Enhanced Skin Cancer Classification Using a Deep CNN with Integrated Transformer Layer

Abdul Rahaman Shaik, P. Rajesh Kumar

***Abstract*—Skin cancer remains a critical public health issue worldwide, with early detection significantly improving patient outcomes. This study introduces an enhanced deep learning approach, combining a custom Convolutional Neural Network (CNN) with an Integrated Transformer Layer (ITL) to boost classification accuracy for skin lesions. Our model is trained and tested on the HAM10000 dataset, comprising 10,015 dermoscopic images of various skin lesion types. Accuracy improved considerably by fine-tuning hyperparameters and refining the model using normalization strategies, dropout-based regularization, data–level augmentation, and techniques to address class imbalance.. The addition of a transformer layer facilitates the capture of long-range dependencies, yielding superior classification performance and achieving an impressive 97% accuracy. This advanced model demonstrates strong potential as a diagnostic tool for dermatologists, supporting timely and precise skin cancer detection.**

***Index Terms*— Neural Network, Hyperparameters, Transformer Layer, Skin Cancer Classification,**

I. Introduction

Skin cancer has emerged as a major global health concern, with rising incidence rates across the world. Early and accurate identification of skin lesions is essential for optimizing treatment outcomes and enhancing patient prognosis. Dermoscopy, a widely adopted, non-invasive imaging technique, has proven invaluable in aiding dermatologists with early-stage skin cancer detection. However, manual interpretation of dermoscopic images can be labor-intensive and is prone to observer variability. Advances in deep learning have shown promise in automating this process, potentially providing robust and consistent skin lesion classification.

Even though CNNs are proven choice for image analysis tasks, their performance can be further improved. In this study, we propose an advanced classification model that integrates a deep CNN with a transformer layer, aimed at enhancing classification accuracy for skin lesions. The HAM10000 dataset, with 10,000 dermoscopic images spanning seven lesion types, provides the core data on which the model is trained and assessed. To boost our model's learning stability and generalization, we employ techniques such as batch normalization for stable training and dropout to reduce overfitting. Additionally, data augmentation involving rotations, scaling and flipping enriches the training data, enabling the model to better capture diverse lesion characteristics. To address the class imbalance, we apply up sampling and down sampling strategies, ensuring an equitable representation of all lesion types.

The inclusion of the transformer layer introduces the capability to capture long-range dependencies within the image data, further refining the model's classification accuracy. The experimental results highlight the effectiveness of our enhanced model, which attains superior accuracy and outperforms our previous CNN-only architecture. Our model demonstrates significant potential as a valuable diagnostic aid for dermatologists, contributing to prompt and accurate skin cancer classification and ultimately enhancing patient care.

Manuscript received December 11, 2024; revised August 29, 2025.

Abdul Rahaman Shaik is a PhD student of the ECE Department, Andhra University College of Engineering, Andhra University, Visakhapatnam, Andhra Pradesh, India (corresponding author: phone: +91-9491185747; e-mail: abdulrahman.s@vishnu.edu.in).

P. Rajesh Kumar is a Professor in the Department of ECE, Andhra University College of Engineering, Andhra University, Visakhapatnam, Andhra Pradesh, India (e-mail: rajeshauce@gmail.com).

II. Literature Review

Tschandl et al..[1] address the challenges of limited and non-diverse dermatoscopic image datasets in automated skin lesion diagnosis by releasing the HAM10000 dataset. This dataset, consisting of 10,015 images from varied populations and sources, was curated with semi-automated workflows and neural network assistance to enhance quality and diversity. It is available via the ISIC archive for academic machine learning research. In [2], which is our previous work, we applied machine learning (ML) algorithms on HAM10000 image set and explored the accuracy of various ML techniques. We observed that Random Forest with Pricipal Component Analysis produced the best results with 92% accuracy. Xu et al. [3] developed a CNN with two-branch encoder and integrated it with a transformer to optimize the process of extracting features from images. The CNN branch consists of four layers that capture localized features through progressive down sampling, while the transformer branch consists four layers of attention mechanisms to grasp global context. This approach achieved an accuracy of 92.79% on HAM image set. Chao Xin et al. [4] introduced Skin-Trans, an enhanced vision transformer (ViT) designed for skin cancer classification. Their approach involves a three-step process: establishing a ViT model to assess Skin-Trans, implementing multi-scale patch embeddings via overlapping sliding windows, and applying contrastive learning to improve feature distinction. They marked an accuracy of 94.3% on HAM image set. Karthik et al. [5] combined the Swin Transformer with the Dense Group Shuffle Non-Local Attention (DGSNLA) Network. The DGSNLA is composed of DenseNet169, Group Shuffle

Depth-wise blocks, and an enhanced non-local attention block. This fusion of deep features enhances both global and local feature representation and resulted in an accuracy of 94.21%. Zhiwei Qin et al. [6] presented a Generative Adversarial Network (GAN) based data augmentation method to improve skin lesion classification. They developed a modified style-based GAN for generating high-quality synthetic images, and a transfer learning-based classifier is trained on this data. Tested on the HAM image set, the method achieved 95.2% accuracy. Mirco Gallazzi et al. [7] proposed a framework utilizing the self-attention mechanism of Transformer models to capture spatial relationships across image regions, bypassing the need for handcrafted features or heavy pre-processing. Their transformer-based architecture demonstrated strong performance in lesion classification. The framework achieved an accuracy of 86.37%.

