applied sciences

MDPI

Article

Enhancing Skin Lesion Classification Performance with the ABC Ensemble Model

Jae-Young Choi [1], Min-Ji Song [2] and You-Jin Shin [2,*]

[1] Department of Mathematics, The Catholic University of Korea, Bucheon 14662, Republic of Korea; sju7428@catholic.ac.kr

[2] Department of Data Science, The Catholic University of Korea, Bucheon 14662, Republic of Korea; mingee0701@gmail.com

* Correspondence: yj.shinn@catholic.ac.kr

Abstract: Skin cancer is one of the most easily developed cancers and is continuously seeing an increased incidence rate. In this study, we propose a novel ABC ensemble model for skin lesion classification by leveraging the ABCD rule, which is commonly used in dermatology to evaluate lesion features such as asymmetry, border, color, and diameter. Our model consists of five distinct blocks, two of which focus on learning general image characteristics, while the remaining three focus on specialized features related to the ABCD rule. The final classification results are achieved through a weighted soft voting approach, where the generalization blocks are assigned higher weights to optimize performance. Through 15 experiments using various model configurations, we show that the weighted ABC ensemble model outperforms the baseline models, achieving the best performance with an accuracy of 0.9326 and an F1-score of 0.9302. Additionally, Grad-CAM analysis is employed to assess how each block in the ensemble focuses on distinct lesion features, further enhancing the interpretability and reliability of the model. Our findings demonstrate that integrating general image features with specific lesion characteristics improves classification performance, and that adjusting the soft voting weights yields optimal results. This novel model offers a reliable tool for early skin lesion diagnosis.

Keywords: ensemble; ABCD rule; vision transformer; data balancing

check for updates

Citation: Choi, J.-Y.; Song, M.-J.; Shin, Y.-J. Enhancing Skin Lesion Classification Performance with the ABC Ensemble Model. *Appl. Sci.* **2024**, *14*, 10294. https://doi.org/10.3390/app142210294

Academic Editor: Pedro Couto

Received: 29 September 2024
Revised: 24 October 2024
Accepted: 5 November 2024
Published: 8 November 2024

1. Introduction

Cancer is one of the leading causes of disease-related deaths in humans. It is a type of disease characterized by the irregular cell growth cycle of a group of cells. Normal cells replicate themselves in a regular cycle, grow through the process of division, and regulate cell growth through cell death. Unlike normal cells, malignant tumors grow very fast and invade surrounding tissues, and are called cancer [1]. Cancer can occur in any part of the body and can spread to other areas. When it affects the outermost layer of the skin, it is referred to as skin cancer. Skin cancer is primarily caused by chronic UV damage to skin cells from exposure to UV rays [2]. It is one of the few cancers with a steadily increasing incidence rate. If detected early, skin cancer can be successfully treated in nearly 95% of cases [3]. However, treatment becomes more challenging in later stages. Early detection and regular self-examination are vital, yet distinguishing malignant tumors from benign ones is challenging, as they often look similar to the naked eye.

Skin lesions are diagnosed by medical domain experts, such as dermatologists, but this relies on the individual's subjective expertise and judgment. It was found that the disagreement among individual scholars regarding the classification of skin cancer reached 25% [4]. To reduce subjectivity and provide more robust diagnoses, AI-based models, including machine learning and deep learning algorithms, are being utilized to support specialists in strengthening their decision-making. Most existing AI-based skin lesion diagnosis methods are based on spatial information. In particular, Convolution Neural Networks (CNNs) have achieved notable success in medical analysis. Various CNN-based models have been developed, and ensemble models that combine the strengths of different architectures have also been introduced. In 2023, an ensemble model combining DenseNet121 and MobileNet achieved promising results [5]. While several ensemble-based skin lesion classification models have been proposed, most models rely heavily on CNN architecture and their convolutional operations. Furthermore, ensemble research utilizing the Transformer algorithm, which has recently appeared and is achieving excellent performance in various fields, has been limited. Therefore, this study incorporates one of the Vision Transformer (ViT) series, the ViT-B/16 model, leveraging the attention mechanism of the Transformer in the ensemble model. However, since the ViT-B/16 model is known for its high performance on large datasets, it did not outperform CNN models on the HAM10000 dataset, which is relatively small, limiting its effectiveness in building an ensemble model [6]. To address this issue, we enhanced the performance by applying data augmentation, allowing ViT-B/16 to learn the patterns better.

Next, we apply the ABCD rule, a clinical criterion for assessing skin cancer, to our model. The ABCD rule is a commonly used tool in dermatology to help identify suspicious moles or skin lesions that could indicate melanoma, a serious form of skin cancer. This method assesses the following features: Asymmetry (A) assesses whether the two halves of a lesion are symmetrical. Border (B) checks if the edges of the lesion are irregular or well-defined. Color (C) examines whether the lesion displays uneven color patterns or a mix of multiple shades. Diameter (D) helps to determine if the lesion is notably large, which may indicate malignancy. Previous studies often employed methodologies that excel in general image classification but did not account for the medical characteristics of skin lesions. In this research, we integrate various data preprocessing methods that align with the medical ABCD rule, ensuring that these important dermatological factors are reflected in the model. Note that, due to the nature of our data, the actual diameter of the lesion is unknown, so D (diameter) is not considered.

In summary, we propose a novel ABC ensemble model for skin lesion classification. This model incorporates various preprocessing techniques that reflect the ABCD rule, along with ensemble models of ViT-B/16, DenseNet121, EfficientNet-B0, MobileNetV2, and ResNet50. Our model is trained and tested on the HAM10000 dataset [7], a widely used benchmark, and aims to classify a total of seven types of skin lesions. The novelty of our research is as follows:

- We propose a novel ABC ensemble model for skin lesion classification by leveraging the ABCD rule, which reflects the clinical criteria used to assess lesions. The model consists of five distinct blocks, including two generalization blocks that learn overall image characteristics and three ABC-related blocks that focus on the asymmetry, border, and color features of skin lesions. Each block undergoes different preprocessing methods before passing through five classifiers: ViT-B/16, DenseNet121, EfficientNet-B0, MobileNetV2, and ResNet50.
- To demonstrate the effectiveness of our model, we conduct experiments using 15 different model configurations. These configurations include models without preprocessing, models with various preprocessing methods, and ensemble models with and without weights. The results consistently show that the ABC ensemble model outperforms the other configurations in terms of accuracy, recall, precision, and F1-score, highlighting its superior performance in skin lesion classification.
- Our model adopts a weighted soft voting approach, where higher weights are assigned to the generalization blocks that analyze overall image features. By doubling the weight assigned to the generalization blocks, we achieve improved classification performance. We experiment with various weight combinations to determine the optimal setup, showing that this weighted voting method enhances the overall accuracy and reliability of the model.
- To further validate the effectiveness of our model, we perform Grad-CAM analysis, which allows us to visualize the regions of the skin lesion that each block focuses on

during classification. The different preprocessing methods in our model highlight both general features and specific lesion characteristics such as asymmetry, border, and color. This analysis shows the interpretability and reliability of the model, ensuring that it is not only effective but also transparent in its decision-making process.

The structure of this paper is as follows: Section 2 presents an overview of research on skin lesion classification using machine learning or deep learning approaches. Section 3 introduces the dataset, overall model architecture, data preprocessing, model detailed structure, ensemble classifier, and experiments. Section 4 reports on the classification performance and analyzes the ABC ensemble model across various experiments. Section 5 covers the limitations and retention future directions of our research. Finally, Section 6 summarizes our conclusions.

2. Related Work

Computational methodologies have been used for a long time to assist in the diagnosis of medical conditions. In the 1980s, computer-aided diagnostic (CAD) systems were developed to help dermatologists overcome challenges in diagnosing skin diseases, including skin cancer [8]. After 2000, machine learning-based approaches began to emerge. Techniques such as Support Vector Machines (SVM), which aim to find the optimal boundary between classes for data classification, k-nearest neighbor (k-NN), a non-parametric method that classifies data based on the closest neighbors in the feature space, and logistic regression, a statistical method used for binary classification by estimating the probability of a given input belonging to a particular class, were explored for diagnostic support. However, these methods did not achieve satisfactory detection performance due to high intra-class variation (where the lesions of the same classes can appear very different) and low inter-class variation (where the lesions of different classes can appear similar) [9–11].

Recently, various automation methods have been developed to classify skin cancer using skin images. Most of the previous studies have focused on developing various artificial intelligence algorithms to automate the diagnosis of different types of skin cancer, and studies on the detection, segmentation, and classification of skin lesions associated with skin cancer have been conducted using the CNN-based architecture. He et al. [12] presented a ResNet model that solved the problem of losing gradients as the depth of the existing CNN model became deeper using a residual block. Alwakid et al. [13] used CNN models and modified ResNet50, applying them to the HAM10000 dataset. They improved image quality using ESRGAN and addressed the class imbalance problem using data augmentation. They obtained results with 86% and 85.3% accuracy using CNN and ResNet50 models. Huang et al. [14] presented DenseNet in 2017, a deep learning model in which each layer receives inputs from all previous layers to optimize information flow and gradient delivery, and prevents overfitting by reusing characteristics with high parameter efficiency. Howard et al. [15] presented MobileNet in 2017, a lightweight network model for learning in resource-constrained situations. Using depthwise separable convolution operations instead of standard CNN convolutions helped to reduce the number of training parameters. In 2019, Tan et al. [16] presented an EfficientNet that can achieve higher performance with fewer operations than other models using a complex scaling method that simultaneously scales the width, depth, and resolution of the model. In 2022, Ali et al. [17] trained EfficientNet B0-B7 on the HAM10000 dataset by performing transfer learning on pre-trained ImageNet weights and fine-tuning the convolutional neural networks. The best model in this paper, EfficientNet B4, achieved a Top-1 accuracy of 87.91%. Also in 2022, Yin et al. [18] added the MD-Net module to the DenseNet model for the ISIC 2019 dataset, achieving accuracies of 83.5%, 84.1%, and 83.8% for DenseNet121, DenseNet169, and DenseNet201. In 2023, Wang et al. [19] performed a skin cancer classification task by extending the dataset and DMN using CycleGaN. Furthermore, to alleviate the gradient vanishing problem, he proposed a DMN to add connections and branches to the MobileNetV2 classification network, achieving an accuracy of 87.07%, which is higher than the 82.58% accuracy of the basic MobileNetV2 model.

Deep learning algorithms that are not CNN-based are also used for image processing. In 2017, Vaswani et al. [20] introduced the Transformer algorithm. This work revolutionized natural language processing and machine learning by introducing a novel architecture based entirely on self-attention mechanisms, eliminating the need for recurrent neural networks (RNNs). In 2020, Dosovitskiy et al. [21] extended this architecture to image processing with the development of Vision Transformer (ViT). By applying the Transformer, originally designed for natural language tasks, to image classification, ViT achieved competitive performance with convolutional neural networks (CNNs), especially when trained on large datasets. In 2024, Song et al. [22] evaluated the performance of the Vision Transformer (ViT) and Swin Transformer in classifying mobile-based oral cancer images. The results indicated that the Swin Transformer achieved higher accuracy compared to the ViT model, and both Transformer models outperformed the conventional convolutional model, VGG19.

