Skin Lesion Diagnosis with Vision and Swin Transformers: A Multi-Model Approach

Kashish Verma[1*], Surbhi Bharti[1], Saachi Verma[1†], Prachi Singla[1†], Pankaj Gupta[1†], Ashwni Kumar[1†]

[1*]Department of Electronics and Communication Engineering, Indira Gandhi Delhi Technical University for Women (IGDTUW), Kashmere Gate, New Delhi, 110006, Delhi, India.

[†]These authors contributed equally to this work.

Abstract

Skin cancer is still one of the most common cancers around the world and early and accurate diagnosis is an important factor for a patient's prognosis. The research presented in this study investigates the viability of three transformer based deep learning architecture, Vision Transformer (ViT), Swin Transformer, and Medical Vision Transformer (MedViT), for automated classification of dermatological lesions. The auto-classification models were trained and validated on two distinct dermatological datasets, HAM10000 and MSLDv2.0. Both datasets consist of 14 lesion classes. A comprehensive pre-processing, including image normalization, image resizing, and data augmentation techniques, was performed to ensure model generalizability. A comparative performance assessment of the auto-classification technologies against traditional CNN architectures, DenseNet201 and ResNet152, demonstrated that the Swin Transformer auto-classified with the most accuracy at 93.18%, followed by MedViT and ViT, respectively. Overall, the Swin Transformer, ViT, and MedViT performs better than conventional CNN (Convolutional Neural Network) architectures for auto-classifying skin lesions, and demonstrates how attention based architectures in dermatological imaging could be utilized for clinical decision-support systems for early skin cancer detection.

Keywords: Skin cancer classification, Vision Transformer (ViT), Swin Transformer, MedViT, Deep learning

1 Introduction

Among neoplastic disorders, cutaneous malignancies account for one of the most prevalent neoplasms worldwide, affecting people of all cutaneous phenotypes, as well as demographic variables such as age and sex [1]. Among the various manifestations, attention should be given to melanoma as it is particularly invasive and metastatic with delayed diagnosis [2]. The melanocytes are cells that are responsible for cutaneous pigmentation, from which this malignancy arises and is mostly caused by excess ultraviolet (UV) radiation exposure derived from both solar and artificial sources through the use of tanning apparatus [3]. Genomic alterations induced by UV radiation undermine the normal cellular regulatory mechanisms, resulting in dysregulated cell proliferation and neoplastic formation [4]. Early detection of melanoma is very important, as five-year survival rates above 90% have been correlated with early therapeutic intervention [5]. However, with delayed diagnosis, the probability of lymphatic and distant organ metastasis greatly increases, survival outcomes are further compromised, and therapeutic approaches are further complicated [6]. Such important factors reinforce the necessity for highly precise, efficient, and automated diagnostic methods in order to improve early detection and the outcome of the patients.

In the field of automated cutaneous malignancy classification, artificial intelligence (AI) and deep learning (DL) architectures have brought many profound influences on the ways medicine analyzes medical images, where modern advances in those fields have intersected with those techniques [7]. Over the course of the field, computer-aided diagnostic (CAD) methodologies that aid in greater detection precision and reduce or alleviate human interpretive errors have achieved widespread implementation after applying machine learning algorithms on comprehensive dermoscopic image repositories [8]. In recent years, majority of the approaches have relied on CNN, including ResNet152, DenseNet201 and Xception, among others, which have achieved good results in skin lesion categorization [9, 10].

In this effort, the automated classification of cutaneous lesions is explored within ViT, Swin Transformer, and MedViT architectures [11]. Research is done with a large dataset comprising HAM10000 (2019) and MSLDv2.0, which cover 14 different lesion typologies [12]. These datasets are integrated together for training of models across heterogeneous and representative specimens, enhancing the model's generalization capabilities with respect to a wide spectrum of patients. We implement sophisticated preprocessing of images such as normalization, dimensional standardization, and augmentation techniques to yield better data quality and prevention against overfitting. Additionally, we assess the discriminative power of transformer-based models to distinguish between malignant and benign manifestations of dermatologic lesions from the standpoint of comparison with conventional CNN-based networks DenseNet201 and ResNet152.

According to the empirical evidence, Swin Transformer and MedViT architectures outperform statistically significant from the conventional ViT and the CNN architecture. These superior models have a better diction power between the diagnostic categories, more robust feature extractions and better classification accuracy metrics.

Through these architectural innovations, these techniques enable a methodology, computational efficiency, and diagnostic precision for detecting cutaneous malignancy in its early stages. The contribution of the present investigation is significant to the development of artificial intelligence augmented clinical decision support systems through its implementation of these sophisticated algorithmic frameworks. Technological interventions of such kind have the potential to drastically improve the precision of dermatological diagnosis, likely with secondary positive feedback effects on the therapeutic interventions, patient outcomes and mortality indices of the dermatologic malignancies.

Furthermore, the subsequent organizational framework of this manuscript is structured along the following structural delineation: Section 2 comprises an overall exhaustive review of the existing behavior in the extant scholarly literature regarding the application of deep learning methodological approaches in cutaneous malignancy classification paradigms. Section 3 describes the material used in the corpus of dermatological imagery, preprocessing algorithmic procedures, architectural configurations of computational models, and training protocols. The empirical outcomes of Section 4 are evaluated in detail, and ViT, Swin Transformer, and MedViT architectural frameworks are compared with each other using a comparative evaluation in terms of their diagnostic efficacy. Section 5 concludes with the manuscript and finally combines the principal findings, critiques methodological constraints, and outlines the prospective directions for future scientific inquiry in this domain.

2 Related Work

B. Li et al. introduced BiDFNet, which combines a convolutional and a transformer-based pathway toward a better skin lesion segmentation [13]. In the model, the global context and fine-grained details are fused with a bidirectional fusion strategy, bypassing the limitation of pure transformer or CNN-based models. It was evaluated using the ISIC2017 dataset; it outperformed baseline models on the Dice Similarity Coefficient (DSC) and Jaccard Index based on using complementary representations in feature space.

In the work by L. Zheng et al., they proposed MHAUFormer, composed of multi-head attention and U-Net-like architecture to capture local dependencies at different scales across multiple scales [14]. Specifically, it aims at improving the low-level feature retention while extracting high-level semantic features important for skin lesion detection. On datasets such as ISIC2018 and PH2, this model was able to better segment the tissue as well as delineate the boundaries of the tissue compared to traditional U-Net and Transformer models.