Chiyu Liu et al. [8] introduced "Reswin," a fusion model combining a CNN (3D-ResNet), with a vision transformer network (Video Swin Transformer). This hybrid model uses a soft voting approach to enhance classification performance. Reswin achieved an accuracy of 90.99%. Hao, Shengnan, et al. [9] presented a novel fusion model, ConvNeXt-ST-AFF, which combines the strengths of CNN CNN-based model and a transformer-based model. The output from these two models is merged through a special fusion method using submodules for improved representation, resulting in an accuracy of 92.16%. In [10] a new transformer-based model is proposed for diabetic retinopathy classification with multiple instance learning. In this method high resolution retinal images are segmented into 224×224 patches, from which features are extracted using Vision Transformer (ViT). Inter-instance features are also captured using another custom block, incorporating global information into the model. This system marked an accuracy of 93.1%. Gou, Quandeng, et al. [11] introduced an innovative hierarchical multi-category framework that integrates multi-scale CNNs to discern features across varying resolutions with the transformer architecture's strength in modeling global dependencies. The model leverages the hierarchical structure of the transformer to enhance understanding of complex image relationships. This integration allows for more effective feature representation and improved classification performance with an accuracy of 94.63%. Saxena et al. [12] proposed maximum sensitivity neural network and experimented with various segmentation algorithms like clustering, watershed, and thresholding followed by nodule extraction and classification and achieved an accuracy of 96%. Oktavian et al. [13] proposed a CNN using ResNet-18 with transfer learning from ImageNet and weighted loss functions to address imbalanced datasets. The Mish activation function was also tested, resulting in an accuracy of 88.30%, demonstrating improved model performance. Wei Dai et al. [14] introduced a model called HierAttn, which utilizes a hierarchical attention approach. This model incorporates a deep supervision methodology to effectively learn both local and global characteristics by leveraging multi-level, multi-path attention strategies under a unified training loss framework. This model is evaluated on the ISIC2019 dermoscopy dataset and it achieved an accuracy of 96.7%. In [15] the authors presented an External Attention Transformer (EAT) model that leverages external attention mechanisms for efficient and precise breast cancer image classification. Achieving a remarkable 99% accuracy on the BreakHis dataset, the EAT model demonstrates both high performance and computational efficiency. Remya et al. [16] inducted a groundbreaking framework that integrates channel attention, and ROI (Region Of Interest) techniques with a ViT for precise skin condition detection, including skin decease. Combining computer vision with patient metadata, it achieved an impressive 99% accuracy on a comprehensive dermoscopic image dataset.

Vachmanus et al. [17] proposed DeepMetaForge, a deep-learning framework for skin cancer detection using metadata-enriched images. Leveraging BEiT which is a ViT pre-trained through images that are masked as its encoding backbone, the framework achieved a high accuracy of 99.3%. Ferdous et al. [18] introduced LCDEiT, a data-efficient image transformer designed for small datasets. It achieved linear computational complexity by leveraging external attention and a teacher-student approach. The transformer-based student model is applied to MRI brain tumour classification and recorded an accuracy of 98.11%. Hossain et al. [19] proposed IVX16, a transfer learning-based model that combines the three top-performing architectures with explainable AI for enhanced interpretability. Developed to find brain tumours and categorize them, IVX16 scored an impressive accuracy of 96.94%. Lingbo Huang et al. [20] explored transformer-based foundation models, specifically the VFM (Vision Foundation Model) and LFM (Language Foundation Model) models, for hyperspectral image (HSI) classification. To enhance traditional HSI classification, a spectral-spatial VFM-based transformer (SS-VFMT) is proposed, integrating spectral-spatial data into the pretrained foundation transformer, achieving competitive accuracy. Xiaoxiao Li et al. [21] addressed the gap in deep learning for disease detection by focusing on clinically meaningful features for diagnosing skin lesions rather than relying on artifacts. The proposed pipeline is designed to uncover novel biomarkers which are not part of current clinical criteria but they are valuable to dermatologists. The model achieved an accuracy of 85%. Haroon Rashid et al. [22] used GANs to generate realistic dermoscopic images for data augmentation in skin lesion classification. By enhancing the training set with synthetic images, the approach aims to address the challenge of limited dataset sizes. The model recorded an accuracy of 86.1%. Rahi et al. [23] discussed a model that evaluates multiple neural network architectures to identify the most effective approach for detecting five primary skin diseases. Initially, they built a custom CNN model with keras sequential API and achieved approximately 80% accuracy. The performance is enhanced by using pre-trained models such as VGG11, ResNet50 and DenseNet121. With ResNet50 they recorded the highest accuracy at 90% highlighting its superior effectiveness. F. Zhao et al. [24] introduced a convolution transformer fusion splicing network (CTFSN) for hyperspectral image (HSI) classification, combining local and global information via addition and channel stacking. A residual splicing convolution block is proposed for shallow feature

preservation, while a convolutional transformer fusion block (CTFB) enhances local and global feature capture. A dual-branch fusion module then merges these features, achieving competitive accuracy across diverse datasets. Ali Jamali et al. [25] presented PolSARFormer, a Vision Transformer (ViT)-based framework that integrates 3-D and 2-D CNNs with local window attention (LWA) for effective classification of polarimetric synthetic aperture radar (PolSAR) data that produced good accuracy.

III. Methodology

In our prior work [2], we developed a Deep CNN (DCNN) with three convolutional layers. Each layer is followed by batch normalization (BN) and dropout layers and the DCNN achieved an accuracy of 96%. The optimizer used is Adam with an LR (learning rate) of 0.1 and a DR (Dropout Rate) of 0.2. In the present work, we propose a methodology to further enhance the accuracy by adding a transformer layer to the model. The resulting model is termed as a Deep CNN with Transformer layer (DCNNT). The addition of a transformer layer aims to capture broader contextual dependencies and enhance classification accuracy on the HAM10000 dataset. Instead of randomly selecting hyperparameters, we adopt a rigorous iterative tuning approach, as before, with the transformer layer added to further enhance the model's capability to interpret geometric relationships within skin lesion images. Python is used for development, with Kaggle facilitating training and testing. The integration of the transformer is designed to boost the DCNNT's performance, bringing increased robustness and precision to skin lesion classification.

A. Data Set

The HAM10000 is a repository of dermatoscopic skin lesion images, and it is widely recognized as a valuable resource in dermatology research. Comprising a total of 10,015 images collected from various clinical settings and hospitals, this dataset was created to advance the field of skin cancer classification, particularly in identifying melanoma. Its primary objective is to support and accelerate research efforts by providing a diverse and comprehensive collection of skin lesion images. Fig. 1. illustrates sample images from this dataset, highlighting its variety and clinical relevance.

