

Article

SkinSwinViT: A Lightweight Transformer-Based Method for Multiclass Skin Lesion Classification with Enhanced Generalization Capabilities

Kun Tang ^{1,†} , Jing Su ^{1,*,†}, Ruihan Chen ^{1,2} , Rui Huang ¹, Ming Dai ^{1,*} and Yongjiang Li ¹

¹ School of Mathematics and Computer, Guangdong Ocean University, Zhanjiang 524008, China

² Artificial Intelligence Research Institute, International (Macau) Institute of Academic Research, Macau 999078, China

* Correspondence: jingsu@gdou.edu.cn (J.S.); daiming@gdou.edu.cn (M.D.)

† These authors contributed equally to this work.

Abstract: In recent decades, skin cancer has emerged as a significant global health concern, demanding timely detection and effective therapeutic interventions. Automated image classification via computational algorithms holds substantial promise in significantly improving the efficacy of clinical diagnoses. This study is committed to mitigating the challenge of diagnostic accuracy in the classification of multiclass skin lesions. This endeavor is inherently formidable owing to the resemblances among various lesions and the constraints associated with extracting precise global and local image features within diverse dimensional spaces using conventional convolutional neural network methodologies. Consequently, this study introduces the SkinSwinViT methodology for skin lesion classification, a pioneering model grounded in the Swin Transformer framework featuring a global attention mechanism. Leveraging the inherent cross-window attention mechanism within the Swin Transformer architecture, the model adeptly captures local features and interdependencies within skin lesion images while additionally incorporating a global self-attention mechanism to discern overarching features and contextual information effectively. The evaluation of the model's performance involved the ISIC2018 challenge dataset. Furthermore, data augmentation techniques augmented training dataset size and enhanced model performance. Experimental results highlight the superiority of the SkinSwinViT method, achieving notable metrics of accuracy, recall, precision, specificity, and F1 score at 97.88%, 97.55%, 97.83%, 99.36%, and 97.79%, respectively.

Keywords: skin lesions; ISIC2018; transformer; data enhancement; multiclassification



Citation: Tang, K.; Su, J.; Chen, R.; Huang, R.; Dai, M.; Li, Y. SkinSwinViT: A Lightweight Transformer-Based Method for Multiclass Skin Lesion Classification with Enhanced Generalization Capabilities. *Appl. Sci.* **2024**, *14*, 4005. <https://doi.org/10.3390/app14104005>

Academic Editor: Jongsung Lee

Received: 27 March 2024

Revised: 2 May 2024

Accepted: 6 May 2024

Published: 8 May 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Skin diseases represent a pervasive health concern across all age cohorts. Among the primary types of skin cancer, namely melanoma and non-melanoma, melanoma exhibits a higher mortality rate and is considered the most malignant form [1]. Timely detection of melanoma substantially elevates the 5-year survival rate to 95%, in stark contrast to a dismal 20% without intervention [2]. As incidence and mortality rates ascend, the imperative of early detection becomes increasingly conspicuous. Presently, machine learning and deep learning methodologies enable automated diagnosis of skin lesions via high-resolution dermoscopic images. To facilitate global research endeavors in skin cancer detection and analysis, the International Skin Imaging Collaboration (ISIC) orchestrates the annual ISIC Grand Challenge [3].

With continual advancements in medical technology, there is burgeoning recognition of artificial intelligence and machine learning's potential in skin lesion identification and diagnosis [4]. Particularly in the realm of skin cancer detection, the intricate and diverse nature of skin ailments poses challenges to traditional diagnostic modalities grounded in subjective clinical assessment, heightening the propensity for misdiagnosis. Hence,

the development of automated and precise dermatological lesion identification systems assumes paramount importance [5]. Computer-aided diagnosis (CAD) systems have made substantial strides in identifying and assessing various malignancies [6], spanning lung cancer [7], breast cancer [8], thyroid cancer [9], brain cancer [10], and liver cancer [11], among others. In the domain of skin cancer detection, CAD system implementation becomes indispensable, enhancing efficiency, curtailing time and costs, and compensating for the scarcity of dermatologists.

Recent years have witnessed rapid advancements in dermatological lesion recognition, attributed to significant strides in CAD systems, owing to the rapid evolution of machine learning techniques [12]. Traditional machine learning approaches, such as support vector machines (SVM) and random forest models, offer interpretability and small sample learning advantages [13]. These methods rely on features designed by domain experts and furnish explanatory lesion attributes, which facilitate medical practitioners' comprehension of the model's decision-making process. Conversely, classic deep learning methods, exemplified by ResNet, MobileNet, and VGGNet models [14], excel in discerning deep abstract features and in exploiting transfer learning for skin disease lesion identification. They adeptly capture intricate characteristics such as lesion texture, shape, and structure, leveraging pre-trained models on extensive image datasets to enhance model efficacy. Deep learning methodologies predicated on transformers, exemplified by Vision Transformer (ViT) and Swin Transformer (SwinViT) [15], offer automatic feature learning, context modeling, scalability, and resilience in skin disease lesion recognition. These methodologies effectively encompass contextual information within images, spanning local and global contexts via a self-attention mechanism, affording a deeper understanding of the lesion relationship.

Nonetheless, a notable challenge stems from the substantial inter-class resemblance observed in numerous skin lesion images, rendering the identification of unique and discernible custom features arduous. Convolutional neural networks (CNNs) may conflate low-level features in such scenarios, resulting in crucial information loss. Moreover, CNNs exhibit constraints in capturing global contextual information, potentially omitting vital details. To surmount these challenges, this study proposes a deep learning framework employing a local–global hierarchical attention mechanism grounded in transformers for multi-category skin lesion classification diagnosis. The study encompasses seven skin lesion types: melanoma (MEL), melanocytic nevus (NV), basal cell carcinoma (BCC), actinic keratosis (AKIEC), benign keratosis (BKL), dermatofibroma (DF), and vasculopathy (VASC), as delineated in Figure 1. The principal objective of this framework is to augment the accuracy, precision, and resilience of multiclass skin lesion classification. Our main contributions to this endeavor are as follows:

- (1) Data augmentation techniques are employed to address training data imbalances and generate a more representative set of skin lesion image samples. This methodology enables the model to glean robust representations of lesion features.
- (2) A novel local–global hierarchical attention mechanism is introduced to capture pivotal features across different abstraction levels. By amalgamating local and global features across multiple levels, the model harnesses both intricate and contextual information to amplify feature representation and diagnostic precision. This hierarchical attention mechanism furnishes adaptability in handling features across distinct abstraction levels, enabling the model to tackle complex diagnostic tasks adeptly.
- (3) This study proposes an encoder–decoder framework based on the Transformer model, which leverages its scalability. The model's representation capacity is bolstered by modulating the encoder and decoder layer counts alongside attention head numbers. Pre-training and fine-tuning strategies are additionally employed to refine the performance of skin lesion identification tasks.
- (4) Comparative experiments against state-of-the-art methodologies corroborate the superior performance of our proposed method in multiclass skin lesion classification. The experimental results demonstrate substantial enhancements in accuracy, precision, specificity, and F1 index, confirming the effectiveness of our method.

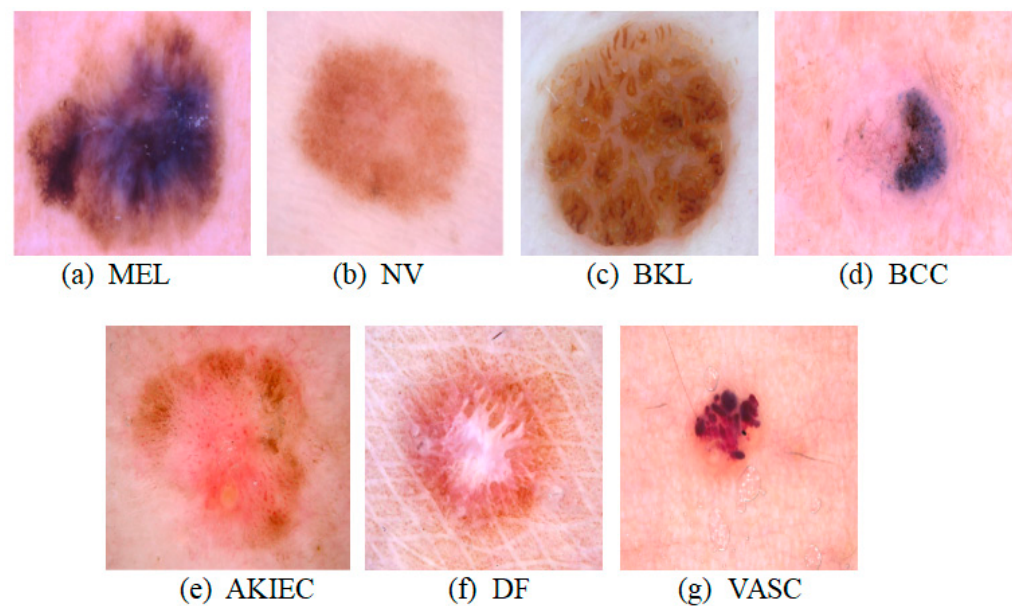


Figure 1. Illustrative Instances of Diverse Skin Lesions from ISIC2018.