A hybrid model of an attention mechanism and Swin Transformer block to be applied on the U-Net backbone to provide better segmentation accuracy has been presented by S. Zhang et al.; we call our model Att-SwinU-Net [15]. The model employs a

hierarchical structure with shifted window attention for maintaining spatial information on different resolutions. Att-SwinU-Net was evaluated on ISIC2018 and Derm7pt datasets, and it achieved the improvement of sensitivity and IoU, especially boosting the performance in the diagnosis of irregular and fuzzy lesion borders. J. Wu et al. [16]develop the model SLMNet based on spatial layout modeling and context refinement for dermoscopic image segmentation. It employs a layout-aware attention module to aid the understanding of lesion shapes and boundaries at pixel-level precision. To benchmark our model, SLMNet was tested on the ISIC 2018 and PH2 datasets and obtained a better precision versus recall trade-off well beyond the lesion type used for training.

Figured out in the study of R. Huang et al. is SECAUNet — consisting of Squeeze and Excitation (SE) blocks and channel attention mechanisms within the U-Net structure [17]. At a channel level, the proposed network is built to dynamically recalibrate feature response based on the channel and enhance focusing on lesion-relevant features. On the ISIC2016 and ISIC2018 datasets, it was evaluated and gave state-of-the-art results on the Dice score and specificity.

TransUNet++ is a nested transformer-CNN hybrid model proposed by J. Liu et al. that improved the skip connection strategy from TransUNet for multi-scale lesion segmentation [18]. As such, it puts particular emphasis on hierarchical semantic guidance and inter-scale feature fusion. To capture such complex lesion structure, TransUNet++ achieved consistent improvements in segmentation metrics of Dice and Hausdorff distance on ISIC2018 and PH2 datasets.

Y. Chen et al. presented SwinUnet3+, which is a hierarchical encoder–decoder model composed of Swin Transformers and the full-scale skip connection of UNet3+ [19]. With such architecture, the context and the spatial resolution are preserved at many scales, and it is effective for the detailed segmentation. The model compared favorably on ISIC2018 and PH2 datasets to its CNN-based predecessors, especially for the cases with small lesions or low contrast areas.In this case, the DEU-Net was developed by A.

Wang et al., which merged dual encoders (one of the CNN-based type and the other of the transformer-based type) via a dynamic fusion module for capturing joint local texture and global context [20]. In addition, it applies dynamic attention for more effective adaptation to lesion sizes and shapes. The model has been tested on ISIC2018, PH2, and Derm7pt, and it shows a high segmentation accuracy and generalizability.

Combining a multi-scale transformer with an attention-based CNN module, HMT-Net, introduced by K. Guo et al. [21] improves the boundary detection and context awareness. In the architecture, there are multiple attention fusion layers that further refine segmentation by iterative updates. Dry HMT-Net also achieved some improvements over previous methods in terms of the Dice coefficient and boundary recall while remaining inferior to the HeNORM subset when tested on the ISIC2018 and ISBI2016 datasets, especially for thin or irregular lesions.

M. Zhang et al. proposed DualFormer, a dual-pathway transformer model based on dual attention between channel and spatial attention on two transformer branches to further improve representation ability [22]. It is specially architected to have high segmentation resolution and low latency inference. On multiple benchmarks of ISIC2018,

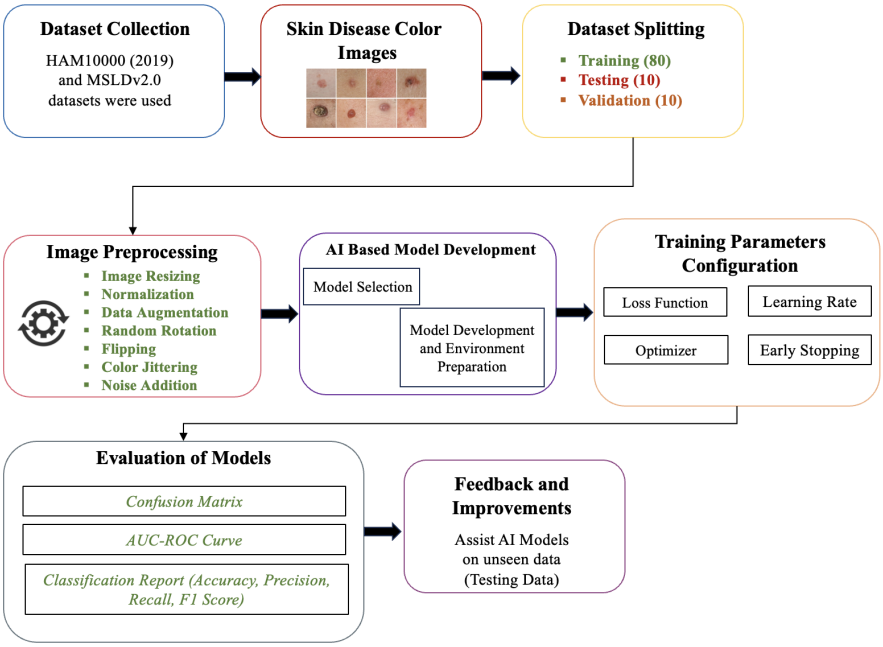

Fig. 1: The Proposed Methodology- Overall Abstract View

PH2, and Derm7pt, it was evaluated with the highest IoU and Dice scores among recent transformer-based models, evidence that it could potentially be used in real-time AI-assisted dermatology.

3 Methodology

In this study, we explore the application of ViT, Swin Transformer, and MedViT for classifying skin lesions. By using self-attention mechanisms, we aim to enhance feature extraction and improve classification accuracy. The workflow of our methodology can be visualized as five important stages: dataset collection, preprocessing, architectures of the model, training strategy, and finally evaluation metrics. For clarity and comprehensiveness in comparing the performance of the transformer-based models, we developed a systematic pipeline.

Figure 1 represents a clear visual representation of the step-by-step workflow involved in this methodology.

3.1 Dataset Collection and Preprocessing

3.1.1 Dataset Description

The dataset utilized for this study consists of a combination of HAM10000 (2019) and MSLDv2.0, providing an extensive array of 14 different types of skin lesions. These datasets were selected because of the variability in appearance of the lesions, the difference in skin tones, and the equal representation of benign and malignant lesions.

To assess the effectiveness of the model, the dataset was partitioned into three subsets: 80% for training the model, 10% for validation (used for hyperparameter optimization and early stopping), and 10% for testing (used for the final assessment of the model).

