Biomedical Signal Processing and Control 104 (2025) 107627

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

A novel CNN-ViT-based deep learning model for early skin cancer diagnosis

Ishak Pacal [a,c,*], Burhanettin Ozdemir [b], Javanshir Zeynalov [c], Huseyn Gasimov [c], Nurettin Pacal [d]

[a] Department of Computer Engineering, Faculty of Engineering, Igdir University, 76000, Igdir, Turkey
[b] Department of Operations and Project Management, College of Business, Alfaisal University, 11533 Riyadh, Saudi Arabia
[c] Department of Electronics and Information Technologies, Faculty of Architecture and Engineering, Nakhchivan State University, AZ 7012, Nakhchivan, Azerbaijan
[d] Department of Biology, Faculty of Arts and Sciences, Igdir University, 76000, Igdir, Turkey

ARTICLE INFO

Keywords:
Medical image analysis
Skin cancer detection
MetaFormers
Vision transformer
Focal self-attention

ABSTRACT

Skin cancer is a serious global health issue where early detection is crucial for effective treatment and improved patient outcomes. However, accurate diagnosis is challenging due to the variety of subtypes and imaging complexities. This study introduces an innovative deep learning model based on the MetaFormer architecture, optimized specifically for skin cancer. The Proposed Model features a hybrid design that replaces traditional self-attention methods with novel focal self-attention mechanisms, enhancing its ability to identify critical regions, reduce noise, and extract features more effectively, ultimately boosting diagnostic accuracy. To evaluate the model's generalization capabilities, it was tested on two benchmark datasets: ISIC 2019, which includes a diverse set of dermatological images across eight skin cancer classes, and HAM10000, widely used in dermatological research. The model achieved outstanding results, including an accuracy of 0.9254, precision of 0.9041, recall of 0.8768, and an F1-score of 0.8886 on ISIC 2019, and an accuracy of 0.9501, precision of 0.9470, recall of 0.9211, and an F1-score of 0.9334 on HAM10000. The Proposed Model surpasses existing methods in the field, outperforming ten advanced CNN models and twenty state-of-the-art ViT models under the same training and evaluation conditions. With a lightweight design of just 35.01 million parameters, it is optimized for real-time and mobile applications, making it highly practical for clinical use. Its reliable performance ensures accurate diagnoses, which are essential for early intervention and treatment, addressing a critical need in modern healthcare.

1. Introduction

The skin is the largest organ in the body that interacts with the environment outside and performs many functions [1]. The skin serves as the body's primary defense system, shielding internal organs from physical injuries, harmful microorganisms such as bacteria and viruses, and the damaging effects of ultra-Violette radiation [2]. Beyond its protective functions, the skin plays a critical role in maintaining homeostasis by preventing water loss, regulating body temperature, and facilitating the storage and synthesis of vitamin D [2,3]. Structurally, the skin comprises three distinct layers: the epidermis, dermis, and subcutis, each contributing to its complex functionality. However, disruptions in cellular processes, such as mutations leading to abnormal DNA coding, can result in uncontrolled cell proliferation and the development of skin cancer [4]. This malignancy underscores the importance of understanding the factors contributing to skin health and disease [5].

Skin cancer refers to the abnormal proliferation of the skin cells [3,4]. Although sun exposure is the leading cause of skin cancer, other factors that including genetic disposition, skin type, and environmental factors may be attributed to skin cancer [6]. An estimated 2,001,140 new cases of cancer are expected to occur in the United States in 2024, with 611,720 deaths from the disease predicted to occur. In contrast, an estimated 108,270 cases of skin cancer are also expected to happen in the US that year, with the estimated number of cases causing fatalities at 13,120 [7]. All these numbers bring to light a tremendous public health concern about skin cancer and the importance of early detection and treatment.

A physical examination of the skin diagnoses skin cancers, and if there are any doubtful areas, a biopsy can be performed [8]. In identifying and classifying skin lesions, imaging the skin for skin cancer plays

* Corresponding author.
E-mail addresses: ishak.pacal@igdir.edu.tr (I. Pacal), bozdemir@alfaisal.edu (B. Ozdemir), cavansirzeynalov@ndu.edu.az (J. Zeynalov), huseynqasimov@ndu.edu.az (H. Gasimov), nurettin.pacal@gdir.edu.tr (N. Pacal).

https://doi.org/10.1016/j.bspc.2025.107627
Received 19 July 2024; Received in revised form 23 November 2024; Accepted 24 January 2025
Available online 28 January 2025

a crucial role. The images have a lot of features and patterns, all of which are used in the analysis to differentiate between cancerous and non-cancerous lesions. Skin cancer treatment is done early in the same way treatment is given in other types of cancers. Early detection of cancer helps to detect the disease at earlier stages and start treatment [9,10]. Artificial intelligence and deep learning techniques have been employed in the early cancer diagnosis due to their capability to process large datasets and identify complex patterns [11]. These are applied in various areas, including defense, agriculture, natural language processing, and large language models, among others, which include early diagnostics of cancers [12].

When it comes to the artificial intelligence and deep learning techniques employed in the early cancer diagnosis, convolutional neural networks (CNNs) are the most widely applied deep learning technique for medical image processing and early cancer diagnosis [13]. Capabilities in image data processing and analysis have been noted with CNNs [14]. In medical diagnosis, such a model can be trained to detect cancerous tissues, cells, or lesions. For instance, it could be applied to detecting abnormalities on MRI or CT scans or classifying cancer cells from a histopathology image. The use of CNNs has excellent potential for tasks in the diagnosis of skin cancer, which includes the classification of skin lesions and the detection of cancerous lesions. Furthermore, the next generation of deep learning models, such as the Vision Transformer (ViT), has also presented great potential in diagnosing cancer in medicine [15]. The ViT model proposes something entirely different from traditional CNN-based methods for feature extraction from image data. These models can manage large-scale image data, identify complex patterns, and, to an increasing extent, be used on issues in medical imaging, such as skin cancer detection. In most cases, however, CNNs can afford better performance than ViTs. The earlier the deep learning technologies are used, the more effectively the presence of cancerous lesions can be detected in the case of skin cancers, which leads to more effective processes for the treatment [16].

There have been many surveys and review studies to analyze the application of deep learning algorithms in diagnosing skin cancer and the impact that deep learning makes in this area. Mazhar et al. [17] presented a review that methods based on different kinds of CNNs-based algorithms that the authors of this work discussed for the detection and classification of skin cancer are, in most cases, noninvasive ones and include such standard stages as pre-processing, picture segmentation, feature extraction, and classification. Their paper mostly centered on ANNs, CNNs, KNNs, and RBFNs for lesion image classification. All these algorithms have their own advantages and disadvantages, with CNNs surpassing the rest of the categories by a considerable margin because these network models represent an abridged form of computer vision. Mirikharaji et al. [18] conducted a study using 177 published research articles about deep learning-based skin lesion segmentation to enhance the automated diagnosis of skin cancers. Their review included data input, model design and evaluation aspects, trends discussion, and addressing limitations to allow comparisons. Shah et al. [19] have achieved early skin cancer detection based on different datasets and hybrid models of deep learning techniques. They highlighted an automated skin lesion recognition system to achieve better diagnosis efficiency and discussed its potential application in developing more efficient and accurate systems for detecting skin cancer. Bhatt et al. [19] compared several machine learning algorithms in classifying skin cancers and identified SVM, KNNs, and CNNs as the most yielding ones. They indicated that deep learning techniques are superior to earlier versions of the methods and put heavy stress on the sharing of technical details for results to be replicable. Melarkode et al. [20] proposed research toward providing comprehensive insight for the diagnosis of skin cancer. Zafar et al. [21] provided a systematic review of methods and approaches taken for the analysis of skin lesions, with an increased focus on the challenges of the complex features and the desire to identify the hurdles that will further assist research in the detection of skin cancer.

Consequently, in the current body of research that has been done on the application of deep learning and machine learning techniques in diagnosing skin cancer, much information has been obtained on which algorithms; ANNs, CNNs, and ViTs; work effectively. Most importantly, CNNs are very promising in achieving high accuracy and effectiveness in the early identification and categorization of skin lesions. However, these investigations have also highlighted the importance of dealing with challenges such as dataset variability, model complexity, and the implementation of automated systems that will help streamline the diagnostic process. Conquering these hurdles and pushing the envelope of research on this subject would bring an inch closer to the technology-driven battle against skin cancer, leading to better patient outcomes and healthcare practice. It has also been emphasized, within the literature, that there is very effective early detection of skin cancer with deep learning. This study attempts to overcome challenges like dataset variability and limitation, low performance due to data scarcity, model complexity, and implementation as an automated system.

The contributions of our proposed model to the literature on skin cancer diagnosis are briefly summarized as follows:

- The Proposed Model integrates focal self-attention, significantly enhancing its ability to concentrate on relevant regions within dermatological images. This approach reduces noise and improves feature extraction, leading to superior performance metrics across all compared to traditional self-attention mechanisms.
- By scaling the CAFormer architecture, the Proposed Model achieves a balanced trade-off between parameter count and performance. This careful optimization results in a robust model with 35.01 million parameters, demonstrating exceptional generalization capabilities and high diagnostic accuracy in the autonomous detection of skin cancer.
- The study conducts a thorough evaluation of 30 deep learning models, including 10 leading-edge CNNs and 20 state-of-the-art Vision Transformers, on the ISIC 2019 and HAM10000 datasets. This extensive comparative analysis provides a detailed understanding of each model's effectiveness, highlighting the strengths and limitations of both traditional CNNs and modern ViTs architectures.
- The proposed model establishes a new state-of-the-art benchmark for autonomous skin cancer diagnosis, achieving impressive metrics on the ISIC 2019 dataset (accuracy: 0.9254, precision: 0.9041, recall: 0.8768, F1-score: 0.8886) and the HAM10000 dataset (accuracy: 0.9501, precision: 0.9470, recall: 0.9211, F1-score: 0.9334). These results advance research in dermatological image analysis and related medical fields.

The manuscript has been structured to enhance readability and comprehension. The second section provides an in-depth discussion of the literature related to deep learning and skin cancer, establishing the context for the study. The third section offers a detailed description of the proposed model and an overview of the datasets, ISIC 2019 and HAM10000. The fourth section presents a comprehensive comparison and discussion of the results obtained by the proposed model and state-of-the-art CNN-based and ViT-based models on the ISIC 2019 and HAM10000 datasets, highlighting experimental findings. Finally, the last section concludes the study by summarizing the key findings and their implications.

2. Related Works

Recently, there has been increased research into deep-learning techniques for skin cancer detection. Most of the studies have focused on the application of different methodologies and algorithms to enhance the efficiency and accuracy of the process of diagnosing skin cancer. For example, Mazhar et al. [17] presented a systematic review that compares different deep learning algorithms. Mirikharaji et al. [18]

performed a very detailed analysis of articles on deep learning-based segmentation of skin lesions, discussing characteristics of datasets, model design, and aspects of evaluation. In another research work, Shah et al. [22] have shown how in-depth learning models can detect the presence of skin cancer at an early stage, again reiterating the importance of such automatic systems for lesion recognition [23].

Attallah presented an advanced and explainable artificial intelligence-based CAD system called "Skin-CAD" which is used for the classification of dermatoscopic images of skin cancer. The proposed model accurately classifies photographs into two general classes as benign or malignant and seven subclasses of skin cancer. The maximum accuracy achieved using Skin-CAD was 97.2 % and 96.5 % for Skin Cancer: Malignant vs Benign and HAM10000 datasets respectively [24]. Houssein et al. [25] proposed a new deep convolutional neural network (DCNN) approach to classify skin cancer lesions. The proposed DCNN model was evaluated using two imbalanced datasets, HAM10000 and ISIC-2019. The DCNN model was compared with other transfer learning models including VGG16, VGG19, DenseNet121, DenseNet201 and MobileNetV2, and its performance was evaluated in terms of accuracy, recall, sensitivity, F1 score, specificity, and AUC. The accuracy with the proposed DCNN model reached 98.5 % and 97.1 %. Goceri [26] presented the design of a neural network—a novel with adjustable properties and a convolutional capsule layer. The layers use learnable biases to encode spatial relationships between capsule vectors, allowing the network to keep vector orientations and learn the spatial relations. The study offers the main contributions of suggesting this novel network, its use in multi-class skin cancer classification and comparing it with other capsule networks on seven types of skin cancers. Pacal et al. [11] designed improvements to the Swin Transformer which provided enhanced model accuracy, speed in training, and improved parameter efficiency. The ISIC 2019 skin dataset was used for testing the proposed model and compared with state-of-the-art CNNs and vision transformer models. Akilandasowmya et al. [27] presented a deep hidden features and ensemble classifier-based method for detecting skin cancer, addressing issues related to real-time data streaming and associated dimensionality. Herein, ResNet50 was hybridized with sand cat swarm optimization and an improved harmony search technique. Their method outperforms state-of-the-art classifiers on benchmark datasets and shows promise for early skin cancer diagnosis.