In addition, studies using existing CNN models [23–25] have continued for the skin lesion classification task, and the ensemble method has been used as one of the approaches. The ensemble method gained higher accuracy by utilizing the diversity of individual models. Ensemble models initially gained prominence in other application scenarios before being applied to medical imaging tasks. Rahman et al. [26] proposed a supervised learning framework to assess the data quality in marine sensor networks, addressing the class imbalance issue using an ensemble classifier system. Experimental results showed that the ensemble classifier balanced the classification accuracy between majority and minority classes, achieving 16.24% and 15.29% higher accuracy compared to individual classifiers. Haralabopoulos et al. [27] explored using ensemble learning to improve multilabel binary classification in sentiment analysis. The study compared a baseline stacked ensemble with a proposed weighted ensemble and tested them on two datasets (Semeval2018-Task 1, Toxic Comment). The weighted ensemble outperformed the stacked ensemble in 75% of cases, with performance improvements ranging from 1.5% to 5.4%. Zhao et al. [28] proposed a method to improve textile fabric defect classification accuracy using artificial intelligence. They applied an ensemble learning-based CNN model (ECTFDC) on an enhanced TILDA database to classify fabric defects. Experimental results showed that ECTFDC achieved a precision of 97.8% and a recall of 97.68%, outperforming other models. Vennelakanti et al. [29] presented a method for Traffic Sign Detection and Recognition (TSDR), emphasizing the use of an ensemble of Convolutional Neural Networks (CNN) for sign recognition. The ensemble approach, implemented using TensorFlow, achieved over 99% recognition accuracy on circular signs from the Belgium and German datasets. This use of CNN ensembles significantly enhanced Advanced Driver Assistance Systems (ADAS) for safer driving environments.

In particular, the ensemble method of the CNN model was also used in the medical classification task because it improved overall performance while reducing overfitting and increasing the robustness of predictions [30]. De et al. [31] proposed an ensemble CNN model for the early diagnosis of COVID-19, incorporating DenseNet169, VGG16, and Xception. The ensemble model demonstrated the best performance, achieving an accuracy of 97.7%, a precision of 97.7%, a recall of 97.8%, and an F1-score of 97.7%. The ensemble model outperformed individual CNN models.

Gajera et al. [5] normalized the skin lesion images to select lightweight CNN models and two pre-trained CNN models such as DenseNet121 and MobileNet, resulting in 91.12% accuracy. Charan et al. [32] designed a deep ensemble learning model for skin cancer diagnosis. They utilized the EfficientNet, SENet154, and ResNet models, preprocessing each with a grayscale, segmented mask, and a multiplication image of the original image plus the segmented image, respectively, and then ensembled them to achieve 92.6% accuracy. Mahbod et al. [33] proposed a multi-scale, multi-CNN (MSM-CNN) ensemble approach for skin lesion classification. By combining three CNN architectures (EfficientNet-B0, EfficientNet-B1, and SeReNeXt-50) trained on dermoscopic images at various scales, the ensemble achieved a balanced multi-class accuracy of 86.2% on the ISIC 2018 test set. Lin et al. [34] experimented with a CNN-based ensemble model for automatic skin disease classification using dermoscopic images. Pre-trained models such as EfficientNet, DenseNet121, and ResNet50 were utilized for feature extraction. Meta classifiers were applied for final classification, improving performance. The ensemble method achieved 91% accuracy on the validation set, outperforming individual CNN models like DenseNet121 (87.3%). The study highlighted the benefits of ensemble learning, especially when individual CNNs were undertrained or weak. Image preprocessing, such as white balance, also contributed to the improved results.

3. Materials and Methods

3.1. Dataset

Our research utilizes the HAM10000 (Human Against Machine with 10,000 training images) dataset [7] to train and test the model. This dataset contains 10,015 dermatoscopic images of seven different types of skin lesions, making it ideal for evaluating the performance of our classification model. The seven types of skin lesions are represented in Figure 1 and described below:

- Akiec: Actinic Keratoses and Intraepithelial Carcinoma can be considered precursor cancer types of non-melanoma skin cancer, rather than actual cancer species [35].
- Bcc: Basal cell carcinoma is the most common form of skin cancer, which grows slowly and rarely spreads [36].
- Bkl: Benign keratosis is a variant of seborrheic keratosis, and it is difficult to classify because there are many biologically similar lesions and many lesions with similar morphological characteristics [37].
- Df: Dermatofibroma is a benign skin lesion considered an inflammatory response [38].
- Mel: Melanoma is a malignant tumor derived from melanocytes and can be cured with simple resection if found early [39].
- Nv: Melanocytic nevi exhibit many changes in the benign forms of melanocytes [40].
- Vasc: Vascular skin lesions include cherry hemangiomas, angiokeratomas, and pyogenic granulomas [41].

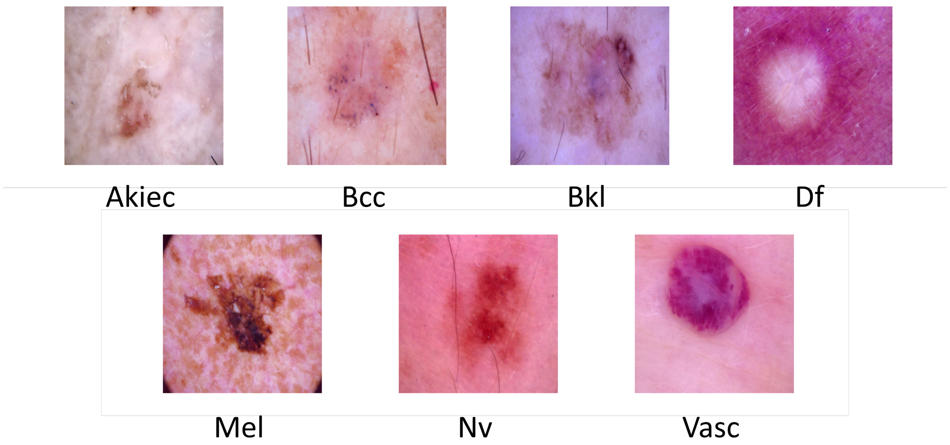
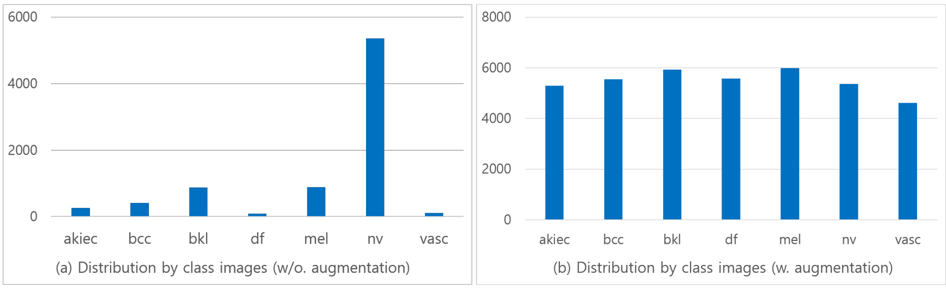


Figure 1. Images representing the seven types of skin lesions from the HAM10000 dataset: Actinic Keratoses (akiec), Basal Cell Carcinoma (bcc), Benign Keratosis (bkl), Dermatofibroma (df), Melanoma (mel), Melanocytic Nevi (nv), and Vascular Lesions (vasc).

The number and percentage of images in each class are shown in Table 1. Approximately 66% of the images belong to the 'nv' class. Notably, the 'vasc' and 'df' categories are highly imbalanced due to a relative lack of images. To address the imbalance among class images, the dataset sizes for classes with fewer images were increased through oversampling, making their numbers similar to the 'nv' class, which has the most images, as shown in Figure 2. This ensured an even balance of images across all classes in the dataset. Basic augmentation methods, including rotation, zoom, horizontal flip, and vertical flip, were applied for oversampling.

Table 1. The number and ratio of images in the 'akiec', 'bcc', 'bkl', 'df', 'mel', 'nv', and 'vasc' classes in the HAM10000 dataset.

Diagnostic Category	Number of Images	Percentage
akiec	327	3.27%
bcc	514	5.13%
bkl	1099	10.97%
df	115	1.15%
mel	113	11.11%
nv	6705	66.95%
vasc	142	1.42%
Total	**10,015**	**100%**



(a) Distribution by class images (w/o. augmentation)

(b) Distribution by class images (w. augmentation)

Figure 2. The difference in image ratios between classes before and after data augmentation. The left figure (**a**) shows the image ratios for each class in the original HAM10000 dataset before oversampling, while the right figure (**b**) shows the image ratios after augmentation.

3.2. Overall Model Architecture

In this section, we present the overall network architecture of our ABC ensemble model for classifying cancer using the HAM10000 dataset. The ABCD rule [42] is a tool used in dermatology to identify skin lesions that may be potential malignancies. Asymmetry (A) evaluates whether the two halves of a lesion are symmetrical. Border (B) checks for irregular or uneven edges around the skin lesion, while Color (C) assesses whether the pigmentation is uneven, with a mix of different shades. Diameter (D) helps to determine if a lesion is larger than 6 mm, which may indicate malignancy [43]. However, diameter (D) is only meaningful when all of the images are captured at a consistent height and angle. In the HAM10000 dataset, as shown in Figure 3, some images are highly magnified, while others are taken from farther away, making it difficult to effectively utilize diameter information. As a result, diameter (D) is excluded from our study. Therefore, our ensemble model focuses on the ABC components rather than the ABCD, and is referred to as the ABC ensemble.

Figure 4 illustrates the overall architecture of our model. First, the ROI of the lesion is extracted from the original image using U-Net by generating a segmentation mask. The details are described in Section 3.3. Next, five different classifiers are trained: two for general classification tasks (Block 1: Generality Training and Block 2: Upsize Training), and three classifiers specifically tailored to the ABCD rule (Block 3: Asymmetry Enhancement Training, Block 4: Border Enhancement Training, and Block 5: Color Enhancement Training). Finally, a weighted ensemble of the five trained classifiers is performed to classify the skin lesions. The details on each of the five classifiers are provided in Section 3.4.