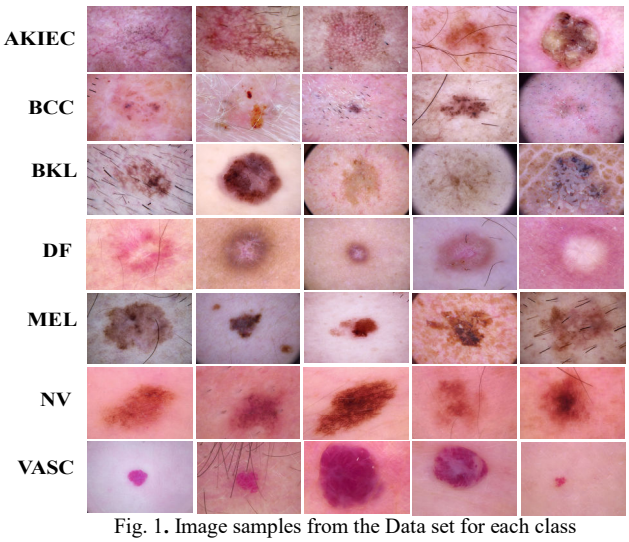

Fig. 1. Image samples from the Data set for each class

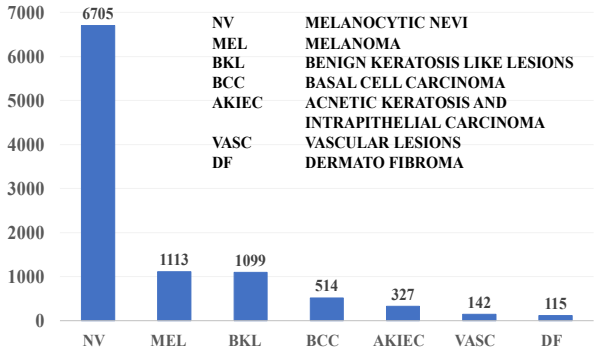

Fig. 2. Unbalanced Data distribution

B. Data Balancing

The seven classes of skin cancer images have an unbalanced distribution, which is shown in Fig. 2, along with counts and class abbreviations for each category. We can observe the imbalance in the dataset. Imbalanced datasets often lead to a model which overly favours the majority class and undermines its ability to accurately predict minority classes. Class imbalance may also lead to overfitting, where the model becomes excessively tailored to the dominant class, resulting in poor generalization on new, unseen data.

Balanced datasets, on the other hand, help maintain equitable decision boundaries, enabling the model to learn meaningful patterns from both majority and minority classes. This improves predictive accuracy and robustness across all categories. Fair evaluation is another critical outcome, as balanced data allows for reliable assessment using various performance metrics. Such metrics provide deeper insights into the model's capability to handle diverse scenarios effectively, fostering better performance interpretation and application across varied real-world contexts.

To maintain equilibrium, classes with fewer images are augmented with additional samples, while classes with a larger number of images are reduced through downsampling. A sample size of 2800 is selected based on iterations done with different sample sizes. Classes 1 to 7, whose count is less than 2800, are oversampled and class 0 with more than 2800 images is downsampled to 2800. The balanced data distribution is shown in the Fig. 3.

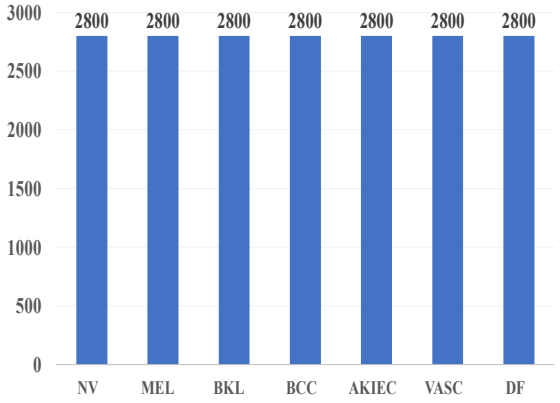

Fig. 3. Balanced Data distribution

C. Filtering and Resizing

Before being fed into the DCNNT, the images undergo 2D bilateral filtering, which is a nonlinear image processing technique widely utilized for tasks such as noise suppression, edge retention, and visual enhancement. This method effectively smoothens the image while safeguarding critical edges and fine details. The term "bilateral" refers to the filter's dual consideration of both spatial closeness and differences in pixel intensity during the processing.

Incorporating 2D bilateral filtering as a preprocessing step enhances the data quality and reliability of the images that are fed into the DCNNT model. After filtering, the images are resized. The original size of the images in the data set is 600X450, but this size is too big and requires more memory and demands more processing time, and hence the images are resized to 75X75. Fig. 4. illustrates the original and resized images.

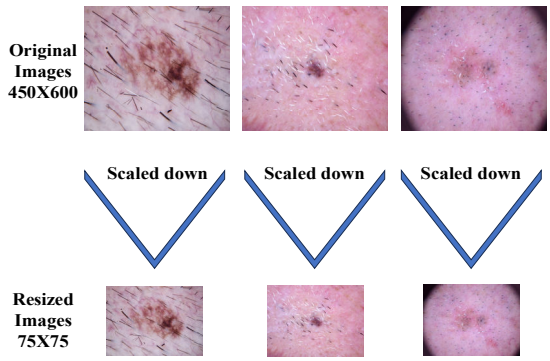

Fig. 4. Original and resized Images

D. Data Augmentation

Data augmentation is implemented dynamically during training to enhance dataset diversity, improve model robustness and prevent overfitting. By generating a variety of transformations on the fly, this approach effectively simulates a larger dataset and helps the model generalize better across unseen data. Additionally, data augmentation plays a vital function in addressing class imbalance by exposing the model to varied representations of less-represented classes during training. The code snippet used to generate augmented images is given below.

Code snippet to generate augmented images:

```
transform = transforms.Compose([
    transforms.RandomRotation(30),
    transforms.RandomHorizontalFlip(),
    transforms.RandomVerticalFlip(),
    transforms.ToTensor(),
    transforms.Normalize(mean=[0.5, 0.5, 0.5], std=[0.5, 0.5, 0.5])
])
image = Image.open('path_to_image.jpg') # Replace with the actual image path
augmented_image = transform(image)
from torchvision.datasets import ImageFolder
from torch.utils.data import DataLoader
dataset=ImageFolder(root='path_to_dataset', transform=transform)
dataloader=DataLoader(dataset,batch_size=32,shuffle=True, num_workers=4)
```

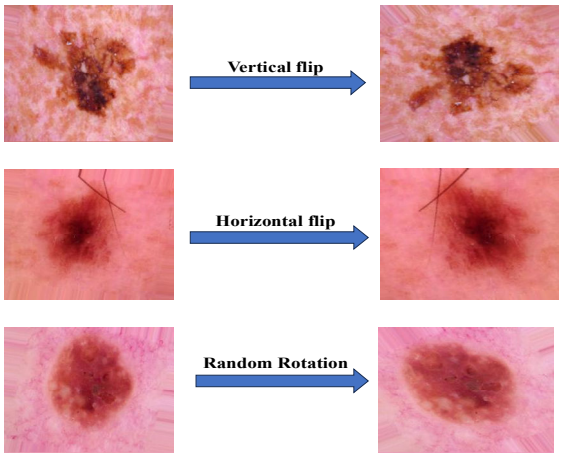

Fig. 5. Original images and their augmented counterparts

As a result of data augmentation, the training process is enriched with diverse samples, reducing the likelihood of the model overfitting to specific patterns or noise in the data. This method not only improves the generalization capabilities of the model but also ensures it can handle variations in real-world scenarios in a better way. Fig. 5. presents examples of original images alongside their augmented counterparts, showcasing the range of transformations applied to enhance the dataset's variability.

