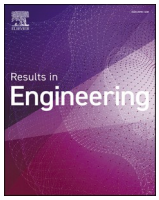
Results in Engineering 25 (2025) 103692

Contents lists available at ScienceDirect

Results in Engineering

journal homepage: www.sciencedirect.com/journal/results-in-engineering



An innovative deep learning framework for skin cancer detection employing ConvNeXtV2 and focal self-attention mechanisms

Burhanettin Ozdemir [a,*], Ishak Pacal [b]

[a] Department of Operations and Project Management, College of Business, Alfaisal University, 11533, Riyadh, Saudi Arabia
[b] Department of Computer Engineering, Faculty of Engineering, Igdir University, 76000, Igdir, Turkey

ARTICLE INFO

Keywords:
Medical image analysis
Skin cancer detection
ConvNeXtv2
Vision transformer
Focal self-attention

ABSTRACT

The skin, the body's largest organ, plays a critical role in protection and regulation, making its health essential. Skin cancer, one of the most prevalent malignancies, continues to rise globally and presents significant risks when diagnosis is delayed. Accurate detection is challenging due to the subtle and overlapping features of skin lesions, often leading to diagnostic errors. Deep learning has emerged as a powerful tool, capable of analyzing complex dermatological data and improving diagnostic accuracy through precise pattern recognition. This study proposes a novel lightweight and mobile-friendly hybrid model that combines ConvNeXtV2 blocks and focal self-attention mechanisms, addressing challenges such as data imbalance and model complexity. The Proposed Model employs ConvNeXtV2 in the first two stages for superior local feature extraction, while focal self-attention in the subsequent stages enhances sensitivity by focusing on diagnostically relevant regions. The Proposed Model was evaluated on the ISIC 2019 dataset, encompassing eight skin cancer classes with significant class imbalances, such as the Melanocytic Nevus class having 51 times more images than the Vascular Lesion class. Despite these disparities, the Proposed Model achieved robust performance across all classes, with 93.60% accuracy, 91.69% precision, 90.05% recall, and a 90.73% F1-score. Compared to baseline models and existing literature, it demonstrated a 10.8% improvement in accuracy over ResNet50 and a 3.3% improvement over the best-performing vision transformer (Swinv2-Base). This innovative design establishes a new benchmark in skin cancer detection, offering accurate, scalable, and generalizable predictions to support early diagnosis and improved clinical outcomes.

1. Introduction

The skin, the largest organ of the human body, serves as a critical interface between the internal systems and the external environment, performing numerous essential functions [1]. Structurally, it is composed of three primary layers: the epidermis, which forms the protective outermost barrier; the dermis, which provides structural support and houses vital components such as blood vessels and sensory receptors; and the hypodermis, the innermost layer responsible for insulation and energy storage [2,3]. Functionally, the skin acts as a protective shield, preventing the entry of harmful microorganisms, safeguarding against physical injuries, and regulating exposure to extreme temperatures [2,4]. Skin cancer specifically refers to the abnormal growth of skin cells and is primarily associated with prolonged exposure to ultraviolet radiation from the sun, although genetic predisposition, skin type, and environmental factors also contribute to its development. In 2024, it is projected that approximately 2001,140 new cancer cases and 611,720 cancer-related deaths will occur in the United States, with 108,270 of these cases and 13,120 fatalities attributed to skin cancer [5]. These statistics highlight the growing public health burden of skin cancer and emphasize the critical importance of early detection and timely intervention to mitigate its impact [6].

Skin cancer ranks among the most prevalent malignancies, posing significant diagnostic challenges due to the subtle and overlapping characteristics of skin lesions [7]. Traditional diagnostic methods, such as manual evaluations and classical machine learning techniques, are hindered by critical limitations, particularly when dealing with the variability and complexity of dermatological datasets [8]. Classical machine learning approaches often depend on extensive manual feature engineering and demonstrate reduced performance under conditions of data imbalance or limited dataset size. These methods frequently fail to detect the nuanced and intricate features required to accurately

* Corresponding author.
E-mail address: bozdemir@alfaisal.edu (B. Ozdemir).

https://doi.org/10.1016/j.rineng.2024.103692
Received 18 October 2024; Received in revised form 1 December 2024; Accepted 8 December 2024
Available online 9 December 2024

distinguish between benign and malignant lesions, resulting in subpar outcomes for accuracy, sensitivity, and specificity in practical applications [8]. Moreover, their reliance on static, predefined feature sets constrain their adaptability to diverse and complex datasets, rendering them less effective for large-scale or generalized diagnostic scenarios.

The progress of artificial intelligence, particularly deep learning, has revolutionized the field of medical image analysis [9,10]. Unlike traditional methods, deep learning excels at managing intricate and high-dimensional datasets through its end-to-end learning capabilities, removing the dependency on manual feature extraction [11]. This innovation has cemented deep learning as a key technology in medical imaging, driving advancements in early diagnosis, disease monitoring, and personalized treatment planning [12,13]. By learning directly from raw image inputs, these models deliver exceptional performance across tasks such as image segmentation, classification, and anomaly detection [14,15]. Their precision and efficiency have reshaped radiology, pathology, and related domains, enhanced diagnostic accuracy while minimizing human error [16–18]. As a result, deep learning now serves as a foundational element in medical image processing, fostering data-driven approaches that are transforming modern healthcare [19–21]. These models can autonomously extract detailed and meaningful features from extensive and heterogeneous datasets, addressing longstanding challenges such as data imbalance, variability, and limited dataset availability. Unlike classical approaches, deep learning significantly enhances diagnostic accuracy, sensitivity, and specificity, establishing itself as a cornerstone of modern medical diagnostics, especially in the detection of skin cancer [8].

Convolutional neural networks (CNNs) have emerged as the standard for medical image analysis, excelling in the identification of intricate patterns within image data [22]. CNNs are particularly effective for tasks such as classifying skin lesions and identifying cancerous abnormalities. Furthermore, the introduction of Vision Transformers (ViTs) represents a breakthrough in deep learning, enabling the analysis of large-scale image datasets and uncovering even more complex relationships within the data [23]. Together, these advances in deep learning provide a powerful, scalable, and reliable framework for developing automated tools for skin cancer detection, addressing the critical shortcomings of traditional diagnostic methods [24].

Several studies have extensively explored the application of deep learning in diagnosing skin cancer, highlighting its transformative potential in medical image analysis [25]. These investigations have demonstrated the efficacy of deep learning models, particularly CNNs, in tasks such as lesion classification, segmentation, and feature extraction. By utilizing noninvasive techniques, these models streamline the diagnostic process, offering improved accuracy and efficiency over traditional methods [26]. Research has also emphasized the importance of addressing limitations such as dataset variability, class imbalance, and reproducibility challenges to advance the development of reliable diagnostic tools. The integration of hybrid models and the exploration of novel architectures further underscore the growing potential of deep learning in enhancing skin cancer detection and classification, paving the way for more robust, scalable, and automated systems.

In the extensive exploration of deep learning and machine learning applications for diagnosing skin cancer, substantial progress has been made in understanding the efficacy of various algorithms, including artificial neural networks (ANNs), CNNs, and ViTs. Among these, CNNs and ViTs have consistently emerged as promising tools, demonstrating remarkable accuracy and efficiency in detecting and classifying skin lesions at early stages.

However, despite these advancements, the literature highlights persistent challenges, such as limited data availability, significant class imbalances, and difficulties in generalizing models across diverse datasets. The ability of models to adapt to variations in image resolution and lesion characteristics remains a critical obstacle. Additionally, integrating deep learning models into practical diagnostic workflows is complicated by high computational demands and the need for real-time processing [8]. This study seeks to address these limitations by introducing a lightweight, mobile-friendly hybrid model that combines ConvNeXtV2 blocks and focal self-attention mechanisms. By addressing issues such as dataset variability, class imbalance, and model complexity, the proposed framework offers an efficient and scalable solution for automated skin cancer detection. The contributions of the Proposed Model are outlined as follows:

1. The model integrates ConvNeXtV2 blocks for enhanced local feature extraction and focal self-attention mechanisms to prioritize diagnostically significant regions. This design effectively captures both intricate details and broader contextual information from dermatological images.
2. Proposed Model with advanced data augmentation and transfer learning techniques enable the model to handle substantial class imbalances within the ISIC 2019 dataset, such as the Melanocytic Nevus class containing 51 times more images than the Vascular Lesion class. This ensures robust and consistent performance across all skin cancer classes.
3. The Proposed Model achieves a notable 93.60% accuracy, 91.69% precision, 90.05% recall, and 90.73% F1-score, outperforming literature and baseline models with a 10.8% accuracy improvement over ResNet50 and a 3.3% improvement over the best-performing ViT model (Swinv2-Base). Despite its high performance, the model remains lightweight with only 36.44 million parameters, making it well-suited for real-time and mobile applications.
4. The Proposed Model sets a new standard in the field of skin cancer detection by addressing critical challenges identified in previous research. It provides a scalable, accurate, and generalizable diagnostic solution, advancing early detection and improving clinical outcomes.
5. The study provides a comprehensive evaluation of more than 25 advanced deep learning models, including 10 cutting-edge CNNs and more than 15 state-of-the-art ViTs, on the ISIC 2019 dataset. This detailed comparison helps identify the strengths and weaknesses of each architecture, offering valuable insights into the effectiveness of both traditional CNNs and modern ViT-based models for skin cancer detection.

By addressing these longstanding limitations and utilizing the strengths of both CNNs and ViTs, this study significantly contributes to the development of automated skin cancer diagnostic tools. The Proposed Model offers a practical and efficient framework that can seamlessly integrate into clinical practice, enhancing patient care and early intervention strategies.

2. Related works

Skin cancer remains a widespread health concern globally, emphasizing the need for advanced diagnostic methods to enhance early detection and treatment outcomes. The advent of deep learning has significantly transformed medical image analysis by providing sophisticated tools to handle complex and large-scale datasets. This section organizes the existing research on deep learning for skin cancer detection into three main categories: CNN-based models, which are effective in identifying localized features in lesions; transformer-based models, which excel at capturing global contextual information; and hybrid approaches, which merge the benefits of both methods to achieve superior results [27]. Each subsection examines the major contributions, challenges, and developments in these areas.

2.1. CNN-based models

CNNs have become foundational in medical image analysis, particularly for detecting and classifying skin lesions. Their layered structure enables them to extract local features like edges, textures, and shapes,

making them particularly adept at identifying patterns relevant to diagnosing skin cancer [28]. Popular architectures, including ResNet, VGG, and DenseNet, have demonstrated their efficacy across various tasks. This section explores the advancements achieved with CNN-based approaches, emphasizing their strengths and the ongoing challenges in addressing data diversity, class imbalance, and generalization.

Goceri [29] presented the design of a neural network 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. Akilandasowmya et al. [30] 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. 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. Sethanan et al. [31] 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. [32], in which deep learning have 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. Shukla et al. [33] introduced a hybrid approach combining a CNN with transfer learning and a random forest classifier for skin cancer detection. The model was tested on two datasets of benign and malignant skin moles, achieving a classification accuracy of up to 90.11%. The results demonstrate the model's feasibility and effectiveness for skin cancer classification.

Gilani et al. [34] 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 [35] 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 meta-data 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. [36] 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 CNN demonstrated superior performance, achieving a 91% accuracy rate after 100 epochs. Tabrizchi et al. [37] used the improved version of the trendy CNN architecture called VGG-16 to train their model. Their proposed method experimentally showed better accuracy. Chaturvedi et al. investigated an automated method for skin cancer classification using a MobileNet model pretrained on 1280,000 images from the 2014 ImageNet Challenge and fine-tuned on 10,015 dermoscopy images from the HAM10000 dataset. The model achieved an overall accuracy of 83.1% for seven classes, with top 2 and top 3 accuracies of 91.36% and 95.34%, respectively. The weighted averages for precision, recall, and F1-score were 89%, 83%, and 83%. This method shows potential in aiding dermatology specialists with critical decision-making stages [38]. Attallah [26] introduces SCaLiNG, a CAD tool designed to overcome diagnostic constraints by utilizing three compact CNNs and Gabor Wavelets (GW) to extract a comprehensive feature vector of spatial, textural, and frequency attributes. SCaLiNG processes images by decomposing them into directional sub-bands using GW, then trains CNNs on these sub-bands and the original image. By fusing attributes from CNNs trained on both the original images and GW-derived sub-bands, SCaLiNG enhances diagnostic accuracy through a more thorough representation of features.