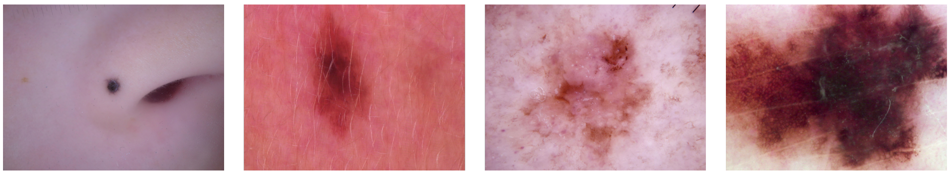


Figure 3. Four images of skin lesions captured from different angles and distances. Due to the variations in measurement angles and distances, the actual diameter of the lesions cannot be determined.

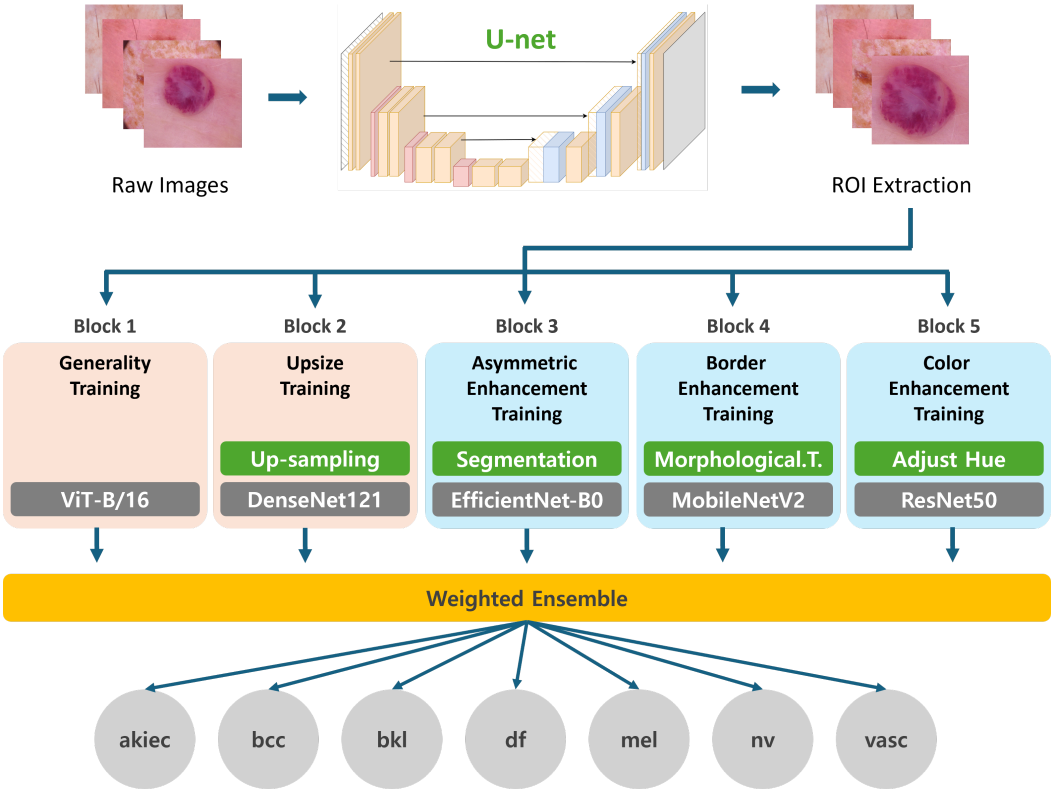


Figure 4. The overall architecture of the ABC ensemble model, consisting of five blocks: Generality Training, Upsize Training, Asymmetry Enhancement Training, Border Enhancement Training, and Color Enhancement Training. The two orange blocks learn the general features of the image, and the three blue blocks learn the features reflecting the ABCD rule.

3.3. Data Preprocessing

The raw data are subjected to basic preprocessing steps. The raw images in the HAM10000 dataset capture not only the lesion but also the surrounding skin, which often includes noise such as wounds or age spots. To avoid the effect of these noises, we apply a U-net model [44] to extract the Region of Interest (ROI) from the original image, isolating the lesion from the surrounding area.

The U-net model applies two convolution layers with a 3 × 3 size kernel, and applies max pooling with stride of 2 to halve the image size and double the number of channels to extract the feature map by repeating downsampling. After that, the convolution layer is applied to increase the robustness of the model, and the feature map is upsampled to the resolution of the original image through a deconvolution layer with a stride of 2. This process allows U-net to reconstruct the image. After training the U-net, we can obtain the segmentation mask of the augmented dataset using the trained U-net model and generate a segmentation mask of 0 s and 1 s as shown in Equation (1) [45]. Using the obtained segmentation mask, we extract the ROI of the lesion, including its location and size, as defined in Equations (2) and (3). Then, the images are upsized by focusing only on the ROI to match the input dimensions required by each classifier, as shown in

Figure 5. Among the five blocks of our model, only the Upsize Training block produces preprocessed images of 384 × 384, while the remaining four blocks produce preprocessed images of 256 × 256.

$$B(x,y) = \begin{cases} 1 & \text{if } I(x,y) > 0.05 \\ 0 & \text{if } I(x,y) \leq 0.05 \end{cases}, \tag{1}$$

where $B(x,y)$ is the pixel value of the binarized image, representing the result of converting the original image's pixel values into black and white based on a threshold, and $I(x,y)$ is the pixel value at position (x,y) in the original image.

$$x_{min} = min(x \mid B(x,y) \in 1),\ y_{min} = min(y \mid B(x,y) \in 1), \tag{2}$$

where x_{min}, y_{min} are the coordinates of the pixels in B with a value of 1. x_{min}, y_{min} represent the top-left corner of the ROI.

$$x_{max} = max(x \mid B(x,y) \in 1),\ y_{max} = max(y \mid B(x,y) \in 1), \tag{3}$$

where x_{max}, y_{max} are the coordinates of the pixels in B with a value of 1. x_{max}, y_{max} represent the bottom-right corner of the ROI.

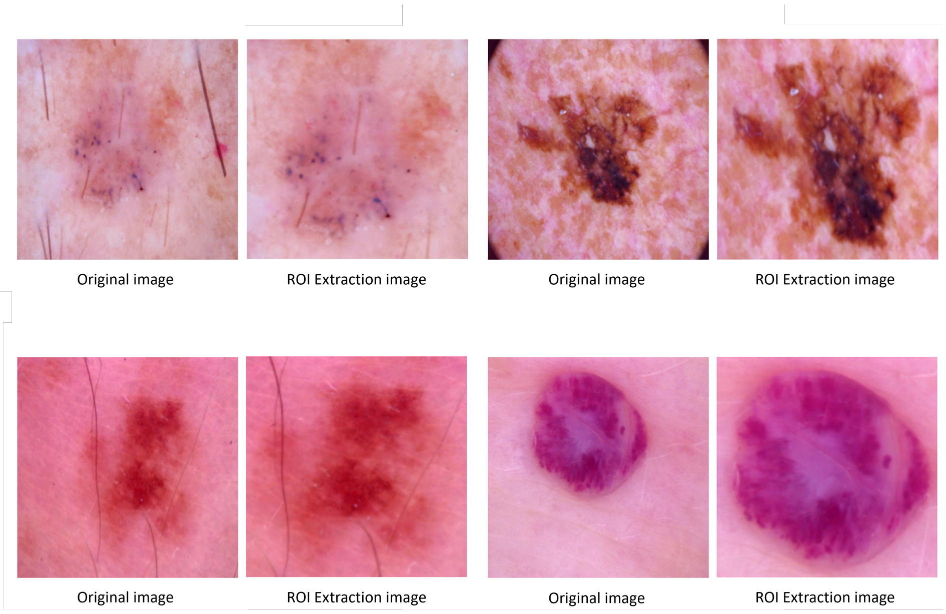


Figure 5. Four different images applying ROI extraction. The image on the left is the original image, and the image on the right is a sample image with ROI extraction applied.

3.4. Five Blocks of ABC Ensemble Model

This section explains the five blocks, designed to enhance learning by reflecting the characteristics of the skin lesions. After preprocessing to extract the ROI using the U-net model, the data are passed to the five modules. As mentioned in the Section 3.2, our model consists of a total of five blocks: two general blocks and three ABC-related blocks. These five modules consist of Generality Training (Block 1), Upsize Training (Block 2), Asymmetry Enhancement Training (Block 3), Border Enhancement Training (Block 4), and Color Enhancement Training (Block 5).

Generality Training (Block 1): The Generality Training block represents the most basic model, which does not consider the ABCD rule. Since no specific techniques related to the ABCD rule are applied, this model learns the overall characteristics of the data in terms of general image features. The detailed structure of the model is described in Figure 6. The classifier for the Generality Training block is Vision Transformer-B/16 (ViT-B/16) [21]. The ViT model vectorizes an image of size $C(Channel) \times H(Height) \times W(Width)$ by cutting it into N patches of size (P, P) and flattening them into N $(1 - D)$ vectors of dimension $P^2 \times C$. It is a model that extracts features with a transformer encoder consisting of multi-head self-attention and proceeds with classification through a fully connected layer. We also tune the parameters of the Transformer encoder to achieve high performance on the relatively small-sized HAM10000 dataset.

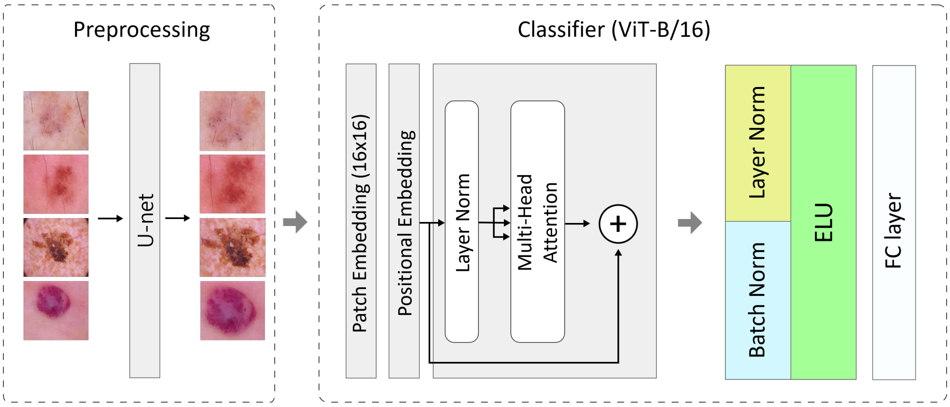


Figure 6. Structure of the Generality Training model (Block 1), including the Vision Transformer (ViT-B/16).

Upsize Training (Block 2): In Upsize Training, unlike the others, a 384 × 384 image is generated through preprocessing and fed into a classifier. The detailed structure of the model is described in Figure 7. By utilizing a larger image size than other inputs, the model can capture finer details and more intricate patterns in the skin lesions, improving feature extraction and learning richer spatial relationships. This allows the model to potentially enhance its performance in detecting subtle irregularities, which may be missed with smaller images. The classifier of the Upsize Training adopts the DenseNet121 model. The DenseNet121 model is a deep learning model developed by connecting feature maps of all layers. In the case of the existing ResNet model, to solve the loss of information between layers, a feature map was added, but in the case of DenseNet, the feature maps are concatenated and the output of the previous layer is used as the input again to extract the result through the Dense block [14].