E. DCNNT Architecture

The proposed DCNNT architecture builds upon the earlier DCNN design by integrating a Transformer layer after the convolutional stages, enhancing its capacity to encode wider contextual dependencies and intricate geometrical relationships in the data. This hybrid approach leverages the convolutional layers' feature extraction capabilities and the transformer layer's strength in learning global context. The architecture consists of four convolutional stages followed by a transformer layer and dense layers.

Convolution Layers:

The initial four stages of the model extract hierarchical features from the input images using convolution layers (Conv2D). The first stage uses 256 filters of size 3×3, activated by the ReLU function. These filters detect basic patterns such as edges and textures. Subsequent stages progressively reduce the number of filters to 128, 64, and 32, focusing on deeper feature representations with improved efficiency in computation.

Max-Pooling Layers:

Each stage incorporates a Max-Pooling layer with a 2×2 kernel to down sample the spatial dimensions. This operation outputs the single greatest value from the region from non-overlapping windows, reducing spatial size and introducing translational invariance.

Batch Normalization and Dropout Layers:

Batch normalization (BN) is a technique designed to improve the stability and efficiency of training deep neural networks. During training, the distribution of activations within the network can shift, a phenomenon known as internal covariate shift. This shift can slow down training and make the optimization process more challenging. BN

addresses this by normalizing the activations within a layer, ensuring they have a consistent mean and variance.

This normalization process happens independently for each feature in a mini-batch. By normalizing the data, BN allows the network to converge faster and reduces sensitivity to initialization, enabling the use of higher learning rates. Additionally, BN introduces two learnable parameters, a scaling parameter (γ) and an offset (β). The network can adjust or even revert the normalization through these parameters whenever it improves performance. These parameters enhance the model's flexibility, ensuring that normalization does not limit its ability to learn complex patterns. Batch normalization also acts as a form of regularization, helping to reduce overfitting in some cases by adding controlled randomness to the training phase.

Dropout is another regularization technique that helps prevent overfitting by introducing randomness into the training process. During training, dropout randomly "drops" or nullifies a fraction of neurons in a layer by setting their outputs to zero. This prevents the network from becoming overly reliant on specific neurons and forces it to learn more robust and generalizable features.

The dropout rate determines the proportion of neurons to drop and is typically set between 0.2 and 0.5, depending on the level of detail in the model and the richness of the dataset. In essence, dropout creates an ensemble-like effect, as each forward pass effectively trains a slightly different subset of the network. During inference, dropout is turned off, and the outputs of all neurons are scaled to account for the absence of dropout, ensuring consistent predictions.

Transformer Layer:

The core enhancement of the DCNNT architecture is the integration of a transformer layer after flattening the output from the convolutional stages. This layer processes the feature vectors using self-attention and feed-forward mechanisms to model global dependencies.

Multi-Head Self-Attention:

The first step in the Transformer layer involves multi-head attention, responsible for generating attention scores to identify relationships between different feature vectors. The attention mechanism is expressed as:

$$Attention(Q,K,V) = softmax(QK^T/\sqrt{d_k})V \quad (1)$$

- Q (Query), K (Key), and V (Value) are matrices derived from the input.
- d_k is the dimensionality of the keys.
- softmax is applied row-wise to normalize the scores.

Multi-head attention computes multiple such attention mechanisms in parallel to capture diverse patterns.

Feed-Forward Network (FFN):

After attention, a feed-forward network (FFN) applies two linear transformations with a ReLU activation in between:

$$FFN(x) = ReLU(xW1 + b1)W2 + b2 \quad (2)$$

where W1, W2 and b1, b2 are learnable parameters. This layer processes the output from attention, enriching the representation further. Layer normalization is applied to stabilize training. Residual connections are added around the self-attention and FFN sub-layers to prevent vanishing gradients and allow better gradient flow.

$$x' = LayerNorm(x + Attention(x)) \quad (3)$$

$$y = LayerNorm(x' + FFN(x')) \quad (4)$$

Dense (Fully-Connected) Layers and SoftMax:

The output of the Transformer layer is averaged along the sequence dimension, reducing it to a single vector. Next, it is processed by two dense layers, the first consisting of 64 units and the second yielding 7 outputs. A SoftMax activation generates probabilities for each class, enabling multi-class classification.

The DCNNT architecture combines the strengths of convolutional and Transformer-based processing, offering enhanced feature extraction and robust contextual modelling, critical for tasks like skin lesion classification.

Key Aspects of the Proposed Architecture:

Input Layer

➔ Input Image: **75×75×3**

First Convolutional Block

➔ Conv2d (3→256), BatchNorm2d, ReLU

➔ MaxPool2d (2×2), Dropout (0.2)

➔ Output: **37×37×256**

Second Convolutional Block

➔ Conv2d (256→128), BatchNorm2d, ReLU

➔ MaxPool2d (2×2), Dropout (0.2)

➔ Output: **18×18×128**

Third Convolutional Block

➔ Conv2d (128→64), BatchNorm2d, ReLU

➔ MaxPool2d (2×2), Dropout (0.2)

➔ Output: **9×9×64**

Fourth Convolutional Block

➔ Conv2d (64→32), BatchNorm2d, ReLU

➔ MaxPool2d (2×2), Dropout (0.2)

➔ Output: **4×4×32**

Flattening

➔ Flatten 4×4×32 → **512**

➔ Shape: (batch_size, 512)

Reshaping for Transformer

➔ Reshape (batch_size, 16, 32)

➔ 16 = 4×4 patches, 32 = embed_dim (channels)

Transformer Encoder Block

➔ Multi-Head Attention (input: 16×32)

➔ Layer Norm

➔ Feed-Forward (ff_dim=64)

➔ Layer Norm
➔ Output: (batch_size, 16, 32)

Global Pooling
➔ Mean Pool over sequence
➔ Output: (batch_size, 32)

Fully Connected (FC) Layers
➔ Linear (32 → 64), ReLU
➔ Linear (64 → 7)
➔ Output: (batch_size, 7)

Output
➔ Final classification scores (7 classes)