Chen et al. [28] proposed MDFNet, a clinical model intending to fuse data from skin images with clinical knowledge to enhance the diagnosis. Testing shows an accuracy of 80.42 % for MDFNet, which is a 9 % improvement over using only medical images. This underscores the distinct fusion capabilities of MDFNet, suggesting it may be helpful in diagnosing melanoma, reinforcing decision-making, and refining clinical effectiveness. They also indicate that their data fusion technique may be applied to other illnesses for value in intelligent diagnostic strategies. Teodoro et al. [29] presented EfficientAttentionNet, a CNN structure utilized to identify skin lesions, such as melanoma and non-melanoma, from their early stages. This method involves image preprocessing to remove hair, balancing the sample classes using generative adversarial network (GAN), generating masks with a U-Net model. This model showed solid results and provided a baseline for upcoming studies in skin lesion classification. Sethanan et al. [30] published their research in designing an accurate system for skin cancer classification using image segmentation with CNNs. Their system classified different types of skin cancer effectively at a remarkable rate of over 99.4 %, validated using feedback from medical experts. The system scored 96.85 % on usability, denoting a very high level of user satisfaction. A new methodology has been created by Tembhurne et al. [31], in which deep learning has been integrated. This approach exploited advanced neural networks in feature extraction processes, along with conventional mechanisms. The results indicated a high accuracy rate of 93 %, where recall rates reached 99.7 % for benign cases and 86 % for malignant cases. Hybrid deep architectures in skin cancer detection have been explored by Diwan et al. [32] for CNNs. The proposed design, inspired by the pre-trained model and three main principles, employing multiple, smaller convolutional filters, including skip connections to address the vanishing gradient issue, and cyclic learning rate annealing sets an up-to-date new benchmark on the HAM10000 dataset. Gilani et al. [16] implemented advanced DNN using surrogate gradient descent in classifying 3670 images of melanoma and 3323 non-melanomas from the ISIC 2019 dataset. The proposed spiking VGG-13 model was able to classify the images with an accuracy of 89.57 % and an F1 score of 90.07 %, outperforming even the full-size VGG-13 and AlexNet with fewer parameters.

Qureshi and Roos [33] introduced a new architecture for an ensembled CNN to address some critical challenges related to the working of small and imbalanced datasets. They collectively used the force of models pre-trained on general data together with data-specific CNN models along with metadata to outperform seven benchmark techniques, including recent techniques based on CNNs, on a dataset of dermoscopic images from 2056 patients across different evaluation metrics. Viknesh et al. [34] utilized various CNNs, such as AlexNet, LeNet, and VGG-16, for the analysis of medical images. They integrated the most accurate model into web and mobile applications and investigated the impact of model depth and dataset size on performance. Additionally, they utilized support vector machines with default RBF kernels to classify images into benign, malignant, or normal categories, achieving an accuracy of 86.6 %. The CNNs demonstrated superior performance, achieving a 91 % accuracy rate after 100 epochs. Tabrizchi et al. [35] used the improved version of the trendy CNN architecture called VGG-16 to train their model. Their proposed method experimentally showed better accuracy. Dahou et al. [36] proposed a skin cancer detection model that used the MobileNetV3 architecture for extracting relevant features from images and an optimized feature selection model employing the modified Hunger Games Search (HGS) algorithm along with Particle Swarm Optimization (PSO) and Dynamic-Opposite Learning (DOLHGS). The system performed with an accuracy of 88.19 % on the ISIC-2016 dataset and 96.43 % on the PH2 dataset, thus effectively diagnosing skin cancers.

3. Methods and materials

In this section, we employ the ISIC 2019 and HAM10000 datasets, one of the datasets for the ISIC challenges and the most extensive publicly available dataset with the most diverse classes. Our approach integrates up-to-date deep-learning models, including an original model based on the MetaFormer architecture [37,38]. The proposed model achieves high sensitivity and specificity in detecting and classifying skin cancers by utilizing robust vision transformers, advanced data augmentation, and transfer learning strategies. To ensure reproducibility and encourage further research in the domain of cancer-related diseases, we provide detailed implementation and training methods.

3.1. Datasets

In this study, we used the ISIC 2019 and HAM10000 datasets, which are among the most popular and publicly available skin cancer datasets, to reveal the true performance of the proposed model as well as other CNN and ViT models. The first dataset we employed, the ISIC 2019 dataset, is widely recognized by researchers and is one of the richest publicly accessible datasets. International Skin Imaging Collaboration (ISIC) 2019 dataset is an essential repository for research in deep learning and artificial intelligence in skin cancer diagnosis and classification [39]. The dataset contains not only dermoscopic images, but also demographic information of the patients and rich clinical metadata associated with diagnoses of skin lesions. Generally, ISIC 2019 is utilized for scientific research in tasks related to early melanoma and skin cancer detection and diagnosis using its training, validation, and testing subsets. Many researchers, therefore, draw on this database for the development and validation of a vast number of techniques for training and

testing deep learning methods, pushing the current state-of-the-art. A few sample images of the classes within the ISIC-2019 dataset are depicted in Fig. 1.

The ISIC 2019 dataset comprises a total of 25,331 labeled images, each belonging to one of eight distinct skin lesion classes: Melanoma (MEL), Melanocytic Nevus (NV), Basal Cell Carcinoma (BCC), Actinic Keratosis (AK), Benign Keratosis (BKL), Dermatofibroma (DF), Vascular Lesion (VASC), and Squamous Cell Carcinoma (SCC). Within this dataset, the images range in resolution from 576 × 768 to 1024 × 1024 pixels, spanning 101 different resolutions. All images are colorful and consist of three-color channels. Notably, there is a significant data imbalance among the classes; for example, the number of images in the NV class is approximately 51 times greater than those in the VASC class. This imbalance presents a challenge for model training and classification accuracy, necessitating careful consideration in the development of algorithms and evaluation strategies.

The second dataset, HAM10000, is a well-recognized and widely used dataset for skin cancer research, consisting of 10,015 images across seven classes. The dataset includes 327 images of AK, 514 images of BCC, 1,099 images of BKL, 115 images of DF, 1,113 images of MEL, 6,705 images of NV, and 142 images of VASC. Unlike the ISIC 2019 dataset, HAM10000 does not include any images from the SCC class. As a result, for experiments involving the SCC class, we exclusively used images from the ISIC 2019 dataset, as shown in Fig. 1. By combining these two datasets, we established a robust evaluation framework that captures the diversity of ISIC 2019 while complementing it with the unique structure of HAM10000, enabling a more comprehensive analysis.

3.2. Deep learning approaches

Deep learning is revolutionary at the forefront of artificial intelligence as it was known earlier and brings a new paradigm characterized by the development of sophisticated algorithms capable of ingesting and understanding voluminous data sets [13]. In the face of all domains affected in disruptive ways by this technological upheaval, computer vision emerges with practical applications. Leading this transformational wave is CNN, known for its ability to outperform in most computer vision tasks [40]. Concerning image data analysis, their architecture is such that interconnected layers collaborate seamlessly to decipher and extract the complex features that naturally embed visual stimuli. These layers reduce dimensionality and ferret out salient features by applying convolutional filters and pooling techniques.

CNNs process input images with a hierarchical feature extraction strategy, progressively distilling increasingly abstract representations to attain a deep and nuanced understanding of underlying image content [41]. Recent research trends respond by introducing ViT models. While vision transformers bear many similarities to CNNs, the key distinction lies in the ability of vision transformers not to employ conventional convolutional layers but to use positional embeddings and self-attention mechanisms [42]. This innovative approach of mixing local and global information makes ViTs solid competitors for the tasks that require an overall understanding of the scenes. Their rise represents a landmark in deep learning algorithms, providing a complementary method to capture subtle details over different scales [43]. While CNNs continue to dominate in many fields, ViTs very clearly offer a unique niche with the potential to enlarge the scopes of computer vision and artificial intelligence.

3.3. Proposed model

In this section, we introduce a new model for diagnosing skin cancer from dermatological images, built on the MetaFormer architecture. The journey of MetaFormers began with PoolFormer, its very first version, which set the foundation by showing how a simple, yet effective design could achieve impressive results [38]. Since then, variants like IdentityFormer, Conformer, CaFormer, RandFormer, and others have emerged, each tailored for specific tasks while staying true to the core principles of MetaFormers. MetaFormers are designed to be highly adaptable, and capable of working with different data types and tasks without needing complex customizations. They allow for easy integration of components like attention mechanisms, spatial MLPs, or other token mixers, making them versatile and efficient. One of the key strengths of MetaFormers lies in their structure. Elements like residual connections and channel MLPs deliver high performance even with basic operators such as pooling. For example, PoolFormer, the first MetaFormer, showed that the overall framework matters more than the specific token mixer. It achieved state-of-the-art results in tasks like classification, detection, and segmentation, outperforming advanced models like DeiT and ResMLP on the ImageNet-1 K dataset, all while

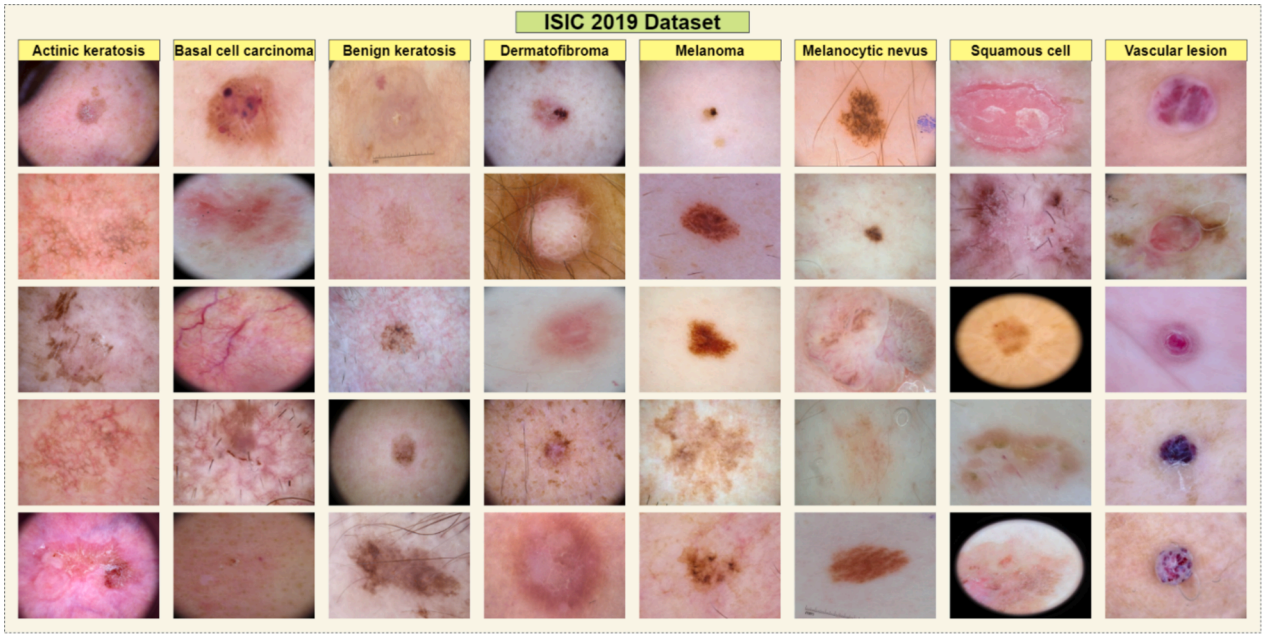

Fig. 1. Some sample images by class from the ISIC 2019 dataset.

using fewer parameters and less computational power. The other variants, with their unique token mixers, ranging from identity mapping to convolutional layers and random token sampling, continue to build on this success. Despite their differences, the consistent use of MetaFormer's core design ensures top performance across applications. These findings highlight the potential for further advancements in the MetaFormer framework, paving the way for more efficient and accurate models. Fig. 2 showcases our Proposed Model (proposed MetaFormer-based model), designed to bring simplicity and adaptability to the task of diagnosing skin cancer.