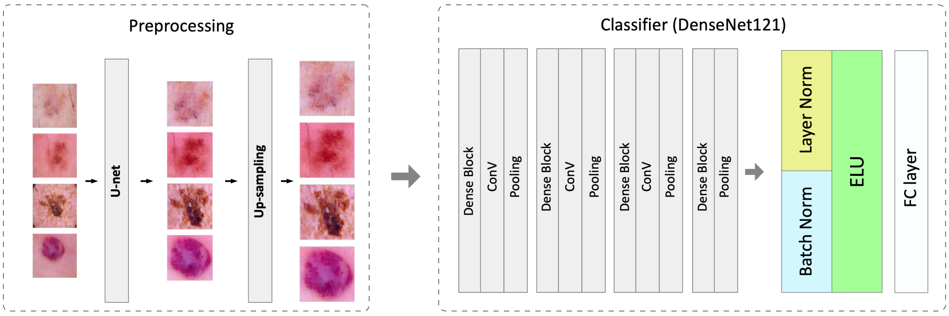


Figure 7. Structure of the Upsize Training model (Block 2), including the Upsampling and DensetNet121.

Asymmetry Enhancement Training (Block 3): In Asymmetry Enhancement Training, segmentation is performed to allow the model to focus on asymmetry in the lesion, highlighting the shape of the lesion. As shown in Figure 8, a segmented image generates a segmentation mask of an image using the U-net model for ROI extraction in the preprocessing step. In the generated segmentation mask, as expressed in Equation (4), the segmentation mask part is dot-produced for the original image, and only the part with the segmentation mask of 1 is derived and preprocessed into a segmented image. Through this process, the model can more accurately capture its symmetry, since we can identify the lesion's shape clearly, as shown in Figure 9.

$$I_{ROI}(x, y) = I(x, y) \cdot B(x, y), \quad (4)$$

where $I_{ROI}(x,y)$ is the pixel value of the region of interest in the original image. $I(x,y)$ is the pixel value of the original image at position (x,y). $B(x,y)$ is the pixel value of the binarized image at position (x,y), which acts as a mask. It is 1 if the pixel is within the ROI and 0 otherwise.

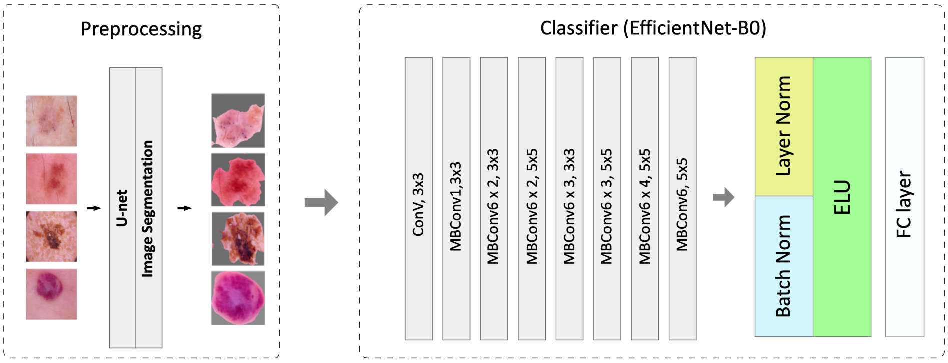
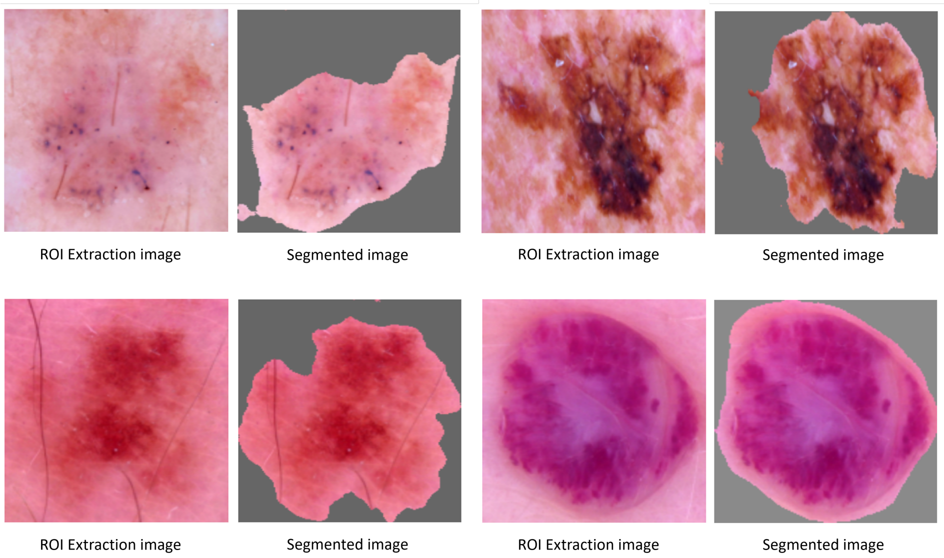


Figure 8. Structure of the Asymmetry Training model (Block 3), including the Segmentation and EfficientNet-B0.


ROI Extraction image

Segmented image

ROI Extraction image

Segmented image

ROI Extraction image

Segmented image

ROI Extraction image

Segmented image

Figure 9. Sample segmented images. The image on the left is the original image with ROI extraction applied to it, and the image on the right is the ROI extracted image with segmented mapping applied to it.

The classifier for Asymmetry Enhancement Training uses the EfficientNet-B0 model, pre-trained on ImageNet. EfficientNet-B0 is built by stacking MBConv blocks [16], which are optimized using the squeeze-and-excitation method to improve performance by recalculating the weight between channels [46]. Among the EfficientNet family, EfficientNet-B0 is the model with the smallest number of parameters, and it is defined by a width-coefficient of 1.0 and a depth-coefficient of 1.0.

Border Enhancement Training (Block 4): Border Enhancement Training involves blurring the overall color of an image so that the model does not rely on color features, allowing it to focus more on the edges instead. As depicted in Figure 10, Morphological Transformation applies erosion and dilation to the image data by pooling the image data through a kernel-sized filter. We apply a Morphological Transformation using a kernel size of 6, as shown in Equation (5). This process involves performing erosion on the original image using max pooling. Then, it was expanded again using Equation (6) to resemble the original image [47]. As the erosion and dilation progresses, the overall color of the image becomes unclear and blurred as shown in Figure 11, allowing the model to learn by focusing on the shape and border of the lesion part rather than the color feature.

$$E(I) = -max(-I, k), \tag{5}$$

where $E(I)$ represents the eroded image obtained by applying the erosion operation on the original image I. I is the original image that is being processed. k is the kernel used for the erosion operation. k defines the shape and size of the neighborhood over which the maximum is calculated.

$$D(E(I)) = max(E(I), k), \tag{6}$$

where $D(E(I))$ represents the dilated image, which is obtained by applying the dilation operation on the eroded image $E(I)$. k is the kernel used for the dilation operation.

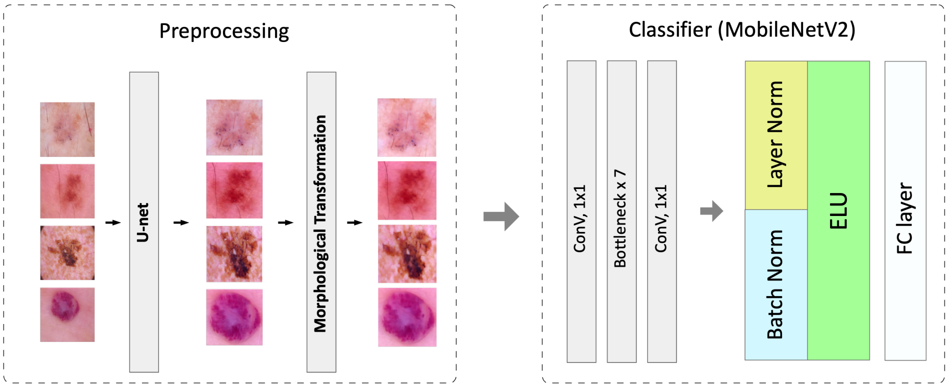


Figure 10. Structure of the Border Training model (Block 4), including the Morphological Transformation and MobileNetV2.

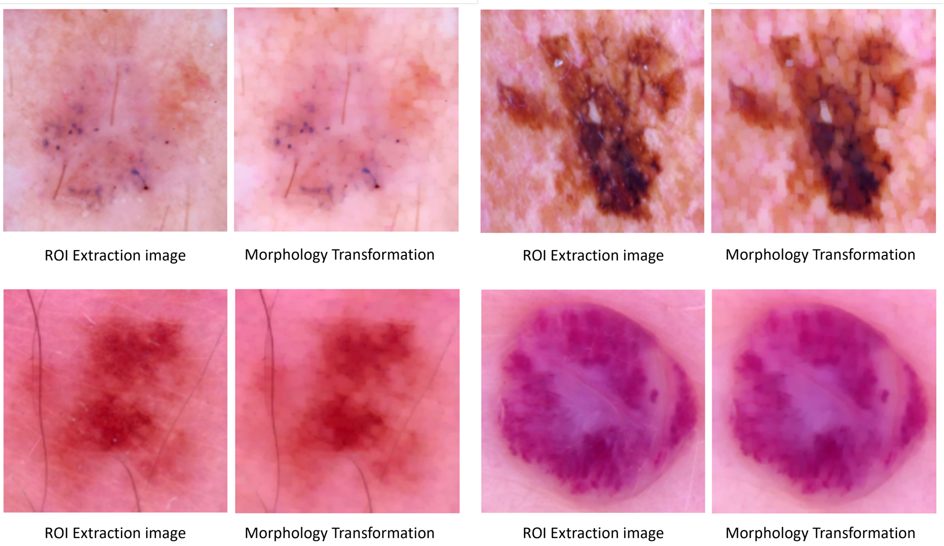

ROI Extraction image
Morphology Transformation
ROI Extraction image
Morphology Transformation
ROI Extraction image
Morphology Transformation
ROI Extraction image
Morphology Transformation

Figure 11. Four different sample images applied Morphological Transformation. The image on the left shows the original image with ROI extraction applied, while the image on the right displays the Morphological Transformation applied to the same image.

As the classifier of Border Enhancement Training, we use the MobileNetV2 model, pre-trained in Imagenet. The MobileNetV2 model is configured using an Inverted Residual block using bottleneck depth-separable convolution, which can achieve high performance even with low computational requirements [15].

