Journal of Innovative Image Processing (ISSN: 2582-4252)
www.irojournals.com/iroiip/

A Hybrid ConvNeXtV2–MaxViT Framework with CNN-based Feature Refinement for Skin Lesion Classification

Bipin P.R.[1], Anoop V.[2], Ramu R.[3], Santhi K.[4], Upendra Kumar[5], Sai Kiran Oruganti[6]

[1]Post Doctoral Research Fellow, [6]Professor, Lincoln University College, Malaysia.
[2]Professor, Department of Artificial Intelligence and Data Science, Jyothi Engineering College, Cheruthuruthi, Kerala, India.
[3]Associate Professor, [4]Assistant Professor, Department of Electronics and Communication Engineering, Adi Shankara Institute of Engineering and Technology, Kerala, India.
[5]Assistant Professor, Department of Computer Science and Engineering, Institute of Engineering and Technology, Uttar Pradesh, India.

E-mail: [1]pdf.bipin@lincoln.edu.my, [2]vanoop@gmail.com, [3]mail2ramureghu@gmail.com, [4]santhik.ece@adishankara.ac.in, [5]upendra.ietlko@gmail.com, [6]saisharma@lincoln.edu.my

Orcid ID: [1]0000-0003-0433-4502, [2]0000-0003-1960-8190, [3]0000-0002-5076-0820, [4]0009-0004-2215-0779, [5]0000-0003-3792-7945, [6]0000-0003-4601-2907

Abstract

Accurate and prompt detection of skin disorders, particularly malignant skin cancers like melanoma, is considered crucial for effective treatment and improved clinical outcomes. It is a difficult task for even experienced dermatologists to correctly distinguish between similar skin lesions. Deep neural architectures, like Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), can be used for the automated classification of abnormalities in dermoscopic images. ViTs require a large amount of data for optimal generalization, while CNNs are less effective at identifying global patterns. A hybrid model that overcomes the limitations of CNNs and ViTs is proposed in this work. A CNN-based feature refinement module is included in the proposed system to improve lesion-focused features while suppressing irrelevant background information. A dual-path classification algorithm utilizing ConvNeXtV2 for efficient local feature identification and MaxViT to model broader contextual relationships is then employed. The proposed architecture was evaluated on the HAM10000 dataset and validated on the ISIC dataset. The proposed model outperforms single CNN, ViT, and classical CNN-ViT combination models, based on the experimental results. The architecture discussed here achieves an accuracy of 96.8%, an AUC of 0.978, and a balanced F1-score of 0.965 on the HAM10000 dataset, while demonstrating competitive performance when validated on the ISIC dataset. The effect of CNN-based feature refinement has also been studied. These results demonstrate the effectiveness of combining CNN-based feature refinement with multi-scale feature identification to develop robust and accurate systems for skin disease classification.

Keywords: Skin Disease, Convolutional Neural Network, Vision Transformer, ConvNeXtV2, MaxViT.

DOI: https://doi.org/10.36548/jiip.2026.3.001
Received: 03.04.2026, received in revised form: 07.05.2026, accepted: 26.05.2026, published: 05.06.2026

1. Introduction

Skin diseases, especially melanoma, are dangerous to human life and are considered the most common types of skin cancer worldwide [1]. For early and automated detection of skin diseases, dermoscopic imaging is used. The diagnosis of skin disorders, even though subjective, involves the differentiation of skin lesions into malignant or benign. This leads to the design and development of an automated skin disease identification system to analyze skin conditions and provide recommendations on which patients can rely. It is usually a difficult task, even for highly trained medical practitioners, to differentiate between different skin conditions. This leads to an intelligent, computer-aided diagnostic system that is able to identify various skin disorders with a high degree of accuracy and consistency.

Convolutional Neural Networks (CNNs), a popular deep learning architecture, may be employed skin disease classification [2,3]. CNNs are highly effective in finding patterns like texture, color and shape in skin images, which are useful in the classification of skin images. CNNs have a layered architecture consisting of multiple convolution, pooling and dense layers arranged sequentially. They are highly effective in identifying different patterns in an image, which avoids the need for manually selecting required features for automated classification [4,5]. While they are suitable for identifying local patterns, they often fail in determining global patterns, which is very critical in tasks like skin lesion classification [6].

Vision Transformers (ViT) may be used as an alternative to CNNs for the same purpose [7,8]. ViTs are attention-based architectures capable of identifying global contextual patterns in input images, making them suitable for localizing skin disorders.

- CNNs use neighborhood-based processing in the input image, whereas ViTs find relationships across the entire image in one processing step. ViTs require large datasets for training, which is often difficult in medical imaging due to the limited availability of labeled datasets. In practice, CNNs can be used to extract low-level local patterns, while ViTs are very useful for identifying broader contextual dependencies.
- Hybrid CNN-ViT architectures that leverage the advantages of both CNN and ViT have been proposed [9]. In these hybrid architectures, CNNs extract low-level features such as edges and textures, while ViTs extract broader-level dependencies [10].

2. Literature Review

N. Codella et al. [1] introduced an automated melanoma detection system along with a standard evaluation framework that used a publicly accessible dataset. The performance of the proposed system is compared with the assessment of expert dermatologists. It highlighted the significance of consistent evaluation metrics in both lesion segmentation and classification tasks. This system suffers from a lack of real-time applicability and variation in imaging conditions, which leads to limited generalizability of the proposed technique in diverse clinical environments.

A. Esteva et al. [2] demonstrated the use of CNNs in skin disease detection by classifying input skin images as benign or malignant. This technique used a publicly accessible dataset containing 12,9000 skin lesion images for training the CNN. This research highlighted

the use of artificial intelligence in diagnosing skin conditions. Advantages of the proposed system include its clinical relevance and scale. Limitations of the system are restricted interoperability of CNN models and insufficient diversity in patient populations. Therefore, careful and intensive validation of the proposed system is required before its clinical usage.

J. Kawahara et al. [3] demonstrated skin lesion classification using deep feature extraction from pre-trained CNNs. Accuracy was improved by utilizing high-level feature representations from multiple network layers. Their framework used high-level representations from multiple layers, enabling accurate differentiation among lesion types. Even with limited labelled data and transfer learning, this approach achieved good performance. Robustness will be reduced because this technique depends on handcrafted processing pipelines and offers limited consideration of lesion context.

E. Nasr-Esfahani et al. [4] used convolutional neural networks (CNNs) for melanoma detection. This model integrated both segmentation and classification modules with colour and shape-based features. With minimal feature refinement, this method achieved good performance. This method is simple and can be applied to real-world skin images. Though the proposed methodology provided the required performance, its capability to generalize well is limited because of limited dataset diversity and variability in imaging environments.