As shown in Fig. 2, the Proposed Model is based on the CAFormer architecture, with an optimized structure that replaces standard self-attention mechanisms with focal self-attention. MetaFormer architectures, such as ConvFormer, IdentityFormer, and RandFormer, are known for their architectural flexibility, allowing for different layer configurations within the same block type. For instance, ConvFormer and IdentityFormer implement such designs effectively. CAFormer, however, adopts a unique approach by combining CNN-based ConvFormer layers in the first two stages with Transformer layers in the last two stages, repeating this pattern multiple times. Our proposed model builds upon the CAFormer architecture by integrating ConvFormer and Transformer blocks into a hybrid structure, further optimized with focal self-attention to enhance its effectiveness for skin cancer diagnosis. Experimental studies and research indicate that models of intermediate size, which are neither too small nor too large, yield the best performance in skin cancer classification. Based on this principle, we scaled the depth of the CAFormer-S18 model from [3,3,9,3] to [4,4,12,4], resulting in the improved CAFormer-S24 (Proposed Model) model. Additionally, we enhanced stages three and four of our Proposed Model with focal self-attention mechanisms to extract more complex features and boost performance. By combining these structural optimizations and advanced attention mechanisms, the Proposed Model addresses the challenges of skin cancer diagnosis with exceptional scalability, efficiency, and accuracy.

3.3.1. Caformer architecture

The CAFormer architecture is designed as a hybrid model that should bring out all the strengths of CNNs and Self-Attention mechanisms for an effective and powerful framework in image classification [37]. As seen in Fig. 2, this architecture consists of four main stages: patch embedding, depth-wise separable convolutions, and self-attention blocks, followed by a classification head. By providing input image x the input runs through a convolutional layer with kernel size and stride equal to *4 x 4* that formulated in Eq.1.

$$PatchEmbed(x) = Conv2D(x, kernel\_size = 4, stride = 4) \quad (1)$$

Stage 1 and Stage 2 both use depth-wise separable convolutions to do the processing on 56 × 56 and 28 × 28 feature maps, respectively. The operations are defined as Eq.2 and Eq. (3).

$$Depthwise(x) = ReLU(Conv2D(x, kernel\_size = 3, groups = C, padding = 1)) \quad (2)$$

$$Pointwise(x) = ReLU(Conv2D(x, kernel_{size} = 1)) \quad (3)$$

where C denotes the number of channels. These stages efficiently capture local features through convolutions. Stages 3 and 4 incorporate self-attention mechanisms, which allow the model to capture long-range dependencies within the feature maps. The self-attention operation is formulated as Eq. (4)

$$SelfAttention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (4)$$

where Q, K, and V represent the query, key, and value matrices, and d_k is the dimensionality of the key vectors. Finally, the classification head consists of global average pooling, followed by a fully connected layer and SoftMax activation, expressed as Eq. (5).

$$Output(x) = Softmax(FC(GlobalAvgPool(x))) \quad (5)$$

In addition, we first augment the third and fourth stages of the CAFormer architecture with focal self-attention mechanisms to extract complex features that boost performance. We then scale the CAFormer-S24 model toward depths [4,4,12,4] to get the ideal scaling of the model for more challenging tasks, such as skin cancer classification. Upscaling the model increases its capacity to learn from intricate patterns and features in the data, making it more effective in handling high-resolution images and complex classification tasks. To sum up, desirable in this situation would be to increase the number of layers at each stage so that the network would learn increasingly complex and more abstract features, which are required for tasks like the classification of skin cancer, where differences between classes are hard to perceive. Second, integrating focal self-attention mechanisms helps the model learn to capture both local and global dependencies within the image efficiently, thereby bypassing the limitations of traditional self-attention mechanisms, which suffer from high computational and memory costs.

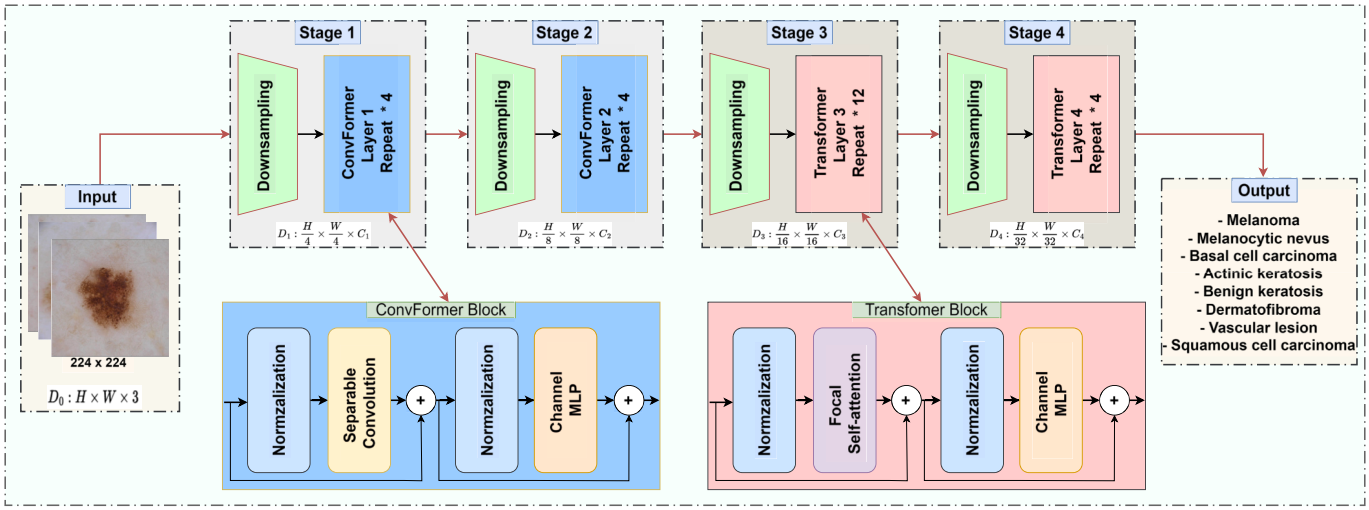

Fig. 2. The detailed structure of the proposed model for the autonomous diagnosis of skin cancer.

3.3.2. Focal self-attention mechanism

One of the main innovations of this model is focal self-attention, designed to render it more efficient in capturing both short- and long-range dependencies by combining fine-grained local attention with coarse-grained global attention. This double form of attention ensures the maintenance of high computational efficiency in capturing a broad range of dependencies within the input data. Local attention focuses on tokens within a local window. It calculates the attention weights for nearby tokens and is particularly effective in capturing local patterns and features. For skin cancer diagnosis, the focal self-attention mechanism was introduced to address the unique challenges inherent in skin cancer image classification. These images often exhibit high intra-class variability, where lesions within the same class can differ significantly in appearance due to variations in size, texture, and color. Additionally, there is low inter-class variability, as lesions from different classes may share overlapping features, making differentiation particularly challenging. Furthermore, fine-grained details, such as irregular borders or specific color patterns, play a critical role in accurate classification but are frequently overshadowed by global features in standard attention mechanisms. By selectively attending to both critical regions and broader contextual information, the focal self-attention mechanism enhances the model's expressive capability, enabling it to effectively discern subtle differences between skin cancer classes. This design is especially valuable for datasets like ISIC 2019 and HAM10000, where accurate classification relies on capturing both local and global patterns. The formulation for local attention is Eq. (6).

$$LocalAttention(Q,K,V) = Softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{6}$$

where Q, K, and V are the query, key, and value matrices, respectively, and d_k is the dimensionality of the key vectors. This mechanism ensures that each token attends to its local neighborhood, thus capturing the local context effectively. In turn, global attention mechanisms focus on the representations of tokens in a higher, more coarse-grained context. This mechanism can capture broader context information by aggregating features from different regions, thus keeping the computational cost lower while reasoning about the general image structure. Global attention is formulated similarly to local attention but is applied to pooled features as formulated in Eq. (7).

$$GlobalAttention(Q,K,V) = Softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{7}$$

Here, K and V represent pooled features from different regions, allowing the model to attend to a summarized representation of distant tokens. The combined focal self-attention mechanism integrates both local and global attention to leverage their respective strengths. This integration is formulated as Eq. (8).

$$\begin{aligned} FocalSelfAttention(Q,K,V) =& LocalAttention(Q,K,V) \\ &+ GlobalAttention(Q,K,V) \end{aligned} \tag{8}$$

By incorporating both types of attention, the model can effectively capture fine-grained details while also understanding the broader context. This dual attention mechanism allows the model to process high-resolution images efficiently and capture complex patterns and dependencies within the data. In the CAFormer-S24 model, focal self-attention is integrated into the third and fourth stages. The third stage processes 14×14 feature maps and includes twelve layers of focal self-attention, while the fourth stage processes 7×7 feature maps and includes four layers of focal self-attention. This setup ensures that the model can effectively capture both local and global dependencies at different scales, enhancing its performance in tasks such as skin cancer classification. The resulting architecture includes an initial patch embedding layer, depth-wise separable convolution layers in the first two stages, focal self-attention layers in the third and fourth stages, and a classification head. This configuration allows the CAFormer-S24 model to maintain computational efficiency while achieving superior performance by capturing a wide range of dependencies in the input data. The total estimated parameters for the CAFormer-S24 model with focal self-attention are approximately 35.01 million, making it a robust and efficient model for complex medical image classification tasks such as skin cancer classification.

4. Results and discussions

This section encompasses the experimental setup where experimental results are obtained, including data preprocessing, data augmentation, transfer learning, performance metrics, and results and comparisons pertaining to deep learning models.

4.1. Experimental setup

This research is done on an Ubuntu 24.04-based Linux system powered with state-of-the-art hardware: a single NVIDIA RTX 3090 graphics card, and an Intel i7-14700 K CPU coupled with 64 GB DDR5 RAM. The experiments have been conducted through model training and testing various deep learning models against the latest PyTorch framework, underlaid with a base of support from NVIDIA CUDA, to ensure similar computation conditions across all evaluations. Standard data augmentation techniques, like scaling, smoothing, mix-up, color jitter, and flipping, have been systematically applied to the training data to ensure optimal model stability and performance. The use of transfer learning with pre-trained weights from the ImageNet dataset helped speed up the process of convergence and generalization to increase accuracy. Almost all the models used a fixed input resolution for the training and validation datasets 224×224. In our experiments, we employed default hyperparameters provided by the 'timm' library for each model to ensure standardization and reproducibility. These hyperparameters, including learning rate = 0.01, learning rate base = 0.1, momentum = 0.9, optimizer = SGD, weight decay = 2.0e − 05, warmup epochs = 5, and warmup learning rate = 1.0e − 05, were applied consistently across all models. To further ensure fair comparisons, we used identical data splits (70 % training, 20 % validation, 10 % testing) and ImageNet-based normalization for preprocessing. Additionally, dynamic learning rate schedulers, such as cosine annealing, were employed to accommodate architectural differences. Pretrained weights from ImageNet were utilized to accelerate convergence and enhance generalization. This uniform setup was chosen to highlight the inherent performance differences between models while maintaining comparability and reproducibility in our evaluations.

4.2. Data preprocessing and data augmentation

Data preprocessing plays a crucial role in optimizing deep learning models. It typically involves key steps such as splitting the data into training, validation, and test sets, normalizing the data, reducing noise, and handling outliers. Unlike many previous studies, we adopted a three-set division, comprising training, validation, and test subsets, instead of the more conventional two-set split or cross-validation approach. This method is particularly important as it directly impacts the deep learning model's performance and marks a significant departure from traditional practices. Such a splitting strategy is essential for accurately evaluating model performance and mitigating the risk of overfitting. The class distributions for the training, validation, and test sets of both the ISIC 2019 and HAM10000 datasets are provided in Table 1, ensuring transparency and reproducibility.

Table 1 provides a detailed breakdown of the number of images in the training, validation, and test sets for both the ISIC 2019 and HAM10000 datasets used to evaluate the proposed model. The datasets were divided into three subsets: 70 % for training, 20 % for validation, and 10 % for testing. This division ensures a robust evaluation framework by allowing the model to learn from a large training set while

Table 1
Number of images for three subsets of the ISIC 2019 and HAM10000 datasets.

Class	Total	Training set (%70)	Validation set (%20)	Test set (%10)
ISIC 2019 dataset				
Actinic Keratosis (AK)	867	607	173	87
Basal Cell Carcinoma (BCC)	3,323	2,326	665	332
Benign Keratosis (BKL)	2,624	1,837	525	262
Dermatofibroma (DF)	239	167	48	24
Melanoma (MEL)	4,522	3,165	904	453
Melanocytic Nevus (NV)	12,875	9,012	2,575	1,288
Squamous Cell Carcinoma (SCC)	628	440	126	62
Vascular Lesion (VASC)	253	177	51	25
Total	25,331	17,731	5,067	2,533
HAM10000 dataset				
Actinic Keratosis (AK)	327	229	65	33
Basal Cell Carcinoma (BCC)	514	360	103	51
Benign Keratosis (BKL)	1,099	769	220	110
Dermatofibroma (DF)	115	80	23	12
Melanoma (MEL)	1,113	779	223	111
Melanocytic Nevus (NV)	6,705	4,694	1,341	670
Vascular Lesion (VASC)	142	99	28	15
Total	10,015	7,010	2,003	1,002

reserving sufficient data for validation and unbiased testing. For the ISIC 2019 dataset, the total number of images is 25,331, with the largest class being "NV" (12,875 images) and the smallest class being "DF" (239 images). Similarly, the HAM10000 dataset contains 10,015 images, with "NV" (6,705 images) as the largest class and "VASC" (142 images) as the smallest. The class distribution reflects the natural imbalance often found in medical datasets, emphasizing the importance of evaluating the model's ability to handle such disparities effectively.