Color Enhancement Training (Block 5): Color Enhancement Training involves adjusting the image's color to emphasize the lesion's color over its shape when extracting features from skin lesion images. As illustrated in Figure 12, the image data are preprocessed using a color adjustment technique via the Adjust Hue method, with the results shown in Figure 13. In this figure, the lesion's color stands out from the surrounding skin tone, even in datasets where color differences are minimal. This preprocessing helps the model to learn color features more effectively than with the original data [48]. As shown in Equation (7), Adjust Hue converts the existing RGB image data into the HSV color space and normalizes them to a range of [0, 1]. When converting to HSV, a value of [−0.5, 0.5] is added to the Hue value, adjusting the color tone, and the image is then converted back to RGB format. We set the hue adjustment value to $h_f = -0.1$. As expressed in Equation (8), the color tone adjustment was performed by adding h_f to the value H converted from RGB space to HSV space.

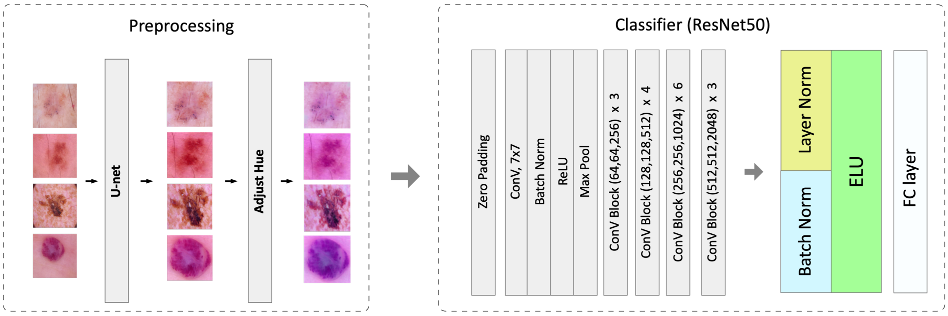


Figure 12. Structure of the Color Training model (Block 5), including the Adjust Hue and ResNet50.

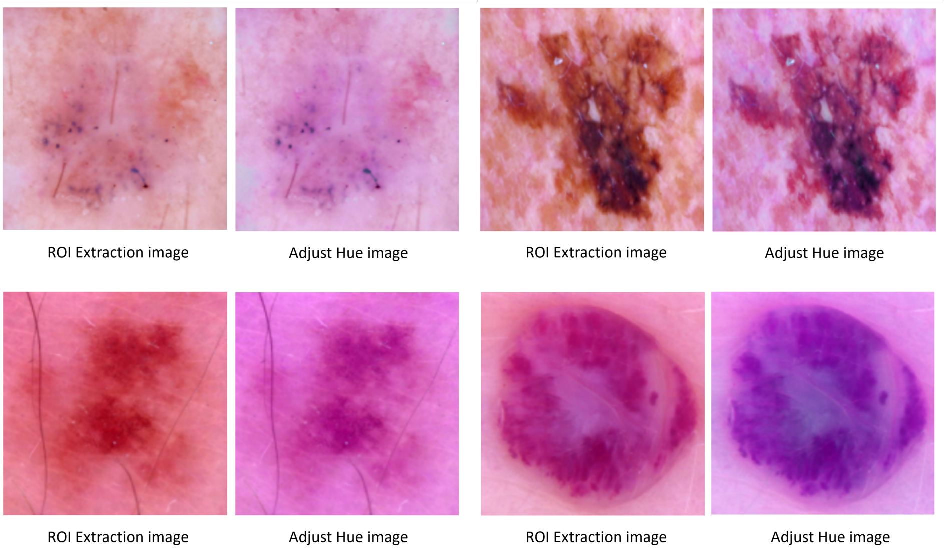


Figure 13. Four different sample applying Adjust Hue. The image on the left shows the original image with ROI extraction applied, while the image on the right displays the same image with Adjust Hue applied after ROI extraction.

$$H = \begin{cases} \frac{G-B}{max\_val-min\_val} \; mod\, 6 & if\ max\_val = R \\ \frac{B-R}{max\_val-min\_val} + 2 & if\ max\_val = G \\ \frac{R-G}{max\_val-min\_val} + 4 & if\ max\_val = B \end{cases}, \tag{7}$$

where H is the hue component of the image in the HSV color space. R, G, B are the red, green, and blue channel values of a pixel in the RGB color space. max_val is the maximum value among RGB values. min_val is the minimum value among RGB values.

$$H' = H + h_f, \tag{8}$$

where H' is the adjusted hue value after applying a hue shift. h_f is the hue adjustment factor, which shifts the hue value to change the color's appearance.

For Border Enhancement Training, we utilize the pre-trained ResNet50 model from ImageNet. ResNet50 consists of 50 layers, constructed by stacking residual blocks made up of a 1 × 1 convolution layer, a 3 × 3 convolution layer, and another 1 × 1 convolution layer. ResNet enhances efficiency and performance in feature extraction by incorporating a residual structure through shortcut connections [12].

3.5. Ensemble Classifier

In this section, we present an ensemble classifier to enhance the performance of the skin lesion classification. Classification is performed via the final fully connected (fc) layer. Our ABC ensemble model includes a total of five classifiers, and each model employs different normalization methods and activation functions. ViT-B/16 adopts GELU, EfficientNet-B0 uses Swish, MobileNetV2 applies ReLU6, and both ResNet50 and DenseNet121 utilize ReLU functions. The inconsistent use of different activation functions and normalization methods across models can degrade performance due to varying scales in each model's predictions. Therefore, instead of using a different final fc layer for each model, we use a classifier, applying the same activation function and normalization across all models. Batch normalization is applied to the batch input, using the mean and variance as shown in Equation (9), while layer normalization applies to the layer input as described in Equation (10).

$$\hat{x}_B = \frac{x_B - E(x_B)}{\sqrt{Var(x_B)}}, \tag{9}$$

where $\hat{x}_B$ is the normalized value of the batch x_B. x_B represents a batch of input values from a dataset. $E(x_B)$ is mean of the batch x_B. $Var(x_B)$ is the variance of the batch x_B.

$$\hat{x}_L = \frac{x_L - E(x_L)}{\sqrt{Var(x_L)}}, \tag{10}$$

where $\hat{x}_L$ is the normalized value of the layer's input x_L. x_L represents the input values to a specific layer in a neural network. $E(x_L)$ is the mean of the input values x_L. $Var(x_L)$ is the variance of the input values x_L.

The normalized features are then passed through the ELU activation function. The selection of ELU as the activation function is based on its ability to combine characteristics of GELU, Swish, and ReLU6, which all have non-linear and linear properties. ELU operates non-linearly for negative values and linearly for positive values, as expressed in Equation (11).

$$ELU(x) = \begin{cases} x & if\ x \geq 0 \\ \alpha(e^x - 1) & if\ x < 0 \end{cases} \tag{11}$$

$ELU(x)$ stands for the Exponential Linear Unit, which is a type of activation function used in neural networks. x is the input value to the activation function. α is a parameter that controls the saturation level for negative inputs, typically set to a constant value.

After deriving the predictions from each model through the ensemble classifier, we conducted the ensemble using a weighted soft voting method. The weights were assigned as follows: [2.0, 2.0, 1.0, 1.0, 1.0] in the order of ViT-B/16, DenseNet121, EfficientNet-B0, MobileNetV2, and ResNet50. Among the five blocks in our ABC ensemble model, a higher weight of 2.0 was applied to the general blocks, which captures the overall characteristics of the data. In contrast, the remaining three blocks, which consider the ABCD rule, were assigned a relatively lower weight of 1.0.

3.6. Experiments

This section summarizes the experiments that effectively demonstrate the performance of our model. It covers the configurations of the 15 experimental models, the training and testing setup, the category-wise classification performance, the Grad-CAM analysis, and the weight optimization.

Experimental Configurations: To demonstrate the superiority of our proposed model, we conducted 15 experiments using different configurations:

- GoogLeNet without preprocessing;
- AlexNet without preprocessing;
- ViT-B/16 without preprocessing;
- DenseNet121 without preprocessing;
- EfficientNet-B0 without preprocessing;
- MobileNetV2 without preprocessing;
- ResNet50 without preprocessing;
- ViT-B/16 with preprocessing (ROI extraction);
- DenseNet121 with preprocessing (ROI extraction, Upsampling);
- EfficientNet-B0 with preprocessing (ROI extraction, Segmentation);
- MobileNetV2 with preprocessing (ROI extraction, Morphological Transformation);
- ResNet50 with preprocessing (ROI extraction, Adjust Hue);
- Ensemble without ABCD;
- Ensemble with weights and without ABCD;
- Ensemble with weights and with ABCD (Ours).

These 15 experiments can be broadly divided into three categories. The first seven evaluate the performance of a single algorithm trained on raw data without preprocessing. These seven algorithms include the five used in our ensemble model (ViT-B/16 [21], DenseNet121 [14], EfficientNet-B0 [16], MobileNetV2 [15], and ResNet50 [12]), as well as two additional algorithms (GoogLeNet [49] AND AlexNet [50]). The next five experiments measure the performance of each of the five algorithms used in the ensemble model, with preprocessing applied (using different techniques depending on the algorithm). The final three experiments involve ensemble models, each differing in the application of weights and the ABCD rule. Detailed results for each configuration are provided in Section 4.1.

Training and Testing Setup: For model training and testing, we divided the HAM10000 dataset into 80% training and 20% testing sets, using 5-fold cross-validation to avoid biased results towards specific test sets. The 10,015 images were split into 8012 training images and 2003 testing images, ensuring no overlap in the test sets for each fold. We trained the model five times and averaged the accuracy, recall, precision, and F1-score across the five folds to measure performance consistently. Table 2 shows the distribution of data per fold during THE K-fold cross-validation of the HAM10000 dataset used in the experiments. For hyperparameters, we set the batch size at 64 and the image size at 256 × 256 pixels. The learning rate was adjusted to 0.0001, and the Adam optimizer was employed for training. We selected 70 epochs after confirming that the model had sufficiently converged. Given that our model has an ensemble structure with multiple internal models, we aimed to keep the number of epochs as low as possible to prevent excessive training time.

Category-wise Classification Performance: We visualized the confusion matrix to assess how effectively the ABC ensemble model performs for each class in skin lesion classification. The confusion matrix allows us to evaluate how closely the model's final predictions align with the actual classes and identify which class the model predicted incorrectly. This not only demonstrates the model's classification performance, but also allows us to see which classes the model has effectively learned and which classes it has struggled to learn adequately. The classification performance results for each of the seven skin lesion classes are presented in Section 4.2.

Table 2. Training and testing data distribution for each fold during the K-fold cross-validation process across lesion class.