The article is structured as follows: starting with the abstract and introduction, Section 2 provides a comprehensive literature review, offering a systematic overview of the current progress and status of related research. Subsequently, Section 3 elaborates on the proposed methodology, detailing the data sources and adopted data augmentation techniques. Furthermore, state-of-the-art models and methods are introduced and presented. In Section 4, this study presents the experimental results and conducts an extensive comparative analysis based on various indicators, including precision, accuracy, and F1 index. The experimental results are thoroughly interpreted to provide a comprehensive evaluation of the proposed method's performance. Finally, the conclusion summarizes the effectiveness and innovativeness of the proposed method.

2. Related Work

In the realm of dermatological image diagnosis, significant advancements have been achieved by scholars and practitioners hailing from both domestic and international spheres. Notably, conventional machine learning methodologies have exhibited promising outcomes in delineating skin lesions through contour analysis. For instance, Chatterjee et al. [16] conducted a study based on the ABCDE criterion [17], encompassing the scrutiny of shape, edge regularity, texture, and color attributes of skin lesions. Leveraging image processing tools, they extracted quantitative features and employed SVM for classification, thereby evaluating the efficacy of the proposed framework. Dhivyaa et al. [18] deployed the region-growing technique for lesion segmentation and feature extraction, followed by decision tree-driven classification, culminating in the selection of the random forest algorithm as the final model. Relative to SVM, random forest demonstrates diminished computational complexity, pliability in design, and proficiency in handling diverse categories of dermatological conditions. Pham et al. [19], in their comparative investigation of melanoma classification, explored diverse data preprocessing methods, feature extraction strategies, and classification algorithms. Their findings underscored that employing linear normalization, HSV feature extraction, and a balanced random forest classifier yielded optimal results on the HAM10000 dataset, achieving an accuracy rate of 74.75%. Despite the strides made by assorted machine learning methodologies in skin classification, wherein some have surpassed the expertise of human practitioners in select cases [20], conventional machine learning algorithms, rooted in statistical principles, exhibit notable performance fluctuations across varied scenarios. Furthermore, these classical algorithms often necessitate intricate feature engineering and lack adaptability.

In recent times, CNNs have garnered considerable traction in the realm of computer vision [21], eclipsing traditional machine-learning approaches in terms of classification and detection prowess. Shen et al. [22] proposed a robust data augmentation technique adaptable to any deep-learning paradigm for skin lesion classification. Their EfficientNetb0 methodology outperformed its counterparts on the ISIC2018 dataset, boasting an accuracy rate of 85.3%. Huang et al. [23] conducted a comparative study of multiple models, affirming that DenseNet [24] exhibited superior performance in benign and malignant binary classification tasks, while EfficientNet [25] excelled in multi-classification endeavors. Liu et al. [26] introduced the multi-level relationship capture network (MRCN), employing a region correlation learning module to emulate interrelations among distinct significant regions within the central lesion zone. Moreover, a cross-image learning module was employed to emulate profound semantic correlations across multiple images. Rigorous experiments across three arduous datasets validated the exceptional performance of MRCN. Tahir et al. [27] introduced DSCCNet, a deep learning architecture tailored for multi-classification diagnosis of skin cancer using dermoscopic imagery. Outperforming the baseline model, DSCCNet furnishes robust diagnostic assistance to dermatologists and medical practitioners. Nonetheless, CNNs confront certain constraints in skin lesion recognition, such as discerning analogous features, information loss, and constraints in capturing holistic, contextual information. While refinements in deep learning models and alternative methodologies exist to bolster recognition efficacy, sustained inquiry and enhancements are imperative to fortify accuracy and resilience. Consequently, sustained endeavors and exploration remain requisite in the realm of skin lesion identification.

Xie et al. [28] propounded Swin-SimAM, a melanoma detection approach amalgamating SwinViT for feature extraction with the parameter-free attention module SimAM. Demonstrating commendable performance, their method attained an impressive AUC performance of 90% in discriminating melanoma from non-melanoma entities, encompassing nevi and seborrheic keratoses. Eskandari et al. [29] devised a skin lesion segmentation framework predicated on a U-shaped hierarchical Transformer and an inter-scale context fusion (ISCF) methodology. This approach amalgamates each stage adaptively, harnessing attention correlation within the encoder at each juncture. Empirical validations underscored the robust applicability and efficacy of the ISCF model within each stage's context. Khan et al. [30] unveiled the SKINVIT model, founded on Outlook and Transformer architectures. This model adeptly captures both fine-grained and global features to bolster the accuracy of melanoma and non-melanoma classification. Validation across three datasets yielded promising results. However, notwithstanding their high accuracy in classification and detection endeavors, these models necessitate substantial computational resources and time, curtailing their real-time feasibility and scalability.

To address these challenges, this study introduces SkinSwinViT as a lightweight model. Building upon global attention and SwinViT, SkinSwinViT incorporates a local–global hierarchical attention mechanism to capture pivotal features across diverse abstraction tiers. By amalgamating local and global features, the model adeptly harnesses both granular details and contextual comprehension, thereby enriching feature representation and diagnostic accuracy, ultimately enhancing overall performance. Furthermore, SkinSwinViT is engineered to be lightweight, demanding fewer computational resources and time while offering enhanced real-time performance and scalability. The primary objective of this study is to proffer a high-precision skin lesion classification model proficient in automatically and accurately identifying seven types of skin maladies. This model aspires to mitigate skin cancer mortality, alleviate the burden on dermatologists, and narrow the accuracy chasm in early-stage lesion diagnosis. Through refined feature representation capabilities and heightened accuracy, SkinSwinViT furnishes dependable support for the identification and treatment of incipient skin lesions, thereby endowing dermatologists with more reliable diagnostic assistance.

3. Materials and Methods

This section elucidates the SkinSwinViT methodology and presents the dermoscopic image dataset utilized for the classification of seven distinct lesion types.

3.1. Dataset

A renowned dataset sourced from ISIC2018 [31] serves as the foundation for evaluating the proposed model. The ISIC2018 Task 3 dataset encompasses 7 distinct categories: melanoma (MEL), melanocytic nevus (NV), basal cell carcinoma (BCC), actinic keratosis (AKIEC), benign keratosis (BKL), dermatofibroma (DF), and vasculopathy (VASC). Comprising a total of 10,015 samples, the dataset exhibits an inherent imbalance in the distribution across various skin lesion categories. Notably, the images maintain a uniform color depth of 24 bits and possess dimensions of 600×450 pixels. The dataset distribution is delineated as follows: MEL (1113 samples), NV (6705 samples), BCC (514 samples), AKIEC (327 samples), BKL (1099 samples), DF (115 samples), and VASC (142 samples). Refer to Figure 1 for illustrative samples showcasing different manifestations of skin lesions.

In this experiment, the ISIC2018 dataset undergoes partitioning into distinct training and testing subsets, adhering to an 80% allocation for training and a 20% allocation for testing, as shown in Table 1.

Table 1. Distribution of samples on ISIC2018 dataset.

Category	NV	MEL	BKL	BCC	AKIEC	VASC	DF	Total
Training set	5364	890	880	411	262	113	92	8012
Testing set	1341	223	219	103	65	29	23	2003

3.2. Data Augmentation

Considering that the imbalance of different categories of samples has a significant impact on model performance, this study performs data augmentation processing on the existing dataset. This study performs several different transformations, such as random rotation within 180° , horizontal or vertical translation, random scaling, and vertical or horizontal flip operations, on the training set samples. An example of a data augmentation image is shown in Figure 2.

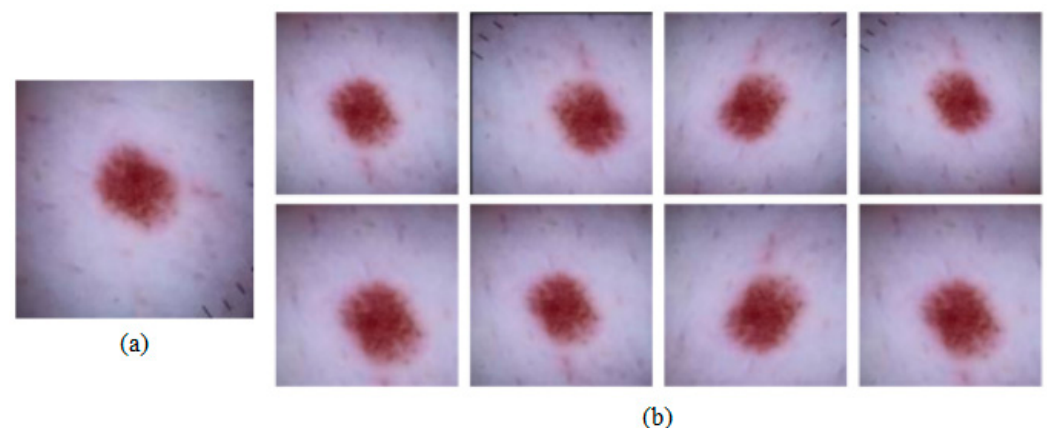


Figure 2. Example of data-enhanced image: (a) Original image in the basic training set; (b) Image after data enhancement.

Data augmentation was implemented on the remaining six categories of samples, excluding melanocytic nevus (NV) samples, thereby mitigating the risk of overfitting attributed to inadequate sample sizes and enhancing the model's generalization capacity. The quantity of data pertaining to melanocytic nevus remained consistent, comprising

5364 images. The resulting training dataset size subsequent to data augmentation is presented in Table 2.