Online data augmentation is an effective approach for addressing class imbalance and improving model performance. By introducing various transformations, such as rotation, flipping, scaling, and noise addition, data augmentation increases the diversity within the dataset. This enables the model to learn from a broader range of scenarios, enhancing its ability to generalize to unseen data. The proposed model benefits from this technique by utilizing hybrid data effectively and achieving better generalization compared to other methods. Additionally, data augmentation ensures that each class has an equal impact on model performance, resulting in more balanced and reliable outcomes. For this study, we systematically applied standard augmentation techniques, including scaling, smoothing, mix-up, color jittering, and flipping, across all models during training. These methods enriched the dataset, increased variability, and improved the model's robustness against overfitting. Data augmentation is a fundamental component of deep learning that equips models with enhanced performance and adaptability for various tasks and domains, especially in cases with limited or highly variable datasets.

4.3. Results

In this section, experiments and model performance are shown for a total of thirty models: ten leading-edge CNNs and twenty state-of-the-art ViTs. It is evaluated on the ISIC 2019 and HAM10000 datasets for which it differentiates the three subsets of data: training, validation, and test. Unlike studies that focus solely on validation, this research highlights the demonstration of the generalization ability of each model to previously unseen test data, especially for applications in skin cancer detection. Assessment of the models using these new test data ensures a more accurate gauge of their effectiveness and performance in the real world. The comprehensive evaluation starts with training, in which serious optimization of each model takes place using established techniques, including data augmentation, scheduling learning rate, and advanced regularization methods. During the validation phase, the first indication of model performance is given for hyperparameter tuning and providing a strategy for early stopping to avoid overfitting. However, the critical focus of this study is the test phase, in which the models are evaluated using an independent, unseen dataset to determine their robustness and generalization capabilities.

Notable architectures among the 10 include ResNet50 [44], VGG16 [45], DenseNet169 [46], Inceptionv4 [47], MobileNetv3-Large [48], EfficientNetv2-Medium [49], RepGhostNet-100 [50], InceptionNext-Base [51], EfficientNet-B6 [52], and ConvNexT-Base [53]. These have been selected because they perform quite well on medical image analysis tasks and span a wide range of architectural characteristics. The 20 image transformers employed in this study encompass a variety of advanced models, including Mixer-B16 [54], PoolFormer-M36 [38], FocalNet-Base [55], MobileViT-Small [56], DeiT3-Base [57], Swin-Base [58], Swinv2-Base [59], BeiTv2-Base [60], MaxViT-Base [61], RepViT-m1 [62], ConViT-Base [63], FastViT-ma36 [64], NextViT-Base [65], CrossViT-Base [66], Tiny-ViT-21 m [67], ConvMixer-768 [68], CAFormer-S18 [37], CAFormer-S36, CAFormer-M36, and CAFormer-B36. These transformers are renowned for their ability to handle complex image recognition tasks, making them highly suitable for skin cancer detection.

4.3.1. Results for ISIC 2019 dataset

The performance results of the Proposed Model, along with ten CNN-based and sixteen ViT-based models, on the publicly available ISIC 2019 dataset are presented in Table 2. All models were evaluated exclusively on a pre-allocated test dataset, ensuring an objective comparison of their ability to generalize effectively to real-world data.

Considering the metrics presented in Table 2 and Fig. 3 and Fig. 4, among the CNN models, ConvNeXt-Base demonstrated the highest performance, achieving an accuracy of 0.9025 and a precision of 0.8993. These results indicate its superior ability to distinguish between different classes in the dataset. Conversely, ResNet50 exhibited relatively lower performance, with an accuracy of 0.8535 and a precision of 0.7865, highlighting the limitations of traditional CNN architectures in

Table 2
Experimental results of the deep learning-based models on ISIC 2019 dataset.

Model	Accuracy	Precision	Recall	F1-score
ResNet50	0.8535	0.7865	0.7667	0.7749
VGG16	0.8674	0.8404	0.7830	0.8090
DenseNet169	0.8723	0.8388	0.7922	0.8127
Inceptionv4	0.8863	0.8369	0.7983	0.8157
MobileNetv3-Large	0.8701	0.8328	0.7848	0.8058
EfficientNetv2-Medium	0.8985	0.8716	0.8489	0.8594
RepGhostNet-100	0.9001	0.8874	0.8292	0.8550
InceptionNext-Base	0.8863	0.8563	0.8225	0.8373
EfficientNet-B6	0.8989	0.8732	0.8524	0.8615
ConvNexT-Base	0.9025	0.8993	0.8160	0.8517
Mixer-B16	0.8883	0.8725	0.8128	0.8389
PoolFormer-M36	0.8891	0.8596	0.8114	0.8339
FocalNet-Base	0.9013	0.8833	0.8545	0.8670
MobileViT-Small	0.8966	0.8473	0.8483	0.8460
DeiT3-Base	0.9056	0.8932	0.8554	0.8735
Swin-Base	0.9056	0.8920	0.8513	0.8703
Swinv2-Base	0.9068	0.9013	0.8690	0.8838
BeiTv2-Base	0.9037	0.8998	0.8581	0.8768
MaxViT-Base	0.9009	0.9040	0.8389	0.8690
RepViT-m1	0.8792	0.8593	0.8062	0.8309
ConViT-Base	0.9017	0.8836	0.8573	0.8682
FastViT-ma36	0.9013	0.8836	0.8478	0.8648
NextViT-Base	0.8752	0.8568	0.8110	0.8320
CrossViT-Base	0.9011	0.8824	0.8528	0.8661
Tiny-ViT-21m	0.8974	0.8653	0.8437	0.8533
ConvMixer-768	0.8760	0.8396	0.7920	0.8125
Proposed Model	**0.9254**	**0.9041**	**0.8768**	**0.8886**

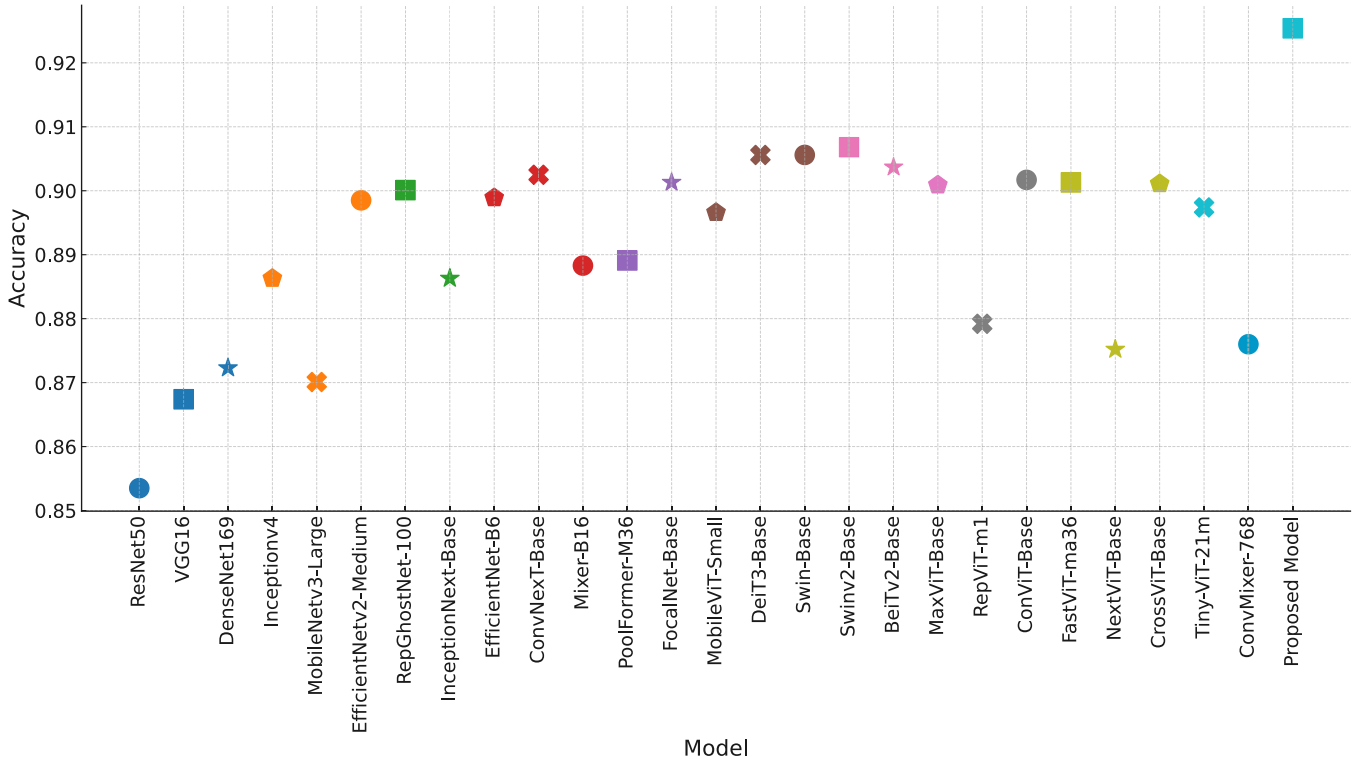

Fig. 3. Accuracy metric for all models, including both CNN and ViT-based models, as well as the Proposed Model on ISIC 2019 dataset.

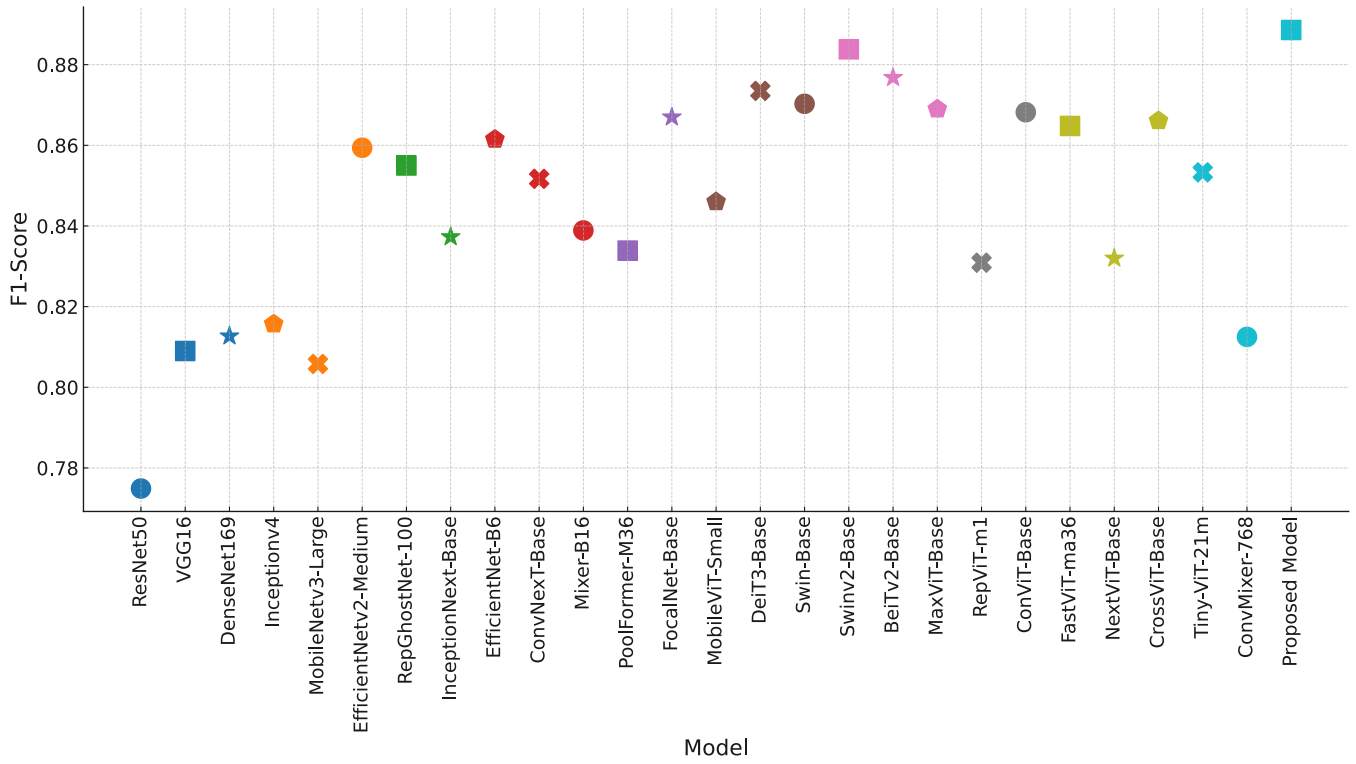

Fig. 4. F1-score metric for all models, including both CNN and ViT-based models, as well as the Proposed Model on ISIC 2019 dataset.

addressing the complexities of dermatological image analysis.