Class	Fold 1		Fold 2		Fold 3		Fold 4		Fold 5	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
akiec	262	65	262	65	261	66	261	66	262	65
bcc	411	103	411	103	411	103	411	103	412	102
bkl	879	220	879	220	879	220	880	219	879	220
df	92	23	92	23	92	23	92	23	92	23
mel	890	223	890	223	891	222	891	222	890	223
nv	5364	1341	5364	1341	5364	1341	5364	1341	5364	1341
vasc	114	28	114	28	114	28	113	29	113	29
sum	8012	2003	8012	2003	8012	2003	8012	2003	8012	2003

Grad-CAM Analysis: We perform A Grad-CAM analysis to understand how our ABC ensemble model interprets and processes skin lesions. This demonstrates that our model is not simply performing well by chance, but is actually making correct and meaningful decisions. Grad-CAM (Gradient-weighted Class Activation Mapping) [51] is a technique used to visualize which areas of an input image a neural network focuses on when making a decision. It works by using the gradients of the target output flowing into the final convolutional layer to produce a coarse localization map, highlighting the important regions in the image. This helps to identify which parts of the image influence the model's decision the most. In Grad-CAM visualizations, areas with larger gradients indicate regions that have a greater influence on the model's decision. These regions are typically displayed in warmer colors like red or yellow, while areas with smaller gradients, where the model focuses less, appear in cooler colors like blue. The heatmap is then superimposed onto the original image, showing which parts of the lesion the model focuses on during classification.

To demonstrate that our model's performance is not due to chance, we generate Grad-CAM visualizations for each of the five modules in the ABC ensemble model. The results are presented in Section 4.3.

Weight Optimization: In our ABC ensemble model, the probability values predicted by the five blocks are combined using the soft voting method in the ensemble classifier to obtain the final classification result. During soft voting, we assign double weight to the two blocks (Block 1 and Block 2) that analyze the features of a general image. This approach is based on the assumption that analyzing the general characteristics of the images is most critical, and incorporating the additional information derived from the ABCD rule would enhance classification performance. Despite our assumptions about the importance of generalization ability, our model contains two generalization blocks and three ABC-related blocks, meaning that there is one fewer generalization block.

To balance this, we assign a weight of 2 to each generalization block and a weight of 1 to each ABC-related block. To validate this weighting optimization assumption, we conduct an experiment by testing different weight combinations. The experiment includes three cases: (1) no weights are applied (i.e., all weights set to 1), (2) only one block is assigned double weight (an experiment across the five blocks), and (3) only the generalization blocks are assigned a weight of 2. The detailed results of these experiments are provided in Section 4.4.

4. Results

In this section, we evaluate the performance of our proposed ABC ensemble model and validate the results. Section 4.1 presents the overall performance of skin lesion classification, while Section 4.2 demonstrates the classification performance of the model for each class.

Section 4.3 uses Grad-CAM analysis to show which characteristics the five blocks within the model focus on. Lastly, Section 4.4 explains the weight optimization in soft voting.

4.1. Overall Classification Performance

The following is an explanation of the performance evaluation results from the skin lesion classification experiment using the HAM10000 dataset. The overall performance for skin lesion classification on the HAM10000 dataset is shown in Table 3. The 'Model' column represents the 15 experimental models, and the 'Preprocessing Method' column indicates the five preprocessing methods: ROI for ROI extraction, Ups. for Upsampling, Seg. for Segmentation, Adj. for Adjust Hue, and Mor. for Morphological Transformation. The best-performing model is highlighted in bold, and the second-best model is underlined. The results for all 15 experimental models are presented as the average accuracy, recall, precision, and F1-score over five experiments, each conducted with 5-fold cross-validation.

Table 3. Overall performance for skin lesion classification on the HAM10000 dataset. All values are the averages of 5 experimental results from 5-fold cross-validation. The best-performing model is highlighted in bold, and the second-best model is underlined.

Model	Preprocessing Method					Accuracy	Recall	Precision	F1-Score
	ROI.	Ups.	Seg.	Adj.	Mor.				
GoogLeNet	-	-	-	-	-	0.8816	0.8779	0.8816	0.8762
AlexNet	-	-	-	-	-	0.8458	0.8424	0.8458	0.8412
ViT-B/16	-	-	-	-	-	0.8937	0.8905	0.8932	0.8900
DenseNet121	-	-	-	-	-	0.9023	0.9001	0.9022	0.8993
EfficientNet-B0	-	-	-	-	-	0.9025	0.9000	0.9024	0.8997
MobileNetV2	-	-	-	-	-	0.8909	0.8878	0.8909	0.8880
ResNet50	-	-	-	-	-	0.8883	0.8854	0.8883	0.8852
ViT-B/16	✓	-	-	-	-	0.9090	0.9059	0.9073	0.9053
DenseNet121	✓	✓	-	-	-	0.8913	0.8900	0.8913	0.8893
EfficientNet-B0	✓	-	✓	-	-	0.9073	0.9057	0.9073	0.9051
MobileNetV2	✓	-	-	✓	-	0.8999	0.8987	0.8999	0.8982
ResNet50	✓	-	-	-	✓	0.8999	0.8982	0.8999	0.8973
Ensemble (w/o ABCD)	-	-	-	-	-	0.9154	0.9137	0.9154	0.9129
<u>Ensemble (w/ weights, w/o ABCD)</u>	-	-	-	-	-	<u>0.9232</u>	<u>0.9223</u>	<u>0.9226</u>	<u>0.9196</u>
Ensemble (w/ weights, w/ ABCD, Ours)	✓	✓	✓	✓	✓	**0.9326**	**0.9316**	**0.9310**	**0.9302**

The model with the best overall performance is the Ensemble (w/ weights, w/ ABCD, Ours), which consistently outperformed all other models across all metrics. Specifically, it achieved an accuracy of 0.9326, a recall of 0.9316, a precision of 0.9310, and an F1-score of 0.9302. This model demonstrates that the combination of weighted soft voting and preprocessing with the ABCD rule leads to superior classification performance. The second-best model is the Ensemble (w/ weights, w/o ABCD), which also used weighted soft voting but did not incorporate the ABCD rule. This model delivered results, with an accuracy of 0.9232, a recall of 0.9223, a precision of 0.9226, and an F1-score of 0.9196. The best model (our ABC ensemble model)'s F1-score of 0.9302 is higher than the second-best model's F1-score of 0.9196, with a difference of approximately 1.06%, showing the significant impact of incorporating the ABCD rule in the classification process. This difference highlights how ABCD-based preprocessing allows for the better classification of lesion characteristics.

Among the seven models without preprocessing (1st model~7th model in Table 3), EfficientNet-B0 showed the best performance, achieving an accuracy of 0.9025 and an F1-score of 0.8997. AlexNet showed the lowest performance, with an accuracy of 0.8458 and an F1-score of 0.8412. As a result, when building the ensemble block, we excluded the lower-performing GoogLeNet and AlexNet models and instead used ViT-B/16, DenseNet121, EfficientNet-B0, MobileNetV2, and ResNet50 to construct the block. Despite its competitive performance, the lack of preprocessing limited its ability to capture finer details and complexities in the dataset. This baseline result emphasizes the importance of preprocessing techniques for improving model accuracy by enhancing the model's focus on relevant features of the data. Among the five models with preprocessing (8th model~12th model in Table 3), EfficientNet-B0 with preprocessing (ROI extraction and Segmentation) again emerged as the top performer. This model achieved an accuracy of 0.9073 and an F1-score of 0.9051, showing a slight improvement in the F1-score compared to its non-preprocessed counterpart. Although DenseNet121 experienced a very slight decrease in performance, other models demonstrated an overall improvement in performance after preprocessing. It demonstrates that preprocessing steps help the model to better identify and focus on the most important aspects of the lesion, ultimately leading to improved performance.

4.2. Category-Wise Classification Performance

In this research, we classify skin cancer images from the HAM10000 dataset into seven classes, and the model's classification performance may differ for each class. Accordingly, we visualized the confusion matrix to understand the model's classification performance for each class on the test dataset, as shown in Figure 14. The y-axis represents the actual classes, while the x-axis indicates the predicted classes by the model. The left confusion matrix (a) shows the raw confusion matrix, displaying the actual counts of correctly and incorrectly classified instances for each class. The diagonal elements represent True Positives (TP), while the off-diagonal elements indicate misclassifications. The right confusion matrix (b) shows the normalized confusion matrix, where the values are represented as percentages, demonstrating the model's classification performance for each class. This helps in understanding the relative performance of the model across different classes, regardless of the imbalance in the number of samples for each class. As demonstrated in Figure 14b, we can see that the True Positive Rate (TPR) for the 'nv' class is the highest at 0.98, followed by the 'vasc', 'bcc', 'bkl', 'mel' and 'akiec' classes, with a TPR of 0.90, 0.87, 0.86, 0.85 and 0.73, respectively. However, the 'akiec' class exhibits a relatively lower TPR of 0.73. This suggests that while our ABC ensemble model did not perform as well in learning features for the 'akiec' class, it successfully captured the features for the remaining six classes. The issue of insufficient learning for the 'akiec' class will be addressed in detail in Section 5.

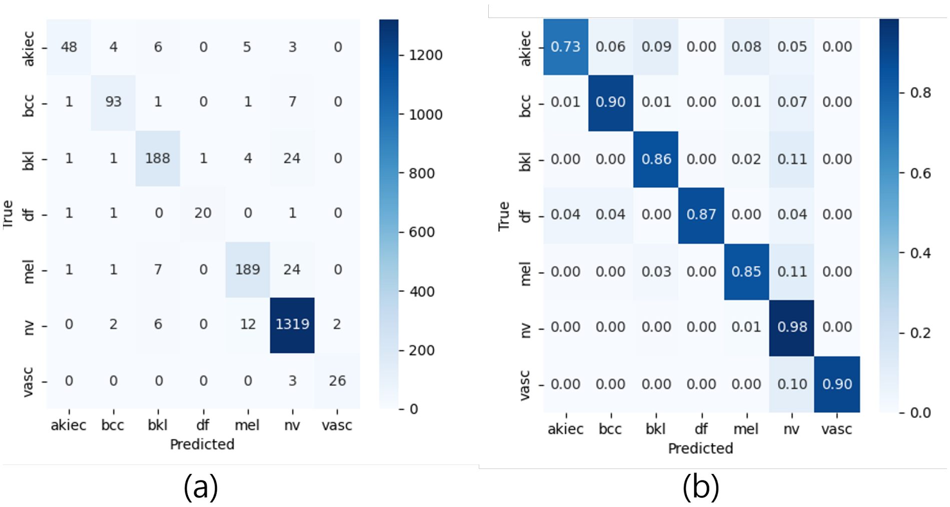


Figure 14. Confusion matrices of the ABC ensemble model for each class of the skin cancer dataset. (**a**) indicates the raw confusion matrix which displays the actual counts of correctly and incorrectly classified instances for each class, while (**b**) shows normalized confusion matrix where the values are represented as percentages, showing the model's classification performance for each class.