3.1.2 Preprocessing Steps

To make the dataset compatible with transformer-based architectures, we performed several preprocessing steps aimed at enhancing the model's generalization and reducing overfitting.

First, all images were resized to 224×224 pixels to ensure consistency across the dataset. Following this, pixel values were normalized to the range [0, 1] through a min-max normalization process (Equation 1), defined as:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

where x is the original pixel value, and $x_{\min}$ and $x_{\max}$ are the minimum and maximum pixel values, respectively.

Furthermore, data augmentation techniques were systematically employed in our experimental framework in order to enhance model generalizability and overcome limited training sample issues. The augmentation protocol included several valid approaches. To tackle orientation differences, we applied random rotations of ±15 degrees on the X and Y axes. To introduce orientation diversity in lesions, horizontal and vertical flipping transformations were adopted, which could potentially reduce positional bias in model training. We altered brightness and contrast to reduce bias in the B and C parameters, respectively, that arise from underexposure. To boost the model's ability to tackle these common issues, Gaussian noise was introduced to the training samples. A lot of augmentation procedures were applied to improve the model's generalization across the different acquisition settings.Figure-1 illustrates the preprocessing pipeline, showcasing original images, normalized outputs, and augmented samples.

3.2 Model Architectures

This study evaluates three transformer-based architectures— ViT, Swin Transformer, and MedViT—to assess their effectiveness in skin lesion classification.

3.2.1 ViT

ViT architecture processes images in a fundamentally different way compared to convolutional neural networks. It works by creating patches of the input image and then processing them using self-attention mechanisms. This architecture divides input images into a series of non-overlapping patches and employs self-attention mechanisms to establish spatial relationships across the entire image domain.

This framework consists of several key components: first, the *patch embedding* step splits the image into fixed 16×16 patches and projects these units into a D-dimensional embedding space. Since transformers lack inherent spatial understanding, learnable *positional encodings* are added to the patch representations to retain critical spatial information.

The fundamental computational building block of the Vision Transformer is the Multi-Head Self-Attention (MHSA) mechanism, as shown in Equation 2:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

Here Q, K, and V refer to the query, key, and value matrices respectively, while d_k stands for the dimensionality of the keys. This mechanism allows the model to learn long-range dependencies and contextual relationships among different regions of the image.

Finally, the architecture concludes with a *classification head*, where the final embedding is passed through a fully connected layer with softmax activation to produce probability distributions for the diagnosis classes.

3.2.2 Swin Transformer

Compared to the conventional ViT model, the Swin Transformer architecture introduces significant enhancements through its shifted window-based self-attention mechanism. This mechanism enables the model to learn hierarchical representations of features while improving computational efficiency.

The framework incorporates several novel components. First, the hierarchical patch merging method aggregates image patches at multiple levels within the network, creating a multiscale representation of visual information. Notably, self-attention is computed over small, non-overlapping windows rather than the entire image, which significantly reduces computational complexity while still capturing expressive features.

To learn global dependencies, the model shifts the windows between consecutive layers, connecting previously non-neighboring patches. Additionally, patch merging layers are used to downsample feature maps in a controlled manner, reducing spatial resolution while preserving hierarchical information across the network.

As a result, the Swin Transformer architecture is capable of processing higher-resolution images more efficiently than other ViT models, while maintaining comparable or even superior performance in dermatological image classification tasks.

3.2.3 MedViT

The MedViT framework is a unique implementation of transformer methods designed specifically for the analysis of medical imaging data. It addresses certain limitations of image classification using current Vision Transformers for low-contrast, texture-rich images that are associated with medical imaging. This customized framework has some very important features.

Most notably, it adds hybrid convolutional-transformer blocks that front-load traditional CNN layers before self-attention blocks, allowing the model to optimally attend to or learn fine-grained spatial features and local textures that are important for diagnosis in dermatology. MedViT includes multi-scale feature fusion that systematically combines information physiologically across resolutions to capture differential detailed examples of pathological changes and contextual anatomical assessment both concurrently.

MedViT implements imaging augmentation methods developed especially for medical imaging and pre-trains the domain-specific pre-training dataset, improving the model's performance on the medical domain-specific diagnostic task when transferring learning. This contrasts with the architect's previously published results on natural images, and when the natural image data would lead to an expected difference in the incident.

In sum, this custom architecture allows MedViT to successfully address stones associated with dermatology images, including subtle color shifts, an irregular edge, and diagnostic textures within which a general-purpose vision thinker system (like CNNs) would fail to resolve..

3.3 Training Procedure

The models were trained using supervised learning techniques with a categorical cross-entropy loss function. Several key parameters and techniques related to training configurations were included to improve the overall performance of the model and to ensure their convergence.

3.3.1 Loss Function

The training process used categorical cross-entropy as the main optimization objective.

Loss function: Categorical cross-entropy loss calculates the distance between the output softmax probabilities and the actual class labels. By penalizing incorrect predictions and also penalizing the confidence in those wrong predictions, it encourages the model to optimize not just for accuracy but also for reliability.

3.3.2 Optimizer

The AdamW optimizer was used for optimization during training since it improves the standard Adam optimization algorithm. One unique aspect of AdamW that differentiates it from others is that it uses weight decay regularization, which directly penalizes large parameter values in order to mitigate overfitting. In this manner, it separates weight decay from its gradient updates, enabling better regularization results than what L2 regularization obtains.

3.3.3 Learning Rate

During model training, we selected an initial learning rate of 1×10^{-4} to begin our training process. To mitigate the learning rate during training, we used a learning rate scheduler method called cosine annealing. This approach decrements the learning rate based on a cosine curve, allowing the parameters of the model to be tuned more precisely as the training process progresses, thereby improving the convergence behavior of the model.

3.3.4 Batch Size

We conducted the training of the neural networks using a batch size of 32 images. The batch size represents a trade-off computer scientists and engineers often have to make, between computational efficiency and training stability. A batch size of 32 images provides enough accuracy to estimate the gradient while also being a reasonable batch size to fit in memory.

3.3.5 Early Stopping & Checkpoints

In order to prevent overfitting and retain the best-performing model states during training we utilized an early stopping mechanism. Early stopping entails the monitoring of the validation performance metrics, ceasing training whenever the performance metric did not improve. Furthermore, checkpoint strategies were used to carry information regarding the storage of the best performing model configurations, whereby the optimal model parameterizations would be stored.

3.4 Evaluation Metrics

The assessment of all the models involved multiple complementary metrics, each of which described different aspects of performance.