Among the ViT models, SwinV2-Base delivered the best overall performance, with an accuracy of 0.9068, precision of 0.9013, recall of 0.8690, and F1-score of 0.8838. This underscores the effectiveness of advanced transformer architectures in capturing the intricate and subtle features of skin cancer images. Similarly, DeiT3-Base achieved commendable results, with an accuracy of 0.9056, precision of 0.8932, recall of 0.8554, and F1-score of 0.8735, further validating the potential of ViT models in the domain of medical image analysis.

The Proposed Model, designed by scaling the CAFormer architecture

and integrating focal self-attention in place of traditional self-attention, achieved superior performance metrics: an accuracy of 0.9254, precision of 0.9041, recall of 0.8768, and F1-score of 0.8886. The enhanced performance of the Proposed Model can be attributed to the integration of focal self-attention, which effectively emphasizes relevant regions of the image while minimizing noise during feature extraction. Additionally, the upscaling of the CAFormer architecture to include more layers optimally balances parameter efficiency and computational performance. With 35.01 million parameters, the Proposed Model demonstrates a carefully optimized design, achieving an excellent trade-off between complexity and efficiency, positioning it as a state-of-the-art solution for dermatological image analysis.

As presented in Table 2, the Proposed Model achieved the highest accuracy at 0.9254, surpassing SwinV2-Base (0.9068) and DeiT3-Base (0.9056). This notable performance highlights the superior generalization capabilities of the Proposed Model, positioning it as the most reliable option for accurate diagnosis in skin cancer screening. For precision, the Proposed Model again led with a score of 0.9041, followed closely by SwinV2-Base at 0.9013 and ConvNeXt-Base at 0.8993. High precision is critical in minimizing false positives, underscoring the effectiveness of the Proposed Model in this regard. In terms of recall, SwinV2-Base demonstrated the highest value at 0.8690, while the Proposed Model followed closely with a recall of 0.8768, outperforming DeiT3-Base at 0.8554. Recall is a particularly significant metric in medical diagnostics, as it reflects the model's ability to correctly identify true positive cases. The superior recall of the Proposed Model underscores its effectiveness in detecting actual instances of skin cancer, which is vital for early intervention. The F1-score, which balances precision and recall, further validated the robust performance of the Proposed Model, achieving a score of 0.8886. SwinV2-Base and DeiT3-Base also demonstrated competitive results, with F1-scores of 0.8838 and 0.8735, respectively. Overall, the Proposed Model's exceptional performance across all metrics underscores its potential as a state-of-the-art tool for accurate and reliable skin cancer diagnosis.

The integration of focal self-attention in the Proposed Model represents a significant advancement in attention mechanisms, enabling the model to prioritize critical regions within the image. This targeted focus enhances feature representation, leading to improved diagnostic accuracy. Furthermore, the scalability of the CAFormer architecture strengthens the model by increasing its capacity to effectively learn and generalize from complex datasets, thereby addressing the challenges inherent in dermatological image analysis. Fig. 5 illustrates the confusion matrix, providing a detailed overview of the class-wise performance and further validating the robustness and reliability of the Proposed Model.

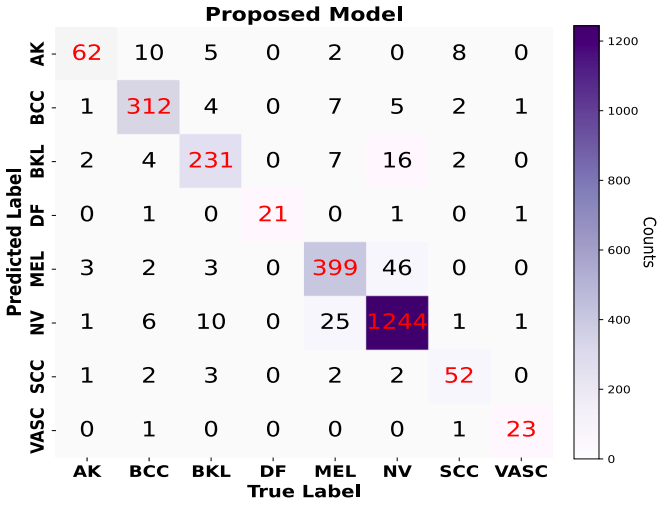

Fig. 5. The confusion matrix showing the class-specific performance of the Proposed Model on ISIC 2019 dataset.

As shown in Fig. 5, the Proposed Model achieves excellent performance in diagnosing NV with 1,244 true positives, precision of 0.9467, recall of 0.9658, and an F1-score of 0.9562, demonstrating its high accuracy in this class. However, its lowest performance is observed in AK with 62 true positives, precision of 0.8857, recall of 0.7126, and an F1-score of 0.7898, indicating challenges in capturing all true cases. Strong results are also evident for BCC; 312 TP and MEL; 399 TP, showcasing robust generalization for critical skin cancers. Lower recall for SCC; 52 TP highlights the need for improved sensitivity to enhance performance across all lesion types.

4.3.2. *Results for HAM10000 dataset*

The performance results of the Proposed Model, along with ten CNN-based and sixteen ViT-based models, on the publicly available ISIC 2019 dataset are presented in Table 3. All models were evaluated exclusively on a pre-allocated test dataset, ensuring an objective comparison of their ability to generalize effectively to real-world data.

Table 3 provides valuable insights into how different deep learning models performed on the HAM10000 dataset. Among the CNN-based models as seen in Fig. 6 and Fig. 7, ConvNeXT-Base stood out with an accuracy of 0.9092 and an F1-score of 0.8472, showing its strong ability to generalize. In contrast, older models like ResNet50 (accuracy: 0.8733, F1-score: 0.7886) and VGG16 (accuracy: 0.8912, F1-score: 0.8119) struggled to handle the complexities of dermatological images effectively.

ViT-based models, on the other hand, demonstrated significantly better performance. Swinv2-Base led the way with an accuracy of 0.9361, precision of 0.9291, recall of 0.9049, and an F1-score of 0.9165, highlighting the strength of advanced transformer architectures in capturing fine-grained details. CrossViT-Base and Tiny-ViT-21 m also delivered impressive results, with F1-scores of 0.8981 and 0.8979, further demonstrating the capabilities of ViT-based approaches.

The Proposed Model surpassed all other models, achieving remarkable results with an accuracy of 0.9501, precision of 0.9470, recall of 0.9211, and an F1-score of 0.9334, as seen in Table 3, Fig. 6 and Fig. 7. This exceptional performance stems from its innovative use of focal self-attention and an optimized CAFormer architecture, which effectively

Table 3
Experimental results of the deep learning-based models on HAM10000 dataset.

Model	Accuracy	Precision	Recall	F1-score
ResNet50	0.8733	0.8194	0.7682	0.7886
VGG16	0.8912	0.8664	0.7744	0.8119
DenseNet169	0.9002	0.8695	0.8267	0.8458
Inceptionv4	0.9042	0.8692	0.8262	0.8427
MobileNetv3-Large	0.9122	0.8546	0.7695	0.8041
EfficientNetv2-Medium	0.9092	0.8581	0.8246	0.8403
RepGhostNet-100	0.9102	0.8791	0.8320	0.8513
InceptionNext-Base	0.9062	0.8720	0.8250	0.8453
EfficientNet-B6	0.8922	0.8295	0.8090	0.8172
ConvNeXT-Base	0.9092	0.8964	0.8240	0.8472
Mixer-B16	0.9032	0.8775	0.8277	0.8507
PoolFormer-M36	0.9162	0.8921	0.8849	0.8862
FocalNet-Base	0.8912	0.8683	0.8019	0.8297
MobileViT-Small	0.9162	0.8861	0.8585	0.8702
DeiT3-Base	0.9162	0.9081	0.8727	0.8892
Swin-Base	0.9301	0.9264	0.8419	0.8755
Swinv2-Base	0.9361	0.9291	0.9049	0.9165
BeiTv2-Base	0.8922	0.8251	0.7704	0.7937
MaxViT-Base	0.9271	0.8999	0.8677	0.8810
RepViT-m1	0.8962	0.8752	0.7800	0.8213
ConViT-Base	0.9172	0.8966	0.8602	0.8751
FastViT-ma36	0.9082	0.8776	0.8262	0.8475
NextViT-Base	0.9062	0.8777	0.8321	0.8512
CrossViT-Base	0.9251	0.9070	0.8902	0.8981
Tiny-ViT-21m	0.9281	0.9085	0.8882	0.8979
ConvMixer-768	0.9132	0.8947	0.8478	0.8664
Proposed Model	**0.9501**	**0.9470**	**0.9211**	**0.9334**

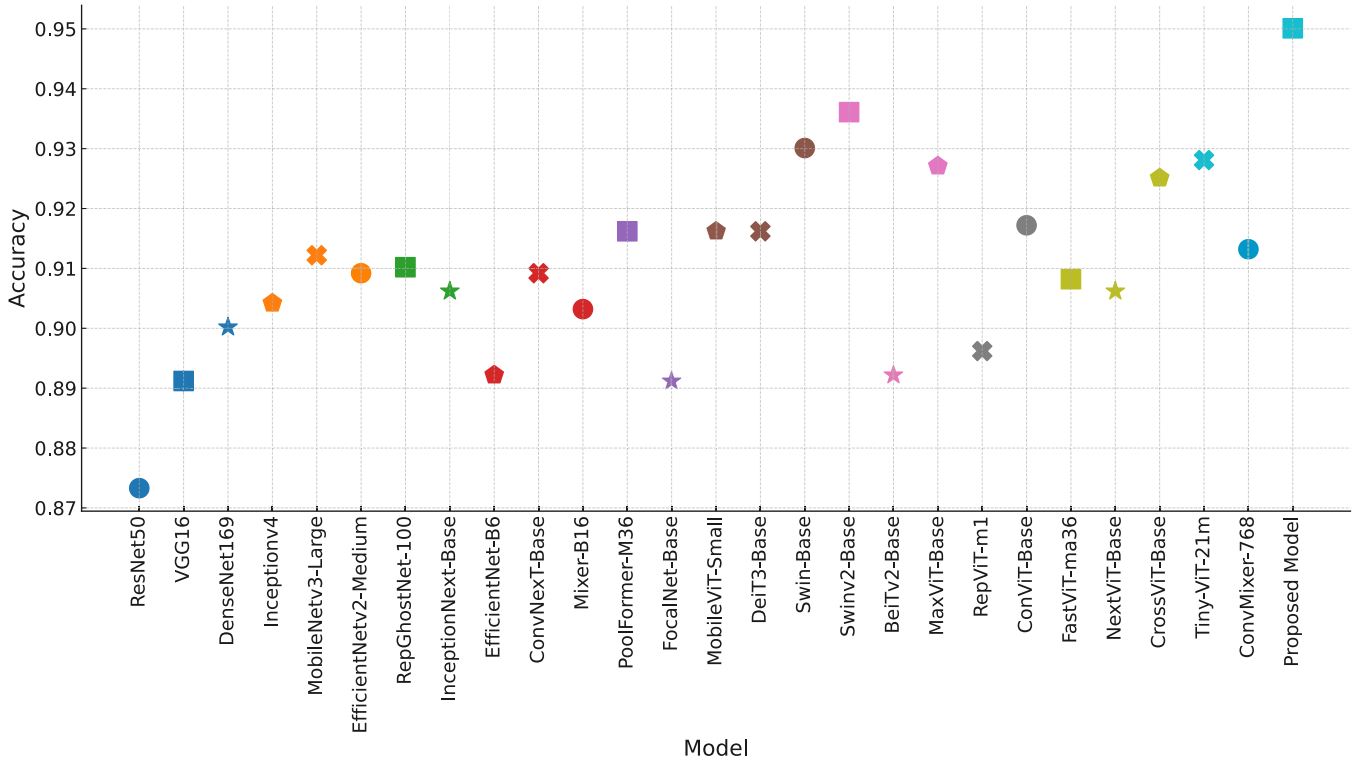

Fig. 6. Accuracy metric for all models, including both CNN and ViT-based models, as well as the Proposed Model on HAM10000 dataset.