The reviewed studies showcase various approaches to enhancing skin cancer detection. Convolutional capsule networks have been developed to improve spatial relationship learning in multi-class classification, while ensemble classifiers hybridizing CNNs with optimization techniques address real-time streaming challenges. Another method fused image and clinical data, achieving a notable accuracy improvement. Segmentation-based systems utilizing CNNs reached remarkable accuracy levels validated by experts, and hybrid methodologies combining advanced and conventional neural networks demonstrated high recall for benign cases. In comparison, our Proposed Model provides a lightweight, computationally efficient architecture tailored specifically for skin cancer detection. By integrating ConvNeXtV2 blocks and focal self-attention mechanisms, it achieves superior accuracy and efficiency, addressing the limitations of computationally intensive methods and enhancing practicality in real-time and resource-constrained environments.

2.2. Transformer-based models

Transformer-based architectures, particularly ViTs, have introduced new opportunities in image analysis by capturing relationships across broader regions of input data. These models utilize attention mechanisms to analyze both fine details and global patterns, making them well-suited for skin cancer detection [39]. This subsection discusses the impact of transformers in medical imaging, focusing on their capacity to improve performance on complex datasets and their potential to overcome limitations associated with traditional CNNs. While transformers deliver remarkable accuracy, their computational requirements often present challenges for practical, real-time applications.

Pacal et al. [40] 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. Their proposed model provided a high accuracy with solved data imbalance problem in ISIC 2019. Xin et al. followed a three-step procedure to verify the efficiency of SkinTrans in skin cancer classification. First, they established a VIT network, then used multi-scale and overlapping sliding windows for image serialization and patch embedding to focus on multi-scale features. Finally, contrastive learning ensured similar data encoded similarly while differentiating different data. Their model achieved 94.3% accuracy on the HAM10000 dataset and 94.1% on a clinical dataset, demonstrating the effectiveness of SkinTrans. The transformer network's success in vision tasks lays a strong foundation for multimodal skin cancer classification, benefiting dermatologists, researchers, and patients [41]. Cai et al. [42] introduced BiADATU-Net, a precision-focused segmentation model for skin cancer images that builds on the Transformer U-Net architecture. This model integrates deformable attention Transformers and bidirectional attention blocks in the encoder to adaptively learn

both global and local features. In the decoder, it incorporates scSE attention modules within the skip connections to enhance feature fusion by capturing image-specific context.

Ramkumar et al. [43] introduced a model combining Residual Learning Machines, Swin Transformers, and Fast Neural Networks (FNN) to handle non-uniformly distributed data effectively and improve diagnostic accuracy. The model was tested on skin cancer datasets, including ISIC-2008, PH-2, and HM007, using metrics like MCC, recall, F1-score, specificity, accuracy, and precision. Comparisons with prior approaches demonstrated its superiority, achieving 98.78% accuracy, 98.7% precision, 98.7% F1-score, 98.64% recall, and an MCC of 0.9863 across multiple trials. Dwivedi et al. [44] introduced a lightweight B-16 Vision Image Transformer (LViT) model for accurate skin lesion classification into various types of skin cancer using transfer learning. Their experiments, conducted on an extensive dataset, demonstrated the model's high accuracy, sensitivity, specificity, and strong generalization to new images. Desale and Patil [45] proposed a comprehensive methodology for skin cancer classification involving multiple stages. Preprocessing combines techniques like piecewise linear bottom hat filtering, adaptive median filtering, Gaussian filtering, and an enhanced gradient intensity method. Segmentation is performed using the self-sparse watershed algorithm, followed by feature extraction using a hybrid Walsh-Hadamard Karhunen–Loeve expansion technique. The final classification employs an improved vision transformer. The approach is validated on the ISIC 2019 database, achieving exceptional results: 99.81% accuracy, 96.65% precision, 98.21% sensitivity and recall, 97.42% F-measure, 99.88% specificity, 98.54% Jaccard coefficient, and 98.89% MCC. Attallah [46] introduced "Skin-CAD," an advanced and interpretable AI-driven computer-aided diagnosis (CAD) system for classifying dermatoscopic images of skin cancer. The system effectively categorizes images into two main groups—benign and malignant—and further distinguishes between seven specific skin cancer subtypes. Skin-CAD achieved a maximum accuracy of 97.2% for distinguishing malignant from benign cases and 96.5% on the HAM10000 dataset.

Various studies have proposed innovative approaches to skin cancer detection using transformer-based and hybrid models. Improvements to Swin Transformer have enhanced accuracy, training speed, and parameter efficiency while addressing data imbalance issues. Other approaches, such as SkinTrans, employed multi-scale features and contrastive learning to achieve over 94% accuracy on multiple datasets. Precision-driven models like BiADATU-Net combined deformable attention and bidirectional attention blocks for superior segmentation performance. Additionally, hybrid frameworks integrating Residual Learning Machines, Swin Transformers, and Fast Neural Networks have demonstrated exceptional accuracy and balanced performance across multiple metrics. Lightweight models like B-16 Vision Image Transformer and advanced methods combining complex preprocessing, feature extraction, and segmentation have also achieved state-of-the-art results, including nearly perfect classification metrics on datasets like ISIC 2019. Compared to these methods, our Proposed Model offers unique advantages. While many transformer-based models provide high accuracy, their computational intensity and complexity reduce their practicality in real-time or resource-limited environments. In contrast, our lightweight model, specifically optimized for skin cancer detection, balances computational efficiency with superior performance, making it a more practical solution for diverse applications.

2.3. Hybrid approaches

Hybrid models, which integrate CNNs and transformers, offer a promising solution by combining their complementary strengths. These models harness the detailed feature extraction capabilities of CNNs with the expansive contextual analysis provided by transformers. As a result, they address issues like data imbalance and variability while maintaining computational efficiency. This section reviews hybrid approaches in skin cancer detection, highlighting their ability to set new benchmarks by balancing precision and recall. Additionally, recent innovations, such as combining ConvNeXt blocks with self-attention mechanisms, are discussed for their role in advancing the field.

Teodoro et al. [47] 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 GAN, generating masks with a U-Net model. This model showed solid results and provided a baseline for upcoming studies in skin lesion classification. Hybrid deep architectures in skin cancer detection have been explored by Diwan et al. [48] 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. Dahou et al. [49] 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. Datta et al. aimed to enhance the classification of skin lesions by implementing a Soft-Attention mechanism to emphasize important features and suppress noise. They compared VGG, ResNet, Inception ResNet v2, and DenseNet architectures with and without Soft-Attention. The networks with Soft-Attention outperformed the baseline by 4.7%, achieving a precision of 93.7% on the HAM10000 dataset and improved the sensitivity score by 3.8%, reaching 91.6% on the ISIC-2017 dataset. The study demonstrates the effectiveness of Soft-Attention in improving skin lesion classification [50].

Monika et al. presented a project focused on detecting and classifying various types of skin cancer using machine learning and image processing tools. The pre-processing stage involves using dermoscopic images, with the Dull Razor method to remove unwanted hair particles and Gaussian and Median filters for image smoothing and noise filtering. Color-based k-means clustering is used for segmentation, and features are extracted using ABCD and GLCM methods. The ISIC 2019 Challenge dataset, consisting of eight types of dermoscopic images, is used for experimental analysis. The Multi-class Support Vector Machine (MSVM) achieved an accuracy of 96.25% [51]. Dorj et al. utilized a pre-trained AlexNet model and an ECOC SVM classifier to classify four types of skin cancer from 3753 images. Their algorithm achieved high average accuracy, sensitivity, and specificity, with maximum values of 95.1% for squamous cell carcinoma, 98.9% for actinic keratosis, and 94.17% for squamous cell carcinoma. Minimum values were 91.8% for basal cell carcinoma, 96.9% for squamous cell carcinoma, and 90.74% for melanoma, demonstrating robust performance across different metrics [52]. Toprak and Aruk [53] introduced a robust model for skin cancer classification that integrates multiple advanced techniques. The approach begins with DeepLabV3+ for precise segmentation of skin lesions in dermoscopic images. Features are extracted using three pretrained models, MobileNetV2, EfficientNetB0, and DenseNet201, ensuring balanced performance and comprehensive feature learning. These features are concatenated and refined using the ReliefF algorithm to select the most relevant ones. Wang et al. [54] proposed a two-stage approach for melanoma detection. In the first stage, Style Generative Adversarial Networks with Adaptive Discriminator Augmentation are used to synthesize realistic and diverse melanoma images, which are combined with the original dataset to create an augmented dataset. In the second stage, a ViT (BatchFormer) is employed with a dual-branch training strategy to extract features and classify skin lesions as melanoma or non-melanoma using the augmented dataset.

Several studies have introduced groundbreaking approaches to enhance skin cancer detection through advanced AI methodologies.

EfficientAttentionNet employs sophisticated preprocessing, GAN-based class balancing, and U-Net-generated masks to establish a solid benchmark for skin lesion classification, setting the stage for future advancements. Hybrid architectures have addressed challenges like vanishing gradients by using smaller convolutional filters, incorporating skip connections, and adopting cyclic learning rate annealing, achieving top-tier performance on datasets such as HAM10000. Other models, such as those built on MobileNetV3 with optimization algorithms like Hunger Games Search and Particle Swarm Optimization, have delivered impressive accuracy on datasets like ISIC-2016 and PH2. Furthermore, soft-attention mechanisms have significantly enhanced classification outcomes by amplifying key features and reducing irrelevant noise, improving precision and sensitivity. In contrast, our Proposed Model provides a streamlined and highly effective approach tailored specifically for skin cancer detection. By integrating ConvNeXtV2 blocks and focal self-attention mechanisms, it achieves exceptional accuracy while maintaining computational efficiency, addressing the shortcomings of resource-intensive and overly complex models. This balance of performance and practicality makes it an ideal choice for real-world applications.

2.4. Problem statement (Formulation)

Accurate and efficient skin lesion classification is critical for early detection and treatment of skin cancers such as melanoma, basal cell carcinoma, and squamous cell carcinoma. Traditional CNNs excel at extracting local features but often struggle with capturing long-range dependencies essential for understanding global image context. Conversely, attention-based models, while effective in capturing global relationships, can be computationally intensive and lack the inductive biases that make CNNs efficient for localized feature learning.

The challenge lies in designing a model that can effectively combine the strengths of these approaches to address the complexity of skin lesion patterns while maintaining computational efficiency. This study addresses this gap by proposing a hybrid deep learning framework that integrates ConvNeXtV2 blocks for localized feature extraction and focal self-attention layers for capturing global dependencies. The architecture is optimized for processing high-resolution dermatoscopic images, ensuring robust performance across diverse lesion types.

3. Methods and materials

In this section, we draw on the ISIC 2019 dataset, one of the most comprehensive and diverse resources available for skin cancer detection. This dataset, which includes a wide variety of skin lesion types, provides an ideal foundation for evaluating the performance of our approach. We employ cutting-edge deep learning techniques, centering on a novel hybrid model that integrates ConvNeXtV2 [55] blocks with focal self-attention mechanisms [56]. This combination allows the Proposed Model to achieve exceptional sensitivity and specificity in identifying and classifying skin cancers. By harnessing the power of ViTs alongside advanced data augmentation and transfer learning, the model effectively captures both fine details and broader contextual patterns in the images. To ensure that our research contributes to the broader scientific community and inspires further work in the detection of cancer-related diseases, we have included comprehensive details on the model's implementation and training processes.

