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1800–1807.
41. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer Normalization. *arXiv* **2016**, arXiv:1607.06450.
42. Hendrycks, D.; Gimpel, K. Gaussian Error Linear Units (Gelus). *arXiv* **2016**, arXiv:1606.08415.
43. Lin, T.Y.; Dollár, P.; Girshick, R.; et al. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.
44. Britt, N.; Sun, H.j. Spatial Attention in Three-Dimensional Space: A Meta-Analysis for the Near Advantage in Target Detection and Localization. *Neurosci. Biobehav. Rev.* **2024**, *165*, 105869.
45. Zhuang, C.; Yuan, X.; Gu, L.; et al. Frequency Regulated Channel-Spatial Attention Module for Improved Image Classification. *Expert Syst. Appl.* **2025**, *260*, 125463.
46. Tschandl, P.; Rosendahl, C.; Kittler, H. The HAM10000 Dataset, a Large Collection of Multi-Source Dermatoscopic Images of Common Pigmented Skin Lesions. *Sci. Data* **2018**, *5*, 180161.
47. Li, H.; Wang, M. Very Deep Convolutional Network for Large-Scale Image Recognition. *Jisuanji Xitong Yingyong Comput. Syst. Appl.* **2021**, *30*, 330–335.
48. Zhang, X.; Zhou, X.; Lin, M.; et al. Shufflenet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6848–6856.
49. Szegedy, C.; Liu, W.; Jia, Y.; et al. Going Deeper with Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
50. Huang, G.; Liu, Z.; Van Der Maaten, L.; et al. Densely Connected Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
51. Liu, Z.; Lin, Y.; Cao, Y.; et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 9992–10002.
52. Ali, K.; Shaikh, Z.A.; Khan, A.A.; et al. Multiclass Skin Cancer Classification Using EfficientNets—A First Step towards Preventing Skin Cancer. *Neurosci. Inform.* **2022**, *2*, 100034.
53. Alhudhaif, A.; Almaslukh, B.; Aseeri, A.O.; et al. A Novel Nonlinear Automated Multi-Class Skin Lesion Detection System Using Soft-Attention Based Convolutional Neural Networks. *Chaos Solitons Fractals* **2023**, *170*, 113409.
54. Alwakid, G.; Gouda, W.; Humayun, M.; et al. Melanoma Detection Using Deep Learning-Based Classifications. *Healthcare* **2022**, *10*, 2481.
55. Abd Elaziz, M.; Dahou, A.; Mabrouk, A.; et al. An Efficient Artificial Rabbits Optimization Based on Mutation Strategy For Skin Cancer Prediction. *Comput. Biol. Med.* **2023**, *163*, 107154–107154.
56. Qian, S.; Ren, K.; Zhang, W.; et al. Skin Lesion Classification Using CNNs with Grouping of Multi-Scale Attention and Class-Specific Loss Weighting. *Comput. Methods Programs Biomed.* **2022**, *226*, 107166–107166.