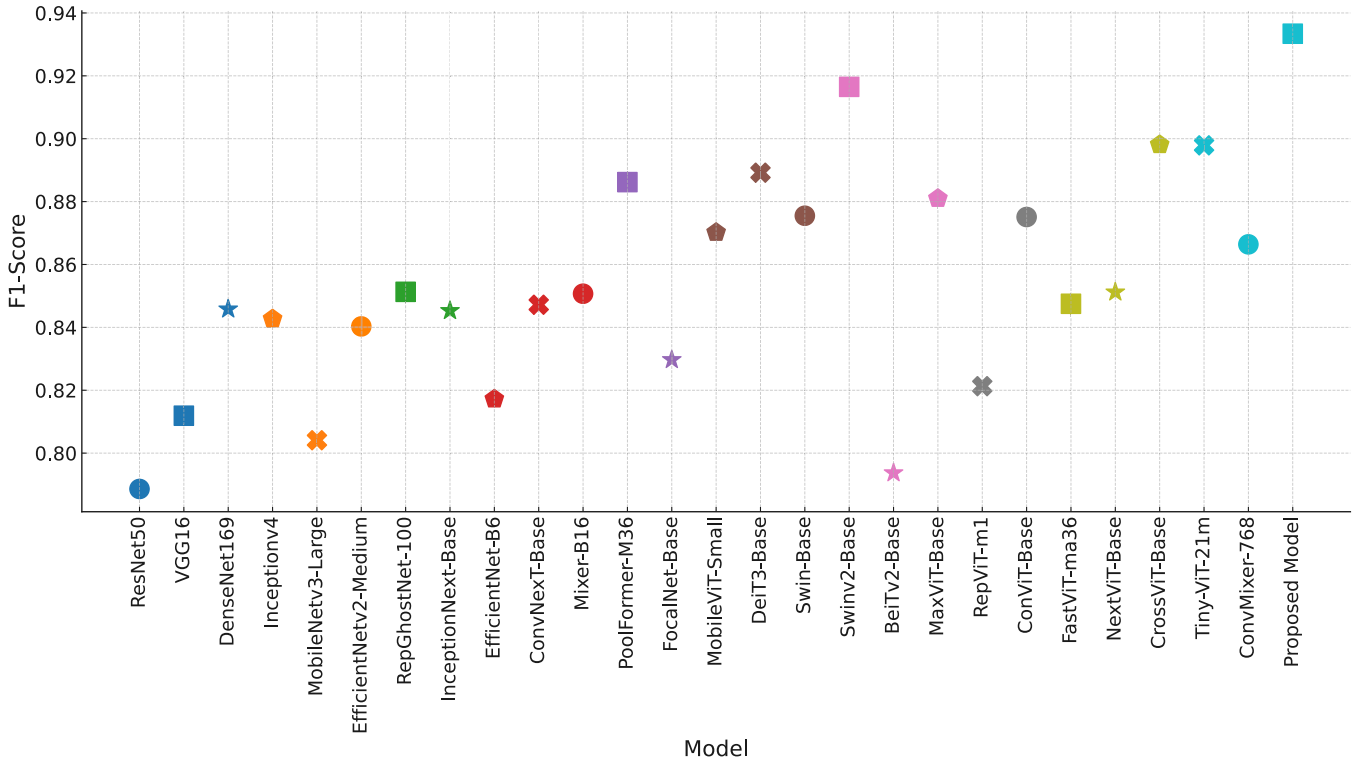

Fig. 7. F1-score metric for all models, including both CNN and ViT-based models, as well as the Proposed Model on HAM10000 dataset.

concentrates on critical image regions while reducing noise. The model also balances complexity and efficiency, boasting 35.01 million parameters, which enhances its overall effectiveness without excessive computational demands.

The findings in Table 3 underline the clear advantages of the Proposed Model, establishing it as a reliable and robust choice for skin lesion classification. These results also highlight the importance of advanced architectures in tackling the challenges of medical image analysis with precision and efficiency. Fig. 8 presents the confusion matrix, providing a detailed overview of the class-wise performance and

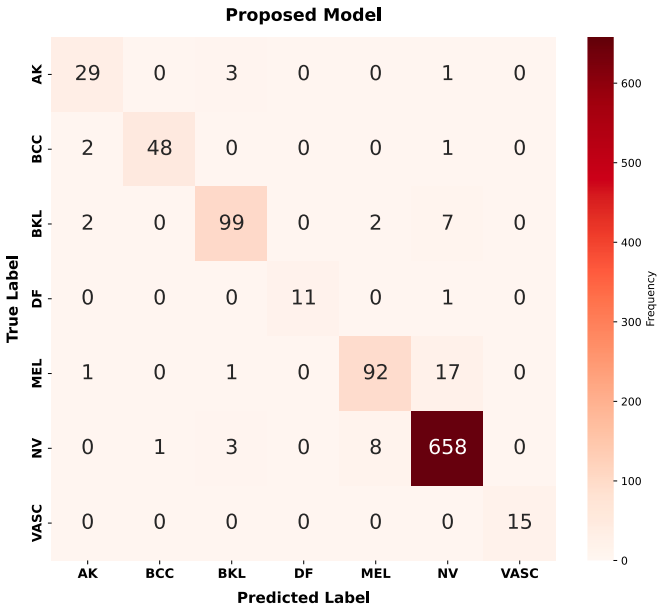

Fig. 8. The confusion matrix showing the class-specific performance of the Proposed Model on HAM10000.

further validating the robustness and reliability of the Proposed Model.

As seen in Fig. 8, the Proposed Model achieved a weighted average precision, recall, and F1-score of 0.9497, 0.9501, and 0.9495, respectively, across 1,002 images. The VASC class excelled with perfect scores (accuracy: 1.000 and 15 TP) across all metrics, likely due to its smaller, well-defined dataset of 15 images. In contrast, the MEL class had a lower F1-score (0.8638) due to reduced recall (0.8288), and with 92 TP, indicating some melanoma cases were misclassified. The NV class stood out with the highest F1-score (0.9712), showcasing its robustness with a larger number of samples.

4.4. Ablation studies

This section presents ablation studies conducted to evaluate the contributions of individual components within the Proposed Model. Specifically, we examine the impact of scaling the CAFormer architecture and substituting standard self-attention mechanisms with focal self-attention. These analyses are designed to elucidate the extent to which these modifications enhance the model's overall efficiency in detecting skin cancer. To achieve this, the model's performance was systematically assessed across different architectural scales, including S18, S36, M36, and B36. Alongside performance metrics such as accuracy and precision, the ablation studies also consider the number of parameters, FLOPs (Floating Point Operations Per Second), and inference time for each scale. This approach provides a holistic evaluation of how scaling affects not only the model's diagnostic capabilities but also its computational efficiency and real-world applicability. These results enable a deeper understanding of the trade-offs between parameter count, computational complexity, and operational speed, contributing to the optimal design of the Proposed Model for practical deployment.

4.4.1. Performance enhancements through scaling and focal self-attention

We systematically assessed the impact of scaling the CAFormer architecture and integrating focal self-attention on model performance. Scaling was implemented by increasing the depth and width of the architecture, enabling the model to capture more complex features and improving its learning capacity. The inclusion of focal self-attention further enhanced the model's ability to prioritize significant regions within the input images. As detailed in Table 4, these advancements contributed to substantial improvements in accuracy across different configurations evaluated using an NVIDIA RTX 3090 GPU, a batch size of 16, and an image size of 224 × 224.

Table 4 provides a comprehensive comparison of the CAFormer variants and the proposed model, evaluated using an RTX 3090 GPU with a batch size of 16 and an image size of 224 × 224. Key metrics include the number of model parameters (in millions, M), computational complexity measured in GFLOPs (giga floating-point operations), inference speed in milliseconds (ms), and classification accuracy on the ISIC 2019 and HAM10000 datasets.

Scaling the CAFormer architecture increases the model's depth and width, enhancing its capacity to learn from complex data and improving performance metrics such as accuracy, precision, recall, and F1-score. For instance, CAFormer-S18 has 24.31 million parameters and a complexity of 3.895 GFLOPs, while CAFormer-B36 scales up to 95.71 million parameters and 22.499 GFLOPs. However, performance gains were not consistent, as the accuracy on the ISIC 2019 dataset improved from 90.56 % in CAFormer-S18 to 90.76 % in CAFormer-S36 but declined to 90.29 % and 89.74 % in CAFormer-M36 and CAFormer-B36, respectively, suggesting diminishing returns with excessive scaling.

Focal self-attention enhances the model's ability to focus on critical regions within an image, significantly reducing noise and improving feature extraction for more accurate skin cancer detection. This mechanism complements scaling by addressing its limitations, enabling the model to better generalize across diverse lesion types. The integration of scaling and focal self-attention in the proposed model results in a synergistic effect, enhancing its generalization capabilities and robustness.

With 35.01 million parameters and a computational complexity of 7.132 GFLOPs, the proposed model strikes a balance between computational cost and performance. Its inference time of 0.324 ms underscores its efficiency. The proposed model achieves the highest accuracy across both datasets, with 92.54 % on ISIC 2019 and 95.01 % on HAM10000, outperforming all other configurations. These results validate the combined effect of scaling and focal self-attention, demonstrating their critical role in improving feature extraction and overall model performance for skin lesion classification.

4.4.2. Generalization capability of the Proposed model

To comprehensively assess the generalization capability of the Proposed Model, a detailed class-wise performance analysis was conducted using the ISIC 2019 and HAM10000 datasets. Evaluating performance at the class level is essential to understanding the model's ability to accurately differentiate between various types of skin lesions. The classification report, presented in Table 5, offers valuable insights into the model's efficacy across different classes, providing a nuanced perspective on its strengths and potential areas for improvement.

The classification performance of the Proposed Model was

Table 4
Comparison of the Proposed Model with other former variants and the effects of scaling and focal self-attention.

Model	Params(M)	FLOPs (GFLOPs)	Inference (ms)	Accuracy ISIC 2019	Accuracy HAM1000
CAFormer-S18	24.31	3.895	0.190	0.9056	0.9232
CAFormer-S36	37.26	7.553	0.339	0.9076	0.9278
CAFormer-M36	53.92	12.745	0.368	0.9029	0.9244
CAFormer-B36	95.71	22.499	0.405	0.8974	0.9202
Scaled Model	31.41	5.523	0.253	0.9104	0.9376
Proposed Model (Scaled Model + Focal Self-attention)	35.01	7.132	0.324	0.9254	0.9501

Table 5
Class-wise performance of the Proposed Model on ISIC 2019 and HAM10000 datasets.

Class	Precision	Recall	F1-score	Number of images
ISIC 2019 dataset				
AK	0.8857	0.7126	0.7898	87
BCC	0.9231	0.9398	0.9313	332
BKL	0.9023	0.8817	0.8919	262
DF	1.000	0.8750	0.9333	24
MEL	0.9027	0.8808	0.8916	453
NV	0.9467	0.9658	0.9562	1288
SCC	0.7879	0.8387	0.8125	62
VASC	0.8846	0.9200	0.9020	25
Macro Average	0.9041	0.8768	0.8886	2533
Weighted Average	0.9251	0.9254	0.9248	2533
HAM10000 dataset				
AK	0.8529	0.8788	08,657	33
BCC	0.9796	0.9412	0.9600	51
BKL	0.9340	0.9000	0.9167	110
DF	1.0000	0.9167	0.9565	12
MEL	0.9020	0.8288	0.8638	111
NV	0.9606	0.9821	0.9712	670
VASC	1.0000	1.0000	1.0000	15
Macro Average	0.9470	0.9211	0.9334	1,002
Weighted Average	0.9497	0.9501	0.9495	1,002

comprehensively evaluated on the ISIC 2019 and HAM10000 datasets, revealing robust and consistent results across various skin lesion classes. A detailed class-wise analysis of the metrics provides valuable insights into the model's strengths and potential areas for refinement.

On the ISIC 2019 dataset, the Proposed Model achieved overall weighted average precision, recall, and F1-score of 0.9251, 0.9254, and 0.9248, respectively, across 2,533 images as seen in Fig. 9. Among the individual classes, the NV class recorded the highest F1-score (0.9562), reflecting the model's strong performance when dealing with a large sample size. Conversely, the AK class exhibited a lower F1-score (0.7898), which can be attributed to its smaller sample size of 87 images, highlighting the impact of class imbalance. Similarly, the DF class achieved perfect precision (1.000) but slightly lower recall (0.8750), indicating the model's ability to minimize false positives while occasionally missing true positive instances.

On the HAM10000 dataset, on the other hand, the Proposed Model achieved weighted average precision, recall, and F1-score of 0.9497, 0.9501, and 0.9495, respectively, across 1,002 images as seen in Fig. 9. The VASC class stood out with perfect scores (1.000) across all metrics, benefiting from a well-defined and smaller dataset of 15 images. In contrast, the MEL class had a relatively lower F1-score (0.8638) due to its lower recall (0.8288), suggesting some melanoma cases were misclassified. Notably, the NV class achieved the highest F1-score (0.9712), reflecting its robustness with many samples.