Fig 6. Shows the functional flow chart of the system

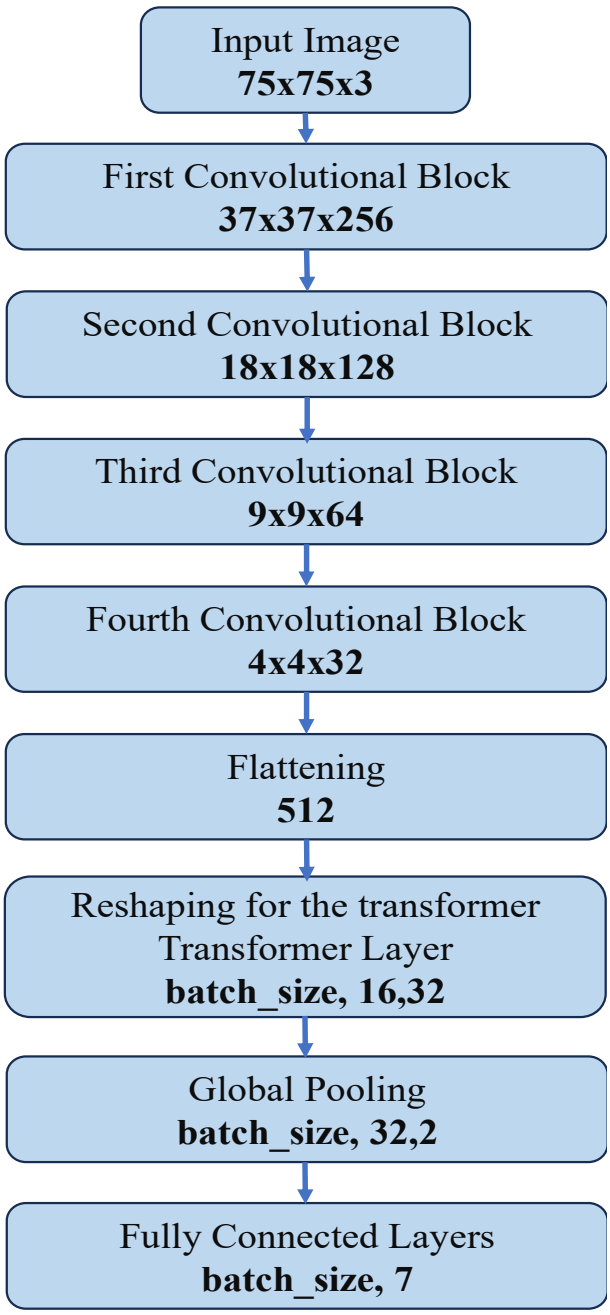

Fig 6. Functional flow chart of DCNNT

F. Tuning Hyper Parameters of DCNNT

Optimizers play a crucial role in training deep learning models like DCNNT. Choosing the appropriate optimizer is essential to balance training efficiency, generalization ability and robustness across different datasets and architectures.

Stochastic Gradient Descent (SGD) is a widely used optimizer for its ability to handle large datasets efficiently. By updating parameters using random subsets of data it reduces memory requirements and accelerates training. However, the inherent noise in its updates can lead to instability and slower convergence. Despite these challenges, SGD remains a popular choice for tasks where gradients are relatively well-behaved, and its performance can be enhanced through techniques such as momentum or learning rate scheduling. Variants like SGD with Nesterov momentum further improve convergence by anticipating future parameter updates. Careful tuning of hyperparameters such as learning rate and batch size is critical to achieving optimal performance, and in many deep learning applications, SGD continues to outperform more complex optimizers when properly configured.

Adagrad, adapts each parameter's learning rate using the past sum of squared gradients. This approach allows it to perform particularly well on sparse data and datasets with widely varying feature scales, making it effective in high-dimensional input spaces. However, its adaptive learning rates decay over time, which can cause the optimization process to stall, especially in deep networks. This diminishing learning rate can prevent models from escaping flat regions in the loss landscape, slowing convergence and limiting Adagrad's applicability to more complex architectures. To address this limitation, variants such as Adadelta and RMSProp have been developed to maintain a more stable learning rate. Nevertheless, Adagrad remains a valuable choice in scenarios where rapid early-stage learning is prioritized over long-term convergence behaviour.

Adam stands out as one of the most effective optimizers for deep networks due to its ability to combine the advantages of both RMSProp and momentum. By maintaining running averages of both the first moment (mean) and second moment (variance) of gradients, Adam provides adaptive learning rates and stable updates. Bias correction ensures these averages are accurate, even in the initial training stages. This makes Adam highly versatile and efficient for a wide range of tasks, particularly in deep architectures like DCNNT, where stability and convergence are critical. Furthermore, Adam's ability to handle sparse gradients and varying learning rates across parameters makes it a preferred choice for many NLP and computer vision applications. With minimal tuning, Adam often delivers competitive performance, making it a strong default optimizer in many deep learning frameworks.

For hyperparameter tuning, several LRs (0.1, 0.01, 0.001, and 0.0001) and DRs (0.2 and 0.3) were tested to find the optimal configuration. SGD and Adagrad provided decent results but fell short in terms of stability and accuracy, particularly on the DCNNT model. The best performance was achieved using Adam with an LR of 0.001 and a DR of 0.2, which effectively balanced overfitting prevention and convergence speed and resulted in an accuracy of 97%. This result is obtained with a train-test split of 80:20.