Table 2. Sample distribution of the training set after data enhancement.

Category	NV	MEL	BKL	BCC	AKIEC	VASC	DF	Total
Training set	5364	5226	5188	5129	5186	5083	5029	36205

Nevertheless, the images within the dataset exhibit variability in size, necessitating their conversion to a uniform size to align with the input requirements of the deep learning model. Accordingly, the images are uniformly scaled to dimensions of 244×244 and subsequently normalized.

3.3. Proposed SkinSwinViT Architecture

The SkinSwinViT model, as proposed in this study, is designed for the purpose of identifying skin disorders classified into seven distinct categories. Figure 3 presents the architectural design of the SkinSwinViT model. Drawing inspiration from SwinViT [32], the model primarily integrates components including data preprocessing, SwinTransformer Block, Global Attention Block, and Classifier Block.

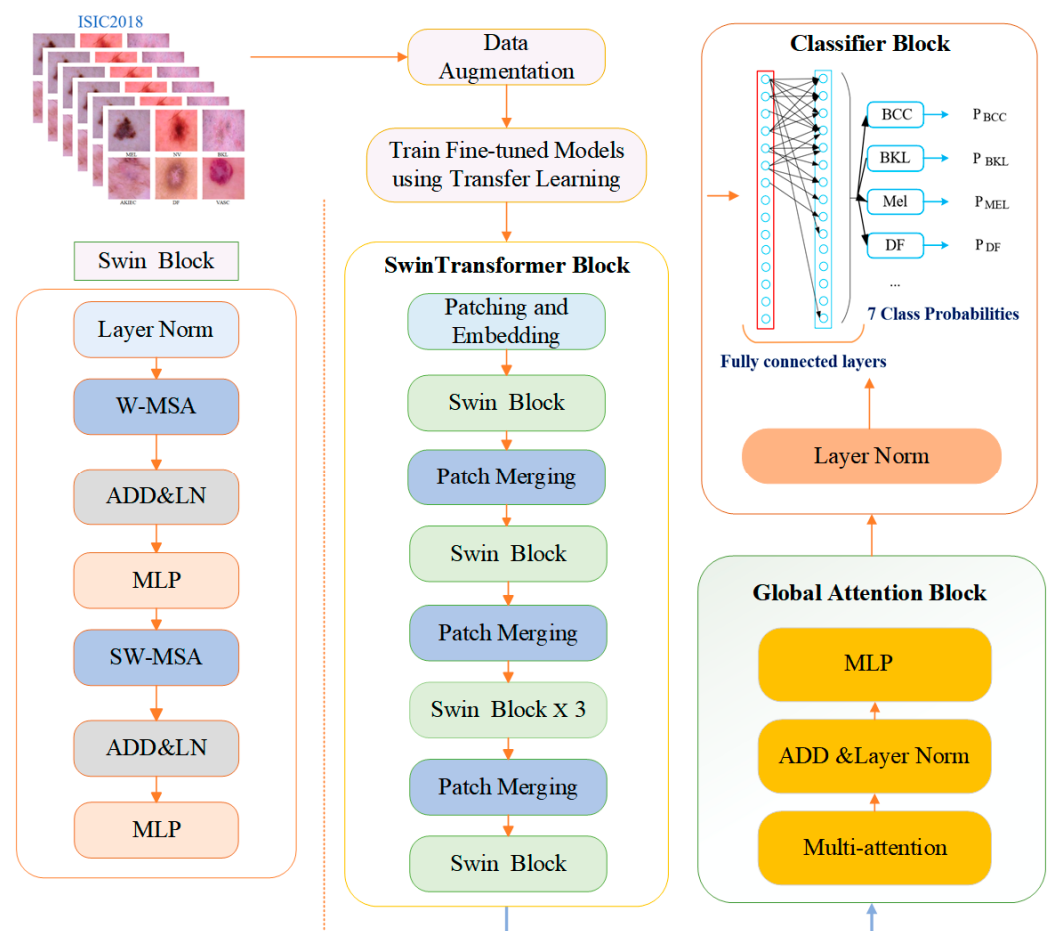


Figure 3. The architecture of SkinSwinViT.

To address the imbalance within the skin lesion dataset during data preprocessing, augmentation techniques are employed to enrich the segmented training set. Subsequently, the parameters of the pre-trained model serve as initial values. Following this, the proposed methodology utilizes a two-layer encoder, comprising SwinTransformer Block and Global Attention Block, to extract nuanced features across various dimensional spaces. Finally, the

proposed method employs a Classifier Block to ascertain the categorical outcome of skin lesion images. The detailed framework is expounded below.

1. SwinTransformer Block: Derived from the SwinViT model, the SwinTransformer Block initially conducts patching to segment skin lesion images into smaller patches, facilitating localized processing and computational efficiency. Embedding preserves spatial relationships among patches, facilitating comprehensive feature capture. Subsequently, Windowed Multi-Head Self-Attention (W-MSA) and Shifted Windowed Multi-Head Self-Attention (SW-MSA) are employed to capture local features and inter-patch relationships. These attention mechanisms enable contextual comprehension of local information, enhancing feature extraction efficacy. By integrating W-MSA and SW-MSA, contextual dependencies between neighboring patches are effectively captured, bolstering feature extraction. Moreover, employing multiple Swin Blocks within the SwinTransformer Block facilitates the learning of complex, abstract-level features. Each Swin Block comprises layers of W-MSA and SW-MSA. Additionally, patch merging reduces image resolution and integrates multi-scale information, enhancing overall comprehension.

Specifically, the Patching and Embedding layer divides the input image $x \in R^{H \times W \times 3}$ into non-overlapping patches of size 4×4 , mapping dimensions to C dimensions to produce an embedded image $x' \in R^{H/4 \times W/4 \times C}$. Subsequently, x' is normalized using Layer Normalization and forwarded to the Swin Block for feature extraction. The SwinTransformer Block consists of four stages, with patch merging operations at the conclusion of the first three stages to alter input feature dimensions. [1, 1, 3, 1] Swin Blocks are utilized across four stages, with channel counts per stage as [C, 2C, 4C, 8C]. The attention mechanism within each Swin Block is detailed as follows:

$$z'_l = F_{W-MSA}(LN(z'_{l-1})) + z'_{l-1} \quad (1)$$

$$z_l = MLP(LN(z'_l)) + z'_l \quad (2)$$

$$z'_{l+1} = F_{SW-MSA}(LN(z_l)) + z_l \quad (3)$$

$$z_{l+1} = MLP(LN(z'_{l+1})) + z'_{l+1} \quad (4)$$

$$z = F_{STB}(x') \quad (5)$$

where z'_l and z_l represent output features of W-MSA and Multi-Layer Perceptron (MLP) modules, respectively. F_{W-MSA} and F_{SW-MSA} denote functions of W-MSA and SW-MSA, respectively. F_{STB} represents the transformation function of the SwinTransformer Block.

2. Global Attention Block: Comprising multi-head self-attention mechanisms, layer normalization, and MLP layers, the Global Attention Block integrates spatial and semantic information. A multi-head self-attention mechanism fuses global information, followed by residual connections and layer normalization for feature representation learning. Subsequently, an MLP layer, comprising two linear layers and a nonlinear activation function (e.g., GELU), integrates feature information from different locations to generate a comprehensive global feature representation. This block enables global contextual comprehension through self-attention calculations on all feature vectors. The algorithm for the global attention mechanism is articulated as follows:

$$Att(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (6)$$

$$\tilde{x} = W(G_{con}Att_i(z, z, z)) + b \quad (7)$$

$$\hat{x} = G_{GlobalAtt}(z) = MLP\left(LN\left(\tilde{x} + z\right)\right) \quad (8)$$

Here, $Att(Q, K, V)$ represents the multi-head self-attention mechanism, Q, K , and V are the query, key, and value matrices, respectively. $\sqrt{d}$ is the dimensionality of the key vectors and $\sqrt{d}$ is used to scale the dot product to prevent gradient vanishing or exploding.

The Softmax function converts the dot product results into a probability distribution. G_{con} signifies the concatenation of multi-head attention mechanisms. W is a weight matrix, and b is a bias term. $G_{GlobalAtt}$ denotes the transformation function of the Global Attention Block. $\tilde{x} + z$ denotes the residual connection. LN stands for Layer Normalization, which is used to stabilize the training process.

3. Classifier Block: Incorporated within the model decoder, the Classifier Block commences with layer normalization to stabilize data distribution and expedite training. Following normalization, the feature vector undergoes dimensionality reduction via adaptive average pooling to compress feature vector length from sequence length to a fixed length 1. This compression retains essential features, facilitating efficient data representation. The data then progresses to a linear transformation layer to learn input-output linear relationships. Ultimately, the linear layer output is processed through a Softmax function, yielding a probability distribution for final classification. This distribution reflects model confidence or likelihood for each class, enabling classification based on the highest probability class.

$$y' = \text{Softmax}(w_c LN(\hat{x}) + b) \quad (9)$$

4. Loss Function: The model's classification loss function utilizes cross-entropy loss, quantifying disparities between model-predicted probability distributions and true label distributions. The loss function is expressed as follows:

$$L = - \sum_{i=1}^M \sum_{j=1}^N y'_j * \log(P_{y'_j}) \quad (10)$$

Here, L denotes the loss value, y'_j represents category j value in the true label (utilizing one-hot encoding), $P_{y'_j}$ symbolizes model-predicted probability value for category j , N represents category count, and M represents sample count. The summation symbol $\sum$ accumulates all categories.