A comparison of macro and weighted averages underscores the model's ability to manage class imbalances effectively. The slightly higher weighted averages in both datasets demonstrate the model's exceptional performance for classes with larger sample sizes (e.g., NV), while the macro averages illustrate balanced performance across all classes. The Proposed Model showcases significant improvements over traditional CNN and ViT architectures. Its ability to achieve high precision and recall across both datasets demonstrates the effectiveness of the integrated focal self-attention mechanism and the scaled CAFormer architecture. Nevertheless, the relatively lower performance in minority classes (AK and SCC in ISIC 2019) suggests the need for further optimization, such as incorporating data augmentation techniques or class-specific weighting strategies to address imbalance issues.

4.5. Discussion

This study evaluated the Proposed Model on two benchmark datasets, ISIC 2019 and HAM10000, highlighting its capability to deliver

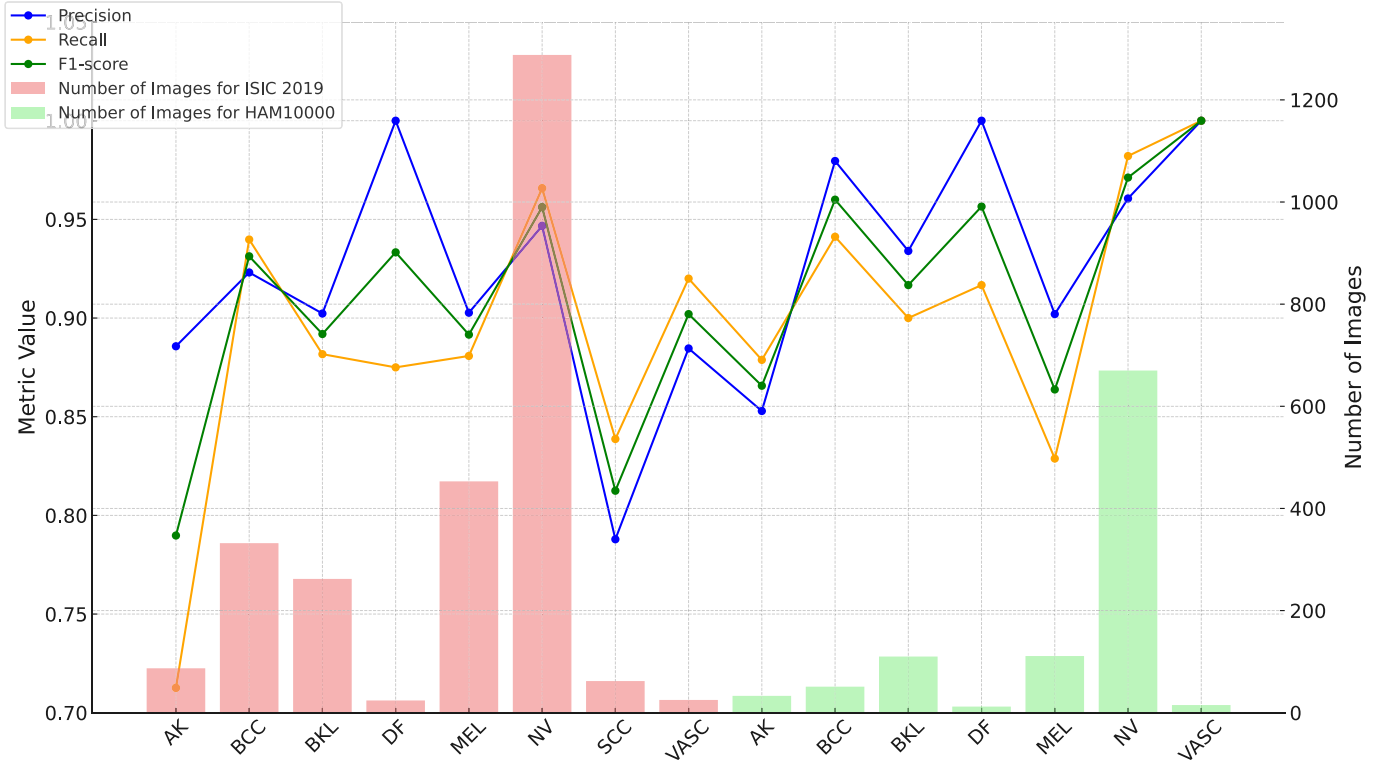

Fig. 9. Number of images and performance metrics for each class in the ISIC 2019 and HAM10000 datasets.

consistent and accurate performance across diverse skin lesion classes. On the ISIC 2019 dataset, the model demonstrated its strength with a weighted average F1-score of 0.9248, excelling in the NV class with an F1-score of 0.9562. However, the performance on minority classes such as AK and SCC was lower, reflecting challenges associated with imbalanced datasets. Similarly, the HAM10000 dataset results revealed strong overall metrics, with a weighted average F1-score of 0.9495. The VASC class achieved perfect precision, recall, and F1-scores due to its distinct features and small but consistent dataset. The MEL class showed slightly lower performance, primarily due to reduced recall, which emphasizes the difficulty of detecting certain melanoma cases.

The comparison of macro and weighted averages across both datasets highlights the Proposed Model's ability to balance performance across classes while maintaining high accuracy in the majority classes. This balance underscores the effectiveness of its architectural innovations, including the scaled CAFormer structure and the incorporation of focal self-attention mechanisms. These features enable the model to identify critical image regions and enhance feature extraction, resulting in reduced noise and improved classification.

Despite its high performance, the model faced limitations in classes with fewer samples, where metrics like recall and F1-score were comparatively lower. These findings suggest the potential benefit of employing data augmentation techniques or optimizing class-specific loss functions to mitigate the impact of class imbalance. Addressing these limitations could further enhance the model's applicability across a wider range of dermatological datasets.

The Proposed Model surpasses the performance of advanced CNN and ViT-based models, demonstrating its superior capability in handling intricate dermatological image analysis. With a lightweight design of 35.01 million parameters, the model is optimized for efficient deployment in real-time and resource-constrained environments. Its computational efficiency, combined with robust diagnostic accuracy, makes it a promising tool for clinical applications in skin cancer detection.

Future work should explore the integration of multimodal data, such as patient demographics and clinical metadata, to enhance classification performance. Moreover, extending the model to other dermatological conditions and validating its effectiveness in real-world clinical environments would broaden its applicability. These directions could solidify the model's role as a transformative tool in dermatological diagnostics.

5. Conclusion

This study focuses on the development of a state-of-the-art deep learning model for early skin cancer diagnosis, utilizing the CAFormer architecture enhanced with focal self-attention mechanisms and framed within the MetaFormer framework. The Proposed Model demonstrated exceptional classification performance on both the ISIC 2019 and HAM10000 datasets, achieving high accuracy, precision, recall, and F1-score metrics. By scaling the CAFormer architecture and integrating focal self-attention mechanisms, the model effectively identifies tumor regions and handles multi-class skin cancer classification challenges. Evaluation on the ISIC 2019 dataset revealed that the Proposed Model surpassed ten advanced CNNs and twenty leading ViTs in classification performance, maintaining robust generalization across classes despite significant class imbalances. Similarly, on the HAM10000 dataset, the model achieved consistently high performance across all metrics, showcasing its adaptability to different datasets with diverse image characteristics and class distributions. This dual evaluation underscores the model's robustness and reliability, which are essential for clinical applications to ensure consistent and trustworthy performance in varied and realistic scenarios. The model's design strikes an optimal balance between computational complexity and performance, enabling it to generalize effectively to unseen data. This characteristic, combined with its lightweight architecture, makes it highly suitable for clinical use, including deployment in resource-constrained environments. These attributes position the Proposed Model as a valuable tool for early skin cancer diagnosis and a new benchmark in AI-driven medical diagnostics.

CRediT authorship contribution statement

Ishak Pacal: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Burhanettin Ozdemir:** Visualization, Validation, Methodology, Investigation, Conceptualization. **Javanshir Zeynalov:** Writing – review & editing, Supervision, Conceptualization. **Huseyn Gasimov:** Writing – review & editing, Supervision, Conceptualization. **Nurettin Pacal:** Writing – review & editing, Validation, Data curation, Conceptualization.

Funding

This work was supported by the Grant provided by TÜSEB under the "2023-C1-YZ" call and Project No: "33934." We would like to thank TÜSEB for their financial support and scientific contributions. Experimental computations were carried out on the computing units at Igdir University's Artificial Intelligence and Big Data Application and Research Center.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

References

[1] U. Leiter, U. Keim, C. Garbe, Epidemiology of Skin Cancer: Update 2019, Adv. Exp. Med. Biol. 1268 (2020) 123–139, https://doi.org/10.1007/978-3-030-46227-7_6.
[2] H.M. Gloster, K. Neal, Skin cancer in skin of color, J. Am. Acad. Dermatol. 55 (2006) 741–760, https://doi.org/10.1016/J.JAAD.2005.08.063.
[3] B.K. Armstrong, A. Kricker, Skin cancer, Dermatol Clin 13 (1995) 583–594, https://doi.org/10.1016/s0733-8635(18)30064-0.
[4] V. Madan, J.T. Lear, R.M. Szeimies, Non-melanoma skin cancer, Lancet 375 (2010) 673–685, https://doi.org/10.1016/S0140-6736(09)61196-X.
[5] H.M. Gloster, D.G. Brodland, The Epidemiology of Skin Cancer, Dermatol. Surg. 22 (1996) 217–226, https://doi.org/10.1111/J.1524-4725.1996.TB00312.X.
[6] R.F. Thomas, J. Scotto, Estimating increases in skin cancer morbidity due to increases in ultraviolet radiation exposure, Cancer Invest 1 (1983) 119–126, https://doi.org/10.3109/07357908309042414.
[7] R.L. Siegel, A.N. Giaquinto, A. Jemal, Cancer statistics, 2024, CA Cancer J Clin (2024) 12–49, https://doi.org/10.3322/caac.21820.
[8] R. Gordon, Skin Cancer: An Overview of Epidemiology and Risk Factors, Semin Oncol Nurs 29 (2013) 160–169, https://doi.org/10.1016/J.SONCN.2013.06.002.
[9] A.F. Jerant, J.T. Johnson, C.D. Sheridan, T.J. Caffrey, Early Detection and Treatment of Skin Cancer, accessed June 23, 2024, Am Fam Physician 62 (2000) 357–368, https://www.aafp.org/pubs/afp/issues/2000/0715/p357.html.
[10] C. Anselmo Lima, M. Sampaio Lima, A. Maria Da Silva, M.A. Prado Nunes, M. M. Macedo Lima, M. Oliveira Santos, D. Lyra, C. Kleber Alves, Do cancer registries play a role in determining the incidence of non-melanoma skin cancers? Eur. J. Dermatol. 28 (2018) 169–176, https://doi.org/10.1684/EJD.2018.3248.
[11] I. Pacal, M. Alaftekin, F.D. Zengul, Enhancing Skin Cancer Diagnosis Using Swin Transformer with Hybrid Shifted Window-Based Multi-head Self-attention and SwiGLU-Based MLP, Journal of Imaging Informatics in Medicine (2024), https://doi.org/10.1007/s10278-024-01140-8.
[12] A. Karaman, I. Pacal, A. Basturk, B. Akay, U. Nalbantoglu, S. Coskun, O. Sahin, D. Karaboga, Robust real-time polyp detection system design based on YOLO algorithms by optimizing activation functions and hyper-parameters with artificial bee colony (ABC), Expert Syst Appl 221 (2023), https://doi.org/10.1016/j.eswa.2023.119741.
[13] Y. Lecun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444, https://doi.org/10.1038/nature14539.
[14] A. Maman, I. Pacal, F. Bati, Can deep learning effectively diagnose cardiac amyloidosis with 99mTc-PYP scintigraphy? J. Radioanal. Nucl. Chem. 2024 (2024) 1–16, https://doi.org/10.1007/S10967-024-09879-8.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, ICLR 2021 - 9th International Conference on Learning Representations (2020). https://arxiv.org/abs/2010.11929v2 (accessed August 7, 2023).

[16] S. Qasim Gilani, T. Syed, M. Umair, O. Marques, Skin Cancer Classification Using Deep Spiking Neural Network, J Digit Imaging 36 (2023) 1137–1147, https://doi.org/10.1007/s10278-023-00776-2.

[17] T. Mazhar, I. Haq, A. Ditta, S.A.H. Mohsan, F. Rehman, I. Zafar, J.A. Gansau, L.P. W. Goh, The Role of Machine Learning and Deep Learning Approaches for the Detection of Skin Cancer, Healthcare (switzerland) 11 (2023), https://doi.org/10.3390/healthcare11030415.