3.1. Dataset

The ISIC 2019 dataset is widely regarded as one of the most comprehensive and valuable publicly available resources for research in skin cancer detection [57]. Developed by the International Skin Imaging Collaboration, this dataset plays a crucial role in advancing deep learning and AI-based solutions for skin cancer diagnosis. In addition to a vast collection of dermoscopic images, it also includes detailed demographic data and clinical metadata tied to skin lesion diagnoses. Researchers frequently use the ISIC 2019 dataset for tasks such as early melanoma detection and skin cancer classification, utilizing its training, validation, and testing subsets. This rich resource has helped drive the development of cutting-edge deep learning techniques, setting new standards in the field. Fig. 1 showcases a few example images from the different classes within the ISIC 2019 dataset.

The ISIC 2019 dataset comprises 25,331 labeled images classified into eight skin lesion categories: 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). These images vary in resolution, ranging from 576×768 to 1024×1024 pixels across 101 sizes, and are fully colored with three channels. A critical limitation of this dataset is the severe class imbalance; for instance, NV images outnumber VASC images by approximately 51 to 1. This disparity complicates model training and affects classification accuracy, necessitating targeted strategies for algorithm development and performance optimization.

3.2. Proposed model

In this study, we propose a hybrid model for skin lesion classification that combines the strengths of convolutional and attention-based architectures. The model utilizes ConvNeXtV2 blocks in its early stages to extract local and hierarchical features, followed by focal self-attention layers that capture long-range dependencies in the later stages. This combination allows the model to benefit from both convolutional inductive biases and the flexibility of transformer-based attention mechanisms. The architecture is designed to effectively classify multiple types of skin lesions, including melanoma, basal cell carcinoma, and squamous cell carcinoma. The overall architecture is composed of four stages. Each stage begins with a downsampling operation, followed by a ConvNeXtV2 block in the first two stages and focal self-attention layers in the last two stages. The input images of size 224×224 pixels are progressively downsampled, reducing the spatial dimensions while increasing the feature channels at each stage. This hybrid structure ensures a balance between computational efficiency and representation learning. The detailed architecture is shown in Fig. 2.

As seen in Fig. 2, the proposed architecture is organized into four stages, each designed to progressively downsample the input dermatologic image and increase the number of feature channels. The overall structure follows a pattern of 4–4-12–4 layers, where the first two stages employ ConvNeXtV2 blocks, and the last two stages use focal self-attention layers. Both components represent some of the most current and popular architectures in deep learning, offering cutting-edge advancements in feature extraction and attention mechanisms, which are highly relevant to the automated diagnosis of skin cancer.

Stage 1 consists of 4 ConvNeXtV2 blocks, which extract local features from the dermatologic image input. ConvNeXtV2 is one of the most recent convolutional architectures, optimized for efficient computation while maintaining strong performance across various computer vision tasks, making it an ideal choice for the precise extraction of skin lesion characteristics. Stage 2 includes another 4 ConvNeXtV2 blocks, further refining these features and reducing the spatial dimensions. Stage 3 employs 12 focal self-attention layers, designed to capture long-range dependencies by focusing on specific regions of the dermatologic image. This layer structure enhances the model's ability to identify subtle and complex patterns that could indicate skin cancer, such as melanoma or basal cell carcinoma. Stage 4 includes 4 additional focal self-attention layers, leading to the final feature representation used for classification.

As shown in Fig. 2, this structure is inspired by the MetaFormer architecture, which integrates convolution and attention mechanisms to offer a more effective model. By combining the local feature extraction capabilities of ConvNeXtV2 blocks with the global attention modeling of focal self-attention, the architecture achieves a balance between

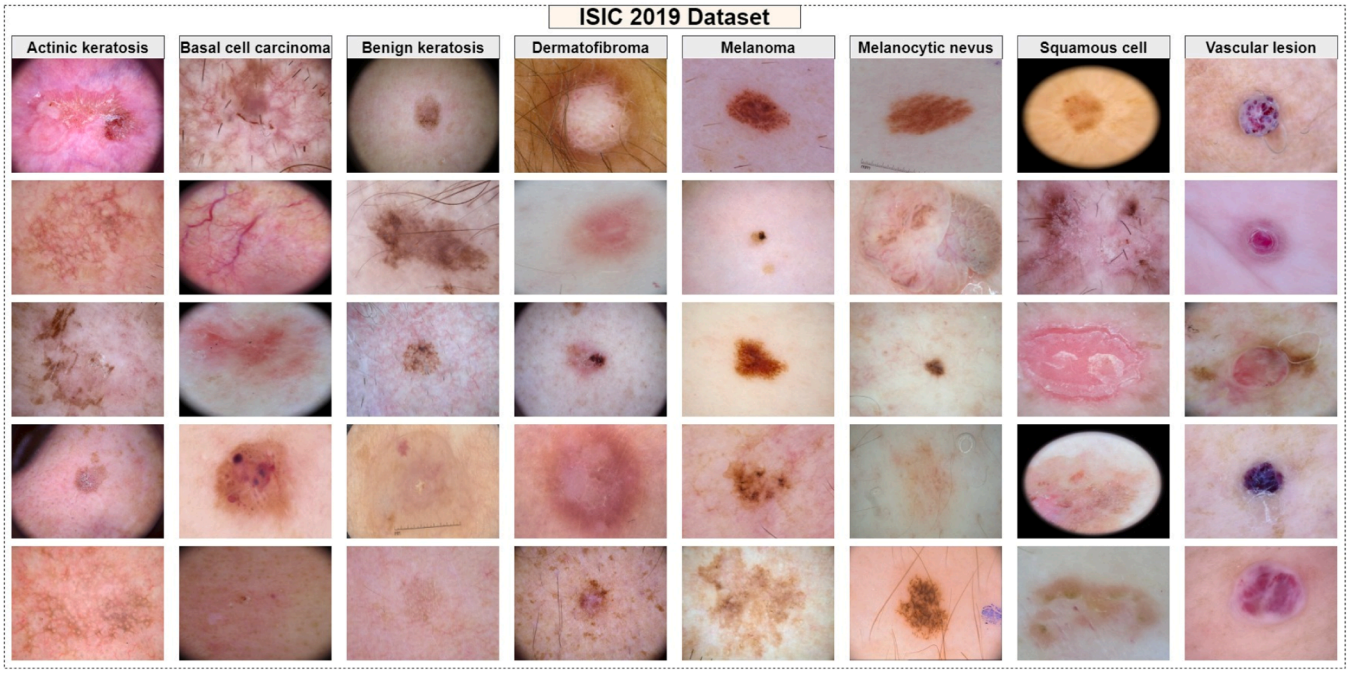


Fig. 1. Some sample images by class from the ISIC 2019 dataset.

parameter efficiency and computational complexity, benefiting from the most state-of-the-art techniques in both domains. This balance is crucial for the autonomous diagnosis of skin cancer, where rapid and accurate decision-making is essential.

Moreover, this design is not only computationally efficient but also optimized in terms of performance. The reduced parameter count, coupled with the effective handling of both local and global information, allows the model to achieve superior classification accuracy while maintaining lower computational costs. This makes it an excellent candidate for real-time, autonomous diagnosis of skin cancer, as the model can efficiently process a large volume of dermatologic images while maintaining high diagnostic accuracy. In comparison to conventional models, our approach offers a more optimized solution, excelling in terms of parameter efficiency, computation, and classification accuracy. By carefully balancing the number of convolutional and attention-based layers, we ensure that the model is capable of generalizing effectively across a diverse set of skin lesion types. The integration of ConvNeXtV2 and focal self-attention, two of the most widely adopted architectures, further strengthens the model's robustness, allowing it to autonomously detect and classify skin cancers with high precision and reliability using dermatologic images. The Proposed Model integrates ConvNeXtV2 blocks and focal self-attention mechanisms, combining their strengths to achieve superior performance in dermatological image analysis is depicted in Algorithm 1.

As seen in Algorithm 1, the proposed algorithm outlines the architecture of the model, emphasizing its sequential and modular design. It begins with the local feature extraction stages (Stages 1 and 2) using ConvNeXtV2 blocks, which excel in identifying fine-grained features such as edges, textures, and other localized patterns within skin lesion images. These blocks utilize depthwise convolutions and expanded kernel sizes, combined with layer normalization, to enhance spatial processing and stability. The next two stages (Stages 3 and 4) incorporate focal self-attention mechanisms, which refine the features further by focusing on diagnostically significant regions. This mechanism assigns higher importance to key regions while maintaining global context, ensuring sensitivity to critical features like irregular pigmentation and shape anomalies. Finally, the processed features are passed through a fully connected classification layer, where softmax activation is applied to generate class probabilities. This structured approach ensures the model effectively addresses challenges such as class imbalance, data variability, and the need for robust and accurate predictions in skin cancer detection.

3.2.1. *ConvNeXtv2 block*

ConvNeXtV2 Blocks form the foundation of the model's feature extraction pipeline. These blocks are designed to utilize the efficiency of convolutional operations while incorporating advanced features inspired by ViTs. A key innovation in ConvNeXtV2 is the use of expanded kernel sizes, which enable the model to capture broader spatial relationships in images, a critical requirement for identifying irregularities in skin lesions. Additionally, dynamic activation functions enhance the adaptability of the model, allowing it to process diverse lesion patterns more effectively. To further stabilize the training process, layer normalization is employed, ensuring robust performance across varying batch sizes. These capabilities make ConvNeXtV2 particularly effective in the initial stages of the model, where detailed local feature extraction is essential. As seen in Fig. 2, at the core of each ConvNeXtV2 block is the depthwise convolution operation, which significantly reduces the computational cost by applying convolution operations independently on each channel of the input feature map. Given an input feature map $X \in R^{H\times W\times C}$, the depthwise convolution with a kernel K of size k x k is applied separately to each channel, which can be expressed as:

$$X^c_{\text{out}} = K^c_{\text{depthwise}} * X^c \quad \forall c \in [1, C] \tag{1}$$

Here, X^c_{out} represents the output for the c-*th* channel, and this operation allows the model to capture spatial features more effectively while reducing computational demands. Following the depthwise convolution, the block applies Layer Normalization (LN), which stabilizes the training process by normalizing across the feature map dimensions. The normalized output is then passed through the GELU activation function, a smoother activation function that allows for better gradient flow, leading to improved convergence in deep models. The normalization process is represented as:

$$\widehat{X} = \frac{X - \mu}{\sigma} \tag{2}$$

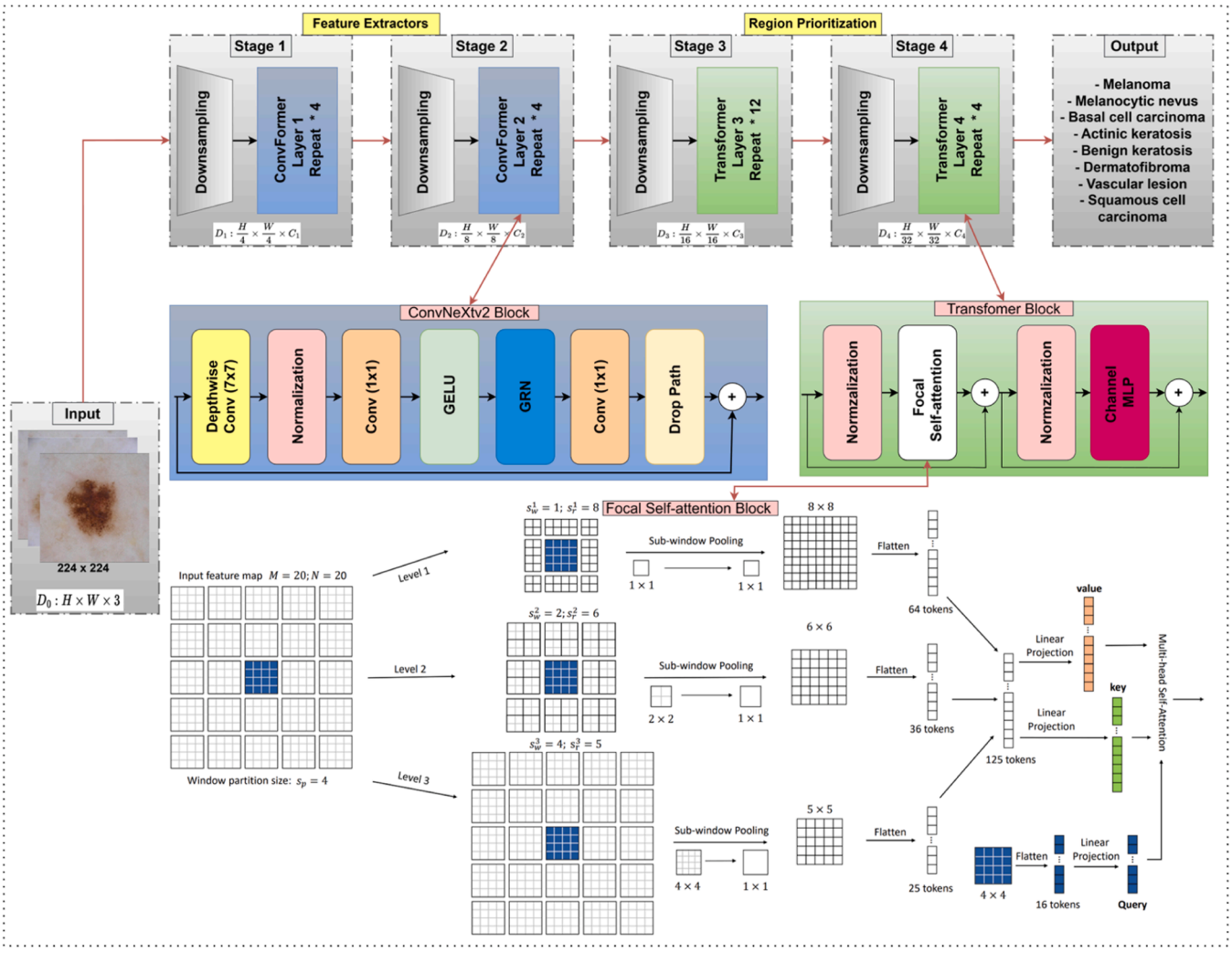


Fig. 2. The detailed structure of the Proposed Model for the autonomous diagnosis of skin cancer.

Algorithm 1
Proposed model algorithm: ConvNeXtV2 + Focal self-attention.

Require: Input Image I of size 224×224
Ensure: Classification Output O
1: **Stage 1: Local Feature Extraction**
2: Perform initial convolution on I using ConvNeXtV2 with depthwise convolutions
3: Apply layer normalization for stability
4: Extract local features F_1
5: **Stage 2: Enhanced Local Feature Extraction**
6: Use an additional ConvNeXtV2 block to process F_1
7: Expand kernel size to capture broader spatial relationships
8: Obtain enhances features F_2
9: **Stage 3: Global and Hierarchical Attention**
10: Apply focal self-attention on F_2
11: Assign higher weights to diagnostically significant regions
12: Integrate global context with local attention to form F_3
13: **Stage 4: Prioritizing Key Regions**
14: Use another focal self-attention block on F_3
15: Focus attention on subtle but crucial patterns
16: Refine features to produce F_4
17: **Final Classification Layer**
18: Flatten F_4 and pass through a fully connected layer
19: Perform softmax activation for class probabilities
20: Output O
21: **return** O

where μ and σ are the mean and standard deviation of the input X, calculated over the feature map dimensions. This normalization helps the model handle different batch sizes and stabilizes training. A key feature of ConvNeXtV2 is the introduction of Global Response Normalization (GRN), which further optimizes feature map responses by reducing inter-channel redundancy. In medical imaging tasks like skin cancer detection, this is particularly beneficial, as it ensures that each channel in the feature map contributes meaningfully to the final prediction. The GRN is defined as:

$$X_{\text{GRN}} = \gamma \cdot \frac{X}{\| X\|_2} + \beta \tag{3}$$

where γ and β are learnable parameters, and $\| X\|_2$ represents the L2-norm of the input X across the feature channels. This operation enhances the model's ability to differentiate subtle variations in dermatoscopic images, which is crucial for distinguishing between benign and malignant lesions. Finally, the ConvNeXtV2 block includes residual connections to help mitigate the vanishing gradient problem. These connections allow the input to bypass the convolutional operations and be added directly to the output, forming a residual function. The final output of the ConvNeXtV2 block can be written as:

$$X_{\text{final}} = X_{\text{GRN}} + X_{\text{input}} \tag{4}$$

This residual connection not only speeds up the training process but also improves the accuracy of the model by ensuring that important information from earlier layers is retained. In the context of skin cancer diagnosis using dermatoscopic images, the ConvNeXtV2 block plays a vital role. Its ability to extract both local and global features from the input image helps capture important visual cues such as texture, color, and border irregularities. These features are crucial for detecting various types of skin cancer, including melanoma and basal cell carcinoma. By integrating ConvNeXtV2 blocks into the overall architecture, we ensure that the model is both computationally efficient and highly accurate, making it well-suited for real-world applications like autonomous skin cancer diagnosis.

3.2.2. Focal self-attention mechanism

Focal self-attention mechanisms, employed in the later stages of the model, enhance the features extracted by ConvNeXtV2 through a hierarchical attention strategy. Local attention focuses on small, diagnostically critical regions, such as irregular borders or abnormal pigmentation, while global attention captures the broader context, including spatial relationships and lesion structure. This combination ensures the prioritization of key regions while maintaining computational efficiency. This dual attention approach is essential for analyzing dermatological images, where both fine details like pigmentation and larger features like lesion shape are critical for skin cancer diagnosis. Focal self-attention targets nearby regions with precision while incorporating broader, coarse-scale information, allowing the model to optimize accuracy and efficiency simultaneously. By seamlessly combining local and global attention, the mechanism captures both short-range details and long-range dependencies. Local attention calculates weights for nearby tokens within a defined window, effectively recognizing localized patterns. The mathematical formulation for local attention is provided in Eq. (5).

$$\mathit{LocalAttention}(Q, K, V) = \mathit{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{5}$$

where Q, K, and V are the query, key, and value matrices, respectively, and d_k is the dimensionality of the key vectors. This approach guarantees that each token focuses on its immediate surroundings, allowing the model to capture local details with precision. At the same time, global attention zooms out to consider a more generalized, high-level context, aggregating features from various regions of the image. By doing so, the model efficiently captures broader structural information while keeping computational demands in check. Global attention operates similarly to local attention but is applied to pooled features, as described in Eq. (6), ensuring the model balances detailed analysis with an understanding of the overall image structure.

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

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 utilize their respective strengths. This integration is formulated as Eq. (7).

$$\begin{aligned}\mathit{FocalSelfAttention}(Q, K, V) = {} & \mathit{LocalAttention}(Q, K, V) \\ & + \mathit{GlobalAttention}(Q, K, V)\end{aligned} \tag{7}$$

The focal self-attention mechanism enhances computational efficiency by reducing the number of tokens processed at each stage while maintaining a broad receptive field. This design makes it particularly suitable for real-time dermatological image analysis in autonomous skin cancer diagnosis, where precision and speed are equally critical. By balancing the capture of fine, high-resolution details with broader contextual understanding, this mechanism excels in complex tasks such as skin lesion classification, where both local features and global patterns are crucial for accurate diagnosis. As illustrated in Fig. 2, the Focal Self-Attention mechanism operates across multiple levels of detail to effectively capture both localized and global patterns. For example, a 20 $\times$ 20 input feature map is divided into 5 $\times$ 5 windows, each measuring 4 $\times$ 4. Taking the central 4 $\times$ 4 query window as a focus, the mechanism gathers surrounding tokens at varying granularities. At the first level, it analyzes the closest 8 $\times$ 8 region, capturing detailed, fine-grained information. At the second level, it broadens its scope to nearby 2 $\times$ 2 sub-windows, forming a 6 $\times$ 6 region. At the third level, it extends attention to the entire feature map, pooling larger 4 $\times$ 4 sub-windows for a global perspective. Tokens from these three levels are combined to compute the keys and values for the 16 query tokens in the blue window. This multi-level attention approach effectively integrates close-up details with broader context, allowing the model to focus on the most diagnostically significant parts of the data. By doing so, it improves both efficiency and accuracy, making it particularly well-suited for skin cancer detection, where fine details and overall lesion structure are equally important.

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 and training details

The experiments in this study were conducted on a high-performance Linux system running Ubuntu 22.04. The system was equipped with state-of-the-art hardware, including an NVIDIA RTX 4090 GPU and an Intel i9–14900 K processor, supported by 64 GB of DDR5 RAM. All deep learning models were implemented and evaluated using PyTorch, utilizing NVIDIA CUDA for efficient and consistent computational performance. To ensure optimal training, data augmentation techniques were applied extensively to enhance the diversity of the training dataset and improve model generalization. These techniques included geometric transformations (scaling, flipping), smoothing, color jitter, and MixUp. Transfer learning was also utilized by initializing the models with pre-trained weights from the ImageNet dataset, enabling faster convergence and better performance, particularly on a relatively imbalanced dataset like ISIC 2019.

The input resolution for all images was standardized to 224 × 224 pixels, a commonly used size in deep learning frameworks. This resolution was chosen to balance computational efficiency with the retention of essential diagnostic features. Standardizing at 224 × 224 ensured compatibility with pre-trained models and architectures, such as ConvNeXtV2 and focal self-attention mechanisms, which are optimized for this resolution. Preliminary evaluations confirmed that this resolution maintained sufficient detail for distinguishing between benign and malignant lesions, even for minority classes. It also reduced computational overhead, making the model suitable for real-time applications and deployment in resource-limited environments. Uniform training parameters were applied across all experiments to maintain consistency and reproducibility. Key hyperparameters included a learning rate of 0.01, a base learning rate of 0.1, momentum of 0.9, weight decay of 2.0e−05, the SGD optimizer, 5 warmup epochs, and a warmup learning rate of 1.0e−05. These settings, along with the standardized resolution and augmentation strategies, ensured fair comparisons between models and reliable findings. This comprehensive setup, combining advanced hardware, standardized resolution, robust augmentation techniques, and consistent hyperparameter settings, provided a reliable foundation for training and evaluating the Proposed Model in addressing the challenges of skin cancer detection.

4.2. Data preprocessing and data augmentation

Data preprocessing is a critical step in fine-tuning deep learning models, as it significantly influences their overall performance. This process involves several key tasks, including dividing the dataset into training, validation, and test subsets, normalizing data values, minimizing noise, and addressing outliers. In contrast to traditional approaches that often employ two-way splits or cross-validation, this study adopted a three-way split for the ISIC 2019 dataset, consisting of distinct training, validation, and test sets. This approach enhances the model's evaluation by ensuring more reliable performance metrics and reducing the risk of overfitting. The class distribution for the dataset, summarized in Table 1, highlights the importance of this segmentation strategy in providing a robust foundation for assessing the model's effectiveness.

Table 1 provides an overview of the image counts across three subsets of the ISIC 2019 dataset. The data is split into training, validation, and testing sets at 70%, 20%, and 10%, respectively, aiming for a fair evaluation of model performance. The ISIC 2019 dataset, utilized in this study, comprises eight skin cancer classes with significant class imbalance, such as the NV class containing 51 times more samples than the VASC class. To mitigate this imbalance and enhance model performance, we employed advanced data augmentation techniques. These included geometric transformations such as random rotations, flips, and cropping to simulate variations in lesion orientation and size. Color adjustments, including alterations to brightness, contrast, and hue, were introduced to replicate changes in lighting conditions. Elastic deformations were applied to mimic the natural variability of skin texture and shape. Additionally, advanced methods such as CutMix and MixUp were used to combine samples and labels, fostering better generalization and robustness during training.

Table 1
Number of images for three subsets of the ISIC 2019 dataset.