3.4. Transfer Learning

In conventional machine learning paradigms, the assumption often prevails that the feature spaces of both training and test datasets remain identical. However, starting the training and testing processes from scratch can be highly time-consuming and resource-intensive [33]. To alleviate this predicament, transfer learning (TL) has emerged as a viable strategy, aiming to achieve precise classification outcomes even with constrained training data. Transfer learning involves leveraging knowledge from one model to improve the performance of another model in a target domain. Its primary objective is to enhance efficiency in the target domain. This approach proves particularly effective when dealing with relatively small target domain datasets, as it can leverage datasets from related source domains [34]. By utilizing transfer learning, this study can address such situations more robustly.

Figure 4 illustrates the TL workflow employed in this study, where original deep models trained on ImageNet are fine-tuned on target datasets like ISIC2018. Pre-trained models (AlexNet, VGG-11, GoogleNet, ResNet50, ViT, SwinViT, and SkinSwinViT) leverage knowledge from ImageNet, and their weights are optimized on comprehensive datasets to enhance feature extraction. Additionally, the prediction layer (fully connected layer) is modified in this paper to accommodate the 7 target categories for training.

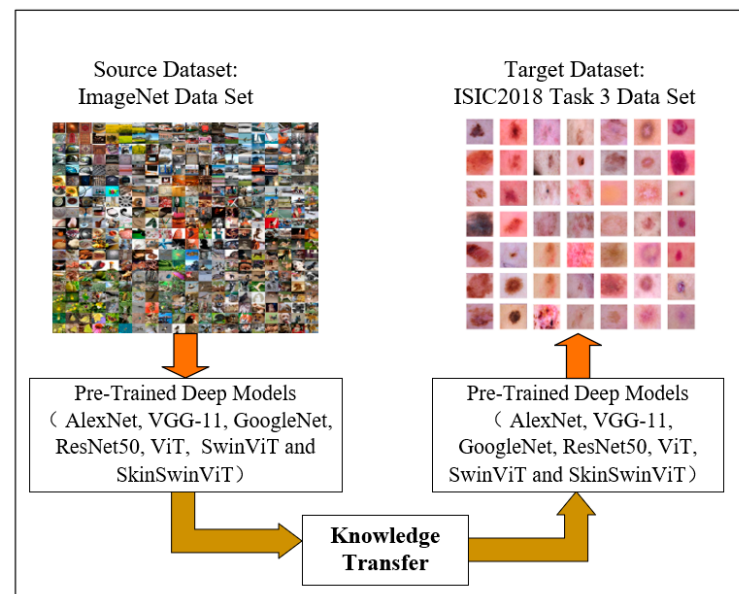


Figure 4. TL-based fine-tuned model training.

3.5. Performance Metrics

To evaluate the performance of individual models such as the proposed SkinSwinViT, this study considers the following performance metrics.

To measure the performance indicators of multi-classification, after data enhancement technology processing, the number of pictures in each category is the same, and there is no imbalance in the number of samples, so this paper adapts the Macro-average method. In multiclass classification problems, Macro-average calculates the indicators of each category and then averages the indicators of all categories. This is equivalent to setting the weights of all categories to be consistent. The specific formula is as follows:

$$Accuracy = \frac{1}{n} \sum_{i=1}^n Accuracy_i = \frac{1}{n} \sum_{i=1}^n \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \quad (11)$$

$$Recall = \frac{1}{n} \sum_{i=1}^n Recall_i = \frac{1}{n} \sum_{i=1}^n \frac{TP_i}{TP_i + FN_i} \quad (12)$$

$$Precision = \frac{1}{n} \sum_{i=1}^n Precision_i = \frac{1}{n} \sum_{i=1}^n \frac{TP_i}{TP_i + FP_i} \quad (13)$$

$$Specificity = \frac{1}{n} \sum_{i=1}^n Specificity_i = \frac{1}{n} \sum_{i=1}^n \frac{TN_i}{TN_i + FP_i} \quad (14)$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (15)$$

$Accuracy_i$ represents the accuracy of prediction of a certain category among the seven categories of skin diseases. i represents one of the seven types of skin diseases. The same applies to other indicators, where n represents the number of classification categories. In this article, $n = 7$, and the summation symbol $\sum$ is used to accumulate all categories. The indicators referred to in this study, by default, refer to the above-average indicators.

TP_i (True Positive): Refers to the number of instances that are correctly predicted to be a certain type of skin disease. FN_i (False Negative): Refers to the number of instances that are incorrectly predicted to be not a certain type of skin disease. FP_i (False Positive): Refers to the number of non-skin disease instances that are incorrectly predicted to be a certain type of skin disease. TN_i (True Negative): Refers to the number of cases of

non-skin diseases that are correctly predicted to be non-skin diseases. Recall quantifies the frequency with which a classifier accurately predicts a positive outcome among all samples that should have been classified as positive. High precision indicates the test's ability to precisely identify positive samples, thereby mitigating the occurrence of false positives. Elevated specificity signifies the accurate exclusion of negative samples and a diminished risk of misdiagnosis. The F1 score, as the harmonic mean of precision and recall, offers a comprehensive assessment of a classifier's accuracy in predicting positive instances.

This study employs the ROC (Receiver Operating Characteristic) curve and AUC (Area Under the Curve) value for separate analyses of each category. The ROC curve plots the False Positive Rate (FPR) on the x-axis and sensitivity on the y-axis. Sensitivity measures the correct identification of positive examples, while the false positive rate represents the incorrect identification of negative examples. The formula for calculating the false positive rate is as follows:

$$FPR = \frac{FP_i}{TN_i + FP_i} \quad (16)$$

The AUC is a metric that measures the overall performance of a classification model by calculating the area under the ROC curve. A higher AUC indicates better classification effectiveness for the model. The AUC value typically falls between 0.5 and 1.0, where 0.5 represents a random classifier, and 1.0 represents a perfect classifier.

4. Experimental Results and Analysis

4.1. Experimental Setup

The proposed SkinSwinViT framework is instantiated within the Anaconda environment, utilizing Python 3.9, with the installation of essential libraries such as Pytorch, Scikit-Learn, Matplotlib, and Numpy on the Linux operating system. The system configuration encompasses an Intel Platinum-8350C processor operating at 2.6 GHz, complemented by 32 GB DDR4 RAM (4 modules) and four NVIDIA RTX 3090 graphics processing units. Training of the SkinSwinViT framework is conducted on the ISIC2018 dataset, augmented as outlined in the data enhancement section to forestall overfitting, bolster classifier efficiency for unobserved images, and mitigate sample imbalance. Furthermore, three optimization strategies, namely Adam, AdamW, and SGD, are employed, with the optimal strategy selected to update SkinSwinViT parameters during training iterations. Concluding specifications entail an epoch count set to 100, with batch size values explored within the range [4, 8, 16, 32], with superior performance observed when the batch size is set to 4 and a default learning rate of 0.00001.

4.2. Experimental Results

In this section, the comprehensive performance comparison of the proposed SkinSwinViT model is carried out and compared with other CNN methods in terms of accuracy measures. The study evaluated the performance of various models, such as AlexNet, VGG-11, GoogleNet, ResNet50, ViT, SwinViT, and our SkinSwinViT model.

Table 3 illustrates that the proposed SkinSwinViT model exhibits superior performance on the dataset, achieving an accuracy of 0.9788, representing a 3.4% improvement over the second-best model, SwinViT, which attained an accuracy of 0.9444. This notable enhancement underscores the efficacy of SkinSwinViT in the classification of skin diseases, evidencing its aptitude in such tasks. Moreover, SkinSwinViT attains the highest recall rate at 0.9775, surpassing SwinViT's rate of 0.9391, indicative of its superior ability to identify positive samples and heightened sensitivity in disease detection. Notably, SkinSwinViT demonstrates commendable precision, specificity, and F1 score, registering values of 0.9783, 0.9936, and 0.9779, respectively. The elevated precision, specificity, and F1 score affirm the model's proficiency in accurately discerning positive samples within skin disease classification, reflecting a well-balanced performance. Despite a marginal increase in parameter count compared to ResNet50 and SwinViT, SkinSwinViT exhibits exceptional performance and notable generalization capabilities, effectively mitigating this numerical

discrepancy. Its remarkable adeptness in image comprehension and feature extraction is complemented by commendable generalizability. With a parameter size of 31 million, SkinSwinViT emerges as a lightweight model with considerable efficacy.

Table 3. Performance comparison between the training set SkinSwinViT and other models.

Model	Accuracy	Recall	Precision	Specificity	F1 Score	Params(M)
AlexNet	0.8953	0.8860	0.8897	0.9253	0.8877	61
GoogLeNet	0.8943	0.8939	0.8940	0.9236	0.8939	200
VGG-11	0.9301	0.9238	0.9266	0.9661	0.9251	133
ResNet50	0.9243	0.9191	0.9215	0.9618	0.9202	25
ViT	0.8031	0.7876	0.7934	0.8379	0.7899	86
SwinViT	0.9444	0.9391	0.9418	0.9612	0.9404	27
Our SkinSwinViT	0.9788	0.9775	0.9783	0.9936	0.9779	31

From Figure 5, it is apparent that with an increasing number of iterations, the model's accuracy on the training set exhibits a gradual ascent followed by stabilization, concurrently with a gradual decline and stabilization in the loss function. This trend indicates an augmentation in the model's learning capacity, leading to improved fitting to the training data. However, an exclusive focus on the model's performance solely on the training set may precipitate overfitting issues. Overfitting delineates a scenario wherein a model performs well on the training set but performs poorly when presented with unseen data due to its excessive complexity or propensity to glean excessive information from noise and intricacies within the training set, consequently resulting in diminished generalization performance.