M. A. Al-Masni et al. [5] developed a multi-path CNN-based framework for the classification of dermoscopic images, which extracts both global and local patterns in the input images. This methodology offered competitive performance compared to other traditional architectures in shape preservation and boundary localization but its higher computational complexity limits its ability to operate in real-time scenarios.

A. Abdelhafeez et al. [6] proposed a methodology that integrates deep feature fusion with a neutrosophic logic approach for the automated classification of skin lesions. This architecture reduces the uncertainty in skin image classification and improves diagnostic robustness. It demonstrates better performance in multi-class classification, but its scalability and adaptability are limited because of its dependence on handcrafted logic components.

A. Dosovitskiy et al. [7] used a Vision Transformer (ViT) instead of a CNN for skin disease detection. Vision Transformers make use self-attention over image patches instead of the convolutions used in CNNs. ViTs are capable of producing better performance than CNNs, but their performance depends on the availability of a large dataset.

G. M. S. Himel et al. [8] proposed a Vision Transformer-based model tailored for segmenting and classifying dermoscopic images. Utilizing the self-attention mechanism, the system was effective in capturing long-range dependencies critical for focusing on clinically relevant lesion regions. Evaluated on ISIC datasets, it demonstrated competitive accuracy and improved interpretability.

P. Pacal et al. [9] proposed an early skin cancer detection model combining CNNs with Vision Transformers. A balanced architecture is the model's key strength and has improved classification accuracy, but the increased computational complexity in the design may affect its deployment on low-resource devices.

S. Khan [10] combined CNNs with Transformers for skin lesion classification. This hybrid model uses CNNs for region-specific feature extraction and transformers for situation-aware attention modeling. This algorithm resulted in better classification accuracy but its real-

time usage is limited because of its computational overhead and large delay in generating inference.

N. Gessert et al. [11] demonstrated how CNNs can be integrated with patch-based attention mechanisms for skin cancer segmentation. This model offered better interoperability and improved efficiency. This method enhanced region-specific learning without incorporating a full transformer architecture. Disadvantages include patch selection, sensitivity and complexity.

Z. Liu et al. [12] proposed ConvNeXt, an architecture redesigned with transformer-inspired elements such as large kernel sizes and depth-wise convolutions. The model effectively bridges the efficiency of convolutional neural networks (CNNs) with the representational power of transformers and showed good performance. This method is not effective on smaller datasets because of the overfitting problem.

S. Woo et al. [13] developed an enhanced version of ConvNeXt, which is known as ConvNeXtV2 which incorporates masked autoencoder pretraining. This model enhanced the feature extraction capability of ConvNeXt through supervised learning. ConvNeXtV2 model is efficient and scalable but its usage is limited in situations where the computational resources and data availability are limited.

Ozdemir and P. Pacal [14] used an architecture that combines ConvNeXtV2 with self-attention to improve effectiveness in skin irregularity detection. This system provided competitive performance compared to CNN-based methods as a result of the inclusion of local attention, which led to the localization of clinically important regions. However, this method suffers from high complexity and complex tuning requirements.

Z. Tu et al. [15] introduced MaxViT which captures global patterns with the help of block-wise and grid-wise attention mechanisms. MaxViT offers better performance when compared with normal ViTs and CNN-based algorithms. This architecture has good classification accuracy but suffers from overfitting when trained on small dermoscopic image datasets.

H.M. Unver et al. [16] proposed an architecture for skin image classification using ensemble learning. Multiple CNN algorithms are used and their results are aggregated to produce the final classification result. This ensemble learning based architecture provided better robustness and greater performance. While this architecture gives stable performance the delay in getting the final results limits its real-time usage.

P. Tschandl et al. [17] introduced a dataset, named HAM10000, with annotated skin images from various clinical settings. This dataset incorporates a large collection of skin lesions obtained from multiple sources under various conditions. It helps to improve the training diversity of machine learning algorithms and improve model generalization. This dataset includes rare lesion varieties, which makes it more popular and relevant. Demerits associated with the dataset are inconsistencies in lighting conditions across images, and reliance on expert annotations affects the generalization of the model's performance.

T. Brinker et al. [18] experimentally analyzed the use of CNNs in skin disease classification applications. This study demonstrated the requirement for a larger dataset for improved performance and increased interoperability in AI applied clinical systems.

S. Han et al. [19] employed CNNs for skin cancer detection, specifically on facial images. It incorporates spatial information to improve lesion localization accuracy. This method is limited to a specific lesion type and anatomical region, which may restrict its applicability.

3. Methodology

The framework proposed here integrates a CNN-based feature refinement technique, a hybrid dual path CNN and ViT based feature extraction algorithm and a feature fusion mechanism. The proposed hybrid model uses ConvNeXtV2 as the CNN branch and MaxViT as the ViT branch. The CNN-based feature refinement generates intermediate feature representations which are subsequently processed in parallel by both branches. The proposed hybrid technique is illustrated in Figure 1.

3.1 CNN-based Feature Refinement

Raw dermatoscopic images contain artifacts such as varying illumination and body hair which are undesirable as they can interfere with the classification of skin images as normal or cancerous. As a remedy, a CNN-based feature refinement module is introduced in this research work. This stage performs a learned transformation of input images into enhanced feature representations, emphasizing lesion-specific structures while suppressing irrelevant artifacts.

We used MobileNetV2, known for its low-resource requirement and competitive performance, as a lightweight feature refinement module rather than a conventional preprocessing module. MobileNetV2 efficiently captures both shallow and moderately deep representations and is effective in identifying texture gradients, edge contours, and localized contrast patterns commonly observed in skin lesions.

Only shallow to intermediate layers of MobileNetV2 have been retained with the intention of performing feature refinement concentrating on structural and textural representations rather than high-level semantic abstraction, as illustrated in Figure 2. This truncation ensures that the model emphasizes mid-level representations like borders and textures but avoids semantic abstraction layers which are needed for classification tasks. To make MobileNetV2 suitable for feature refinement, we intentionally eliminate its deeper classification layers. The layers removed are:

- Inverted residual blocks beyond Block 6, which typically generate 96 or more channels
- The final convolution layer containing 1280 channels.
- The global average pooling layer.
- The classification (fully connected) head.

These layers are removed as they are mainly intended for feature abstraction and the separation of different classes. Including these layers would risk discarding fine-grained lesion information by learning overly semantic representations that are not necessary for feature refinement. We enhance the lower and structural characteristics by keeping only mid and lower-level blocks.

The output feature dimensions at each stage are progressively reduced from 224×224×3 to 28×28×32 through convolutional and inverted residual operations, ensuring efficient spatial compression while preserving discriminative features.