3.4.1 Accuracy

Accuracy is a simple metric that describes the proportion of instances from both classes (i.e., positive and negative) that are identified correctly. Accuracy is a general and intuitive indicator of model performance and is especially useful when classes do not exhibit large distribution differences.

3.4.2 Precision, Recall, and F1-Score

Precision quantifies the ratio of true-positive predictions to all positive predictions and is an indicator of how well the model is able to make a positive classification without false-positives. Precision is a critical factor when the penalties or implications of false-positives are potentially costly.

Recall (or sensitivity) measures the proportion of actual positive cases that have been identified by the model as positive. Recall measures how well the model is able to detect all relevant participants instances/subjects related to the evaluation, and is an important consideration when the cost of missing a positive result is a concern.

The F1-Score is the harmonic mean of precision and recall. The F1-Score gives the overall result that considers both false-positive and false-negative cases. The F1-Score is especially valuable when class imbalance is occurring in the sample or when both precision and recall are deemed equally important.

3.4.3 AUC-ROC Curve

The AUC-ROC (Area under the Receiver Operating Characteristics)curve quantifies the ability of the model to discriminate between classifications over multiple thresholds. The measure will calculate how well the model distinguishes the classes regardless of any actual threshold decision point. The higher the score approaches 1, the better the model ranks a positive instance higher than the negative instances in terms of classification. The AUC provides a measure of classification not just limited to the actual threshold decision point.

4 Evaluation and Discussions

This section provides a comprehensive look at the experimental results from the training and evaluation of the ViT, Swin Transformer, and MedViT models in classifying skin lesions. We evaluated their performance through various metrics, such as accuracy, loss, precision, recall, F1-score, and the Area Under the AUC-ROC curve. A thorough comparison of these metrics helps us understand how well each model can differentiate between different types of skin lesions. Moreover, this section also discusses the training convergence behavior of each model, their ability to generalize to unseen data, and the overall effectiveness of transformer-based architectures in medical image analysis [23].

4.1 Evaluation Metrics

For our evaluation of transformer-based model performance, we employed benchmark metrics to measure both discriminative performance and misclassification rates for false predictions[24].

4.1.1 Accuracy

Accuracy (A), as defined in Equation 3 [25] at which the instances were correctly classified, and is calculated as:

$$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

where TP (True Positives) represents the number of correctly classified malignant lesions, TN (True Negatives) represents the number of correctly classified benign lesions, FP (False Positives) is the count of benign lesions misclassified as malignant, and FN (False Negatives) is the count of malignant lesions misclassified as benign.

4.1.2 Loss Function

The loss function (L), defined in Equation 4, measures the classification mistakes by comparing the label with the cross-entropy:

$$L = -\sum [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \tag{4}$$

where y_i is the true label and p_i is the predicted probability. Models converge better for optimization when they have lower loss [26].

4.1.3 Precision

Precision (P), as shown in Equation 5, measures how many of the positive cases were correct:

$$P = \frac{TP}{TP + FP} \tag{5}$$

This metric indicates how many of the lesions categorized as malignant were, in fact, malignant .

4.1.4 Recall (Sensitivity)

Recall or Sensitivity (R), defined in Equation 6, measures how well the model can find all positive cases:

$$R = \frac{TP}{TP + FN} \tag{6}$$

Higher Recall scores indicate better detection of malignant lesions—a significant determinant in dermatological diagnosis .

Tables 1, 2, and 3 provide a summary of the evaluation metrics for each model and confirm the superiority of ViT and Swin Transformer in the classification of skin lesions.

Table 1: Classification Report for ViT

Class	Precision	Recall	F1-Score	Support
Actinic keratoses	0.82	0.78	0.80	88
Basal cell carcinoma	0.91	0.94	0.93	333
Benign keratosis-like lesions	0.84	0.79	0.81	263
Chickenpox	1.00	0.98	0.99	113
Cowpox	0.99	0.99	0.99	99
Dermatofibroma	0.92	0.88	0.90	25
HFMD	1.00	1.00	1.00	242
Healthy	1.00	0.99	0.99	171
Measles	0.99	1.00	1.00	83
Melanocytic nevi	0.91	0.95	0.93	1288
Melanoma	0.83	0.80	0.81	453
Monkeypox	0.99	0.99	0.99	426
Squamous cell carcinoma	0.86	0.73	0.79	64
Vascular lesions	0.90	0.92	0.91	26
Accuracy			0.92	
Macro Avg	0.94	0.90	0.92	3674
Weighted Avg	0.92	0.92	0.92	3674

Table 2: Classification Report for Swin Transformer

Class	Precision	Recall	F1-Score	Support
Actinic keratoses	0.78	0.80	0.79	88
Basal cell carcinoma	0.90	0.93	0.91	333
Benign keratosis-like lesions	0.85	0.78	0.81	263
Chickenpox	1.00	0.97	0.99	113
Cowpox	1.00	0.99	0.99	99
Dermatofibroma	0.95	0.80	0.87	25
HFMD	0.99	1.00	0.99	242
Healthy	0.99	1.00	1.00	171
Measles	1.00	1.00	1.00	83
Melanocytic nevi	0.92	0.95	0.93	1288
Melanoma	0.82	0.78	0.80	453
Monkeypox	0.99	1.00	0.99	426
Squamous cell carcinoma	0.87	0.73	0.80	64
Vascular lesions	0.93	0.96	0.94	26
Accuracy		0.92		
Macro Avg	0.93	0.90	0.92	3674
Weighted Avg	0.92	0.92	0.92	3674

Table 3: Classification Report for MedViT

Class	Precision	Recall	F1-Score	Support
Actinic keratoses	0.41	0.32	0.36	88
Basal cell carcinoma	0.61	0.80	0.69	333
Benign keratosis-like lesions	0.53	0.37	0.43	263
Chickenpox	0.88	0.94	0.91	113
Cowpox	0.94	0.95	0.94	99
Dermatofibroma	0.88	0.28	0.42	25
HFMD	0.95	0.98	0.95	242
Healthy	0.98	0.98	0.98	171
Measles	0.94	0.98	0.96	83
Melanocytic nevi	0.81	0.88	0.84	1288
Melanoma	0.59	0.53	0.56	453
Monkeypox	0.97	0.92	0.94	426
Squamous cell carcinoma	0.50	0.16	0.24	64
Vascular lesions	0.77	0.77	0.77	26
Accuracy		0.78		
Macro Avg	0.76	0.70	0.71	3674
Weighted Avg	0.77	0.78	0.77	3674

4.2 Accuracy and Loss Comparison

Regarding analysis of deep learning models, the most important part of evaluation is the training and validation accuracy trends and convergence behavior in relation to loss function. Accuracy and loss curves at multiple epochs are plotted to check how well each model learns from data and generalises to unseen examples.