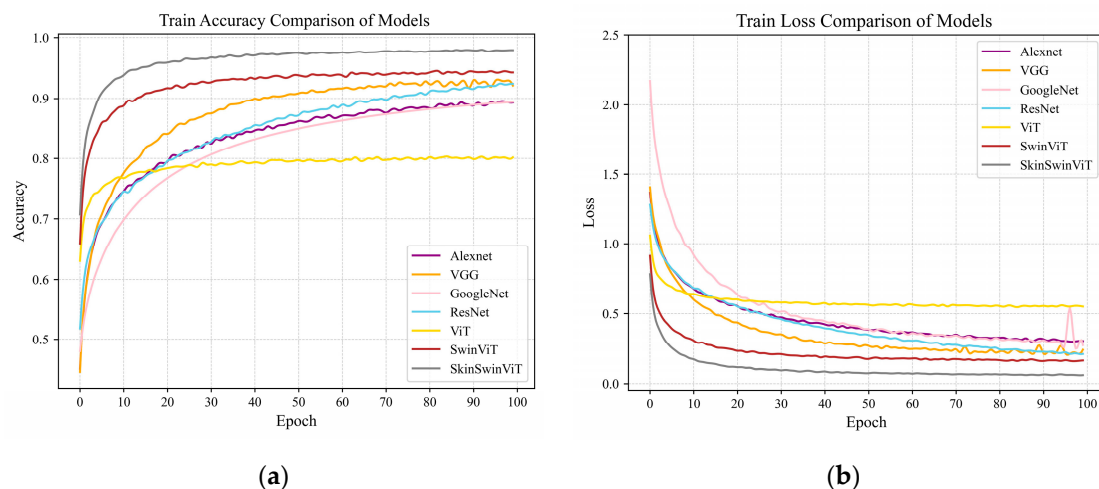


Figure 5. Comparison of train Accuracy and Loss results of multiple models in the training set. (a) Comparison of train Accuracy results of multiple models; (b) Comparison of train Loss results of multiple models.

Figure 6 delineates a discernible trend wherein, with an increasing number of iterations, the model's accuracy on the testing set exhibits a gradual augmentation followed by stabilization concurrently with a gradual reduction and stabilization in the loss function. This observation signifies a progressive enhancement in the model's predictive prowess and generalization capabilities concerning unfamiliar data during the testing process. Additionally, comparative analysis of various metrics on the testing set reveals improved performance in contrast to the training set, thereby indicating the model's superior performance on the testing set while mitigating overfitting. Notably, as the testing loss diminishes, the model's alignment with the testing data augments. Furthermore, with the model's improvement, the testing loss demonstrates a decline, underscoring the model's heightened generalization performance.

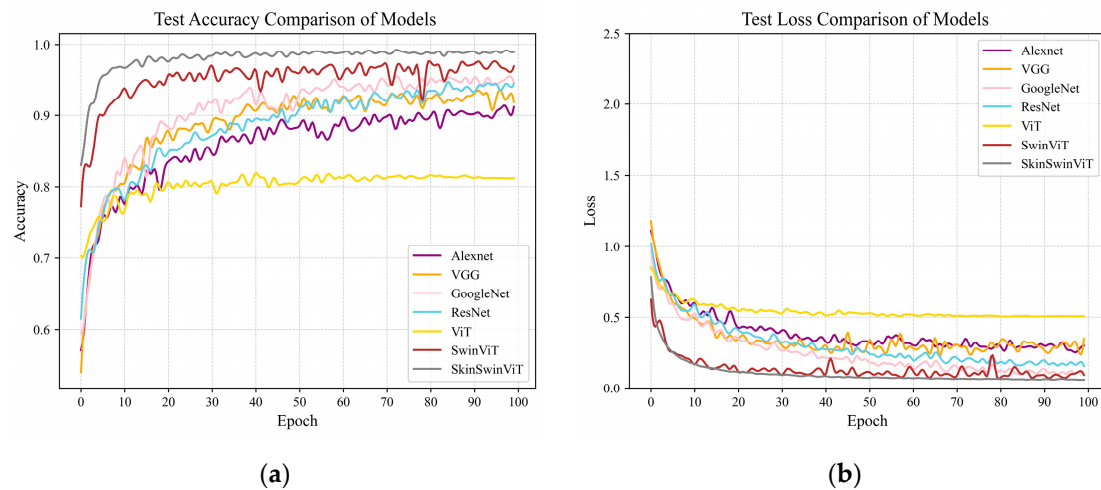


Figure 6. Comparison of test Accuracy and Loss results of multiple models in the testing set: (a) Comparison of test Accuracy results of multiple models; (b) Comparison of test Loss results of multiple models.

Table 4 unequivocally demonstrates the superior performance of the SkinSwinViT model over other models on the testing set, attaining exemplary accuracy of 0.9906, recall of 0.9906, precision of 0.9916, specificity of 0.9995, and F1 scores of 0.9910. Notably, SkinSwinViT exhibits a notable improvement of more than 1.55% across all metrics compared to SwinViT. This discernible superiority underscores the robust classification and generalization capabilities inherent in the SkinSwinViT model.

Table 4. Comparison between testing set SkinSwinViT and other models.

Model	Accuracy	Recall	Precision	Specificity	F1 Score
AlexNet	0.9095	0.9072	0.9018	0.9330	0.9036
GoogLeNet	0.9540	0.9540	0.9543	0.9738	0.9541
VGG-11	0.9354	0.9331	0.9295	0.9596	0.9307
ResNet50	0.9456	0.9418	0.9430	0.9634	0.9421
ViT	0.8156	0.8027	0.8021	0.8513	0.7998
SwinViT	0.9751	0.9752	0.9738	0.9846	0.9744
Our SkinSwinViT	0.9906	0.9906	0.9916	0.9995	0.9910

In summary, the SkinSwinViT model proposed within this study emerges as the frontrunner on the evaluated dataset, boasting high-performance metrics. Moreover, the model parameters are judiciously configured, thereby consuming fewer resources. This exemplary performance underscores the model's efficacy in skin disease classification tasks, positioning it as a promising candidate for an effective CAD system for skin disease diagnosis. The experimental findings furnish compelling evidence in support of further research and application in the realms of dermatological diagnosis and treatment.

4.3. SkinSwinViT Performance Analysis

This section presents the performance analysis of the proposed SkinSwinViT model on the considered dataset. Recognizing that accuracy alone does not provide a comprehensive assessment of a model's performance, a confusion matrix was generated to elucidate the model's performance across multiple classes.

Figure 7 presents the confusion matrix comparing SkinSwinViT and SwinViT. It highlights SkinSwinViT's precise performance across the majority of categories, effectively classifying a substantial portion of the samples. Remarkably, the integration of the global attention module significantly enhances classification accuracy while mitigating misclassification. This module facilitates the model in discerning and optimizing pivotal features,

thereby improving differentiation between categories. Despite occasional misclassifications observed in the NV, MEL, and BKL classes, possibly attributable to class similarity or inherent noise within the samples, the overall high performance of SkinSwinViT remains evident.

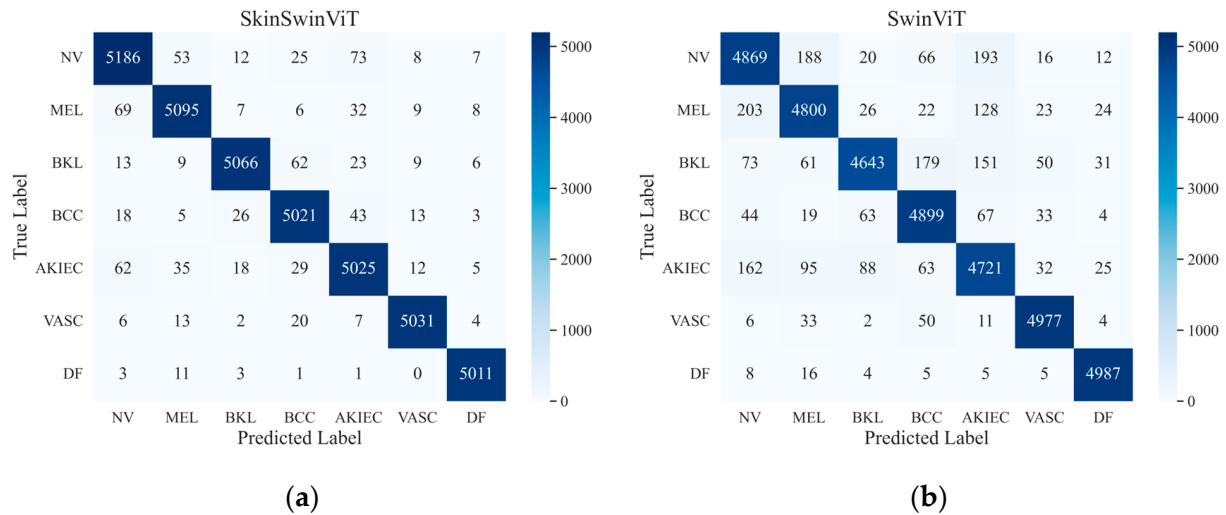


Figure 7. Confusion matrix of SkinSwinViT and SwinViT in the training set. (a) Confusion matrix of SkinSwinViT; (b) Confusion matrix of SwinViT.