Stage 1: Initial Feature Capture using Convolution

A convolutional layer with 32 filters, a kernel size of 3×3, and a stride of 2 will process the input image. This stage reduces the image resolution by half while retaining the spatial relationship among its pixels. This step lays the foundation for lesion-centric enhancement by localizing low-level intensity variations, initial contours, and pigmentation boundaries. The ReLU6 activation function is employed to maintain bounded non-linearity, and batch normalization is used to stabilize and accelerate convergence.

Stage 2: Edge Contrast Enhancement via Inverted Residual Block 1

Images, after initial feature identification, will pass through the first inverted residual block, which has 16 output channels and an expansion factor of 1. Abrupt transitions in texture and intensity traits, usually present around the border of the lesion, will be captured in this step. A bottleneck structure is also introduced in this block to improve feature efficiency and reduce background interference.

Stage 3: Artifact Suppression via Down-sampling in Blocks 2 and 3

Further abstraction and down-sampling will be introduced by blocks 2 and 3. These two blocks have 24 output channels with strides of 1 and 2 for blocks 2 and 3, respectively. These layers help to concentrate on high-contrast lesion structures while eliminating finer-scale noise, such as hairs and subtle skin creases.

Stage 4: Texture Feature Refinement in Blocks 4–6

This step captures regional and textural features, with 32-channel output. These blocks gradually reduce the spatial size, thereby enhancing features like globules, pigmentation networks, and asymmetries, which are key indicators in skin disease diagnosis

Stage 5: Generation and Resizing of the Feature Map

The final feature map from the output of the feature refinement module will have dimensions of 28×28×32. This feature map will then be transformed into a format compatible with downstream architectures using a two-step process:

1. Spatial Upsampling: Bilinear interpolation upsamples the feature map to a size of 224×224×32, without changing the spatial continuity and structural information.
2. Channel Projection: The 224×224×32 feature map will then be transformed into a 224×224×3 representation with the help of a 1 x 1 convolution layer which ensures its compatibility with ConvNeXtV2 and MaxViT input requirements (Here there are three channels, i.e., C=3).

The above transformation embeds the enhanced structural and textural information into a spatial format suitable for downstream processing

This transformation preserves diagnostically relevant features without degrading classification performance. The subsequent ConvNeXtV2 and MaxViT modules further refine these representations, compensating for any minor information loss.

The entire feature refinement process can be represented as in Equation (1)

$$F = f_{CNN}(I) \tag{1}$$

Where F represents the feature enhanced output after feature refinement and I indicate the input image. F is then coupled to both ConvNeXtV2 and MaxViT for subsequent feature extraction.

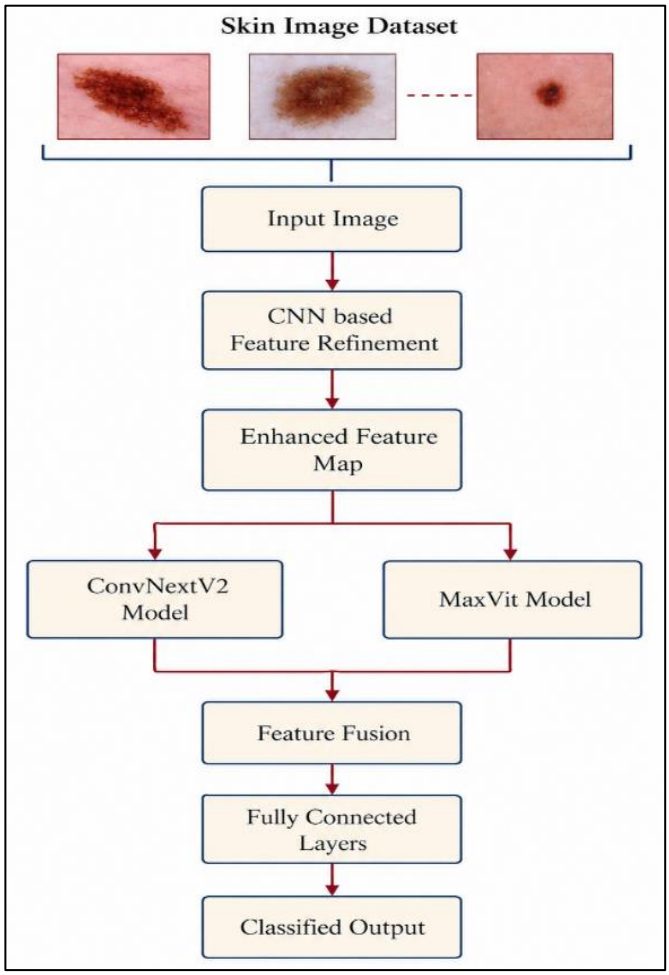

Figure 1. Proposed Hybrid Algorithm

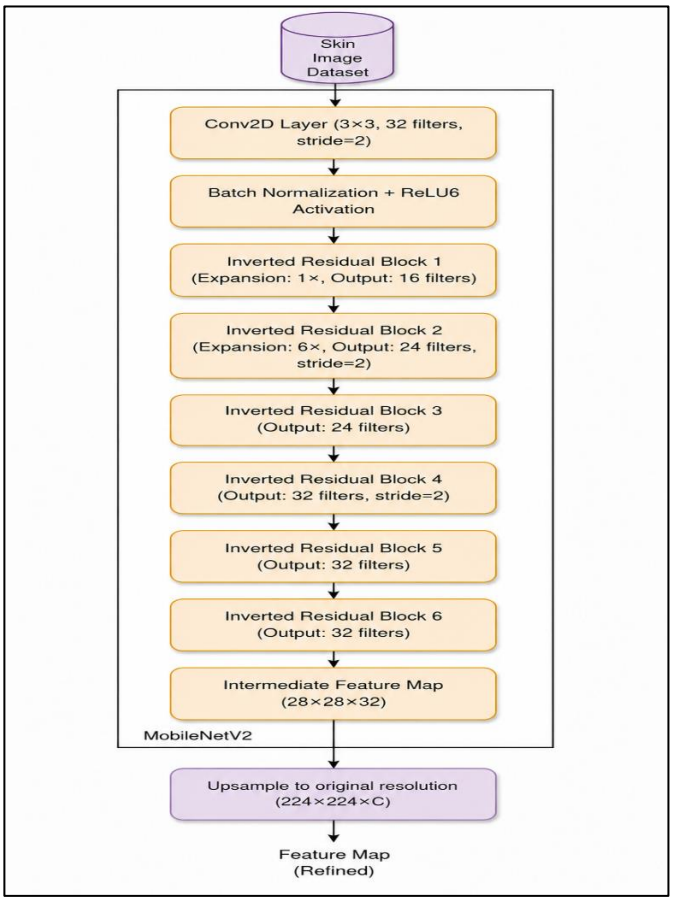

Figure 2. CNN-based Feature Refinement

3.2 ConvNeXtV2 for Local Feature Extraction

The local feature extraction will be carried out with the help of ConvNeXtV2, a CNN model. It incorporates some of the design aspects of vision transformers like depth wise convolutions, normalization strategies and large kernel sizes [12,13]. As a result, this model helps in classifying skin diseases effectively by extracting low-level to mid-level spatial patterns like pigmentation distribution, textural variations and lesion borders. The local feature extraction procedure is mathematically represented as:

$$F_{local} = f_{ConvNeXtV2}\,(F) \tag{2}$$

Where F_{local} in Equation (2) indicates the local feature map obtained with the help of ConvNeXtV2 and $f_{ConvNeXtV2}$ is the function to represent forward pass of ConvNeXtV2 architecture.