[18] Z. Mirikharaji, K. Abhishek, A. Bissoto, C. Barata, S. Avila, E. Valle, M.E. Celebi, G. Hamarneh, A survey on deep learning for skin lesion segmentation, Med Image Anal 88 (2023) 102863, https://doi.org/10.1016/j.media.2023.102863.

[19] H. Bhatt, V. Shah, K. Shah, R. Shah, M. Shah, State-of-the-art machine learning techniques for melanoma skin cancer detection and classification: a comprehensive review, Intelligent Medicine 3 (2023) 180–190, https://doi.org/10.1016/j.imed.2022.08.004.

[20] N. Melarkode, K. Srinivasan, S.M. Qaisar, P. Plawiak, AI-Powered Diagnosis of Skin Cancer: A Contemporary Review, Open Challenges and Future Research Directions, Cancers (Basel) 15 (2023), https://doi.org/10.3390/cancers15041183.

[21] M. Zafar, M.I. Sharif, M.I. Sharif, S. Kadry, S.A.C. Bukhari, H.T. Rauf, Skin Lesion Analysis and Cancer Detection Based on Machine/Deep Learning Techniques: A Comprehensive Survey, Life 13 (2023) 1–18, https://doi.org/10.3390/life13010146.

[22] A. Shah, M. Shah, A. Pandya, R. Sushra, R. Sushra, M. Mehta, K. Patel, K. Patel, A comprehensive study on skin cancer detection using artificial neural network (ANN) and convolutional neural network (CNN), Clinical EHealth 6 (2023) 76–84, https://doi.org/10.1016/j.ceh.2023.08.002.

[23] O. Attallah, Skin cancer classification leveraging multi-directional compact convolutional neural network ensembles and gabor wavelets, Scientific Reports | 14 (123AD) 20637. https://doi.org/10.1038/s41598-024-69954-8.

[24] O. Attallah, Skin-CAD: Explainable deep learning classification of skin cancer from dermoscopic images by feature selection of dual high-level CNNs features and transfer learning, Comput Biol Med 178 (2024) 108798, https://doi.org/10.1016/J.COMPBIOMED.2024.108798.

[25] E.H. Houssein, D.A. Abdelkareem, G. Hu, M.A. Hameed, I.A. Ibrahim, M. Younan, An effective multiclass skin cancer classification approach based on deep convolutional neural network, Cluster Comput (2024), https://doi.org/10.1007/s10586-024-04540-1.

[26] E. Goceri, Classification of skin cancer using adjustable and fully convolutional capsule layers, Biomed Signal Process Control 85 (2023) 104949, https://doi.org/10.1016/j.bspc.2023.104949.

[27] G. Akilandasowmya, G. Nirmaladevi, S.U. Suganthi, A. Aishwariya, Skin cancer diagnosis: Leveraging deep hidden features and ensemble classifiers for early detection and classification, Biomed Signal Process Control 88 (2024) 105306, https://doi.org/10.1016/J.BSPC.2023.105306.

[28] Q. Chen, M. Li, C. Chen, P. Zhou, X. Lv, C. Chen, MDFNet: application of multimodal fusion method based on skin image and clinical data to skin cancer classification, J Cancer Res Clin Oncol 149 (2023) 3287–3299, https://doi.org/10.1007/s00432-022-04180-1.

[29] A.A.M. Teodoro, D.H. Silva, R.L. Rosa, M. Saadi, L. Wuttisittikulkij, R.A. Mumtaz, D.Z. Rodríguez, A Skin Cancer Classification Approach using GAN and RoI-Based Attention Mechanism, J Signal Process Syst 95 (2023) 211–224, https://doi.org/10.1007/s11265-022-01757-4.

[30] K. Sethanan, R. Pitakaso, T. Srichok, S. Khonjun, P. Thannipat, S. Wanram, C. Boonmee, S. Gonwirat, P. Enkvetchakul, C. Kaewta, N. Nanthasamroeng, Double AMIS-ensemble deep learning for skin cancer classification, Expert Syst Appl 234 (2023) 121047, https://doi.org/10.1016/j.eswa.2023.121047.

[31] J.V. Tembhurne, N. Hebbar, H.Y. Patil, T. Diwan, Skin cancer detection using ensemble of machine learning and deep learning techniques, Multimed Tools Appl 82 (2023) 27501–27524, https://doi.org/10.1007/s11042-023-14697-3.

[32] T. Diwan, R. Shukla, E. Ghuse, J.V. Tembhurne, Model hybridization & learning rate annealing for skin cancer detection, Multimed Tools Appl 82 (2023) 2369–2392, https://doi.org/10.1007/s11042-022-12633-5.

[33] A.S. Qureshi, T. Roos, Transfer Learning with Ensembles of Deep Neural Networks for Skin Cancer Detection in Imbalanced Data Sets, Neural Process Lett 55 (2023) 4461–4479, https://doi.org/10.1007/s11063-022-11049-4.

[34] C.K. Viknesh, P.N. Kumar, R. Seetharaman, D. Anitha, Detection and Classification of Melanoma Skin Cancer Using Image Processing Technique, Diagnostics 13 (2023), https://doi.org/10.3390/diagnostics13213313.

[35] H. Tabrizchi, S. Parvizpour, J. Razmara, An Improved VGG Model for Skin Cancer Detection, Neural Process Lett 55 (2023) 3715–3732, https://doi.org/10.1007/s11063-022-10927-1.

[36] A. Dahou, A.O. Aseeri, A. Mabrouk, R.A. Ibrahim, M.A. Al-Betar, M.A. Elaziz, Optimal Skin Cancer Detection Model Using Transfer Learning and Dynamic-Opposite Hunger Games Search, Diagnostics 13 (2023) 1–20, https://doi.org/10.3390/diagnostics13091579.

[37] W. Yu, C. Si, P. Zhou, M. Luo, Y. Zhou, J. Feng, S. Yan, X. Wang, MetaFormer Baselines for Vision, IEEE Trans Pattern Anal Mach Intell 46 (2022) 896–912, https://doi.org/10.1109/TPAMI.2023.3329173.

[38] . Yu, M. Luo, P. Zhou, C. Si, Y. Zhou, X. Wang, J. Feng, S. Yan, MetaFormer Is Actually What You Need for Vision, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2022-June (2021) 10809–10819. https://doi.org/10.1109/CVPR52688.2022.01055.

[39] N.C.F. Codella, D. Gutman, M.E. Celebi, B. Helba, M.A. Marchetti, S.W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, A. Halpern, Skin Lesion Analysis Toward Melanoma Detection: A Challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), Hosted by the International Skin Imaging Collaboration (ISIC), Proceedings - International Symposium on Biomedical Imaging 2018-April (2017) 168–172. https://doi.org/10.1109/ISBI.2018.8363547.

[40] I. Pacal, MaxCerVixT: A novel lightweight vision transformer-based Approach for precise cervical cancer detection, Knowl Based Syst 289 (2024), https://doi.org/10.1016/j.knosys.2024.111482.

[41] I. Kunduracioglu, I. Pacal, Advancements in deep learning for accurate classification of grape leaves and diagnosis of grape diseases, J. Plant Dis. Prot. (2024), https://doi.org/10.1007/s41348-024-00896-z.

[42] I. Pacal, Enhancing crop productivity and sustainability through disease identification in maize leaves: Exploiting a large dataset with an advanced vision transformer model, Expert Syst Appl 238 (2024), https://doi.org/10.1016/j.eswa.2023.122099.

[43] I. Pacal, A novel Swin transformer approach utilizing residual multi-layer perceptron for diagnosing brain tumors in MRI images, Int. J. Mach. Learn. Cybern. (2024), https://doi.org/10.1007/s13042-024-02110-w.

[44] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-Decem (2016), pp. 770–778, https://doi.org/10.1109/CVPR.2016.90.

[45] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, 2015, pp. 1–14.

[46] G. Huang, Z. Liu, L. van der Maaten, K.Q. Weinberger, Densely Connected Convolutional Networks, (2016). http://arxiv.org/abs/1608.06993.

[47] C. Szegedy, S. Ioffe, V. Vanhoucke, A.A. Alemi, Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning, 31st AAAI Conference on Artificial Intelligence, AAAI 2017 (2016) 4278–4284. https://doi.org/10.1609/aaai.v31i1.11231.

[48] A. Howard, M. Sandler, B. Chen, W. Wang, L.C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, Q. Le, H. Adam, Searching for mobileNetV3, in, in: Proceedings of the IEEE International Conference on Computer Vision, Institute of Electrical and Electronics Engineers Inc, 2019, pp. 1314–1324, https://doi.org/10.1109/ICCV.2019.00140.

[49] I. Pacal, O. Celik, B. Bayram, A. Cunha, Enhancing EfficientNetv2 with global and efficient channel attention mechanisms for accurate MRI-Based brain tumor classification, Cluster Comput (2024), https://doi.org/10.1007/s10586-024-04532-1.

[50] C. Chen, Z. Guo, H. Zeng, P. Xiong, J. Dong, RepGhost: A Hardware-Efficient Ghost Module via Re-parameterization, (2022). http://arxiv.org/abs/2211.06088.

[51] W. Yu, P. Zhou, S. Yan, X. Wang, InceptionNeXt: When Inception Meets ConvNeXt, (2023). http://arxiv.org/abs/2303.16900.

[52] M. Tan, Q. V. Le, EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, 36th International Conference on Machine Learning, ICML 2019 2019-June (2019) 10691–10700. https://arxiv.org/abs/1905.11946v5 (accessed February 2, 2024).

[53] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A ConvNet for the 2020s, (2022). http://arxiv.org/abs/2201.03545.

[54] I. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, M. Lucic, A. Dosovitskiy, MLP-Mixer: An all-MLP Architecture for Vision, accessed June 23, 2024, Adv Neural Inf Process Syst 29 (2021) 24261–24272, https://arxiv.org/abs/2105.01601v4.

[55] J. Yang, C. Li, X. Dai, J. Gao, Focal Modulation Networks, accessed June 23, 2024, Adv Neural Inf Process Syst 35 (2022), https://arxiv.org/abs/2203.11926v3.

[56] S. Mehta, M. Rastegari, MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer, 3 (2021). http://arxiv.org/abs/2110.02178.

[57] H. Touvron, M. Cord, H. Jégou, DeiT III: Revenge of the ViT, Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 13684 LNCS (2022) 516–533. https://doi.org/10.1007/978-3-031-20053-3_30.

[58] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.

[59] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, B. Guo, Swin Transformer V2: Scaling Up Capacity and Resolution, (2021). http://arxiv.org/abs/2111.09883.

[60] H. Bao, L. Dong, S. Piao, F. Wei, BEiT: BERT Pre-Training of Image Transformers, (2021). http://arxiv.org/abs/2106.08254.

[61] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, Y. Li, MaxViT: Multi-axis Vision Transformer, Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 13684 LNCS (2022) 459–479. https://doi.org/10.1007/978-3-031-20053-3_27.

[62] A. Wang, H. Chen, Z. Lin, J. Han, G. Ding, RepViT: Revisiting Mobile CNN From ViT Perspective, n.d. https://github.com/pytorch/vision/tree/main/references/classification.

[63] S. d'Ascoli, H. Touvron, M. Leavitt, A. Morcos, G. Biroli, L. Sagun, ConViT: Improving Vision Transformers with Soft Convolutional Inductive Biases, (2021). https://doi.org/10.1088/1742-5468/ac9830.

[64] P.K.A. Vasu, J. Gabriel, J. Zhu, O. Tuzel, A. Ranjan, FastViT: A Fast Hybrid Vision Transformer using Structural Reparameterization, (2023). http://arxiv.org/abs/2303.14189.

[65] J. Li, X. Xia, W. Li, H. Li, X. Wang, X. Xiao, R. Wang, M. Zheng, X. Pan, Next-ViT: Next Generation Vision Transformer for Efficient Deployment in Realistic

Industrial Scenarios, (2022). https://arxiv.org/abs/2207.05501v4 (accessed June 23, 2024).
[66] C.F. Chen, Q. Fan, R. Panda, CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification, in: Proceedings of the IEEE International Conference on Computer Vision, 2021, pp. 347–356, https://doi.org/10.1109/ICCV48922.2021.00041.
[67] K. Wu, J. Zhang, H. Peng, M. Liu, J. Fu, L. Yuan, TinyViT: Fast Pretraining Distillation for Small Vision Transformers, n.d.
[68] A. Trockman, J.Z. Kolter, Patches Are All You Need?, (2022). https://arxiv.org/abs/2201.09792v1 (accessed June 23, 2024).