Table 5 highlights the notable performance of the SkinSwinViT model, particularly evident in the DF category, where it demonstrates exceptionally high accuracy. Furthermore, the SkinSwinViT model consistently surpasses the SwinViT model across all individual categories, underscoring its enhanced classification capabilities. Specifically, it exhibits superior accuracy in distinguishing NV, MEL, BKL, and other sample types.

Table 5. Performance of model on various skin diseases in the training set.

Model	Metrics	NV	MEL	BKL	BCC	AKIEC	VASC	DF
SwinViT	Accuracy	0.9078	0.9189	0.8951	0.9556	0.9108	0.9793	0.9909
	Recall	0.9078	0.9188	0.8950	0.9553	0.9105	0.9798	0.9909
	Precision	0.9078	0.9315	0.9122	0.9405	0.9133	0.9758	0.9869
	Specificity	0.9628	0.9649	0.9637	0.9616	0.9633	0.9761	0.9783
	F1 Score	0.9078	0.9251	0.9035	0.9478	0.9119	0.9776	0.9889
SkinSwinViT	Accuracy	0.9672	0.9726	0.9544	0.9847	0.9688	0.9918	0.9952
	Recall	0.9673	0.9728	0.9543	0.9840	0.9686	0.9913	0.9952
	Precision	0.9697	0.9724	0.9646	0.9798	0.9686	0.9907	0.9952
	Specificity	0.9939	0.9949	0.9973	0.9971	0.9941	0.9985	0.9991
	F1 Score	0.9684	0.9725	0.9595	0.9823	0.9686	0.9913	0.9952

Figure 8 presents the ROC curve of SkinSwinViT and SwinViT. A comparison between the two models reveals significant improvements in the AUC values across the majority of categories. This noteworthy finding substantiates the efficacy of integrating global attention into the model, leading to enhanced classification performance within each category. The remarkable achievements of the SkinSwinViT model in the multiclass classification of skin diseases bear substantial importance in aiding the diagnosis and classification of such conditions.

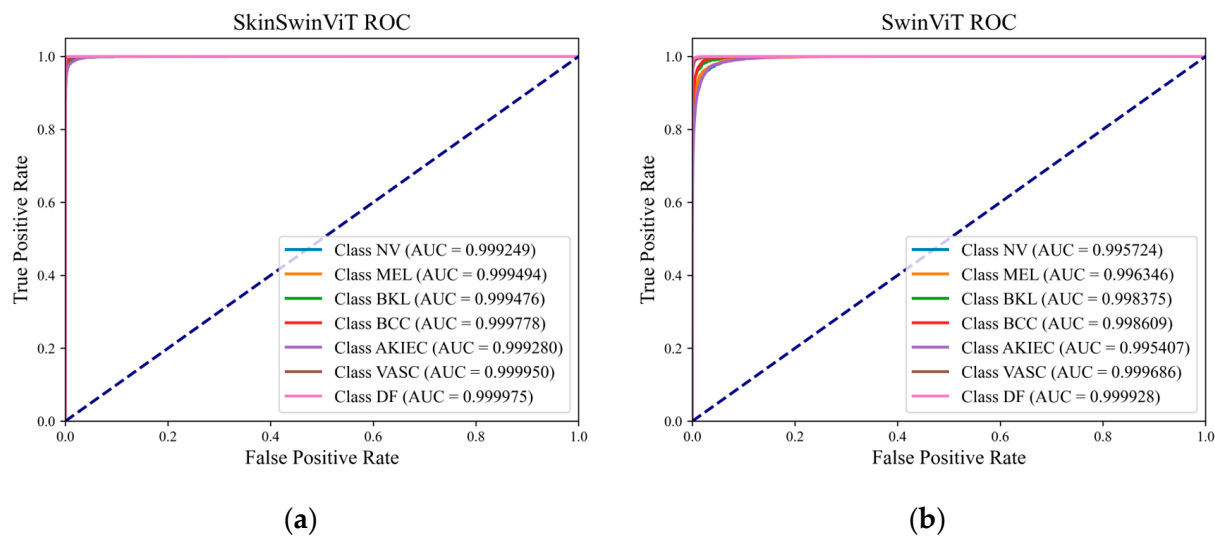


Figure 8. ROC curve of SkinSwinViT and SwinViT in the training set. (a) ROC curve of SkinSwinViT; (b) ROC curve of SwinViT.

4.4. Ablations

The ablation analysis comprises several facets: (1) Pre-training Model Impact on SkinSwinViT; (2) training the model with different optimizers to determine the optimal value; (3) the impact of augmented and unaugmented data on the proposed method; (4) the effect of global attention mechanism on the proposed method.

1. Pre-training Model Impact on SkinSwinViT: Table 6 illustrates the performance comparison across three distinct configurations of SkinSwinViT: SkinSwinViT_A, devoid of pre-training; SkinSwinViT_L, integrating ImageNet pre-training and training all layers; and SkinSwinViT, incorporating ImageNet pre-training with only the fully connected layer trained.

Table 6. Performance Evaluation of Pre-trained Models.

Measures	Accuracy	Recall	Precision	Specificity	F1 Score
SkinSwinViT_A	0.8621	0.8618	0.8601	0.8932	0.8609
SkinSwinViT_L	0.9763	0.9751	0.9761	0.9908	0.9756
SkinSwinViT	0.9786	0.9773	0.9781	0.9928	0.9776

The experimental findings reveal that SkinSwinViT_A achieves a mere 86.21% accuracy on the limited sample dataset, indicating subpar performance. Transformer-based models typically demand extensive data for effective training and exhibit diminished performance in small-scale tasks. To address this, and inspired by SpotTune's approaches, this study freezes the feature extraction layer and exclusively trains the fully connected layers [35]. Notably, both methods yield comparable accuracy, with training solely the fully connected layer being more resource-efficient.

2. Optimizer Selection: The proposed model is trained with various optimizers to evaluate classification performance. Table 7 presents the classification performance of SkinSwinViT trained with various optimizers.

Table 7. Performance Evaluation under Different Optimizers.

Optimizer	Accuracy	Recall	Precision	Specificity	F1 Score
SGD	0.9543	0.9488	0.9501	0.9774	0.9494
Adam	0.9673	0.9646	0.9647	0.9862	0.9646
AdamW	0.9773	0.9763	0.9772	0.9933	0.9768

Following 100 epochs of training, the Adam optimizer outperforms SGD. Adaptive optimizers such as Adam are more conducive to Transformer models, as advocated by the Swin Transformer authors. AdamW, integrating Weight Decay, facilitates more rapid decay of parameters with excessive values.

3. Impact of Data Augmentation: Data augmentation is employed to rectify dataset imbalance. Table 8 showcases the performance of SkinSwinViT with and without augmentation.

Table 8. Performance of using data augmentation or not.

Dataset	Accuracy	Recall	Precision	Specificity	F1 Score
With Augmentation	0.9783	0.9774	0.9781	0.9931	0.9777
Without Augmentation	0.9466	0.9456	0.9462	0.9669	0.9460

The results demonstrate that the accuracy of SkinSwinViT with augmentation on the dataset increases by 3.17% compared to SkinSwinViT without augmentation. Additionally, this study reveals that horizontal flipping minimally affects the accuracy of image classification, possibly due to the dataset being captured from various angles during sampling.

4. The impact of global attention on the proposed SkinSwinViT: The study compares the performance of SkinSwinViT with and without the global attention module. Table 9 outlines the differences in performance metrics.

Table 9. Comparison of metrics on whether to use global attention.

SkinSwinViT	Accuracy	Recall	Precision	Specificity	F1 Score
With Global Attention	0.9788	0.9775	0.9783	0.9936	0.9779
Without Global Attention	0.9444	0.9391	0.9418	0.9612	0.9404

The integration of the global attention module into the SkinSwinViT model yields superior metrics on the dataset. The experimental results indicate that the global attention mechanism enables better modeling of global context features, enhancing the model's deep representation capabilities and enabling more accurate feature identification and understanding in images.

4.5. Comparison with State-of-the-Art Methods

This study undertakes a comparative evaluation of the accuracy of the proposed framework against state-of-the-art methodologies on the ISIC2018 Task 3 dataset. The pertinent outcomes are delineated in Table 10. Crucially, all enumerated SOTA methodologies utilize the identical experimental dataset, ensuring equitable and precise comparisons. Our proposed method attains a remarkable accuracy of 97.8% on the ISIC2018 dataset, surpassing the performance of alternative techniques in terms of accuracy (Acc), precision (Pre), and specificity (Spe). This enhancement underscores the superiority of our SkinSwinViT method over extant state-of-the-art technologies.

Table 10. Comparison of models trained on the ISIC2018 Task 3 dataset.