4.3 AUC-ROC Curve Analysis

The AUC-ROC curve is an essential metric to denote the model ability to discriminate between the classes. It means an increased capability of the model to distinguish malignant and benign lesions. As shown in Figures 2, 3, and 4, the Swin Transformer gives the highest AUC compared to MedViT and the ViT generally has the lower overall classification performance.

4.3.1 Observations from AUC-ROC Curves

Our test results show how different transformer-based models perform in skin image analysis. The Swin Transformer stands out achieving the highest Area Under the Curve (AUC) scores among the models we tested. This model's layered approach and its shifted window attention feature seem to capture lesion details at various scales better than other methods. MedViT's mixed approach shows clear improvements over the basic ViT. This is likely because it combines convolutional layers with self-attention mechanisms allowing it to capture both small-scale texture details and big-picture context that are key to accurate diagnosis. Even with these steps forward, the standard ViT's lower AUC points to its struggles with the complex patterns and subtle edge features often seen in skin lesions. This suggests that using transformer models without tweaking them for specific medical tasks might not be the best choice for detailed medical image analysis.

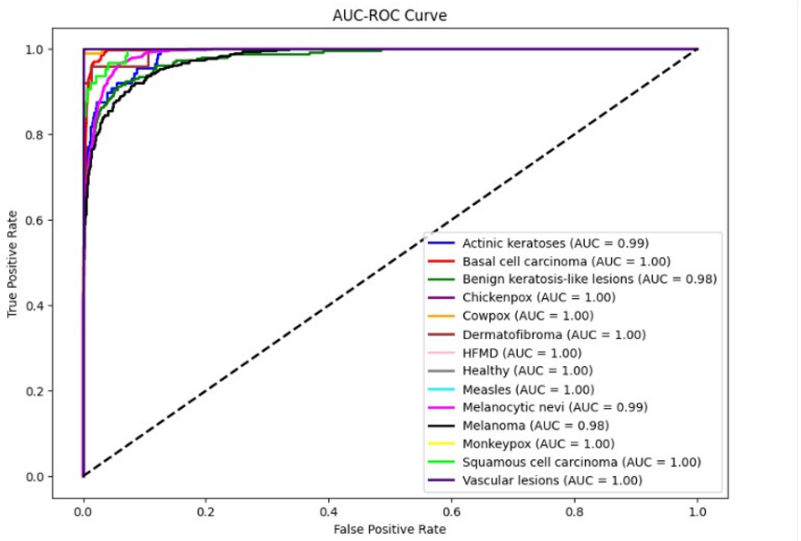

Fig. 2: AUC-ROC Curve for ViT

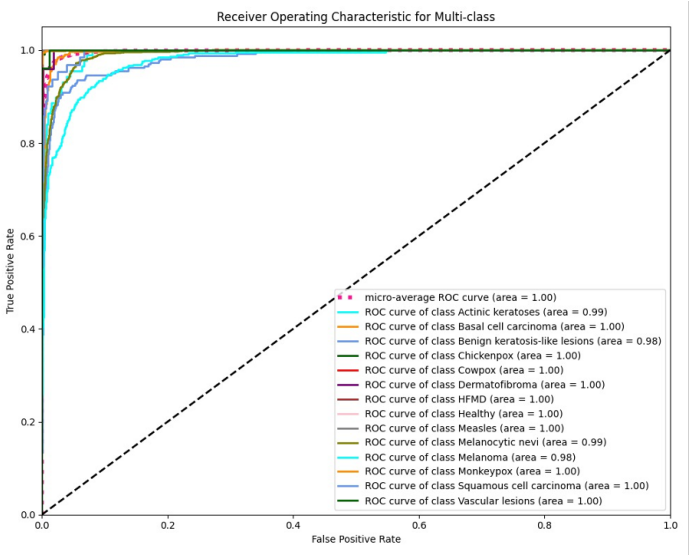

Fig. 3: AUC-ROC Curve for Swin Transformer

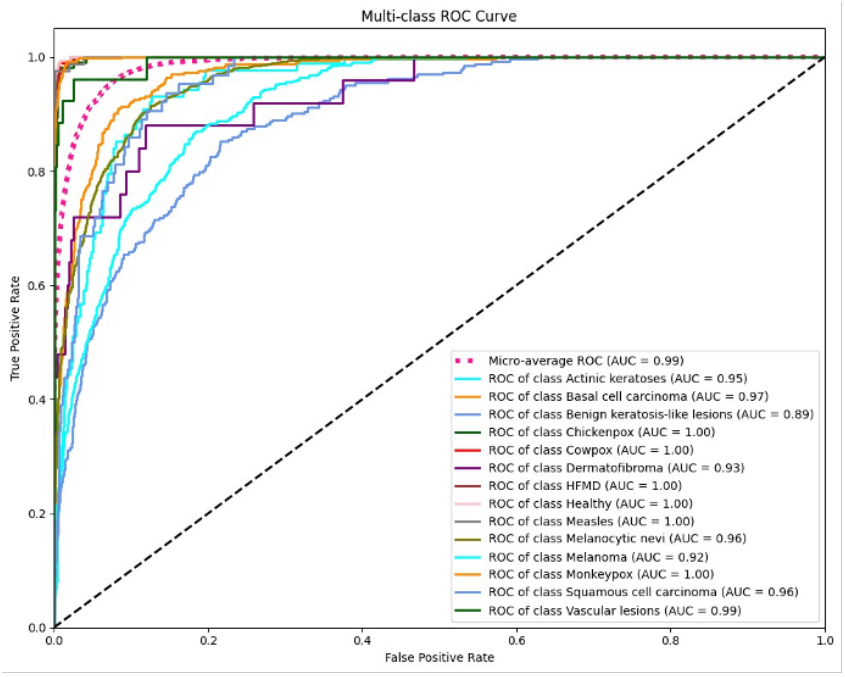

Fig. 4: AUC-ROC Curve for Med ViT

4.4 Confusion Matrix Analysis

The confusion matrix provides a summary of every model classification performance with respect to different types of skin lesions. The distribution shows the actual occurrence of TP, FP, TN, and FN in all categories.

Observations from Confusion Matrices

The performance of each model was further analyzed through its confusion matrix, which provides insights into class-wise predictions and misclassifications. These are illustrated in Figures 5, 6, and 7.