4.3. Grad-CAM Analysis

Our model applies different preprocessing methods across the five blocks of the ensemble model, based on the clinical guideline, the ABCD rule. Block 1 and Block 2 from Figure 4 are designed to capture general image characteristics, Block 3 focuses on the symmetry of the lesion, Block 4 blurs color information to emphasize the lesion's border, and Block 5 concentrates on the color of the lesion. To ensure that each preprocessing method works as intended, we performed Grad-CAM analysis, with the results shown in Figure 15. This figure shows the original images in the first row and their corresponding Grad-CAM visualizations in the second row for the five blocks.

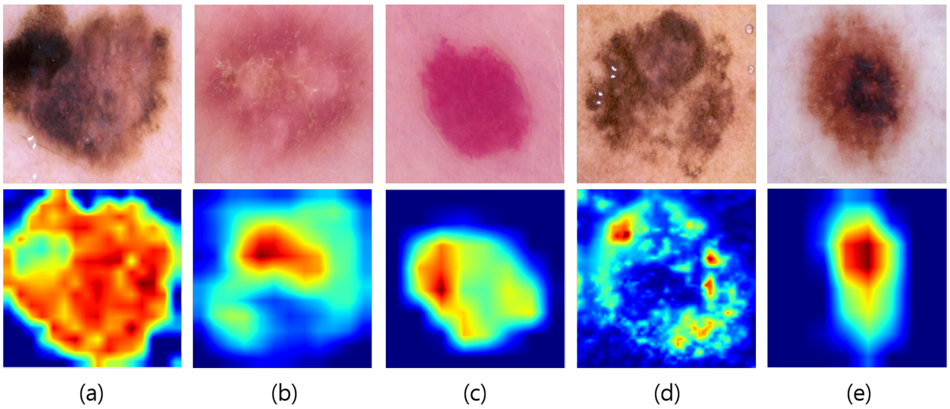


Figure 15. Grad-CAM visualizations for different five blocks: (**a**) Generality Training, (**b**) Upsize Training, (**c**) Asymmetry Enhancement Training, (**d**) Border Enhancement Training, (**e**) Color Enhancement Training.

Grad-CAM visualizations use colors to indicate the model's focus. Warmer colors such as red and yellow highlight areas that receive more attention, while cooler colors like blue indicate less focus.

In Figure 15a for the Generality Training, the model demonstrates consistent attention across the entire lesion area. This indicates that Generality Training is effectively interpreting the overall information of the image without bias, ensuring that the model captures the full context of the lesion. In the case of Upsize Training, the Grad-CAM focuses on the primary area of the lesion while still considering the overall lesion as shown in Figure 15b. By learning the main features of the lesion without any specific enhancement, this approach compensates for any information loss in other feature-enhanced models. Unlike the ViT-B/16 model, which processes lesions using patch embeddings, Upsize Training improves the overall performance of the ensemble by concentrating on the most important parts of the lesion. In Figure 15c for the Asymmetry Enhancement Training, we can see that the model concentrates on the shape of the most critical area within the lesion. This helps transmit the core lesion area to the classifier, enabling it to learn and become more sensitive to shape asymmetry. In Figure 15d for the Border Enhancement Training, we can find that the model specifically focuses on and learns from the border of the skin lesion. Since this method blurs the overall color features of the image, it ensures that the model concentrates more on the lesion's contour rather than on color information. This shows that the Border Enhancement Training effectively extracts the border feature rather than color or other characteristics. In Figure 15e for the Color Enhancement Training, we can find that the Grad-CAM focuses mostly on the darkest areas of the lesion, indicating that the Color Enhancement Training intensely focuses on the color aspect of the skin lesion. Thus, it shows that this method allows for effective extraction of color-related features in skin lesions.

In summary, Grad-CAM allowed us to observe that each of the five different preprocessing methods focuses on different characteristics of the lesion. This suggests that our model's use of various preprocessing techniques, guided by the ABCD rule, effectively contributed to the improvement in performance.

4.4. Weighted Optimization

In our ABC ensemble model, the probability values from the five blocks are combined through soft voting to obtain the final classification. We assign double weight to two generalization blocks, assuming general image features are most crucial, while ABC-related blocks provide additional information. Table 4 presents the results of an experiment conducted to validate this assumption and identify the optimal weight combination for the ensemble model. In this table, the order of the model weights in parentheses in the column 'weight' is as follows: [ViT-B/16 (Block 1), DenseNet121 (Block 2), EfficientNet-B0 (Block 3), MobileNetV2 (Block 4), and ResNet50 (Block 5)]. Excluding the title row, the first row represents the case where no weight is applied, meaning all weights are set to 1. The second to sixth rows represent cases where a weight of 2 is assigned to only one model. The seventh row shows the case where a weight of 2 is assigned to the three ABC-related blocks, while the last row shows the case where a weight of 2 is given to the two generalization blocks. The best performance is highlighted in bold, and the second-best performance is underlined.

Table 4. Performance comparison of ABC ensemble model with different weight combinations. The order of the model weights in the column 'weight' is as follows: [ViT-B/16 (Block 1), DenseNet121 (Block 2), EfficientNet-B0 (Block 3), MobileNetV2 (Block 4), ResNet50 (Block 5)]. The best-performing weight is highlighted in bold and the second-best-performing weight is underlined.

Model	Weights	Accuracy	Recall	Precision	F1-Score
ABC ensemble	[1, 1, 1, 1, 1]	0.9313	0.9304	0.9313	0.9294
<u>ABC ensemble</u>	<u>[2, 1, 1, 1, 1]</u>	<u>0.9320</u>	<u>0.9311</u>	<u>0.9320</u>	<u>0.9302</u>
ABC ensemble	[1, 2, 1, 1, 1]	0.9315	0.9307	0.9315	0.9297
ABC ensemble	[1, 1, 2, 1, 1]	0.9313	0.9307	0.9313	0.9295
ABC ensemble	[1, 1, 1, 2, 1]	0.9294	0.9289	0.9294	0.9276
ABC ensemble	[1, 1, 1, 1, 2]	0.9303	0.9295	0.9303	0.9284
ABC ensemble	[1, 1, 2, 2, 2]	0.9296	0.9289	0.9296	0.9278
ABC ensemble	**[2, 2, 1, 1, 1]**	**0.9326**	**0.9316**	**0.9326**	**0.9308**

The best-performing configuration is the last row, where weights of 2 are assigned to the two generalization blocks (Block 1 and Block 2), resulting in an accuracy of 0.9326, a recall of 0.9316, a precision of 0.9326, and an F1-score of 0.9308. The second-best model, which assigned a weight of 2 to Block 1 (ViT-B/16), also achieved strong performance, with an accuracy of 0.9320, a recall of 0.9311, a precision of 0.9320, and an F1-score of 0.9302, followed by the third-best model, which assigned a weight of 2 to Block 2 (DenseNet121). These results suggest that placing greater emphasis on generalization blocks tends to yield better results. They also indicate that boosting general image feature analysis through additional weighting on the generalization blocks leads to stronger overall performance across the ensemble.

5. Discussion

In our study, we aimed to create a model that fully incorporates the ABCD rule. However, we were unable to measure the D (diameter) of the lesions because the images in the HAM10000 dataset were collected under inconsistent conditions, with varying degrees of angles, distances, and magnifications. To overcome these challenges, one possible improvement would be to collect a dataset where lesions are photographed under standardized conditions, ensuring consistent angles, distances, and lighting. This would allow the model to better utilize diameter information, a critical aspect for skin lesion diagnosis, especially given that lesion size is an important feature in melanoma classification. We expect that integrating diameter as a learnable feature could potentially improve diagnostic accuracy.

In the HAM10000 dataset, certain skin lesion types, such as nevi (nv), are overrepresented compared to rarer lesion types like melanoma (mel) or dermatofibroma (df). This imbalance can lead to the model becoming biased toward the more frequent classes, reducing its sensitivity to detecting the rarer yet clinically significant lesions like melanoma. To mitigate this issue, we applied classic augmentation techniques such as rotation, zoom, horizontal flip, and vertical flip, as described in Section 3.1. However, the excessive oversampling of minority classes using basic augmentation methods could result in repetitive data, which might negatively affect the model by introducing redundancy and overfitting. A more advanced approach, such as incorporating Generative Adversarial Networks (GANs), could be a potential solution. GANs can generate synthetic but diverse images for minor classes, helping the model to learn better without relying on repeated data. Therefore, the model would be exposed to a broader set of lesion features, reducing overfitting and improving generalization.

Another limitation of our study is the lack of data. When the dataset is insufficient, the model may overfit to the training data, leading to poor generalization performance, or it may fail to adequately learn the necessary patterns and features. As shown in Section 4.2, the performance of the ‘akiec’ class was not as high as desired, likely due to the insufficient learning of the lesion characteristics caused by the limited data available for this class. This is further supported by the observation that the ‘nv’ class, which has the most images, achieved noticeably better classification performance compared to the other classes. To address this issue, future studies should aim to collect and integrate additional datasets to improve the model’s predictive performance. Especially in medical applications, the demands for high performance are critical because of the serious implications that incorrect classifications can have. Although our model has shown better classification performance compared to several other models in this study, we acknowledge that there is still room for improvement. Therefore, at the current stage, we think that our model is more suitable as an auxiliary diagnostic tool for medical professionals, rather than as a standalone diagnostic system. However, by expanding the dataset, training the model for more epochs, increasing the current 5-fold cross-validation to 10-fold, and exploring improved methodologies, we expect that we can further enhance the model’s classification accuracy and reliability. The ultimate goal will be to reach a level of performance that instills sufficient clinical confidence for the model to be used as an independent diagnostic tool.

6. Conclusions

Skin cancer is one of the most common and serious cancers worldwide, and its incidence is increasing due to several factors. In this study, we developed an ABC ensemble model for skin lesion classification by leveraging the ABCD rule, which reflects the clinical criteria used to assess lesions. The model consists of five distinct blocks, each focusing on different lesion characteristics through unique preprocessing methods.

To demonstrate the effectiveness of our proposed model, we conducted experiments across 15 different model configurations, showing that our model outperformed others in terms of accuracy, recall, precision, and F1-score. Additionally, we demonstrated the effectiveness of the weight assignments for each model in soft voting through experiments using various weight combinations. By assigning higher weights to the generalization blocks that analyze general image features, we observed notable improvements in classification results. This suggests that emphasizing general image features while integrating ABC-related features yields better performance. We further validated the effectiveness of our model through Grad-CAM analysis. The different preprocessing methods in our model were shown to focus on not only general features but also specific lesion characteristics, including asymmetry, border, and color, enhancing the model’s overall reliability and interpretability.

In conclusion, the ABC ensemble model, with ABCD rule-based preprocessing and weighted voting, demonstrates robust performance in skin lesion classification. We hope that our research contributes to the early detection of skin cancer, ultimately enhancing public health and improving patient outcomes.