3.3 MaxViT for Global Multi-Scale Attention

The Vision Transformer used in this research is MaxViT, which extracts global features from the input skin images. The MaxViT architecture, unlike traditional ViTs, is based on a transformer architecture that combines both grid-based global attention and window-based local attention mechanisms [15]. This dual-attention framework makes it an appropriate choice for capturing both fine-grained patterns and extended structural information. MaxViT provides multi-scale attention by computing local attention within small windows and global attention across the entire spatial domain. This operation is symbolically represented in Equation (3).

$$F_{global} = f_{MaxViT}(F) \tag{3}$$

Where, f_{MaxViT} represents the MaxViT forward pass operation and the resulting feature map extracted is denoted as F_{global}

3.4 Feature Fusion and Classification

After extracting local and global feature maps using ConvNeXtV2 and MaxViT, these features will be combined using a feature-level concatenation operation along the channel dimensions, as represented in Equation (4). This preserves both the local and global feature maps from both branches,

$$F_{fusion} = Concat\ (F_{local}, F_{global}) \tag{4}$$

The direct concatenation produces a more diverse and richer feature embedding, F_{fusion}, that represents the lesion characteristics more effectively. This results in improved classification performance as the concatenated vector includes multi-scale and multi-perspective information.

The fused feature representation is subsequently flattened and passed through fully connected layers for final classification. Categorical cross entropy is used as the objective function, which is usually considered an appropriate choice for classification involving multiple output classes.

4. Results and Discussions

The experimental findings of the proposed Hybrid ConvNeXtV2 and MaxViT model with CNN-based feature refinement in skin disease classification are presented in this section. To ascertain its relative effectiveness, the proposed model was tested along with established baseline methods, including traditional CNN architectures, standalone ViTs, and CNN-ViT hybrid variants. Details of the training process including various hyperparameters, are also discussed in this section. Along with the performance evaluation on the HAM10000 dataset, a cross-dataset validation was also performed using the ISIC dermoscopic image dataset, to assess the robustness and cross-dataset performance of the proposed method.

The proposed model was evaluated using the HAM10000 dermoscopic image dataset [17], which includes a diverse range of pigmented skin lesion types. The dataset consists of 10,015 images with 600 x 450 pixels resolution, categorized into seven clinical classes: melanoma (mel), basal cell carcinoma (bcc), melanocytic nevi (nv), actinic keratoses (akiec), dermatofibroma (df), vascular lesions (vasc), and benign keratosis-like lesions (bkl).

Table 1. Categories of Images

Category	Number of Samples
nv	6705
mel	1112
bkl	1098
bcc	514
akiec	327
vasc	142
df	115

Table 2. List of Hyperparameters

Parameters	Proposed Hybrid Algorithm	CNN	ViT	CNN-ViT Hybrid
Batch Size	32	32	32	32
Learning Rate	0.0001	0.0001	0.0001	0.0001
Epochs	50	50	50	50
Optimizer	Adam	Adam	Adam	Adam
Momentum	0.9	0.9	0.9	0.9
Dropout Rate	0.3	0.3	0.3	0.3

Table 1 displays various image classes present in the HAM10000 dataset. This dataset was divided into training (80%), validation (10%), and test (10%) sets. All images in the dataset will be resized to 224 x 224 pixels and scaled to the range [0,1] during feature refinement. These images are then fed to the model, thereby maintaining uniform input representation across the deep learning framework.

To enhance input quality prior to feature extraction, a CNN-based feature refinement stage (illustrated in Figure 2, Section 3.1) was incorporated. This stage is developed as discussed in section 3.1. Implemented using the TensorFlow framework, this module extracts intermediate feature maps, retaining diagnostically relevant information and regions. These lesion-focused features are then fed into the ConvNeXtV2 and MaxViT branches, which improves the classification performance.

An NVIDIA P100 GPU was used to train the proposed hybrid ConvNeXtV2–MaxViT model, employing mixed-precision techniques to increase computational efficiency. The model was trained end-to-end using a supervised learning approach, where both the feature refinement module and the dual-branch architecture were optimized jointly using categorical cross-entropy loss. Overfitting was mitigated using a combination of dropout regularization (rate = 0.3), batch normalization within convolutional layers, and evaluation under distorted input conditions, which improved model robustness. A learning rate of 0.0001 and the Adam optimizer were used to ensure stable convergence; however, no learning rate scheduling or warm-up strategies were employed as the chosen configuration produced stable training performance. Table 2 lists various parameters used during the training of the proposed algorithm.

The performance of the proposed hybrid ConvNeXtV2–MaxViT algorithm is compared against various baseline models in Table 3. Test images were intentionally distorted using controlled distortions such as Gaussian blur (σ = 1.0), salt-and-pepper noise (density =0.03) and random occlusions using black square patches at random locations. These modifications were made to replicate various challenges in teledermatology, such as motion-induced blur, occlusions caused by hair or artifacts, and noise introduced during image acquisition. Similar experimental settings were applied to different existing algorithms, such as CNN-only, ViT-only, and hybrid CNN–ViT models and their performance parameters are compared in Table 3.

Table 3. Performance Parameters

Model	Accuracy	Precision	Recall	F1-Score	AUC
CNN-Only	91.40%	0.902	0.907	0.904	0.932
ViT-Only	92.10%	0.911	0.915	0.913	0.944
CNN-ViT	94.30%	0.932	0.937	0.934	0.958
Proposed Hybrid (Without feature refinement)	95.20%	0.948	0.945	0.946	0.965
Proposed Hybrid (With feature refinement - Distorted input)	95.68%	0.952	0.949	0.96	0.969

Proposed Hybrid (With feature refinement - Clean input)	96.80%	0.968	0.964	0.965	0.978

ViT-B/16 was used as the baseline model for comparison. This was implemented using PyTorch and the Hugging Face transformers library. Each input image was partitioned into 16 × 16 patches, which were flattened and projected into 768-dimensional embeddings. These embeddings were then forwarded through 12 transformer encoder layers, each consisting of multi-head self-attention and feed-forward sublayers, with GELU activation. The final classification layer was configured with seven neurons specifically for skin disease classification.