Class	Total	Training set (%70)	Validation set (%20)	Test set (%10)
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

Beyond augmentation, the inherent design of the Proposed Model played a crucial role in addressing class imbalance. ConvNeXtV2 blocks, integrated into the initial stages, excel at extracting localized features, ensuring that subtle details from underrepresented classes are effectively captured. The focal self-attention mechanism further enhances the model's sensitivity by focusing on diagnostically significant regions, which is especially beneficial for minority classes. Together, these strategies ensured consistent performance across all classes, enabling the model to robustly handle the challenges posed by class imbalance.

4.3. Results

In this section, we present the experimental results and performance analysis of thirty advanced deep learning models, including ten cutting-edge CNNs and twenty state-of-the-art ViTs. These models are evaluated using the ISIC 2019 dataset, which is carefully divided into training, validation, and test subsets. A key distinction of this study is our focus on how well each model performs on unseen test data, moving beyond the commonly emphasized validation phase. This approach is particularly critical for applications like skin cancer detection, where the ability to generalize effectively to new, real-world cases is paramount. By rigorously assessing the models on previously unseen data, we provide a more accurate measure of their practical utility and diagnostic reliability. The evaluation begins with comprehensive training, where each model is carefully optimized through techniques such as data augmentation, learning rate scheduling, and regularization. The validation phase serves as an important checkpoint, allowing for hyperparameter adjustments and helping to prevent overfitting through early stopping. However, the true test of each model's robustness lies in the independent test phase, which provides a clear indication of how well the models can generalize to data outside of their training environment. This final stage is essential for understanding the models' potential for real-world clinical application.

Notable architectures among the 10 include ResNet50 [58], VGG16 [59], DenseNet169 [60], Inceptionv4 [61], MobileNetv3-Large [62], EfficientNetv2-Medium [63], RepGhostNet-100 [64], InceptionNext-Base [65], EfficientNet-B6 [66], and ConvNexT-Base [67]. 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 [68], PoolFormer-M36 [69], FocalNet-Base [70], MobileViT-Small [71], DeiT3-Base [72], Swin-Base [73], Swinv2-Base [74], BeiTv2-Base [75], MaxViT-Base [76], RepViT-m1 [77], ConViT-Base [78], FastViT-ma36 [79], NextViT-Base [80], CrossViT-Base [81], Tiny-ViT-21 m [82] and ConvMixer-768 [83]. These transformers are renowned for their ability to handle complex image recognition tasks, making them highly suitable for recognition of skin diseases.

The Proposed Model introduces a significant advancement in deep learning by scaling an integrated hybrid architecture that combines ConvNeXtV2 blocks with focal self-attention, replacing the standard self-attention mechanism. This thoughtful integration allows the model to capture fine-grained details while maintaining an understanding of the broader context within dermatological images, resulting in marked improvements in both accuracy and robustness. Throughout this study, we have prioritized optimizing the model to achieve state-of-the-art performance across multiple key metrics. These architectural enhancements have proven particularly valuable, as the model consistently excels in challenging test environments, demonstrating its capacity for generalization. The experimental results, which highlight the performance of various CNN and ViT-based models on the ISIC 2019 dataset, are presented in Table 2, showcasing the effectiveness of the Proposed Model.

Table 2
Experimental results of the deep learning-based models.

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	0.9360	0.9169	0.9005	0.9073

Table 2 provides an evaluation of several deep learning models applied to the ISIC 2019 dataset, focusing on their performance in skin cancer classification through metrics such as accuracy, precision, recall, and F1-score. Complementing this, Fig. 3 illustrates the performance metrics of all models using a radar chart, while Fig. 4 offers an in-depth comparison by highlighting only accuracy and F1-score. Analyzing these figures and the table reveals that the Proposed Model outperforms all others, achieving impressive metrics: 93.60% accuracy, 91.69% precision, 90.05% recall, and a 90.73% F1-score. This exceptional performance demonstrates the model's capability to effectively reduce false positives and false negatives, an essential attribute in medical diagnostics, where accurate identification of conditions can profoundly impact patient care.

Several ViT-based models show strong performance in skin cancer classification. Swinv2-Base and DeiT3-Base achieve high accuracies of 90.68% and 90.56%, respectively, with Swinv2-Base excelling in precision (90.13%) and recall (86.90%). Similarly, DeiT3-Base demonstrates solid generalization with a precision of 89.32% and a recall of 85.54%, though both lag slightly in recall compared to the Proposed Model. Traditional convolutional models like ResNet50 and VGG16, while respectable, are outperformed by transformers. ResNet50 achieves 85.35% accuracy and 76.67% recall, while VGG16 performs slightly better with 86.74% accuracy and 78.30% recall, highlighting their limitations in capturing complex global features. Newer transformer models, including MaxViT-Base and BeiTv2-Base, offer accuracies of 90.09% and 90.37%, respectively, utilizing attention mechanisms to capture local and global features. MaxViT-Base achieves 90.40% precision, though its recall of 83.89% suggests room for sensitivity improvement. While the Proposed Model leads in overall performance, these findings emphasize the growing importance of ViTs in achieving high accuracy and sensitivity for medical imaging, surpassing traditional models in diagnosing skin cancer effectively.

Fig. 3 and Fig. 4 present a radar chart and a line graph comparing the F1-scores and accuracy of deep learning models applied to the ISIC 2019 dataset, including traditional CNNs (ResNet50, VGG16, DenseNet169) and advanced ViT-based models (Swinv2-Base, DeiT3-Base), accordingly. F1-scores and accuracy are shown with blue and red dashed lines, respectively. The Proposed Model achieves the highest F1-score (0.9073) and accuracy (0.9360), effectively balancing precision and recall minimizing false positives and negatives, crucial in medical diagnostics. ViT-based models like Swinv2-Base and BeiTv2-Base also perform strongly, highlighting the advantage of attention mechanisms in capturing complex patterns. In contrast, traditional CNNs show more variability and lower performance. Fig. 5 provides the Proposed Model's confusion matrix, illustrating its superior class-specific performance and establishing it as a benchmark for autonomous skin cancer diagnosis.

As shown in Fig. 5, the Proposed Model achieves exceptional performance in diagnosing nevus (NV, 1247 TP), with a precision of 0.9558, recall of 0.9726, and an F1-score of 0.9641, demonstrating its strong

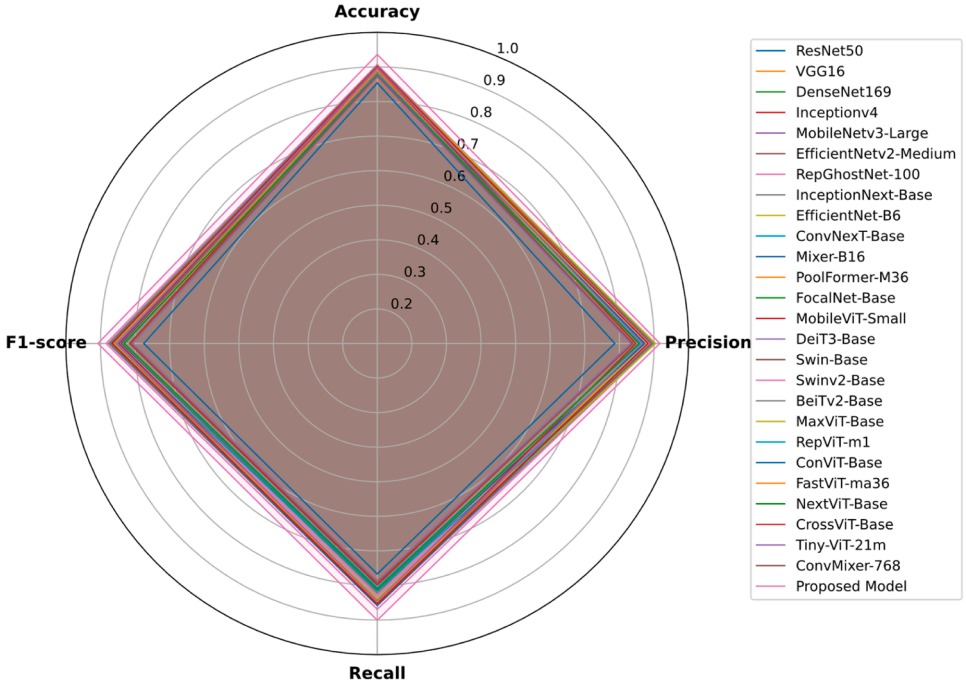


Fig. 3. Performance metrics for all models, including CNN, ViT-based models, and the Proposed Model.

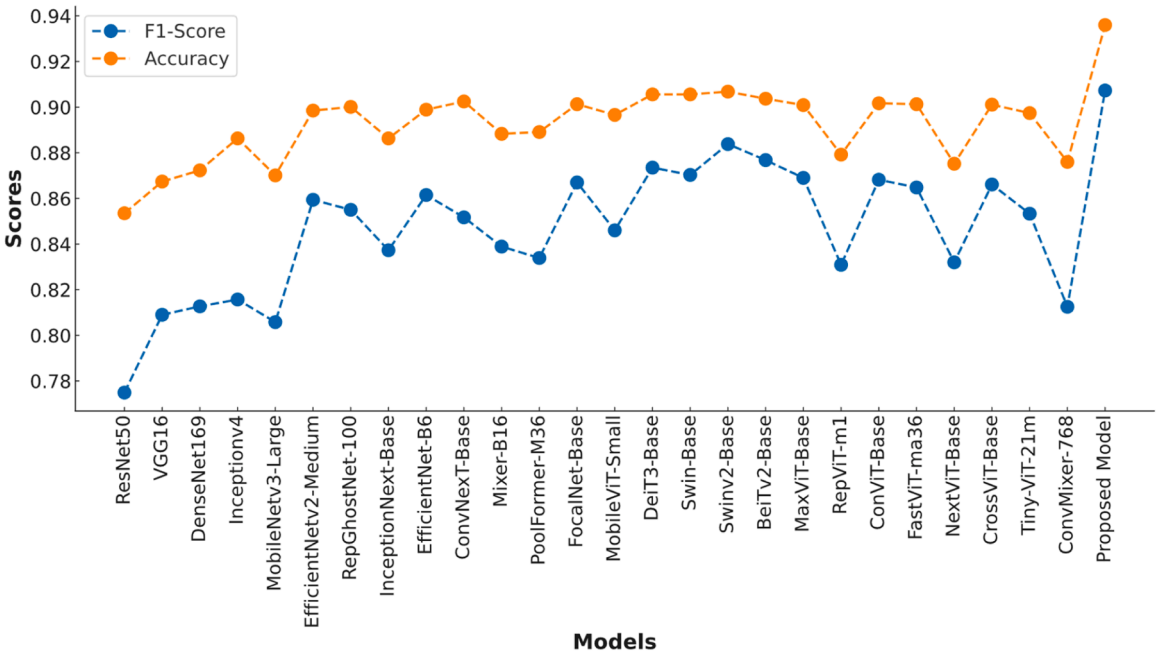


Fig. 4. Accuracy and F1-score metric for all models, including both CNN and ViT-based models, as well as the Proposed Model.