Technique	Acc (%)	Pre (%)	Spe (%)
MobileNet+handcrafted features [36]	92.4	92.1	90.0
ResNet+Inceptionv3 [37]	85.1	79.6	82.91
Hybrid Model+handcrafted features [38]	91.9	94.1	97.7
Deep learning and moth flame optimization [39]	90.6	--	--
A CNN-based pigmented framework [40]	91.5	--	--
A CNN and nature-inspired optimization algorithm [41]	91.7	92.4	--
Two-stream CNN framework [42]	96.5	--	--
BF ² SkNet model [43]	97.1	95.1	--
Proposed SkinSwinViT	97.8	97.8	99.3

5. Discussion

In this study, we propose a novel approach for skin lesion classification, termed SkinSwinViT, which integrates the Swin Transformer with a global attention mechanism to capture fine-grained local and global features within skin lesion images. Our method demonstrates outstanding performance on the ISIC2018 dataset, surpassing prior research efforts. We juxtapose our findings with existing literature data to comprehensively elucidate the progress achieved, as illustrated in Table 10.

Relative to alternative models for skin lesion classification, our proposed SkinSwinViT model exhibits superior advancement. Specifically, compared to the extended hybrid model + handcrafted feature model by Sharafudeen et al. (2023) [38], our SkinSwinViT model significantly enhances predictive accuracy, precision, and specificity while maintaining a more parsimonious architecture, with improvements of 5.9%, 3.7%, and 1.6%, respectively. Furthermore, compared to the outstanding BF² SkNet model by Ajmal et al. (2023) [43], our SkinSwinViT model demonstrates exceptional performance, with increases in accuracy and precision of 0.7% and 2.7%, respectively, under similar complexity conditions. These notable outcomes establish our SkinSwinViT model as the state-of-the-art method in the field of skin lesion classification. Moreover, although our research primarily focuses on demonstrating the effectiveness of the SkinSwinViT model in multiclass skin lesion classification, we believe its potential extends to other medical domains. The model's ability to capture nuanced local features and global contextual information facilitates accurate image classification across diverse diseases.

Nonetheless, our study has certain limitations. The attention mechanism may not fully capture global dependencies, necessitating further refinement and optimization. Furthermore, the model's robustness needs to be evaluated on larger, more complex medical image datasets to ensure generalizability. To effectively address these limitations and challenges, our future research will focus on expanding the classification task and improving the model's capacity to handle larger, more complex medical image datasets. Additionally, we will undertake a comprehensive exploration of cross-modal information fusion techniques and delve into model interpretability and visualization approaches to enhance physicians' trust and acceptance of the model's predictions. The primary objective of our research is to enhance the performance and effectiveness of the algorithm in facilitating accurate and reliable diagnoses by applying it to expansive and diverse datasets of skin lesions, encompassing a broader range of classifications. Moreover, we strive to develop an advanced diagnostic assistance system that embodies practicality, providing invaluable support in the field of dermatology diagnostics. These endeavors will advance the field of dermatology diagnostics, improving the model's acceptability and usability in clinical practice.

6. Conclusions

The classification of dermoscopic images of seven types of skin diseases was performed using a deep convolutional neural network model. By incorporating a layered Transformer architecture and window attention mechanism, the SwinViT model achieved improved prediction accuracy and computing efficiency. However, its ability to capture global dependencies is limited as it primarily focuses on local and adjacent windows' dependencies. In order to overcome this limitation and enhance prediction accuracy, this study proposes an improved lightweight model called SkinSwinViT that introduces a global–local attention mechanism to comprehensively consider information from other locations. Our main conclusion from this endeavor is as follows:

- (1) This study effectively addressed the issues of limited data volume and imbalanced samples by employing data augmentation techniques. The results unequivocally confirmed the effectiveness of data augmentation in improving the performance of the 7-class skin lesion SkinSwinViT recognition model.
- (2) The proposed local–global hierarchical attention mechanism effectively captures crucial features, enhancing feature representation and diagnostic accuracy.

- (3) Leveraging a Transformer-based encoder–decoder framework, this study improves the model’s representation and scalability while optimizing skin lesion recognition performance through pre-training and fine-tuning techniques.
- (4) The experimental results demonstrate the exceptional performance of the SkinSwinViT model in the image classification task of various skin diseases, surpassing SOTA methods in the field. Moreover, the model exhibits advantages such as low computational resource requirements and minimal time consumption, thereby validating the effectiveness of our proposed approach.

Significantly, the SkinSwinViT model not only demonstrates exceptional proficiency in the multiclass classification of skin disease lesions but also possesses the ability to model local features and global context, which can provide valuable assistance for image classification tasks in other medical fields and further promote the development of medical image analysis.

In summary, the model proposed in this study holds substantial promise for enhancing the effectiveness and applicability of skin disease image classification models, effectively alleviating the inherent risks of misdiagnosis and missed diagnosis while concurrently fostering improvements in treatment outcomes and the overall quality of life for patients. Furthermore, this model has the capacity to provide medical professionals with more reliable and accurate diagnostic assistance, driving progress in the field of dermatology.

Author Contributions: Data curation, K.T. and R.H.; Formal analysis, K.T. and Y.L.; Funding acquisition, J.S., R.C. and M.D.; Investigation, K.T. and R.H.; Methodology, K.T. and J.S.; Project administration, M.D.; Resources, K.T., J.S. and R.H.; Software, K.T. and J.S.; Supervision, J.S., R.C. and Y.L.; Validation, K.T., J.S., R.C., M.D. and Y.L.; Visualization, K.T.; Writing—original draft, K.T., J.S. and M.D.; Writing—review and editing, K.T., J.S., R.C. and M.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by a special grant from the program for scientific research start-up funds of Guangdong Ocean University under Grant No. 060302102303, Guangdong Basic and Applied Basic Research Foundation under Grant No. 2023A1515011326, program for scientific research start-up funds of Guangdong Ocean University under Grant No. 060302102101, Guangdong Provincial Science and Technology Innovation Strategy under Grant No. pdjh2023b0247, National College Students Innovation and Entrepreneurship Training Program under Grant No. 202310566022, and Guangdong Ocean University Undergraduate Innovation Team Project under Grant No. CXTD2023014.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used and analyzed during the current study are available from the corresponding author upon reasonable request. The ISIC2018 dataset used in this study is publicly available at <https://challenge.isic-archive.com/data> (accessed on 16 January 2024).

Conflicts of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CNN	Convolutional Neural Network
ISIC	International Skin Imaging Collaboration
CAD	Computer-Aided Diagnosis
SVM	Support Vector Machine
ViT	Vision Transformer
SwinViT	Swin Transformer
MEL	Melanoma
NV	Melanocytic Nevus
BCC	Basal Cell Carcinoma
AKIEC	Actinic Keratosis

BKL	Benign Keratosis
DF	Dermatofibroma
VASC	Vasculopathy
MRCN	Multi-level Relationship Capture Network
ISCF	Inter-scale context fusion
W-MSA	Windowed Multi-Head Self-Attention
SW-MSA	Shifted Windowed Multi-Head Self-Attention
MLP	Multi-Layer Perceptron
TL	Transfer Learning
TP	True Positive
FN	False Negative
FP	False Positive
TN	True Negative
ROC	Receiver Operating Characteristic
AUC	Area Under the Curve
FPR	False Positive Rate
M	Million
Accuracy	Acc
Precision	Pre
specificity	Spe