The hybrid CNN–ViT architecture incorporates the CNN and ViT architectures described previously. Feature representations generated by the CNN and ViT branches were combined through feature-level concatenation, without altering the internal configurations of either network. The hybrid structure produces a fused vector that contains both localized spatial details and broader contextual information from the CNN and ViT, respectively. This fused feature vector will then be forwarded to a fully connected layer with a Softmax activation function for classification.

Along with the evaluation of CNN-only, ViT-only, CNN–ViT, and the proposed architectures, an ablation study was conducted to assess the contribution of the CNN-based feature refinement module. The proposed hybrid model was evaluated under two conditions: with and without the feature refinement stage.

The comparative evaluation of CNN-only, ViT-only, and hybrid CNN–ViT models demonstrates the contribution of each branch. The CNN branch effectively captures local spatial features, while the ViT branch models global contextual dependencies. The improved performance of the hybrid model confirms the complementary nature of both branches.

Figure 3 illustrates the confusion matrix of the proposed hybrid model in the absence of the feature refinement stage, whereas Figures 4 and 5 present the corresponding results after applying feature refinement to clean and distorted (noisy) images, respectively.

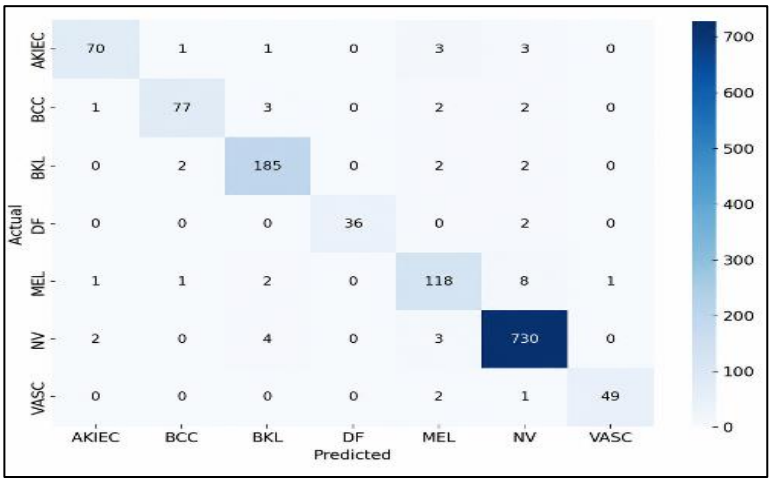

Figure 3. With Noise and Without Feature Refinement

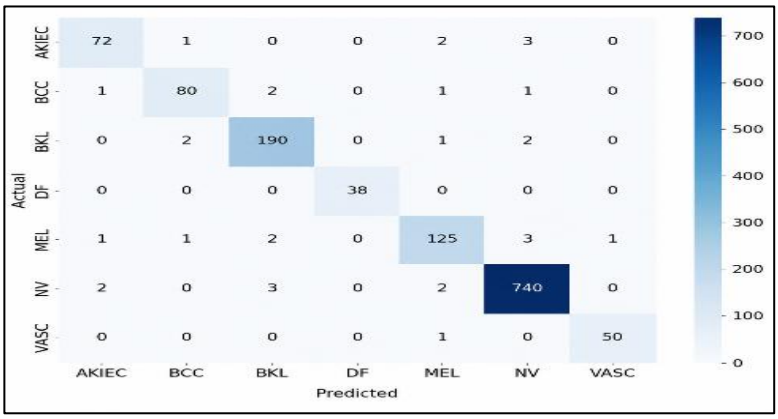

Figure 4. Without Noise and With Feature Refinement

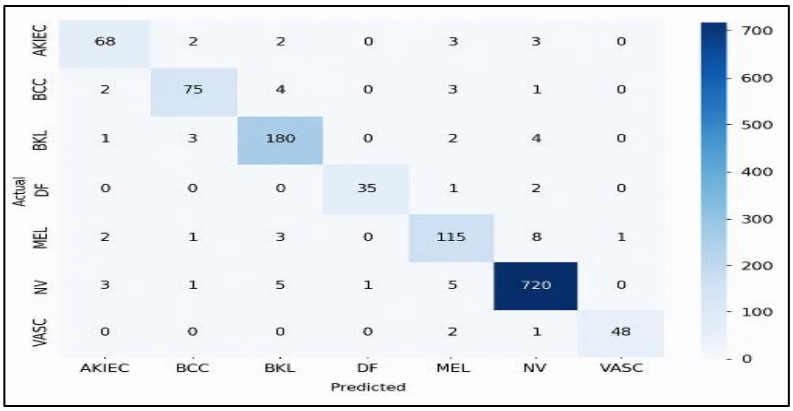

Figure 5. With Noise and With Feature Refinement

The proposed hybrid model achieved a classification accuracy of 96.8% on the HAM10000 dataset, outperforming conventional methodologies, including standalone CNN and ViT models, as illustrated in Table 3 and Figure 6. The CNN-based feature refinement module enabled the ConvNeXtV2 and MaxViT models to capture local and global features more efficiently, resulting in improved overall accuracy.

The proposed hybrid model achieved a precision of 0.968 and a recall of 0.964 on clean data, as illustrated in Table 3 and graphically plotted in Figures 7 and 8. Additionally, the proposed model's precision and recall were 0.952 and 0.949, respectively, under noisy conditions, which indicates its capability to classify various skin disease types.