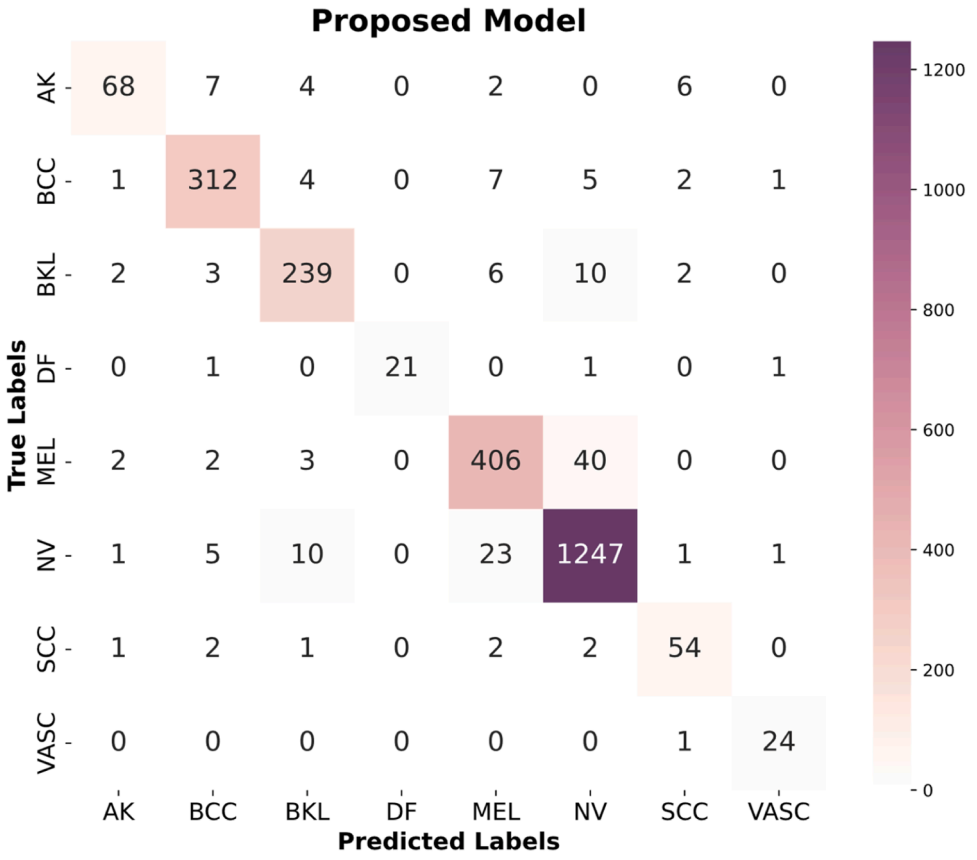


Fig. 5. The confusion matrix showing the class-specific performance of the Proposed Model.

capability to accurately classify this well-defined class. Similarly, high metrics are observed for basal cell carcinoma (BCC, 312 TP) and melanoma (MEL, 406 TP), with robust precision, recall, and F1-scores, reflecting the model's ability to generalize effectively to critical skin cancer types. However, the model exhibits challenges with more complex classes, such as actinic keratosis (AK, 68 TP) and squamous cell carcinoma (SCC, 54 TP). For AK, a precision of 0.9324 and recall of 0.7907 result in an F1-score of 0.8560, indicating a tendency to miss true positive cases. This challenge arises from AK's visual similarities to benign lesions like seborrheic keratosis and early-stage SCC, where

subtle differences in texture and pigmentation complicate differentiation. For SCC, the model achieves a precision of 0.8871, but its recall of 0.7260 and F1-score of 0.7983 highlight difficulties in capturing all true positives. These misclassifications are often due to SCC's overlapping morphological features with BCC and certain melanoma subtypes, further exacerbated by variations in image resolution and quality. These limitations emphasize the need for targeted enhancements to improve sensitivity in underperforming classes. Potential solutions include incorporating advanced data augmentation to simulate subtle variations, applying class-specific weighting to address underrepresented classes, and exploring ensemble learning strategies to strengthen recall.

4.4. Ablation studies

In this section, we conduct ablation studies to rigorously assess the individual contributions of key components within the Proposed Model. We focus on evaluating the effect of model scaling, the integration of ConvNeXtV2 blocks, and the substitution of standard self-attention with focal self-attention. These experiments are designed to isolate the impact of each architectural feature, providing a clearer understanding of how they enhance the model's ability to accurately detect skin cancer. Furthermore, we examine the generalization capacity of the Proposed Model, exploring how these modifications influence its performance across diverse skin lesion types, thereby highlighting the model's robustness and adaptability in clinical diagnostic applications.

4.4.1. Effect of ConvNeXtV2 block and focal self-attention block

In this section, we present the results of ablation studies to evaluate the effect of ConvNeXtV2 blocks and focal self-attention mechanisms on the performance of the Proposed Model. By comparing the model with variations that utilize only ConvNeXtV2 or focal self-attention, we aim to assess the contribution of each component in improving accuracy, precision, recall, and F1-score for skin cancer classification. Table 3 summarizes the performance metrics of these variants, highlighting the importance of both architectural elements.

Table 3 summarizes the results of our ablation study, which evaluates the individual contributions of ConvNeXtV2 and focal self-attention mechanisms in the Proposed Model. This analysis highlights the synergistic effect of these components in enhancing performance while reducing computational complexity. The Focal-Base model, with 89.80 million parameters, achieves an accuracy of 90.88%, precision of 88.65%, recall of 85.61%, and an F1-score of 87.10%. Although its focal self-attention mechanism effectively prioritizes diagnostically relevant regions, it lacks the ability to extract fine-grained local features critical for distinguishing subtle variations in lesions. Conversely, the ConvNeXtV2-Base model, with 89.00 million parameters, achieves an accuracy of 90.96%, precision of 88.46%, recall of 86.32%, and an F1-score of 87.37%, excelling in local feature extraction but falling short in capturing global context, which is essential for differentiating complex lesions. The Proposed Model integrates these mechanisms into a streamlined architecture with only 36.44 million parameters, achieving an accuracy of 93.60%, precision of 91.69%, recall of 90.05%, and an F1-score of 90.73%. This optimized balance between local feature extraction and global contextual awareness allows the model to excel in classifying diverse skin cancer lesions while remaining computationally efficient. The study underscores the complementary roles of ConvNeXtV2 and focal self-attention, with the latter enhancing focus on diagnostically significant regions and the former ensuring robust local feature extraction.

Fig. 6 visualizes the comparative performance of the Proposed Model and its variants, demonstrating its superior diagnostic capability and scalability for real-world applications, including resource-constrained settings. Future work could explore refining focal self-attention or incorporating additional architectural enhancements to address challenges in more complex datasets.

Table 3
Comparison of the Proposed Model with other ConvNextv2 and Focal Self-attention variants.

Model	Params (M)	Accuracy	Precision	Recall	F1-score
Focal-Base	89.80	0.9088	0.8865	0.8561	0.8710
ConvNeXtv2-Base	89.00	0.9096	0.8846	0.8632	0.8737
Proposed Model (Scaled Model + Focal Self-attention + ConvNeXtv2)	36.44	0.9360	0.9169	0.9005	0.9073

4.4.2. Generalization capability of the proposed model

To gain a deeper understanding of how well the Proposed Model generalizes across different skin lesion types, we performed a thorough analysis of its performance on each class within the ISIC 2019 dataset. Evaluating the class-wise performance is essential to grasp how effectively the model can differentiate between various types of skin lesions, which can vary greatly in appearance. This detailed breakdown helps identify any strengths or potential limitations in the model's diagnostic accuracy. The classification report, as shown in Table 4, offers valuable insights into the model's effectiveness across these different classes, revealing its ability to handle both common and rare lesion types with high precision and reliability.

The Proposed Model demonstrates exceptional performance in diagnosing various skin lesion types by integrating ConvNeXtV2 blocks in its initial stages and focal self-attention layers in later stages. The ConvNeXtV2 blocks effectively capture fine-grained local features, crucial for early detection and differentiation, as reflected in the outstanding F1-score of 96.18% for NV (Table 4). Meanwhile, the focal self-attention layers provide a comprehensive understanding of global context, enhancing the model's ability to differentiate visually similar lesions. For instance, the model achieves an F1-score of 90.32% for MEL and performs exceptionally well in DF and BCC, with F1-scores of 93.33% and 93.98%, respectively, showcasing its ability to balance local detail extraction with global contextual analysis.

However, challenges remain for AK and SCC, which are particularly difficult to classify due to overlapping features. AK often resembles benign conditions like seborrheic keratosis or early SCC, causing subtle variations that lead to higher misclassification rates. Similarly, SCC shares morphological similarities with BCC and certain melanoma subtypes, and these issues are exacerbated by inconsistencies in image resolution and lighting. For AK, the model achieves a high precision of 90.67%, but a lower recall of 78.16%, leading to an F1-score of 83.95%. SCC poses similar challenges, with an F1-score of 84.38%, indicating that further enhancements are needed to improve sensitivity for these challenging classes.

Despite these limitations, the Proposed Model demonstrates strong

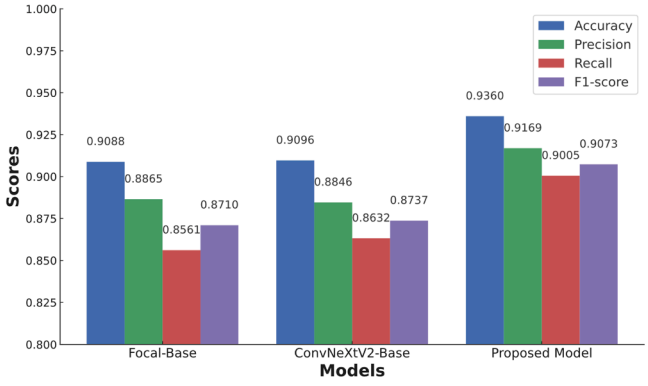


Fig. 6. Comparison of ConvNeXtV2, Focal Self-attention and the Proposed Model.

Table 4
Class-wise performance of the Proposed Model.

Class	Precision	Recall	F1-score	Number of images
AK	0.9067	0.7816	0.8395	87
BCC	0.9398	0.9398	0.9398	332
BKL	0.9157	0.9122	0.9140	262
DF	1.000	0.8750	0.9333	24
MEL	0.9103	0.8962	0.9032	453
NV	0.9556	0.9682	0.9618	1288
SCC	0.8182	0.8710	0.8438	62
VASC	0.8889	0.9600	0.9231	25
Macro Average	0.9169	0.9005	0.9073	2533
Weighted Average	0.9360	0.9360	0.9358	2533

generalization across a wide range of lesion types, achieving a macro average F1-score of 90.73% and a weighted average F1-score of 93.58%. As shown in Fig. 7, these results highlight its robust diagnostic performance while also identifying areas for refinement, particularly for challenging classes like AK and SCC. Strategies such as advanced data augmentation or class-specific optimization could further enhance its capability in addressing these complex cases.

4.5. Discussion

In this study, we introduced a novel deep learning model specifically designed for the early detection of skin cancer. The model combines ConvNeXtV2 blocks for feature extraction in the initial stages with focal self-attention mechanisms in the later stages, allowing it to capture both detailed local features and broader contextual information. This unique architecture enabled the model to achieve impressive results: 92.54% accuracy, 90.41% precision, 87.68% recall, and an F1-score of 88.86%. These metrics reflect the model's ability to reliably differentiate between cancerous and non-cancerous skin lesions, which is vital for timely and effective treatment. What sets this work apart is the extensive evaluation of our model in comparison to other established deep learning models, including traditional CNNs like ResNet50 and VGG16, as well as more advanced ViT-based models such as Swinv2-Base and DeiT3-Base. While CNNs have long been successful in image analysis tasks, our findings indicate that they were consistently outperformed by ViT-based models in this study. However, the Proposed Model, which strategically integrates ConvNeXtV2 with focal self-attention, outperformed all other architectures. This hybrid approach allows the model to accurately capture detailed local patterns while also maintaining a global perspective, which is particularly important in medical imaging where precision is paramount.

Another key aspect of this study is how the model handles the inherent class imbalance in the ISIC 2019 dataset. Some classes, such as NV, contain far more images than others like DF. Despite this imbalance, the model consistently performed well across all classes, demonstrating its robustness and adaptability. This is a crucial feature for real-world applications where imbalanced data is often unavoidable. Furthermore, the model's high performance on an independent test set validates its ability to generalize effectively, offering a reliable indicator of how well it would perform in clinical settings.

Challenges remain in accurately classifying lesion types with overlapping features, such as actinic keratosis and squamous cell carcinoma. These findings underscore the importance of incorporating additional data sources, such as clinical metadata, or exploring ensemble learning methods to improve the model's discriminatory power. Future evaluations on independent datasets and in clinical settings will be crucial to further validate the model's effectiveness and ensure its applicability across diverse populations.