Our test results show how different transformer-based models perform in skin image analysis. The Swin Transformer stands out, achieving the highest AUC scores among the models we tested. The layered approach of this model and its shifted window attention feature seem to capture the lesion details at various scales better than other methods. MedViT's mixed approach shows clear improvements over the basic ViT. This is likely because it combines convolutional layers with self-attention mechanisms, allowing it to capture both small-scale texture details and big-picture context that are key to accurate diagnosis. Even with these steps forward, the standard ViT's lower AUC points to its struggles with the complex patterns and subtle edge features often seen in skin lesions. This suggests that using transformer models without tweaking them for specific medical tasks might not be the best choice for detailed medical image analysis.

A quantitative analysis of confusion matrices shows multiple significant performance differences between the transformer-based architectures in our study. The Swin Transformer's highest true positive rate demonstrates its exceptional sensitivity in detecting malignant lesions, which holds great importance for clinical dermatology because missed diagnoses can lead to serious consequences. The ViT produces numerous false negatives, which cause it to incorrectly identify malignant lesions as benign, likely due to its insufficient capability to analyze detailed texture patterns that reveal minor cancerous changes.

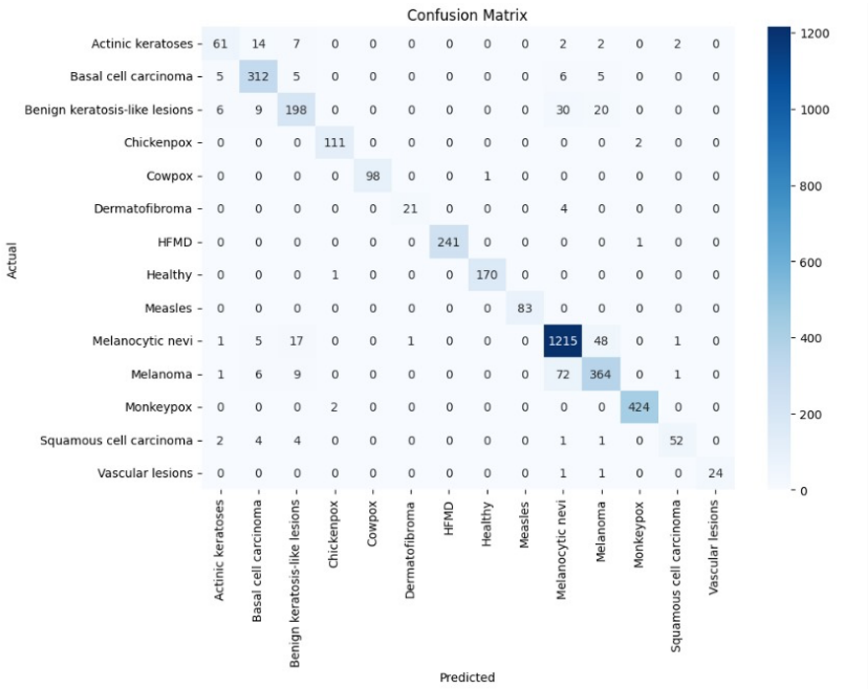

Fig. 5: Confusion Matrix for ViT

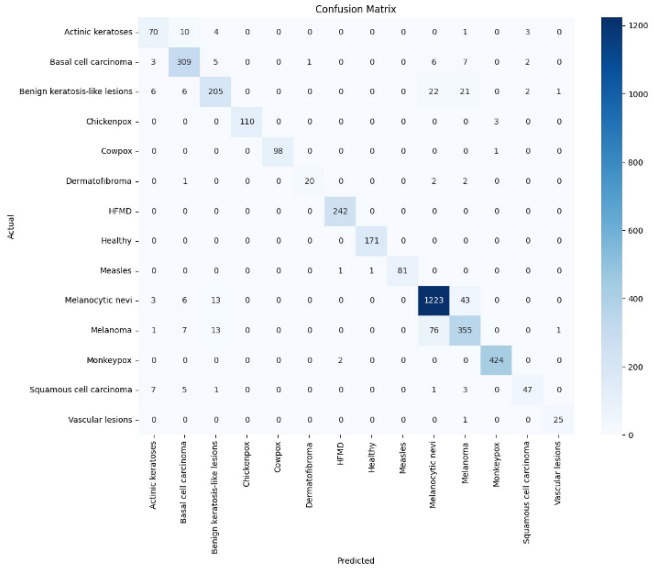

Fig. 6: Confusion Matrix for Swin Transformer

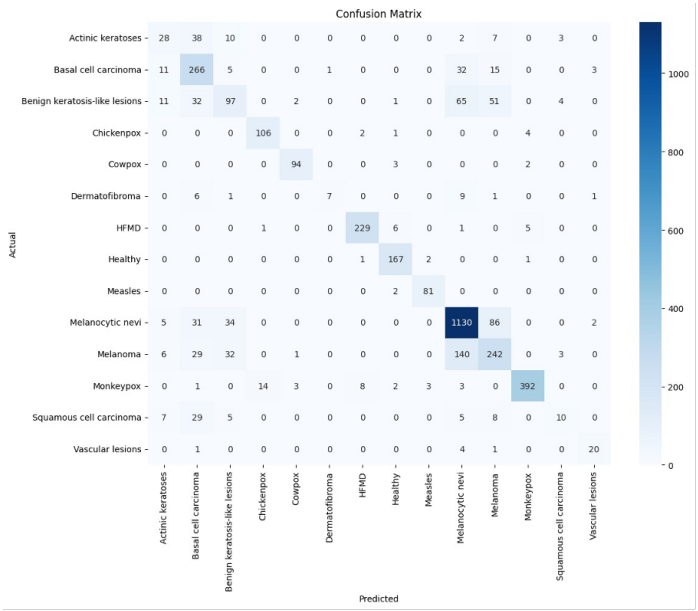

Fig. 7: Confusion Matrix for Med ViT

5 Conclusion

The classification of skin lesions using the Swin Transformer is shown.This is better from ViT and MedViT, as the hierarchical self attention mechanism improves the accuracy and AUC-ROC score. Both MedViT and ViT were competitive to a hybrid CNN transformer approach, but ViT suffered from visual similarity of lesions. Thus, this demonstrates the possibility of using transformer based models in AI in dermatology, but again, before using them in real-world scenarios, further optimisation and validation are needed to ensure the robustness and reliability of medical diagnostics. For future research, the dataset should be more diverse and can be done by integrating data from real world clinical images of different demographics, thereby making the model more robust. For further optimizing the feature extraction from medical images, hybrid CNN transformer architectures may be used. Additionally, hybrid CNN transformer architectures may serve for further optimizing the feature extraction from medical images. Additionally, multimodal learning that uses dermoscopic images with patient metadata might be more holistic diagnostic. Finally, explainable AI techniques for practitioners will be applied to the model interpretability, including Grad-CAM and SHAP. Finally, at last it must be validated in the real world, i.e., in clinical settings, to check the reliability and scalability of AI-driven applications in dermatology.