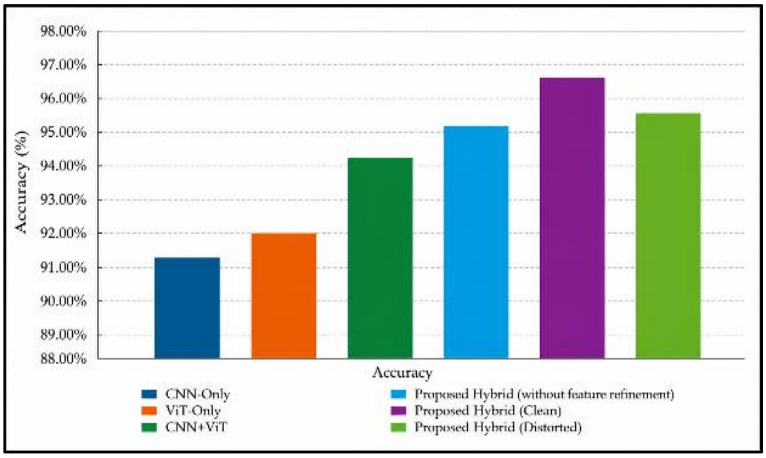

Figure 6. Accuracy

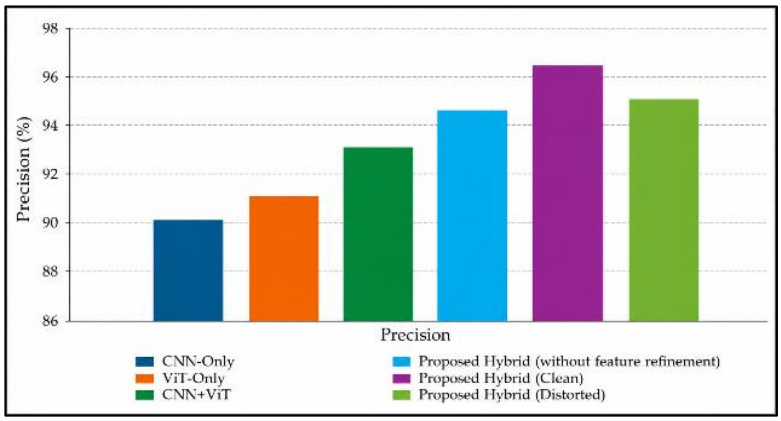

Figure 7. Precision

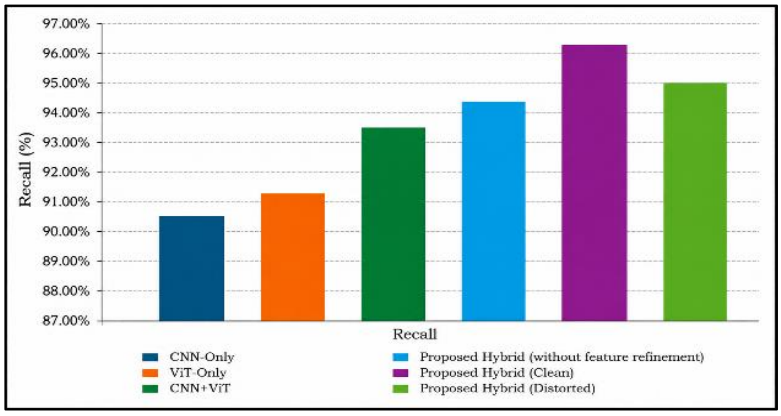

Figure 8. Recall

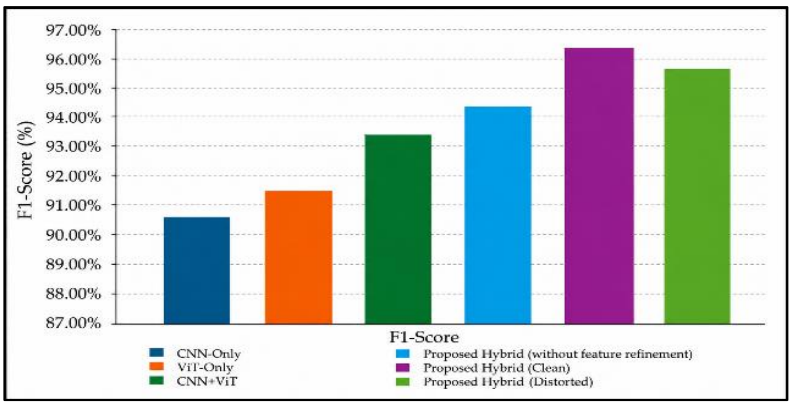

Figure 9. F1-Score

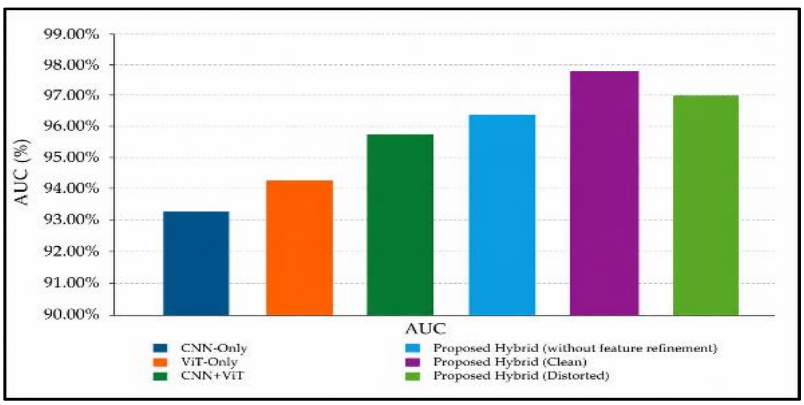

Figure 10. AUC

As shown in Figure 9, the proposed hybrid model achieved F1-scores of 0.965 for clean data and 0.960 under noisy conditions. The proposed model secured an AUC of 0.978, as shown in Figure 10, indicating its ability to classify various skin disease types. Individual AUC-ROC curves have been plotted in Figures 11-16.