One of the most significant advantages of this model is its relatively low complexity compared to other state-of-the-art models. With only 36.44 million parameters, our model is not only lighter but also more efficient, allowing it to handle high-resolution images and complex classification tasks without requiring excessive computational resources. This lightweight architecture makes it highly practical for deployment in resource-constrained environments, such as mobile applications or point-of-care systems, increasing its accessibility in clinical practice.

Beyond skin cancer detection, the modular design of the Proposed Model offers opportunities for adaptation to other medical imaging tasks. Its capacity for multi-modal data integration, such as combining imaging with genetic or histopathological data, highlights its potential to serve as a comprehensive diagnostic tool. Additionally, its lightweight yet high-performing architecture lays the foundation for scaling the model to broader applications without significant computational trade-offs.

The incorporation of focal self-attention in the later stages further

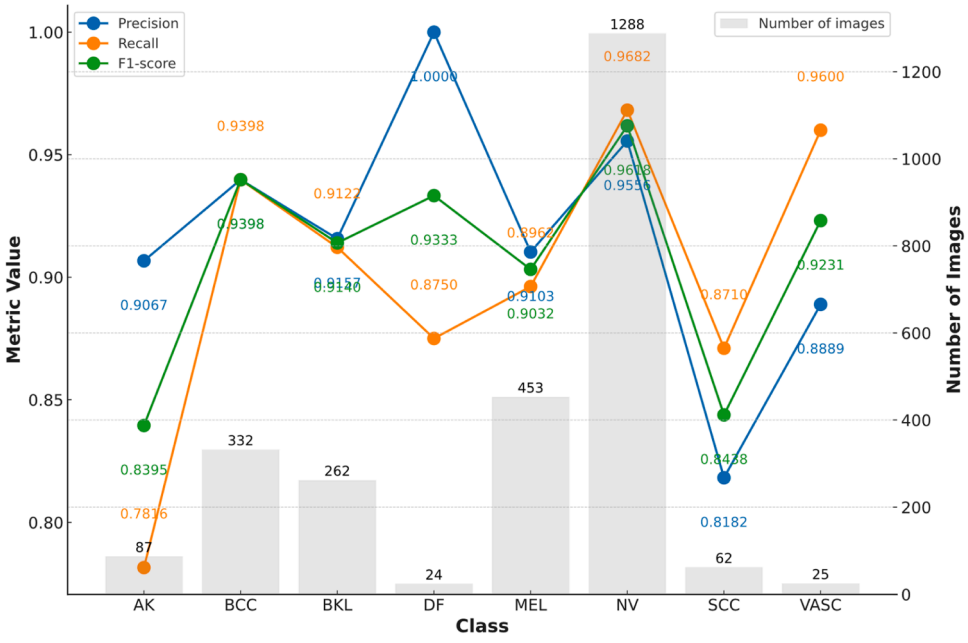


Fig. 7. Class-wise performance of the Proposed Model.

strengthens the model's ability to focus on the most diagnostically relevant areas of an image while filtering out unnecessary background information. In medical image analysis, this capability is essential, as subtle differences in skin lesions can dramatically alter the diagnosis. Focal self-attention enhances the model's precision, especially when dealing with more challenging conditions such as MEL and SCC, where early and accurate detection is critical. The Proposed Model not only surpasses traditional CNN and ViT-based architectures in performance but also offers scalability and computational efficiency with fewer parameters. The model's consistent performance across imbalanced classes, coupled with its robustness in real-world test scenarios, makes it a promising tool for clinical deployment. Future research can expand on this work by exploring its application in other medical imaging domains or refining the model's sensitivity to rarer skin conditions for even greater diagnostic accuracy.

4.6. Limitations and future directions

This study demonstrates significant advancements in skin cancer detection using a hybrid deep learning model that integrates ConvNeXtV2 blocks and focal self-attention mechanisms. However, several limitations must be acknowledged. First, the ISIC 2019 dataset, while comprehensive, exhibits inherent class imbalance, which may affect the generalization of the model to underrepresented lesion types. Although advanced data augmentation techniques were employed to mitigate this imbalance, future studies should explore synthetic data generation methods, such as GAN-based approaches, to further enhance the diversity of minority classes.

Second, the model's performance, while robust across most lesion types, shows lower sensitivity for classes like actinic keratosis and squamous cell carcinoma. These classes often have overlapping features with other skin conditions, making them particularly challenging to classify. To address this, future work could investigate ensemble learning strategies or incorporate domain-specific knowledge through clinical metadata to improve the model's discriminatory power for these difficult cases.

Third, the model's architecture, despite being optimized for mobile and real-time applications, requires computational resources that may be inaccessible in low-resource settings. Future research should prioritize developing lightweight versions of the model using pruning or quantization techniques to ensure broader applicability.

Additionally, this study did not evaluate the model on external datasets or in real-world clinical settings. Future work should include cross-dataset evaluations and clinical validation to confirm the model's robustness and applicability in diverse environments. Exploring federated learning frameworks could also facilitate training on distributed datasets without compromising patient privacy.

Finally, while this study focused on skin cancer detection, the proposed architecture's versatility presents opportunities for adaptation to other medical imaging tasks. Future directions could involve extending this framework to multi-modal data analysis, integrating imaging with clinical and genetic data for holistic diagnostics. By addressing these limitations and pursuing the outlined future directions, the Proposed Model could be further refined, expanding its utility in clinical practice and advancing the state-of-the-art in medical image analysis.

5. Conclusion

This study presents an innovative deep learning model for skin cancer detection, integrating ConvNeXtV2 blocks and focal self-attention mechanisms within a sophisticated hybrid architecture. By combining the strengths of convolutional layers and self-attention mechanisms, the model effectively captures detailed local features and broader contextual patterns in dermatological images. Rigorous evaluations on the ISIC 2019 dataset highlight its superior performance, consistently outperforming traditional CNN and ViT-based models across key metrics such as accuracy (93.60%), precision (91.69%), recall (90.05%), and F1-score (90.73%).

ConvNeXtV2 blocks significantly enhance the model's ability to detect subtle and complex lesion features, crucial for differentiating visually similar skin types. Simultaneously, the focal self-attention mechanism directs attention to diagnostically relevant regions, improving accuracy and sensitivity. With a lightweight architecture of only 36.44 million parameters, the model is optimized for real-time performance and is particularly well-suited for scalable clinical applications. Its compact design makes it ideal for mobile devices and point-of-care systems, offering rapid and cost-effective diagnostic capabilities, especially in resource-limited settings.

Beyond technical innovations, the model effectively addresses challenges like data imbalance, ensuring consistent and reliable classification across diverse skin cancer classes. This robustness ensures its suitability for real-world clinical workflows, where early and accurate detection is vital for enhancing patient outcomes.

The Proposed Model combines exceptional performance, computational efficiency, and scalability, marking it as a transformative tool in skin cancer diagnosis. Its potential clinical applications include accelerated diagnosis, enhanced accessibility to advanced screening technologies, and affordable implementation in underserved areas. Future research could further refine this architecture to enhance its utility in addressing even more complex medical imaging challenges, driving progress in healthcare diagnostics and improving global health outcomes.

Funding

This study was funded by Alfaisal University, which supports research initiatives aimed at advancing knowledge and innovation in alignment with its commitment to academic excellence.

Ethics approval

No ethics approval was required for this work as it did not involve human subjects, animals, or sensitive data that would necessitate ethical review.

Consent to participate

No formal consent to participate was required for this work as it did not involve interactions with human subjects or the collection of sensitive personal information.

Consent to publish

This study did not use individual person's data.

CRediT authorship contribution statement

Burhanettin Ozdemir: Writing – review & editing, Writing – original draft, Validation, Supervision, Software, Resources, Methodology, Investigation, Funding acquisition, Conceptualization. **Ishak Pacal:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Methodology, Funding acquisition, Formal analysis, Conceptualization, Data curation, Investigation, Project administration, Resources.

Declaration of competing interest

The authors declare no competing interests.

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] 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.

[4] B.K. Armstrong, A. Kricker, Skin cancer, Dermatol. Clin. 13 (1995) 583–594, https://doi.org/10.1016/s0733-8635(18)30064-0.

[5] 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.

[6] V. Madan, J.T. Lear, R.M. Szeimies, Non-melanoma skin cancer, The Lancet 375 (2010) 673–685, https://doi.org/10.1016/S0140-6736(09)61196-X.

[7] A.F. JERANT, J.T. JOHNSON, C.D. SHERIDAN, T.J. CAFFREY, Early Detection and Treatment of Skin Cancer, Am. Fam. Physician 62 (2000) 357–368. https://www.aafp.org/pubs/afp/issues/2000/0715/p357.html. accessed June 23, 2024.

[8] 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.

[9] B.K. Chaurasia, H. Raj, S.S. Rathour, P.B. Singh, Transfer learning–driven ensemble model for detection of diabetic retinopathy disease, Med. Biol. Eng. Comput. 61 (2023) 2033–2049, https://doi.org/10.1007/s11517-023-02863-6.

[10] 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.

[11] M. Lubbad, D. Karaboga, A. Basturk, B. Akay, U. Nalbantoglu, I. Pacal, Machine learning applications in detection and diagnosis of urology cancers: a systematic literature review, Neural Comput. Appl. 2 (2024), https://doi.org/10.1007/s00521-023-09375-2.

[12] 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.

[13] A. Kumar, B.K. Chaurasia, Detection of SARS-CoV-2 Virus Using Lightweight Convolutional Neural Networks, Wirel. Pers. Commun. 135 (2024) 941–965, https://doi.org/10.1007/s11277-024-11097-0.

[14] P.B. Singh, P. Singh, H. Dev, B.K. Chaurasia, Glaucoma Classification Using Enhanced Deep Transfer Learning Models with Hybrid ROI Cropped Optic Disc Technique, SN. Comput. Sci. 4 (2023), https://doi.org/10.1007/s42979-023-02163-8.

[15] 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.

[16] P.B. Singh, P. Singh, H. Dev, A. Tiwari, D. Batra, B.K. Chaurasia, Glaucoma Classification using Light Vision Transformer, EAI. Endorsed. Trans. Pervasive Health Technol. 9 (2023), https://doi.org/10.4108/eetpht.9.3931.

[17] H. Ayaz, O. Oladimeji, I. McLoughlin, D. Tormey, T.C. Booth, S. Unnikrishnan, An eXplainable deep learning model for multi-modal MRI grading of IDH-mutant astrocytomas, Result. Eng. 24 (2024), https://doi.org/10.1016/j.rineng.2024.103353.

[18] P. C, A.V. Phamila Y, Deep hybrid architecture with stacked ensemble learning for binary classification of retinal disease, Result. Eng. 24 (2024), https://doi.org/10.1016/j.rineng.2024.103219.

[19] M.A.H. Lubbad, I.L. Kurtulus, Dervis Karaboga, K. Kilic, Alper Basturk, Bahriye Akay Ozkan, U. Nalbantoglu, O. Melis, D. Yilmaz, Mustafa Ayata, Serkan Yilmaz, Ishak Pacal, A Comparative Analysis of Deep Learning-Based Approaches for Classifying Dental Implants Decision Support System, J. Imaging Inform. Med. 2024 (2024) 1–22, https://doi.org/10.1007/S10278-024-01086-X.

[20] I. Leblebicioglu, M. Lubbad, O.M.D. Yilmaz, K. Kilic, D. Karaboga, A. Basturk, B. Akay, U. Nalbantoglu, S. Yilmaz, M. Ayata, I. Pacal, A robust deep learning model for the classification of dental implant brands, J. Stomatol. Oral Maxillofac. Surg. (2024) 101818, https://doi.org/10.1016/J.JORMAS.2024.101818.

[21] 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.