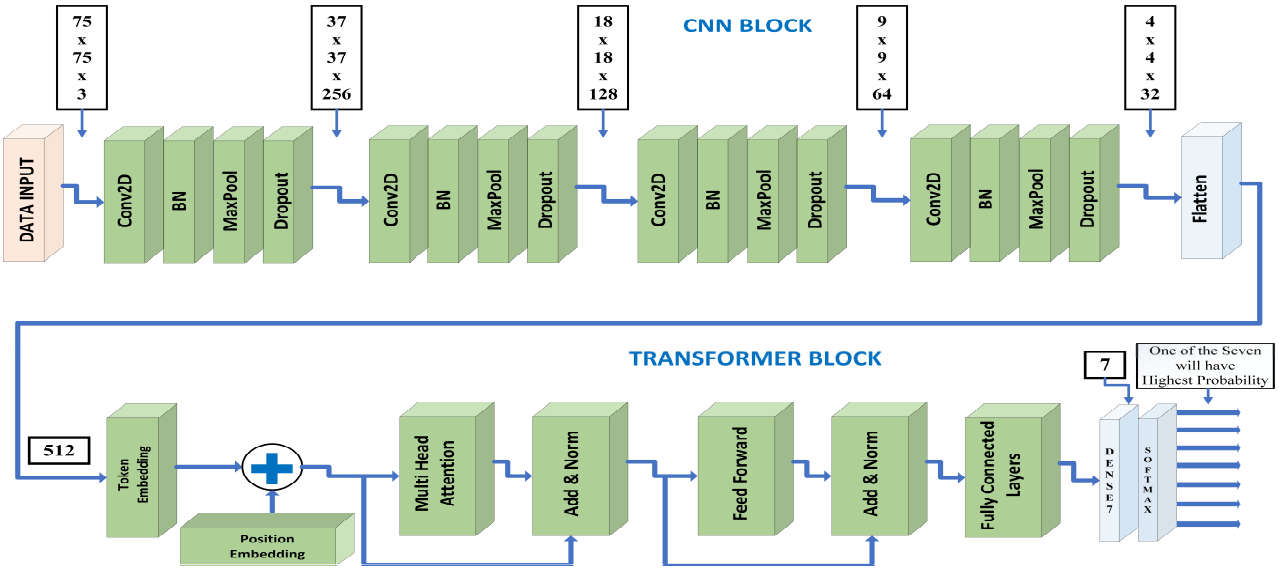

Fig 7. Block Diagram of DCNNT

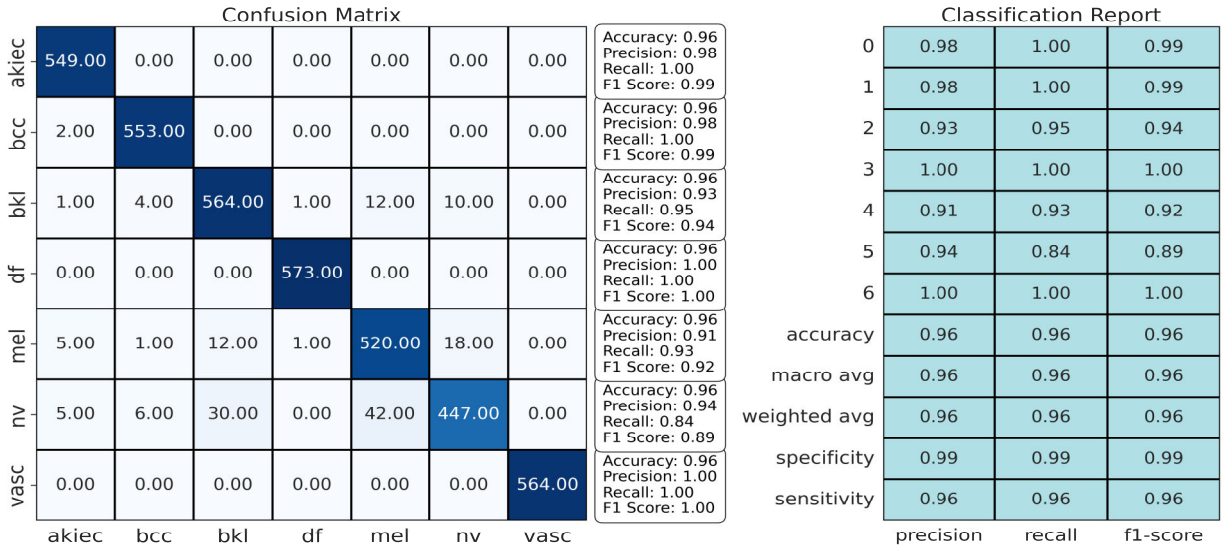

Fig. 8. C-Matrix and C-Report of the model without the transformer layer (DCNN)

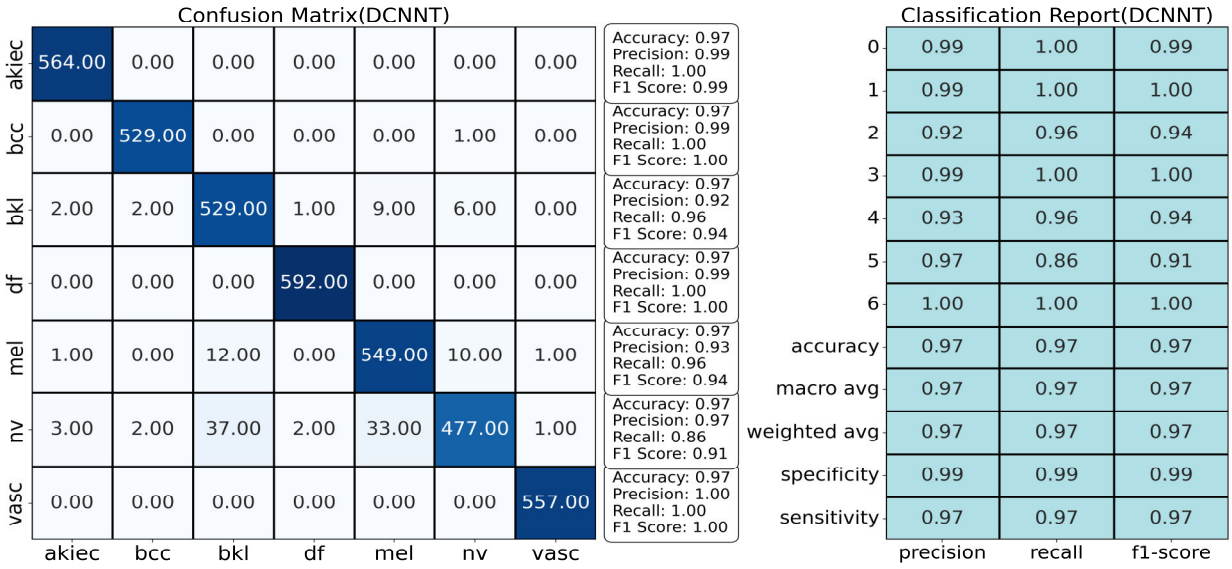

Fig. 9. C-Matrix and C-Report of the model after adding the transformer layer (DCNNT)

TABLE I-IMPACT OF HYPER-PARAMETER TUNING ON THE PERFORMANCE OF DCNNT

Optimizer	LR	DR	Accuracy	Precision	Recall	F-1 Score	AUC (%)	Sensitivity (%)	Specificity (%)
Adam	0.01	0.3	0.65	0.68	0.65	0.65	0.80	0.65	0.92
	0.01	0.2	0.96	0.96	0.96	0.96	0.98	0.96	0.99
	0.001	0.3	0.94	0.94	0.94	0.94	0.96	0.94	0.99
	0.001	**0.2**	**0.97**	**0.97**	**0.97**	**0.97**	**0.98**	**0.97**	**0.99**
	0.0001	0.3	0.90	0.90	0.90	0.90	0.94	0.90	0.98
	0.0001	0.2	0.94	0.94	0.94	0.94	0.97	0.94	0.99
SGD	0.01	0.3	0.88	0.88	0.88	0.88	0.93	0.88	0.98
	0.01	0.2	0.94	0.94	0.94	0.94	0.96	0.94	0.99
	0.001	0.3	0.63	0.70	0.63	0.63	0.79	0.63	0.92
	0.001	0.2	0.78	0.79	0.78	0.77	0.87	0.78	0.96
	0.0001	0.3	0.29	0.33	0.29	0.28	0.59	0.29	0.74
	0.0001	0.2	0.43	0.46	0.42	0.39	0.67	0.42	0.83
Adagrad	0.01	0.3	0.88	0.88	0.88	0.87	0.93	0.88	0.98
	0.01	0.2	0.92	0.92	0.92	0.92	0.96	0.92	0.99
	0.001	0.3	0.61	0.67	0.62	0.61	0.78	0.62	0.91
	0.001	0.2	0.78	0.78	0.78	0.77	0.87	0.78	0.96
	0.0001	0.3	0.45	0.40	0.45	0.41	0.68	0.45	0.83
	0.0001	0.2	0.37	0.42	0.37	0.34	0.63	0.37	0.82