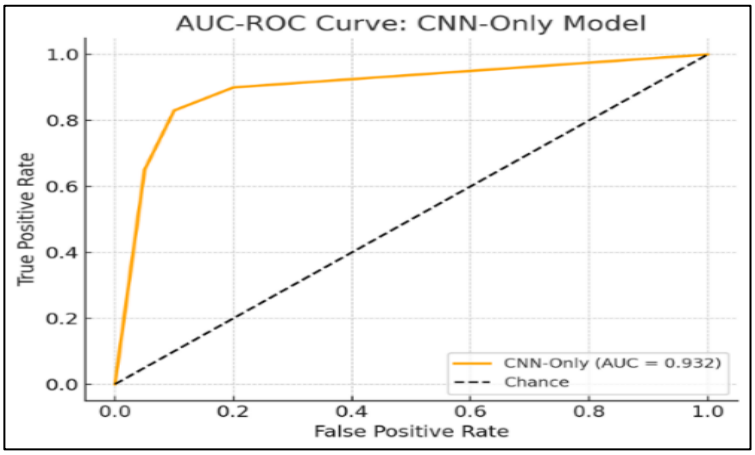

Figure 11. CNN Only

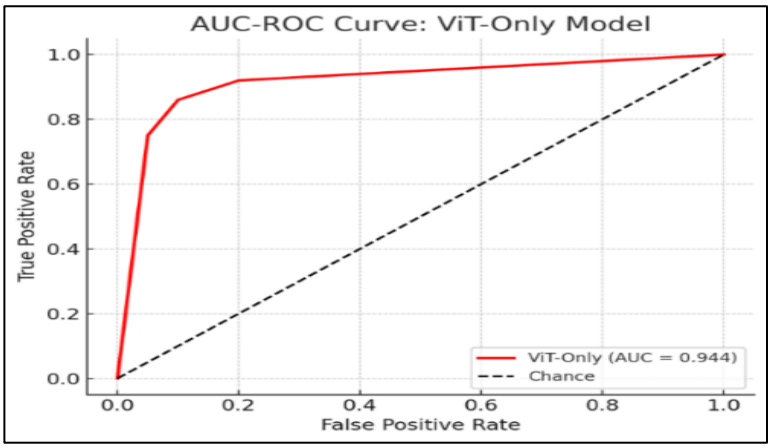

Figure 12. ViT Only

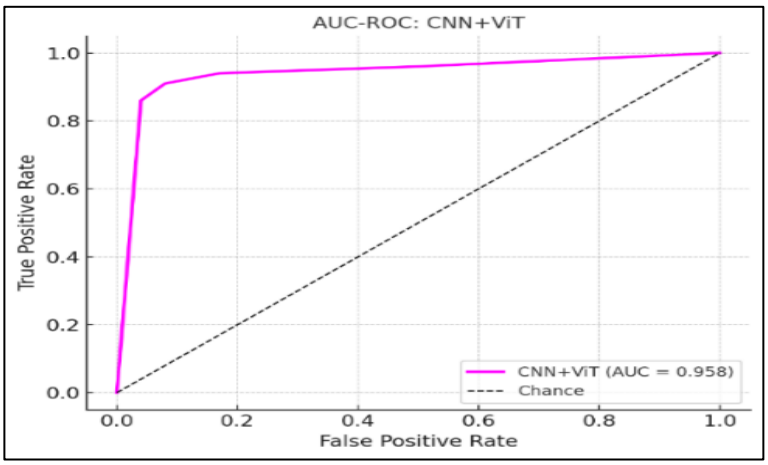

Figure 13. CNN+ViT

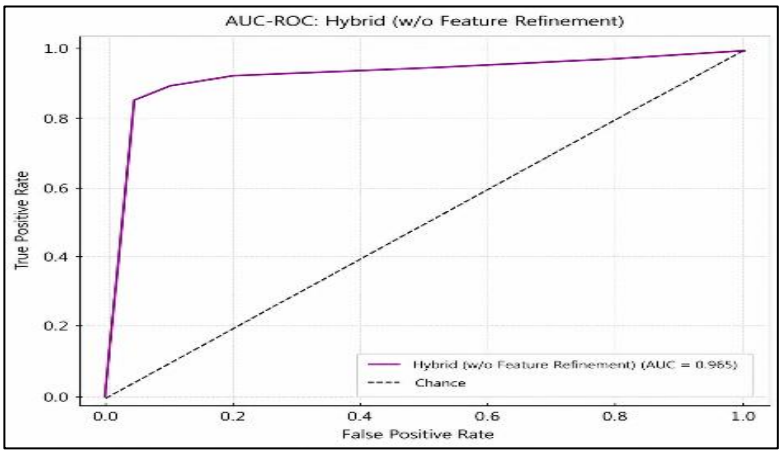

Figure 14. Hybrid Without Feature Refinement

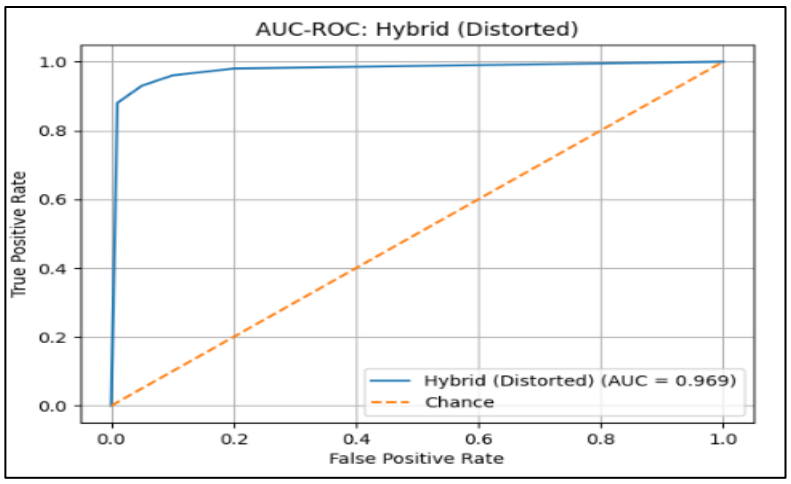

Figure 15. Hybrid Distorted

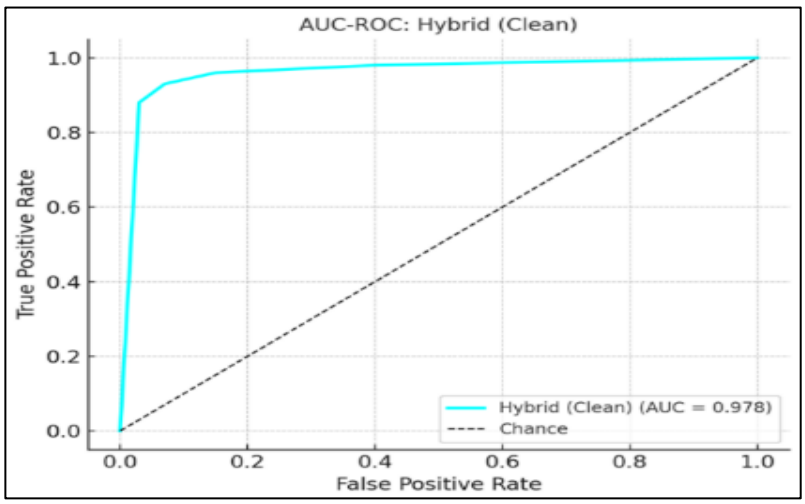

Figure 16. Hybrid Clean

As is evident from the results depicted in Table 3 and the relevant performance curves, the proposed ConvNeXtV2–MaxViT algorithm is able to provide superior performance compared to conventional CNN, ViT, and CNN–ViT models. The CNN-based feature refinement algorithm improved feature extraction by filtering out irrelevant details of the input image while retaining important lesion-specific information.

4.1 Cross-Dataset Validation

To evaluate the effectiveness of the proposed algorithm, cross-dataset validation was conducted using skin images from the publicly available ISIC 2019 dataset. For this external validation, only lesion classes present in both the HAM10000 and ISIC datasets were considered to maintain consistency during evaluation. The ISIC dataset contains skin lesion images of size 768 x 512 pixels, which were then resized to 224 x 224 pixels. CNN-based feature refinement was performed on these images before feeding them into the proposed model for inference. This external validation was implemented using the same hardware and software setup we used while training with the HAM10000 dataset. The cross-dataset validation results obtained are summarized in Table 4.