[22] U. Vignesh, R. Parvathi, K. Gokul Ram, Ensemble deep learning model for protein secondary structure prediction using NLP metrics and explainable AI, Result. Eng. 24 (2024), https://doi.org/10.1016/j.rineng.2024.103435.

[23] M. Agarwal, G. Rani, A. Kumar, P.K. K, R. Manikandan, A.H. Gandomi, Deep learning for enhanced brain Tumor Detection and classification, Result. Eng. 22 (2024), https://doi.org/10.1016/j.rineng.2024.102117.

[24] A. Armghan, J. Logeshwaran, S.M. Sutharshan, K. Aliqab, M. Alsharari, S.K. Patel, Design of biosensor for synchronized identification of diabetes using deep learning, Result. Eng. 20 (2023), https://doi.org/10.1016/j.rineng.2023.101382.

[25] D. Gutman, N.C.F. Codella, E. Celebi, B. Helba, M. Marchetti, N. Mishra, A. Halpern, Skin Lesion Analysis toward Melanoma Detection: a Challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC), (2016). https://arxiv.org/abs/1605.01397v1 (accessed May 5, 2024).

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

[27] S. Haggenmüller, R.C. Maron, A. Hekler, J.S. Utikal, C. Barata, R.L. Barnhill, H. Beltraminelli, C. Berking, B. Betz-Stablein, A. Blum, S.A. Braun, R. Carr, M. Combalia, M.T. Fernandez-Figueras, G. Ferrara, S. Fraitag, L.E. French, F. Gellrich, K. Ghoreschi, M. Goebeler, P. Guitera, H.A. Haenssle, S. Haferkamp, L. Heinzerling, M.V. Heppt, F.J. Hilke, S. Hobelsberger, D. Krahl, H. Kutzner, A. Lallas, K. Liopyris, M. Llamas-Velasco, J. Malvehy, F. Meier, C.S.L. Müller, A. A. Navarini, C. Navarrete-Dechent, A. Perasole, G. Poch, S. Podlipnik, L. Requena, V.M. Rotemberg, A. Saggini, O.P. Sangueza, C. Santonja, D. Schadendorf, B. Schilling, M. Schlaak, J.G. Schlager, M. Sergon, W. Sondermann, H.P. Soyer, H. Starz, W. Stolz, E. Vale, W. Weyers, A. Zink, E. Krieghoff-Henning, J.N. Kather, C. von Kalle, D.B. Lipka, S. Fröhling, A. Hauschild, H. Kittler, T.J. Brinker, Skin cancer classification via convolutional neural networks: systematic review of studies involving human experts, Eur. J. Cancer 156 (2021) 202–216, https://doi.org/10.1016/J.EJCA.2021.06.049.

[28] B.C.R.S. Furriel, B.D. Oliveira, R. Prôa, J.Q. Paiva, R.M. Loureiro, W.P. Calixto, M. R.C. Reis, M. Giavina-Bianchi, Artificial intelligence for skin cancer detection and classification for clinical environment: a systematic review, Front. Med. (Lausanne) 10 (2023) 1305954, https://doi.org/10.3389/FMED.2023.1305954/BIBTEX.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] M.M. Shukla, B.K. Tripathi, T. Dwivedi, A. Tripathi, B.K. Chaurasia, A hybrid CNN with transfer learning for skin cancer disease detection, Med. Biol. Eng. Comput. 62 (2024) 3057–3071, https://doi.org/10.1007/s11517-024-03115-x.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] S.S. Chaturvedi, K. Gupta, P.S. Prasad, Skin Lesion Analyser: an Efficient Seven-Way Multi-class Skin Cancer Classification Using MobileNet, Adv. Intell. Syst. Comput. 1141 (2021) 165–176, https://doi.org/10.1007/978-981-15-3383-9_15.

[39] 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, Intell. Med. 3 (2023) 180–190, https://doi.org/10.1016/j.imed.2022.08.004.

[40] 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, J. Imaging Inform. Med. (2024), https://doi.org/10.1007/s10278-024-01140-8.

[41] C. Xin, Z. Liu, K. Zhao, L. Miao, Y. Ma, X. Zhu, Q. Zhou, S. Wang, L. Li, F. Yang, S. Xu, H. Chen, An improved transformer network for skin cancer classification, Comput. Biol. Med. 149 (2022) 105939, https://doi.org/10.1016/J.COMPBIOMED.2022.105939.

[42] L. Cai, K. Hou, S. Zhou, Intelligent skin lesion segmentation using deformable attention Transformer U-Net with bidirectional attention mechanism in skin cancer images, Skin Res. Technol. 30 (2024), https://doi.org/10.1111/srt.13783.

[43] K. Ramkumar, E.P. Medeiros, A. Dong, V.H. Victor, M.R. Hassan, M.M. Hassan, A novel deep learning framework based swin transformer for dermal cancer cell classification, Eng. Appl. Artif. Intell. 133 (2024), https://doi.org/10.1016/j.engappai.2024.108097.

[44] T. Dwivedi, B.K. Chaurasia, M.M. Shukla, Lightweight vision image transformer (LViT) model for skin cancer disease classification, Int. J. Syst. Assur. Eng. Manag. (2024), https://doi.org/10.1007/s13198-024-02521-6.

[45] R.P. Desale, P.S. Patil, An efficient multi-class classification of skin cancer using optimized vision transformer, Med. Biol. Eng. Comput. 62 (2024) 773–789, https://doi.org/10.1007/s11517-023-02969-x.

[46] 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.

[47] 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.

[48] 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.

[49] 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.
[50] S.K. Datta, M.A. Shaikh, S.N. Srihari, M. Gao, Soft Attention Improves Skin Cancer Classification Performance, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 12929 LNCS, 2021, pp. 13–23, https://doi.org/10.1007/978-3-030-87444-5_2.
[51] M. Krishna Monika, N. Arun Vignesh, C. Usha Kumari, M.N.V.S.S. Kumar, E. Laxmi Lydia, Skin cancer detection and classification using machine learning, Mater. Today Proc. 33 (2020) 4266–4270, https://doi.org/10.1016/J.MATPR.2020.07.366.
[52] U.O. Dorj, K.K. Lee, J.Y. Choi, M. Lee, The skin cancer classification using deep convolutional neural network, Multimed. Tools. Appl. 77 (2018) 9909–9924, https://doi.org/10.1007/S11042-018-5714-1/METRICS.
[53] A.N. Toprak, I. Aruk, A Hybrid Convolutional Neural Network Model for the Classification of Multi-Class Skin Cancer, Int. J. ImAging Syst. Technol. 34 (2024), https://doi.org/10.1002/ima.23180.
[54] R. Wang, X. Chen, X. Wang, H. Wang, C. Qian, L. Yao, K. Zhang, A novel approach for melanoma detection utilizing GAN synthesis and vision transformer, Comput. Biol. Med. 176 (2024), https://doi.org/10.1016/j.compbiomed.2024.108572.
[55] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I.S. Kweon, S. Xie. ConvNeXt V2: co-designing and Scaling ConvNets with Masked Autoencoders, 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 2023, pp. 16133–16142. http://arxiv.org/abs/2301.00808.
[56] J. Yang, C. Li, P. Zhang, X. Dai, B. Xiao, L. Yuan, J. Gao, Focal Self-attention for Local-Global Interactions in Vision Transformers, NeurIPS (2021) 1–21. https://arxiv.org/abs/2107.00641v1. accessed June 23, 2024.
[57] 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), in: Proceedings - International Symposium on Biomedical Imaging 2018-April, 2017, pp. 168–172, https://doi.org/10.1109/ISBI.2018.8363547.
[58] 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.
[59] 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.
[60] G. Huang, Z. Liu, L. van der Maaten, K.Q. Weinberger, Densely Connected Convolutional Networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 2261–2269, https://doi.org/10.1109/CVPR.2017.243.
[61] C. Szegedy, S. Ioffe, V. Vanhoucke, A.A. Alemi, Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning, in: 31st AAAI Conference on Artificial Intelligence, AAAI 2017, 2016, pp. 4278–4284, https://doi.org/10.1609/aaai.v31i1.11231.
[62] 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: 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.
[63] 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.
[64] C. Chen, Z. Guo, H. Zeng, P. Xiong, J. Dong, RepGhost: a hardware-efficient ghost module via re-parameterization, ArXiv (2022). http://arxiv.org/abs/2211.06088.
[65] W. Yu, P. Zhou, S. Yan, X. Wang. InceptionNeXt: when Inception Meets ConvNeXt, 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2024, pp. 5672–5683, https://doi.org/10.1109/CVPR52733.2024.00542.
[66] M. Tan, Q.V. Le, EfficientNet: rethinking Model Scaling for Convolutional Neural Networks, in: 36th International Conference on Machine Learning, ICML 2019 2019-June, 2019, pp. 10691–10700. https://arxiv.org/abs/1905.11946v5. accessed February 2, 2024.
[67] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie. A ConvNet for the 2020s, 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 11966–11976, https://doi.org/10.1109/CVPR52688.2022.01167.
[68] 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, Adv. Neural Inf. Process. Syst. 29 (2021) 24261–24272. https://arxiv.org/abs/2105.01601v4. accessed June 23, 2024.
[69] W. Yu, M. Luo, P. Zhou, C. Si, Y. Zhou, X. Wang, J. Feng, S. Yan, MetaFormer Is Actually What You Need for Vision, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2021, pp. 10809–10819, https://doi.org/10.1109/CVPR52688.2022.01055, 2022-June.
[70] J. Yang, C. Li, X. Dai, J. Gao, Focal Modulation Networks, Adv. Neural Inf. Process. Syst. 35 (2022). https://arxiv.org/abs/2203.11926v3. accessed June 23, 2024.
[71] S. Mehta, M. Rastegari, MobileViT: light-weight, Gen.-Purpose, Mob.-Friendly Vis. Transf. 3 (2021). http://arxiv.org/abs/2110.02178.
[72] H. Touvron, M. Cord, H. Jégou, DeiT III: revenge of the ViT, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 13684 LNCS, 2022, pp. 516–533, https://doi.org/10.1007/978-3-031-20053-3_30.
[73] 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: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.
[74] 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, 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 11999–12009, https://doi.org/10.1109/CVPR52688.2022.01170.
[75] H. Bao, L. Dong, S. Piao, F. Wei, BEiT: BERT Pre-Training of Image Transformers, (2021). ArXiv, http://arxiv.org/abs/2106.08254.
[76] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, Y. Li, MaxViT: multi-axis Vision Transformer, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 13684 LNCS, 2022, pp. 459–479, https://doi.org/10.1007/978-3-031-20053-3_27.
[77] A. Wang, H. Chen, Z. Lin, J. Han, G. Ding, RepViT: revisiting Mobile CNN From ViT Perspective, n.d. ArXiv, https://arxiv.org/abs/2307.09283.
[78] S. d'Ascoli, H. Touvron, M. Leavitt, A. Morcos, G. Biroli, L. Sagun, ConViT: improving vision transformerswith, soft convolutional inductive biases, Stat. Mech. (2022) 114005, https://doi.org/10.1088/1742-5468/ac9830.
[79] P.K.A. Vasu, J. Gabriel, J. Zhu, O. Tuzel, A. Ranjan, FastViT: a Fast Hybrid Vision Transformer using Structural Reparameterization, (2023). IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2023, pp. 5762-5772, https://www.doi.org/10.1109/ICCV51070.2023.00532.
[80] 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). ArXiv, https://arxiv.org/abs/2207.05501v4 (accessed June 23, 2024).
[81] 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.
[82] K. Wu, J. Zhang, H. Peng, M. Liu, J. Fu, L. Yuan, TinyViT: fast Pretraining Distillation for Small Vision Transformers, in: ECCV. Lecture Notes in Computer Science 13681, Springer, Cham, 2022. https://doi.org/10.1007/978-3-031-19803-8_5.
[83] A. Trockman, J.Z. Kolter, Patches Are All You Need?, (2022). Trans. Mach. Learn. Res. https://arxiv.org/abs/2201.09792v1 (accessed June 23, 2024).