IV. RESULTS AND DISCUSSION

The block diagram of the DCNNT is shown in Fig. 7. Various performance metrics are observed by exploring the Confusion Matrix (C-Matrix), Classification Report (C-Report) of DCNNT.

A C-Matrix is a compact table that provides a clear overview of a classification model's performance. It serves as a valuable resource for evaluating the accuracy of predictions and identifying areas where the model may be misclassifying data.

A classification report is a concise report that highlights the performance of a classification model by providing key metrics such as precision, recall and F1-score. It helps assess the model's effectiveness across different classes. Fig. 8 shows the C-Matrix and C-Report of the model without a transformer layer (DCNN) and Fig. 9 shows the C-Matrix and C-Report of the model with an integrated transformer layer (DCNNT). We can observe that the integration of the transformer layer has improved all the performance metrics. The impact of hyperparameter tuning is shown in Table I. We can observe that DCNNT produced the best results. when the optimizer is Adam with an LR of 0.001 and a DR of 0.2.

TABLE II COMPARISON OF METRICS BETWEEN DCNN AND DCNNT

Performance Metric	DCNN	DCNNT
Accuracy	**0.96**	**0.97**
Precision	**0.96**	**0.97**
F-score	**0.96**	**0.97**
Recall	**0.96**	**0.97**
Sensitivity	**0.96**	**0.97**
Specificity	**0.99**	**0.99**

A comparison of metrics without (DCNN) and with a transformer layer (DCNNT) is shown in Table II. We can observe that almost all metrics, including accuracy, were improved in DCNNT compared to DCNN.

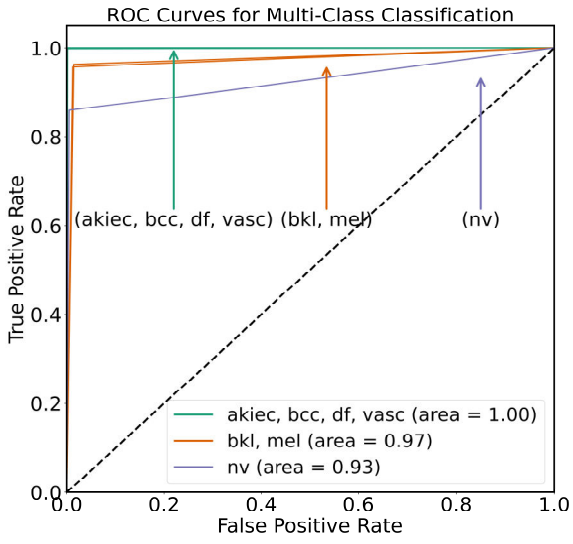

Fig. 10. ROC Curves

An ROC curve is a graphical representation that illustrates the trade-off between the true positive rate (sensitivity) and the false positive rate (specificity) across different thresholds. It provides insights into a model's ability to distinguish between classes, making it particularly useful for binary classification tasks. The area under the curve (AUC) serves as a single metric to summarize overall performance, where a higher AUC indicates better

discrimination capability. Fig. 10 shows the ROC curve and Fig. 11 shows the training and testing accuracy curves.

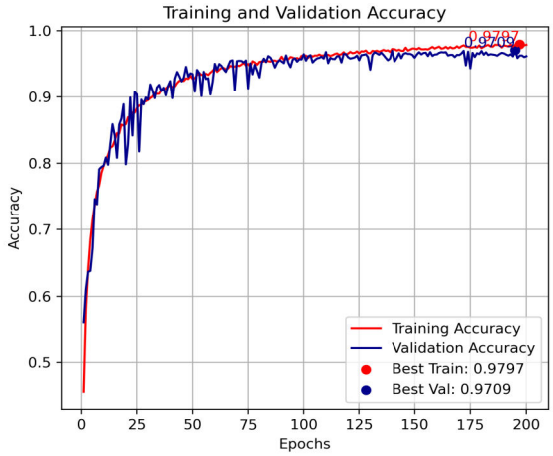

Fig. 11. Training and Testing Accuracy Curves

TABLE III COMPARISON OF PROPOSED MODEL WITH EXISTING MODELS

S.No.	Existing Model (Validated on HAM10000)	Accuracy
1	Xu, Zhijian. et al. [3]	92.79%
2	Xin C. et al. [4]	94.30%
3	Karthik, R. et al. [5]	94.21%
4	Qin, Zhiwei. et al. [6]	95.20%
5	X. Li. et al. [21]	85.00%
6	Rashid. et al. [22]	86.10%
7	Rahi. et. al. [23]	89.00%
8	**Our Proposed DCNNT**	**97.00%**

The consolidated results of the proposed DCNNT model are compared with some of the state-of-the-art models in Table III. We can observe that the proposed DCNNT achieved an accuracy of 97% when Adam is used as an optimizer with LR=0.001 and DR=0.2. The results obtained with our proposed DCNNT model are compared with the existing state-of-the-art models and are presented in Table II.

V. CONCLUSION AND FUTURE WORK

It is evident from the findings that the proposed DCNNT model yields strong and encouraging results, and achieved a promising accuracy of 97% and has outperformed several well-established models.

There is potential to further improve the accuracy by using an ensemble of transfer modelling techniques and our DCNNT. In our future work, we will integrate the DCNNT with models like DenseNet121, INCEPTIONV3 and form an ensemble model that may have the potential of achieving much higher accuracy.

REFERENCES

[1] Tschandl P, Rosendahl C, Kittler H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data. 2018 Aug 14;5(1):1-9.

[2] Abdul Rahaman Shaik, and P. Rajesh Kumar, "Performance Evaluation of Machine Learning Algorithms on Skin Cancer Data Set Using Principal Component Analysis and Gabor Filters," IAENG International Journal of Computer Science, vol. 51, no. 7, pp831-841, 2024.