Author Contributions: Conceptualization, J.-Y.C.; methodology, J.-Y.C.; software, J.-Y.C. and M.-J.S.; resources, Y.-J.S.; validation J.-Y.C.; data curation, J.-Y.C. and M.-J.S.; formal analysis and investigation, M.-J.S.; writing—original draft preparation, M.-J.S.; writing—review and editing, J.-Y.C.; visualization, J.-Y.C.; supervision, Y.-J.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Research Fund of The Catholic University of Korea in 2024 and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00213456).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000, (accessed on 6 November 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Lee, H.; Chen, Y.P.P. Image based computer aided diagnosis system for cancer detection. *Expert Syst. Appl.* **2015**, *42*, 5356–5365. [CrossRef]
2. Feller, L.; Khammissa, R.; Kramer, B.; Altini, M.; Lemmer, J. Basal cell carcinoma, squamous cell carcinoma and melanoma of the head and face. *Head Face Med.* **2016**, *12*, 11. [CrossRef]
3. Basak, H.; Kundu, R.; Sarkar, R. MFSNet: A multi focus segmentation network for skin lesion segmentation. *Pattern Recognit.* **2022**, *128*, 108673. [CrossRef]
4. Haggenmüller, S.; Maron, R.C.; Hekler, A.; Utikal, J.S.; Barata, C.; Barnhill, R.L.; Beltraminelli, H.; Berking, C.; Betz-Stablein, B.; Blum, A.; et al. Skin cancer classification via convolutional neural networks: systematic review of studies involving human experts. *Eur. J. Cancer* **2021**, *156*, 202–216. [CrossRef]
5. Gajera, H.K.; Nayak, D.R.; Zaveri, M.A. M2CE: Multi-convolutional neural network ensemble approach for improved multiclass classification of skin lesion. *Expert Syst.* **2023**, *40*, e13435. [CrossRef]
6. Zhao, Z. Skin cancer classification based on convolutional neural networks and vision transformers. *J. Phys. Conf. Ser.* **2022**, *2405*, 012037. [CrossRef]
7. 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. [CrossRef]
8. White, R.; Rigel, D.S.; Friedman, R.J. Computer applications in the diagnosis and prognosis of malignant melanoma. *Dermatol. Clin.* **1991**, *9*, 695–702. [CrossRef]
9. Blum, A.; Luedtke, H.; Ellwanger, U.; Schwabe, R.; Rassner, G.; Garbe, C. Digital image analysis for diagnosis of cutaneous melanoma. Development of a highly effective computer algorithm based on analysis of 837 melanocytic lesions. *Br. J. Dermatol.* **2004**, *151*, 1029–1038. [CrossRef]
10. Ballerini, L.; Fisher, R.B.; Aldridge, B.; Rees, J. A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions. In *Color Medical Image Analysis*; Springer: Dordrecht, The Netherlands, 2013; pp. 63–86.
11. Celebi, M.E.; Kingravi, H.A.; Uddin, B.; Iyatomi, H.; Aslandogan, Y.A.; Stoecker, W.V.; Moss, R.H. A methodological approach to the classification of dermoscopy images. *Comput. Med. Imaging Graph.* **2007**, *31*, 362–373. [CrossRef]
12. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
13. Alwakid, G.; Gouda, W.; Humayun, M.; Sama, N.U. Melanoma detection using deep learning-based classifications. *Healthcare* **2022**, *10*, 2481. [CrossRef] [PubMed]
14. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
15. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
16. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.

17. Ali, K.; Shaikh, Z.A.; Khan, A.A.; Laghari, A.A. Multiclass skin cancer classification using EfficientNets–a first step towards preventing skin cancer. *Neurosci. Inform.* **2022**, *2*, 100034. [CrossRef]
18. Yin, W.; Huang, J.; Chen, J.; Ji, Y. A study on skin tumor classification based on dense convolutional networks with fused metadata. *Front. Oncol.* **2022**, *12*, 989894. [CrossRef]
19. Wang, H.; Qi, Q.; Sun, W.; Li, X.; Dong, B.; Yao, C. Classification of skin lesions with generative adversarial networks and improved MobileNetV2. *Int. J. Imaging Syst. Technol.* **2023**, *33*, 1561–1576. [CrossRef]
20. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 6000–6010.
21. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
22. Song, B.; Kc, D.R.; Yang, R.Y.; Li, S.; Zhang, C.; Liang, R. Classification of Mobile-Based Oral Cancer Images Using the Vision Transformer and the Swin Transformer. *Cancers* **2024**, *16*, 987. [CrossRef]
23. Kwasigroch, A.; Grochowski, M.; Mikołajczyk, A. Neural architecture search for skin lesion classification. *IEEE Access* **2020**, *8*, 9061–9071. [CrossRef]
24. Rahi, M.M.I.; Khan, F.T.; Mahtab, M.T.; Ullah, A.A.; Alam, M.G.R.; Alam, M.A. Detection of skin cancer using deep neural networks. In Proceedings of the 2019 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), Melbourne, Australia, 9–11 December 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–7.
25. Kawahara, J.; BenTaieb, A.; Hamarneh, G. Deep features to classify skin lesions. In Proceedings of the 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), Prague, Czech Republic, 13–16 April 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 1397–1400.
26. Rahman, A.; Smith, D.V.; Timms, G. Multiple classifier system for automated quality assessment of marine sensor data. In Proceedings of the 2013 IEEE Eighth International Conference on Intelligent Sensors, Sensor Networks and Information Processing, Melbourne, Australia, 2–5 April 2013; IEEE: Piscataway, NJ, USA, 2013; pp. 362–367.
27. Haralabopoulos, G.; Anagnostopoulos, I.; McAuley, D. Ensemble deep learning for multilabel binary classification of user-generated content. *Algorithms* **2020**, *13*, 83. [CrossRef]
28. Zhao, X.; Zhang, M.; Zhang, J. Ensemble learning-based CNN for textile fabric defects classification. *Int. J. Cloth. Sci. Technol.* **2021**, *33*, 664–678. [CrossRef]
29. Vennelakanti, A.; Shreya, S.; Rajendran, R.; Sarkar, D.; Muddegowda, D.; Hanagal, P. Traffic sign detection and recognition using a CNN ensemble. In Proceedings of the 2019 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 11–13 January 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–4.
30. Müller, D.; Soto-Rey, I.; Kramer, F. An analysis on ensemble learning optimized medical image classification with deep convolutional neural networks. *IEEE Access* **2022**, *10*, 66467–66480. [CrossRef]
31. de Jesus Silva, L.F.; Cortes, O.A.C.; Diniz, J.O.B. A novel ensemble CNN model for COVID-19 classification in computerized tomography scans. *Results Control Optim.* **2023**, *11*, 100215. [CrossRef]
32. Charan, D.S.; Nadipineni, H.; Sahayam, S.; Jayaraman, U. Method to classify skin lesions using dermoscopic images. *arXiv* **2020**, arXiv:2008.09418.
33. Mahbod, A.; Schaefer, G.; Wang, C.; Dorffner, G.; Ecker, R.; Ellinger, I. Transfer learning using a multi-scale and multi-network ensemble for skin lesion classification. *Comput. Methods Programs Biomed.* **2020**, *193*, 105475. [CrossRef]
34. Lin, T.C.; Lee, H.C. Skin cancer dermoscopy images classification with meta data via deep learning ensemble. In Proceedings of the 2020 International Computer Symposium (ICS), Tainan, Taiwan, 17–19 December 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 237–241.
35. Akay, B.; Kocyigit, P.; Heper, A.; Erdem, C. Dermatoscopy of flat pigmented facial lesions: Diagnostic challenge between pigmented actinic keratosis and lentigo maligna. *Br. J. Dermatol.* **2010**, *163*, 1212–1217. [CrossRef]
36. Lallas, A.; Apalla, Z.; Argenziano, G.; Longo, C.; Moscarella, E.; Specchio, F.; Raucci, M.; Zalaudek, I. The dermatoscopic universe of basal cell carcinoma. *Dermatol. Pract. Concept.* **2014**, *4*, 11. [CrossRef]
37. Moscarella, E.; Zalaudek, I.; Pellacani, G.; Eibenschutz, L.; Catricalà, C.; Amantea, A.; Panetta, C.; Argenziano, G. Lichenoid keratosis-like melanomas. *J. Am. Acad. Dermatol.* **2011**, *65*, e85–e87. [CrossRef]
38. Zaballos, P.; Puig, S.; Llambrich, A.; Malvehy, J. Dermoscopy of dermatofibromas: A prospective morphological study of 412 cases. *Arch. Dermatol.* **2008**, *144*, 75–83. [CrossRef]
39. Schiffner, R.; Schiffner-Rohe, J.; Vogt, T.; Landthaler, M.; Wlotzke, U.; Cognetta, A.B.; Stolz, W. Improvement of early recognition of lentigo maligna using dermatoscopy. *J. Am. Acad. Dermatol.* **2000**, *42*, 25–32. [CrossRef]
40. Rosendahl, C.; Cameron, A.; McColl, I.; Wilkinson, D. Dermatoscopy in routine practice: 'Chaos and clues'. *Aust. Fam. Physician* **2012**, *41*, 482–487. [PubMed]
41. Zaballos, P.; Carulla, M.; Ozdemir, F.; Zalaudek, I.; Bañuls, J.; Llambrich, A.; Puig, S.; Argenziano, G.; Malvehy, J. Dermoscopy of pyogenic granuloma: A morphological study. *Br. J. Dermatol.* **2010**, *163*, 1229–1237. [CrossRef] [PubMed]
42. Friedman, R.J.; Rigel, D.S.; Kopf, A.W. Early detection of malignant melanoma: The role of physician examination and self-examination of the skin. *CA Cancer J. Clin.* **1985**, *35*, 130–151. [CrossRef]

43. Healthline. What Is the ABCDE Rule for Detecting Skin Cancer? Available online: https://www.healthline.com/health/skin-cancer/abcd-rule-for-skin-cancer (accessed on 6 November 2024).
44. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; Proceedings, Part III 18; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
45. Izadpanahkakhk, M.; Razavi, S.M.; Taghipour-Gorjikolaie, M.; Zahiri, S.H.; Uncini, A. Deep region of interest and feature extraction models for palmprint verification using convolutional neural networks transfer learning. *Appl. Sci.* **2018**, *8*, 1210. [CrossRef]
46. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
47. Srisha, R.; Khan, A. Morphological operations for image processing: Understanding and its applications. *NCVSComs-13* **2013**, *13*, 19.
48. Kang, H.C.; Han, H.N.; Bae, H.C.; Kim, M.G.; Son, J.Y.; Kim, Y.K. HSV color-space-based automated object localization for robot grasping without prior knowledge. *Appl. Sci.* **2021**, *11*, 7593. [CrossRef]
49. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
50. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 84–90. [CrossRef]
51. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.