Acknowledgements. Acknowledgements are not compulsory. Where included they should be brief. Grant or contribution numbers may be acknowledged.

Please refer to Journal-level guidance for any specific requirements.

Declarations

Some journals require declarations to be submitted in a standardised format. Please check the Instructions for Authors of the journal to which you are submitting to see if you need to complete this section. If yes, your manuscript must contain the following sections under the heading 'Declarations':

- Funding
- Conflict of interest/Competing interests (check journal-specific guidelines for which heading to use)
- Ethics approval and consent to participate
- Consent for publication
- Data availability
- Materials availability
- Code availability
- Author contribution

References

[1] M. Bao et al., "ASP-VMUNet: Atrous Shifted Parallel Vision Mamba U-Net for Skin Lesion Segmentation," in *arXiv preprint*, Mar. 2025, doi: 10.48550/arXiv.2503.19427.

[2] T. Dwivedi, B. K. Chaurasia, and M. M. Shukla, "Lightweight Vision Image Transformer (LViT) Model for Skin Cancer Disease Classification," in *International Journal of System Assurance Engineering and Management*, Springer, 2024, doi: 10.1007/s13198-024-02521-6.

[3] R. Sinha, S. R. Dubey, and S. K. Singh, "A Hybrid Ensemble Framework with Deep Learning Models for Skin Cancer Detection Using Dermoscopy Images," in *Multimedia Tools and Applications*, Springer, 2025, doi: 10.1007/s11042-025-20739-9.

[4] A. Pathak, A. Acharya, and S. Panigrahi, "Efficient Hybrid Deep Learning Model for Early Detection of Skin Cancer Using Dermoscopic Images," in *Journal of the Institute of Engineers (India): Series B*, Springer, 2024, doi: 10.1007/s40031-024-00811-6.

[5] A. Kumar, K. R. Kanthen, and J. John, "GS-TransUNet: Integrated 2D Gaussian Splatting and Transformer UNet for Accurate Skin Lesion Analysis," in *arXiv preprint*, Feb. 2025, doi: 10.48550/arXiv.2502.16748.

[6] S. Dey, A. Mukherjee, A. Saha, and R. Sharan, "Enhancing Skin Disease Classification Leveraging Transformer-Based Models and Heatmap Insights," *Computers in Biology and Medicine*, vol. 190, 110007, 2025, doi: 10.1016/j.compbiomed.2025.110007.

[7] M. Rao K, R. Rajesh, A. S. Kumar, and A. M. Khan, "LesionAid: An Explainable ViTGAN Framework for Skin Lesion Generation and Classification," in *Multimedia Tools and Applications*, Springer, 2025, doi: 10.1007/s11042-025-20797-z.

[8] J. Hu et al., "Multi-Scale Transformer Architecture for Accurate Medical Image Classification," in *arXiv preprint*, Feb. 2025, doi: 10.48550/arXiv.2502.06243.

[9] J. Amin et al., "Skin-Lesion Segmentation using Boundary-Aware Segmentation Network and Classification based on a Mixture of Convolutional and Transformer Neural Networks," in *Frontiers in Medicine*, vol. 12, 2025, doi: 10.3389/fmed.2025.1524146.

[10] M. Azhar et al., "Skin Lesion Classification Using CNN and Transformer Networks for Computer-Assisted Diagnosis," in *International Conference on Smart Systems and Advanced Computing (SysCom 2022)*, Mar. 2025, doi: 10.1007/978-3-031-40905-9$_3$.

[11] Y. Gautam et al., "FusionEXNet: An Interpretable Fused Deep Learning Model for Skin Cancer Detection," in *International Journal of Computers and Applications*, vol. 46, no. 9, pp. 743–753, Aug. 2024, doi: 10.1080/1206212X.2024.2385923.

[12] C. Flosdorf et al., "Skin Cancer Detection Utilizing Deep Learning: Classification of Skin Lesion Images Using a Vision Transformer," in *arXiv preprint*, Jul. 2024, doi: 10.48550/arXiv.2407.18554.

[13] M. Akter et al., "An Integrated Deep Learning Model for Skin Cancer Detection Using Hybrid Feature Fusion Technique," in *arXiv preprint*, Oct. 2024, doi: 10.48550/arXiv.2410.14489.

[14] G. M. S. Himel et al., "Skin Cancer Segmentation and Classification Using Vision Transformer for Automatic Analysis in Dermatoscopy-based Non-invasive Digital System," in *arXiv preprint*, Jan. 2024, doi: 10.48550/arXiv.2401.04746.

[15] M. Imran, M. I. Tiwana, M. M. Mohsan, N. S. Alghamdi, and M. U. Akram, "Transformer-based framework for multi-class segmentation of skin cancer from histopathology images," *Frontiers in Medicine*, vol. 11, Apr. 2024, doi: 10.3389/fmed.2024.1380405.

[16] A. Faghihi, M. Fathollahi, and R. Rajabi, "Diagnosis of Skin Cancer Using VGG16 and VGG19 Based Transfer Learning Models," in *arXiv preprint*, Apr. 2024, doi: 10.48550/arXiv.2404.01160.

[17] E. H. I. Eliwa, "Enhancing Skin Cancer Diagnosis Through Fine-Tuning of Pretrained Models: A Two-Phase Transfer Learning Approach," *International Journal of Breast Cancer*, vol. 2025, Article ID 4362941, Feb. 2025, doi: 10.1155/ijbc/4362941.

[18] I. Suneetha, "Hybrid Deep Learning Model for Skin Cancer Classification," in *E3S Web of Conferences*, vol. 591, 2024, doi: 10.1051/e3sconf/202459109010.

[19] R. S. Tripathi and R. Pandey, "Multiclass Skin Cancer Classification Using Ensemble Transfer Learning," in *Proceedings of the 6th International Conference on ICT for Sustainable Development (ICT4SD 2021)*, vol. 320, Springer Singapore, 2024, pp. 353–362, doi: 10.1007/978-981-99-8222-0_31.