References

1. American Cancer Society. Available online: <https://www.cancer.org/cancer/types/melanoma-skin-cancer/about/key-statistics.html> (accessed on 3 March 2024).
2. WHO Newsroom Fact Sheet. Available online: <https://www.who.int/news-room/fact-sheets/detail/cancer> (accessed on 6 March 2024).
3. ISIC Challenge. Available online: <https://challenge.isic-archive.com/> (accessed on 6 March 2024).
4. Zhang, J.; Zhong, F.; He, K.; Ji, M.; Li, S.; Li, C. Recent Advancements and Perspectives in the Diagnosis of Skin Diseases Using Machine Learning and Deep Learning: A Review. *Diagnostics* **2023**, *13*, 3506. [\[CrossRef\]](#)
5. Yu, K.H.; Beam, A.L.; Kohane, I.S. Artificial intelligence in healthcare. *Nat. Biomed. Eng.* **2018**, *2*, 719–731. [\[CrossRef\]](#)
6. Wang, Z.; Luo, Y.; Xin, J.; Zhang, H.; Qu, L.; Wang, Z.; Yao, Y.; Zhu, W.; Wang, X. Computer-aided diagnosis based on extreme learning machine: A review. *IEEE Access* **2020**, *8*, 141657–141673. [\[CrossRef\]](#)
7. Agnes, S.A.; Anitha, J.; Solomon, A.A. Two-stage lung nodule detection framework using enhanced UNet and convolutional LSTM networks in CT images. *Comput. Biol. Med.* **2022**, *149*, 106059. [\[CrossRef\]](#)
8. Chattopadhyay, S.; Dey, A.; Singh, P.K.; Oliva, D.; Cuevas, E.; Sarkar, R. MTRRE-Net: A deep learning model for detection of breast cancer from histopathological images. *Comput. Biol. Med.* **2022**, *150*, 106155. [\[CrossRef\]](#)
9. Abdolali, F.; Kapur, J.; Jaremko, J.L.; Noga, M.; Hareendranathan, A.R.; Punithakumar, K. Automated thyroid nodule detection from ultrasound imaging using deep convolutional neural networks. *Comput. Biol. Med.* **2020**, *122*, 103871. [\[CrossRef\]](#)
10. Kluk, J.; Ogiela, M.R. AI Approaches in Computer-Aided Diagnosis and Recognition of Neoplastic Changes in MRI Brain Images. *Appl. Sci.* **2022**, *12*, 11880. [\[CrossRef\]](#)
11. Xu, S.S.-D.; Chang, C.-C.; Su, C.-T.; Phu, P.Q. Classification of Liver Diseases Based on Ultrasound Image Texture Features. *Appl. Sci.* **2019**, *9*, 342. [\[CrossRef\]](#)
12. Kadhim, Y.A.; Khan, M.U.; Mishra, A. Deep learning-based computer-aided diagnosis (CAD): Applications for medical image datasets. *Sensors* **2022**, *22*, 8999. [\[CrossRef\]](#)
13. Rodriguez-Galiano, V.; Sanchez-Castillo, M.; Chica-Olmo, M.; Chica-Rivas, M. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geol. Rev.* **2015**, *71*, 804–818. [\[CrossRef\]](#)
14. Pan, H.; Pang, Z.; Wang, Y.; Wang, Y.; Chen, L. A new image recognition and classification method combining transfer learning algorithm and mobilenet model for welding defects. *IEEE Access* **2020**, *8*, 119951–119960. [\[CrossRef\]](#)
15. Arkin, E.; Yadikar, N.; Muhtar, Y.; Ubul, K. A Survey of Object Detection Based on CNN and Transformer. In Proceedings of the 2021 IEEE 2nd International Conference on Pattern Recognition and Machine Learning, Chengdu, China, 16–18 July 2021. [\[CrossRef\]](#)
16. Chatterjee, S.; Dey, D.; Munshi, S. Integration of morphological preprocessing and fractal-based feature extraction with recursive feature elimination for skin lesion types classification. *Comput. Methods Programs Biomed.* **2019**, *178*, 201–218. [\[CrossRef\]](#)
17. Tasoulis, S.K.; Doukas, C.N.; Maglogiannis, I. Skin lesions characterisation utilising clustering algorithms. In Proceedings of the 6th Hellenic Conference on AI, Athens, Greece, 4–7 May 2010. [\[CrossRef\]](#)

18. Dhivyaa, C.R.; Sangeetha, K.; Balamurugan, M.; Amaran, S.; Vetriselvi, T.; Johnpaul, P. Skin lesion classification using decision trees and random forest algorithms. *J. Ambient. Intell. Humaniz. Comput.* **2020**, *2020*, 1–13. [\[CrossRef\]](#)
19. Pham, T.C.; Tran, G.S.; Nghiem, T.P.; Doucet, A.; Luong, C.M.; Hoang, V.-D. A Comparative Study for Classification of Skin Cancer. In Proceedings of the 2019 International Conference on System Science and Engineering, Dong Hoi, Vietnam, 20–21 July 2019. [\[CrossRef\]](#)
20. Tschandl, P.; Codella, N.; Akay, B.N.; Argenziano, G.; Braun, R.P.; Cabo, H.; Gutman, D.; Halpern, A.; Helba, B.; Hofmann-Wellenhof, R.; et al. Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: An open, web-based, international, diagnostic study. *Lancet Oncol.* **2019**, *20*, 938–947. [\[CrossRef\]](#) [\[PubMed\]](#)
21. Cong, S.; Zhou, Y. A review of convolutional neural network architectures and their optimizations. *Artif. Intell. Rev.* **2022**, *56*, 1905–1969. [\[CrossRef\]](#)
22. Shen, S.; Xu, M.; Zhang, F.; Shao, P.; Liu, H.; Xu, L.; Zhang, C.; Liu, P.; Yao, P.; Xu, R.X. A Low-Cost High-Performance Data Augmentation for Deep Learning-Based Skin Lesion Classification. *Biomed. Eng. Front.* **2022**, *2022*, 9765307. Available online: <https://spj.science.org/doi/10.34133/2022/9765307> (accessed on 6 March 2024). [\[CrossRef\]](#)
23. Huang, H.W.; Hsu, B.W.Y.; Lee, C.H.; Tseng, V.S. Development of a light-weight deep learning model for cloud applications and remote diagnosis of skin cancers. *J. Dermatol.* **2021**, *48*, 310–316. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; Volume 1, pp. 2261–2269. [\[CrossRef\]](#)
25. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019. Available online: <https://proceedings.mlr.press/v97/tan19a.html> (accessed on 8 March 2024).
26. Liu, Z.; Xiong, R.; Jiang, T. Multi-level Relationship Capture Network for Automated Skin Lesion Recognition. In Proceedings of the Medical Image Computing and Computer Assisted Intervention, Strasbourg, France, 27 September–1 October 2021. [\[CrossRef\]](#)
27. Tahir, M.; Naeem, A.; Malik, H.; Tanveer, J.; Naqvi, R.A.; Lee, S.W. DSCC_Net: Multi-Classification Deep Learning Models for Diagnosing of Skin Cancer Using Dermoscopic Images. *Cancers* **2023**, *15*, 2179. [\[CrossRef\]](#)
28. Wang, Z.; Lu, H.; Jin, J.; Hu, K. Human Action Recognition Based on Improved Two-Stream Convolution Network. *Appl. Sci.* **2022**, *12*, 5784. [\[CrossRef\]](#)
29. Eskandari, S.; Lump, J.; Sanchez Giraldo, L. Skin Lesion Segmentation Improved by Transformer-Based Networks with Inter-scale Dependency Modeling. In Proceedings of the Machine Learning in Medical Imaging, Strasbourg, France, 27 September–1 October 2021. [\[CrossRef\]](#)
30. Khan, S.; Khan, A. SkinViT: A transformer based method for Melanoma and Nonmelanoma classification. *PLoS ONE* **2023**, *18*, e0295151. [\[CrossRef\]](#)
31. ISIC2018 Challenge Datasets. Available online: <https://challenge.isic-archive.com/data/#2018> (accessed on 9 March 2024).
32. Liu, Z.; Lin, Y.T.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, Montreal, Canada, 11–17 October 2021. [\[CrossRef\]](#)
33. Weiss, K.; Khoshgoftaar, T.M.; Wang, D. A survey of transfer learning. *J. Big Data* **2016**, *3*, 1–40. [\[CrossRef\]](#)
34. Akram, T.; Laurent, B.; Naqvi, S.R.; Alex, M.M.; Muhammad, N. A deep heterogeneous feature fusion approach for automatic land-use classification. *Inf. Sci.* **2018**, *467*, 199–218. [\[CrossRef\]](#)
35. Ge, W.; Yu, Y. Borrowing Treasures from the Wealthy: Deep Transfer Learning through Selective Joint Fine-Tuning. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017. [\[CrossRef\]](#)
36. Almaraz-Damian, J.-A.; Ponomaryov, V.; Sadovnychiy, S.; Castillejos-Fernandez, H. Melanoma and Nevus Skin Lesion Classification Using Handcraft and Deep Learning Feature Fusion via Mutual Information Measures. *Entropy* **2020**, *22*, 484. [\[CrossRef\]](#) [\[PubMed\]](#)
37. Shahin, A.H.; Kamal, A.; Elattar, M.A. Deep Ensemble Learning for Skin Lesion Classification from Dermoscopic Images. In Proceedings of the 2018 9th Cairo International Biomedical Engineering Conference, Cairo, Egypt, 15–17 December 2018. [\[CrossRef\]](#)
38. Sharafudeen, M.; S, V.C.S. Detecting skin lesions fusing handcrafted features in image network ensembles. *Multimed. Tools Appl.* **2023**, *82*, 3155–3175. [\[CrossRef\]](#)
39. Khan, M.A.; Sharif, M.; Akram, T.; Damaševičius, R.; Maskeliūnas, R. Skin Lesion Segmentation and Multiclass Classification Using Deep Learning Features and Improved Moth Flame Optimization. *Diagnostics* **2021**, *11*, 811. [\[CrossRef\]](#) [\[PubMed\]](#)
40. Sevli, O. A deep convolutional neural network-based pigmented skin lesion classification application and experts evaluation. *Neural Comput. Appl.* **2021**, *33*, 12039–12050. [\[CrossRef\]](#)
41. Arshad, M.; Khan, M.A.; Tariq, U.; Armghan, A.; Alenezi, F.; Javed, M.Y.; Aslam, S.M.; Kadry, S. A computer-aided diagnosis system using deep learning for multiclass skin lesion classification. *Comput. Intell. Neurosci.* **2021**, *2021*, 9619079. [\[CrossRef\]](#) [\[PubMed\]](#)

42. Khan, A.; Sharif, M.; Akram, T.; Kadry, S.; Hsu, C.H. A two-stream deep neural network-based intelligent system for complex skin cancer types classification. *Int. J. Intell. Syst.* **2022**, *37*, 10621–10649. [[CrossRef](#)]
43. Ajmal, M.; Khan, M.A.; Akram, T.; Alqahtani, A.; Alhaisoni, M.; Armghan, A.; Althubiti, S.A.; Alenezi, F. BF2SkNet: Best deep learning features fusion-assisted framework for multiclass skin lesion classification. *Neural Comput. Appl.* **2023**, *35*, 22115–22131. [[CrossRef](#)]

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.