Table 4. Results of Cross-dataset Validation

Type of evaluation	Testing dataset used	Accuracy	Precision	Recall	F1-Score	AUC
Internal Evaluation	HAM10000	96.8 %	0.968	0.964	0.965	0.978
External Evaluation	ISIC	94.90%	0.944	0.938	0.941	0.962

Results in Table 4 illustrate the classification performance of the proposed model when evaluated using an external dataset, which the model was never trained on. The proposed model maintained competitive classification performance even though there was a slight reduction in the performance parameters.

5. Conclusion

This paper introduced a hybrid CNN-ViT architecture with a CNN-based feature refinement module for skin disease classification. The CNN-based feature refinement enhanced the input image by emphasising lesion-specific regions and removing irrelevant artifacts, which allows the hybrid model to produce accurate classification results. Experimental analysis demonstrated the effectiveness of our proposed methodology, which outperforms the baseline CNNs, ViTs, and CNN-ViT hybrids in AUC, F1-score, and accuracy. The proposed ConvNeXtV2 and MaxViT hybrid model with CNN-based feature refinement has achieved an AUC of 0.978, an accuracy of 96.8 % and an F1-score of 0.965 when trained on the publicly available HAM10000 standard dataset. The proposed hybrid algorithm utilized the effectiveness of CNNs in capturing local patterns and ViTs' ability to capture long-range dependencies. Also, the CNN-based feature refinement module improved the performance of the proposed algorithm by providing images with emphasized lesion-specific details and eliminating irrelevant artifacts. The proposed model also maintained competitive classification ability during cross-dataset validation using the ISIC dataset. Although the proposed framework demonstrated promising performance on the HAM10000 and ISIC datasets, further validation using larger clinical datasets under real-world imaging conditions is necessary before clinical deployment.

References

[1] N. C. F. Codella et al., "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)," 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, USA, 2018, 168-172.

[2] Esteva, Andre, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. "Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks." nature 542, no. 7639 (2017): 115-118.

[3] Kawahara, Jeremy, Aicha BenTaieb, and Ghassan Hamarneh. "Deep Features to Classify Skin Lesions." In 2016 IEEE 13th international symposium on biomedical imaging (ISBI), IEEE, 2016, 1397-1400.

[4] Nasr-Esfahani, Ebrahim, Shadrokh Samavi, Nader Karimi, S. Mohamad R. Soroushmehr, Mohammad H. Jafari, Kevin Ward, and Kayvan Najarian. "Melanoma Detection by Analysis of Clinical Images Using Convolutional Neural Network." In 2016 38th annual international conference of the IEEE engineering in medicine and biology society (EMBC), IEEE, 2016, 1373-1376.

[5] Al-Masni, Mohammed A., Mugahed A. Al-Antari, Mun-Taek Choi, Seung-Moo Han, and Tae-Seong Kim. "Skin Lesion Segmentation in Dermoscopy Images via Deep Full Resolution Convolutional Networks." Computer methods and programs in biomedicine 162 (2018): 221-231.

[6] Abdelhafeez, Ahmed, Hoda K. Mohamed, Ali Maher, and Nariman A. Khalil. "A Novel Approach Toward Skin Cancer Classification Through Fused Deep Features and Neutrosophic Environment." Frontiers in Public Health 11 (2023): 1123581.

[7] Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." arXiv preprint arXiv:2010.11929 (2020).

[8] Himel, Galib Muhammad Shahriar, Md Masudul Islam, Kh Abdullah Al-Aff, Shams Ibne Karim, and Md Kabir Uddin Sikder. "Skin Cancer Segmentation and Classification Using Vision Transformer for Automatic Analysis in Dermatoscopy-Based Noninvasive Digital System." International journal of biomedical imaging 2024, no. 1 (2024): 3022192.

[9] Pacal, Ishak, Burhanettin Ozdemir, Javanshir Zeynalov, Huseyn Gasimov, and Nurettin Pacal. "A Novel CNN-ViT-Based Deep Learning Model for Early Skin Cancer Diagnosis." Biomedical Signal Processing and Control 104 (2025): 107627.

[10] Khan, Somaiya, Athar Shahzad Fazal, Amna Khan, and Ali Khan. "An Automated Skin Lesions Classification Using Hybrid CNN and Transformer Based Deep Learning Model." In Proceedings of the 2023 8th international conference on biomedical imaging, signal processing, 2023, 26-31.

[11] Gessert, Nils, Thilo Sentker, Frederic Madesta, Rüdiger Schmitz, Helge Kniep, Ivo Baltruschat, Rene Werner, and Alexander Schlaefer. "Skin Lesion Classification Using CNNs with Patch-Based Attention and Diagnosis-Guided Loss Weighting." IEEE Transactions on Biomedical Engineering 67, no. 2 (2019): 495-503.

[12] Liu, Zhuang, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. "A Convnet for the 2020s." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, 11976-11986.

[13] Woo, Sanghyun, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. "Convnext v2: Co-Designing and Scaling Convnets with

Masked Autoencoders." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, 16133-16142.

[14] Ozdemir, Burhanettin, and Ishak Pacal. "An Innovative Deep Learning Framework for Skin Cancer Detection Employing ConvNeXtV2 and Focal Self-Attention Mechanisms." Results in Engineering 25 (2025): 103692.

[15] Tu, Zhengzhong, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. "Maxvit: Multi-Axis Vision Transformer." In European conference on computer vision, Cham: Springer Nature Switzerland, 2022, 459-479.

[16] Goyal, Manu, Amanda Oakley, Priyanka Bansal, Darren Dancey, and Moi Hoon Yap. "Skin Lesion Segmentation in Dermoscopic Images with Ensemble Deep Learning Methods." Ieee Access 8 (2019): 4171-4181.

[17] Tschandl, Philipp, Cliff Rosendahl, and Harald Kittler. "The HAM10000 Dataset, a Large Collection of Multi-Source Dermatoscopic Images of Common Pigmented Skin Lesions." Scientific data 5, no. 1 (2018): 180161.

[18] Brinker, Titus Josef, Achim Hekler, Jochen Sven Utikal, Niels Grabe, Dirk Schadendorf, Joachim Klode, Carola Berking, Theresa Steeb, Alexander H. Enk, and Christof Von Kalle. "Skin Cancer Classification Using Convolutional Neural Networks: Systematic Review." Journal of medical Internet research 20, no. 10 (2018): e11936.

[19] Han, Seung Seog, Ik Jun Moon, Woohyung Lim, In Suck Suh, Sam Yong Lee, Jung-Im Na, Seong Hwan Kim, and Sung Eun Chang. "Keratinocytic Skin Cancer Detection on the Face Using Region-Based Convolutional Neural Network." JAMA dermatology 156, no. 1 (2020): 29-37.