[20] B. V. N. Siliveri and N. R. Chittineni, "Skin Cancer Classification using Vision Transformers," in *2024 International Conference on Computer Communication and Informatics (ICCCI)*, IEEE, 2024, pp. 1–6, doi: 10.1109/ICCCI62068.2024.10383242.

[21] S. S. H. S. Islam, S. Shahnawaz and M. M. Rahman, "SCVTNet: Skin Cancer Classification Using CNN and Vision Transformer Network," in *2024 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)*, IEEE, 2024, pp. 1–4, doi: 10.1109/IC4ME260205.2024.10448725.

[22] V. Kumar, P. Gupta, and S. Sharma, "Advanced Transfer Learning Models for Improved Skin Cancer Detection: A Comprehensive Evaluation," *Journal of Biomedical Informatics*, vol. 148, p. 104413, 2024, doi: 10.1016/j.jbi.2024.104413.

[23] R. Bogne Tchema, A. C. Polycarpou, and M. Nestoros, "Skin Cancer Classification Using Machine Learning," *Multimedia Tools and Applications*, vol. 84, pp. 3239–3256, Jan. 2025, doi: 10.1007/s11042-025-20595-7.

[24] A. T. Ibrahim et al., "Categorical classification of skin cancer using a weighted ensemble of transfer learning with test time augmentation," in *Data Science and Management*, Dec. 2024, doi: 10.1016/j.dsm.2024.10.002.

[25] K. Suresh, S. Suresh, A. Suresh, I. Ahmed, and S. S. Basha, "A Deep-Ensemble-Learning-Based Approach for Skin Cancer Diagnosis," in *Electronics*, vol. 12, no. 6, 1342, Mar. 2024, doi: 10.3390/electronics12061342.

[26] Y. Mousa, R. Taha, R. Kaur, and S. Afifi, "Melanoma Classification Using Deep Learning," in *Image and Video Technology*, W. Q. Yan, M. Nguyen, P. Nand, and X. Li, Eds., Lecture Notes in Computer Science, vol. 14403, Springer, Singapore, 2024, pp. 1–12, doi: 10.1007/978-981-97-0376-0-20.

[27] S. Iqbal, M. Zeeshan, M. Mehmood, T. M. Khan, and I. Razzak, "TESL-Net: A Transformer-Enhanced CNN for Accurate Skin Lesion Segmentation," *arXiv preprint*, Aug. 2024, doi: 10.48550/arXiv.2408.09687.

[28] S. Haque, F. Ahmad, V. Singh, D. M. Mathkor, and A. Babegi, "Skin Cancer Detection Using Deep Learning Approaches," *Cancer Biotherapy & Radiopharmaceuticals*, Mar. 2025, doi: 10.1089/cbr.2024.0161.

[29] N. Torbati, A. Meshcheryakova, D. Mechtcheriakova, and A. Mahbod, "A Multi-Stage Auto-Context Deep Learning Framework for Tissue and Nuclei Segmentation and Classification in H&E-Stained Histological Images of Advanced Melanoma," *arXiv*, Mar. 2025, doi: 10.48550/arXiv.2503.23958.

[30] R. Islam, J. K. Dipu, and M. H. Tusar, "Using Computer Vision for Skin Disease Diagnosis in Bangladesh: Enhancing Interpretability and Transparency in Deep Learning Models for Skin Cancer Classification," *arXiv*, Feb. 2025, doi: 10.48550/arXiv.2502.01985.

[31] T. M. Khan et al., "TAFM-Net: A Novel Approach to Skin Lesion Segmentation Using Transformer Attention and Focal Modulation," in *arXiv preprint*, 2024, doi: 10.48550/arXiv.2411.17556.

[32] S. A. Hanum, A. Dey, and M. A. Kabir, "An Attention-Guided Deep Learning Approach for Classifying 39 Skin Lesion Types," *arXiv preprint*, Jan. 2025, doi: 10.48550/arXiv.2501.05991.

[33] M. A. Aljanabi, A. Albukhari, and M. B. Alazzam, "A deep learning model for melanoma classification using multi-input vision transformers," *Biomedical Signal Processing and Control*, vol. 87, 2024, doi: 10.1016/j.bspc.2023.105420.

[34] S. Zou, M. Zhang, B. Fan, Z. Zhou, and X. Zou, "SkinMamba: A Precision Skin Lesion Segmentation Architecture with Cross-Scale Global State Modeling and Frequency Boundary Guidance," *arXiv preprint*, Sep. 2024, doi: 10.48550/arXiv.2409.10890.

[35] I. Matas et al., "AI-Driven Skin Cancer Diagnosis: Grad-CAM and Expert Annotations for Enhanced Interpretability," *arXiv preprint*, Jun. 2024, doi: 10.48550/arXiv.2407.00104.

[36] K. Muhammad, A. S. Khan, O. Rashid, and A. Mahmood, "Skin cancer classification using deep learning and explainable AI with Grad-CAM," *Computers in Biology and Medicine*, vol. 170, 2024, doi: 10.1016/j.compbiomed.2023.107583.

[37] S. P. Angelin Claret, J. P. Dharmian, and A. Muthu Manokar, "Artificial intelligence-driven enhanced skin cancer diagnosis: leveraging convolutional neural networks with discrete wavelet transformation," *Egyptian Journal of Medical Human Genetics*, vol. 25, no. 50, 2024, doi: 10.1186/s43042-024-00522-5.

[38] R. Xu, C. Wang, J. Zhang, S. Xu, W. Meng, and X. Zhang, "SkinFormer: Learning Statistical Texture Representation with Transformer for Skin Lesion Segmentation," *arXiv preprint*, Sep. 2024, doi: 10.48550/arXiv.2409.08652.

[39] P. Zhang and D. Chaudhary, "Hybrid Deep Learning Framework for Enhanced Melanoma Detection," *arXiv preprint*, Jul. 2024, doi: 10.48550/arXiv.2408.00772.

[40] S. Dey, A. Mukherjee, A. Saha, and R. Sharan, "Enhancing Skin Disease Classification Leveraging Transformer-Based Models and Heatmap Insights," *Computers in Biology and Medicine*, vol. 190, p. 110007, 2025, doi: 10.1016/j.compbiomed.2025.110007.

[41] A. Faghihi, M. Fathollahi, and R. Rajabi, "Diagnosis of Skin Cancer Using VGG16 and VGG19 Based Transfer Learning Models," *arXiv preprint*, Apr. 2024, doi: 10.48550/arXiv.2404.01160.