[3] Xu, Zhijian, Xingyue Guo, and Juan Wang, "Enhancing skin lesion segmentation with a fusion of convolutional neural networks and transformer models," *Heliyon* 10.10 (2024).

[4] Xin C, Liu Z, Zhao K, Miao L, Ma Y, Zhu X, Zhou Q, Wang S, Li L, Yang F, and Xu S, "An improved transformer network for skin cancer classification," Computers in Biology and Medicine, 2022 Oct 1;149:105939.

[5] Karthik, R., et al. "A Hybrid Deep Learning Approach for Skin Cancer Classification using Swin Transformer and Dense Group Shuffle Non-Local Attention Network." *IEEE Access* (2024).

[6] Qin, Zhiwei, et al. "A GAN-based image synthesis method for skin lesion classification." *Computer methods and programs in biomedicine* 195 (2020): 105568.

[7] M. Gallazzi, S. Biavaschi, A. Bulgheroni, T. M. Gatti, S. Corchs, and I. Gallo, "A Large Dataset to Enhance Skin Cancer Classification With Transformer-Based Deep Neural Networks," in *IEEE Access*, vol. 12, pp. 109544-109559, 2024.

[8] Liu, Chiyu, and Cunjie Sun. "A Fusion Deep Learning Model of ResNet and Vision Transformer for 3D CT Images." *IEEE Access* (2024).

[9] S. Hao *et al*., "ConvNeXt-ST-AFF: A Novel Skin Disease Classification Model Based on Fusion of ConvNeXt and Swin Transformer," in *IEEE Access*, vol. 11, pp. 117460-117473, 2023.

[10] Y. Yang, Z. Cai, S. Qiu, and P. Xu, "A Novel Transformer Model with Multiple Instance Learning for Diabetic Retinopathy Classification," in *IEEE Access*, vol. 12, pp. 6768-6776, 2024.

[11] Gou, Quandeng, and Yuheng Ren. "Research on Multi-scale CNN and Transformer-based Multi-level Multi-classification Method for Images." *IEEE Access* (2024).

[12] Sugandha Saxena, S.N Prasad, and Deepthi Murthy T S, "Utilizing Deep Learning Techniques to Diagnose Nodules in Lung Computed Tomography (CT) Scan Images," IAENG International Journal of Computer Science, vol. 50, no.2, pp. 537-552, 2023.

[13] Muhammad Wildan Oktavian, Novanto Yudistira, and Achmad Ridok, "Classification of Alzheimer's Disease Using the Convolutional Neural Network (CNN) with Transfer Learning and Weighted Loss," IAENG International Journal of Computer Science, vol. 50, no.3, pp. 947-953, 2023.

[14] W. Dai, R. Liu, T. Wu, M. Wang, J. Yin, and J. Liu, "Deeply Supervised Skin Lesions Diagnosis with Stage and Branch Attention," in *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 2, pp. 719-729, Feb. 2024.

[15] K. Vanitha *et al*., "Attention-Based Feature Fusion with External Attention Transformers for Breast Cancer Histopathology Analysis," in *IEEE Access*, vol. 12, pp. 126296-126312, 2024.

[16] S. Remya, T. Anjali, and V. Sugumaran, "A Novel Transfer Learning Framework for Multimodal Skin Lesion Analysis," in *IEEE Access*, vol. 12, pp. 50738-50754, 2024.

[17] S. Vachmanus, T. Noraset, W. Piyanonpong, T. Rattananukrom, and S. Tuarob, "DeepMetaForge: A Deep Vision-Transformer Metadata-Fusion Network for Automatic Skin Lesion Classification," in *IEEE Access*, vol. 11, pp. 145467-145484, 2023,.

[18] G. J. Ferdous, K. A. Sathi, M. A. Hossain, M. M. Hoque, and M. A. A. Dewan, "LCDEiT: A Linear Complexity Data-Efficient Image Transformer for MRI Brain Tumor Classification," in *IEEE Access*, vol. 11, pp. 20337-20350, 2023.

[19] S. Hossain, A. Chakrabarty, T. R. Gadekallu, M. Alazab, and M. J. Piran, "Vision Transformers, Ensemble Model, and Transfer Learning Leveraging Explainable AI for Brain Tumor Detection and Classification," in *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 3, pp. 1261-1272, March 2024.

[20] L. Huang, Y. Chen, and X. He, "Foundation Model-Based Spectral–Spatial Transformer for Hyperspectral Image Classification," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1-25, 2024, Art no. 5529825.

[21] X. Li, J. Wu, E. Z. Chen, and H. Jiang, "From Deep Learning Towards Finding Skin Lesion Biomarkers," *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Berlin, Germany, 2019, pp. 2797-2800.

[22] H. Rashid, M. A. Tanveer, and H. Aqeel Khan, "Skin Lesion Classification Using GAN based Data Augmentation," *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Berlin, Germany, 2019, pp. 916-919.

[23] M. M. I. Rahi, F. T. Khan, M. T. Mahtab, A. K. M. Amanat Ullah, M. G. R. Alam, and M. A. Alam, "Detection of Skin Cancer Using Deep Neural Networks," *2019 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, Melbourne, VIC, Australia, 2019, pp. 1-7.

[24] H. Yan, E. Zhang, J. Wang, C. Leng, A. Basu, and J. Peng, "Hybrid Conv-ViT Network for Hyperspectral Image Classification," in *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1-5, 2023, Art no. 5506105.

[25] A. Jamali, S. K. Roy, A. Bhattacharya, and P. Ghamisi, "Local Window Attention Transformer for Polarimetric SAR Image Classification," in *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1-5, 2023, Art no. 4004205.

Abdul Rahaman Shaik is a PhD student in the Department of Electronics and Communication Engineering at the College of Engineering, Andhra University, Visakhapatnam, Andhra Pradesh, India. He earned his M.Tech from the National Institute of Technology, Warangal, Telangana, India. At present, he serves as an Associate Professor in the Department of Electronics and Communication Engineering at Vishnu Institute of Technology, Bhimavaram, Andhra Pradesh, India. His research focuses primarily on image processing and computer vision.

Dr. P. Rajesh Kumar obtained his Ph.D. from the College of Engineering, Andhra University, Visakhapatnam, Andhra Pradesh, India, in 2007. He is a Professor in the Department of Electronics and Communication Engineering at the same institution. Over the years, he has taken on leadership roles such as Head of the Department and Assistant Principal at the College of Engineering, Andhra University. He has authored numerous research publications in reputed national and international journals and conferences and has supervised multiple research projects. His areas of interest include digital signal and image processing, computational intelligence, human–computer interaction, and radar signal processing.