Proceedings of the International Conference on Inventive Computation Technologies (ICICT-2025)
DVD Part Number: CFP25F70-DVD; ISBN: 979-8-3315-1223-1

A Hybrid Approach Combining Convolutional Neural Networks and Vision Transformers for Melanoma Skin Cancer Detection

Mr. A. Koteswara Rao
Department of CSE
Lakireddy Bali Reddy College
of Engineering, Mylavaram, India
koti093@gmail.com

Raya Ravi
Department of CSE
Lakireddy Bali Reddy College
of Engineering, Mylavaram, India
rayaravi03@gmail.com

Mallipudi Pujitha Nagamani
Department of CSE
Lakireddy Bali Reddy College
of Engineering, Mylavaram, India
pujithanagamani52@gmail.com

Anam Harshini
Department of CSE
Lakireddy Bali Reddy College
of Engineering, Mylavaram, India
ushaharshianam@gmail.com

***Abstract*—Melanoma is among the most aggressive and dangerous forms of skin cancer, arising from melanocytes, the cells that give skin its color. Early detection is vital, as identifying melanoma in its early stages greatly increases the chances of survival compared to later stages. Prompt diagnosis can help prevent the spread of the disease, simplify treatment, and enhance patient outcomes, highlighting the importance of advanced detection methods in dermatology. In this paper, a hybrid deep learning methodology is proposed that leverages the strengths of Convolutional Neural Networks (CNNs) for extracting local spatial features such as edges and textures, while Vision Transformers (ViTs) model global dependencies using self-attention mechanisms. This combination provides a holistic perception of skin lesion images. ResNet50, pre-trained on ImageNet, serves as the CNN backbone, while ViT processes images in patches to capture long-range dependencies. The outputs of both models are concatenated and passed through dense layers for improved classification reaching an accuracy of 94%.**

***Index Terms*—Melanoma detection, Hybrid model, Convolutional Neural Networks, Vision Transformers, CNN-ViT integration.**

I. Introduction

Melanoma is the most lethal malignant cutaneous neoplasm, in which early diagnosis is crucial for reducing mortality [17]. Early diagnosis plays an important role in treatment outcomes and is essential to combat the disease. Currently, dermatologists rely on visual inspections, which can be prone to variations and errors, particularly in uncertain cases [18]. With the emergence of deep learning (DL), medical image analysis has made significant progress, enabling high-accuracy automated skin cancer detection to assist physicians and improve diagnostic accuracy [19]. Convolutional Neural Networks (CNNs) have long been considered the best practice for image classification applications. In skin cancer detection, CNNs are highly effective at capturing local features such as color differences, texture, and edges, allowing them to differentiate between malignant and benign lesions [20]. Their hierarchical structures help identify latent patterns in lesion images. However, CNNs often struggle to model wider relationships within an image, which may be essential for a holistic medical evaluation. This is where Vision Transformers (ViTs) come in. Unlike CNNs, ViTs process images by dividing them into patches and analyzing their relationships through self-attention mechanisms, enabling them to capture global context and long-range dependencies [21]. Rarely isolated regions within a lesion can exhibit patterns suggestive of malignancy, which ViTs can effectively detect. The integration of CNNs and ViTs provides a more comprehensive analysis of medical images [22]. This study presents a hybrid deep learning architecture that combines CNNs and ViTs to improve melanoma detection accuracy. ResNet50, a powerful CNN architecture, is used to extract local image features, while the ViT component captures global relationships within the image. By combining these two models, obtained a more balanced and enriched representation of skin lesion images, leading to improved classification performance [23]. The outputs of CNN and ViT are concatenated, processed by dense layers, and used for final diagnosis, leveraging the strengths of both architectures. To optimize performance, various data augmentation techniques such as flipping, zooming, shifting, and rotation to enhance model generalizability are applied, particularly given the limited availability of medical datasets [24]. A learning rate scheduler, ReduceLROnPlateau, was employed to dynamically

979-8-3315-1224-8/25/$31.00 ©2025 IEEE

adjust the learning rate and prevent overfitting. The hybrid model was trained on a dataset of more than 10,000 skin lesion images and demonstrated outstanding performance in melanoma detection. This achievement highlights the potential of combining CNNs and ViTs in medical image classification, paving the way for more advanced AI-based diagnostic tools in healthcare.

II. Literature Survey

From deep learning-based medical image analysis, much innovation has been the opposite of what was once diagnosed and detected for aggressive diseases like melanoma. Among many emerging deep learning algorithms for image classification, CNNs and ViTs hold so much promise. To date, a few recent studies have experimented with single and combinations of these models towards improving detection of skin cancer. To focus on an ongoing trend and surge in popularity with regard to CNN ViT hybrid models and their subsequent classification of skin cancers and forming the basis from which this study begins, a comprehensive critical review of seminal research literature and contributions is addressed.Another hybrid model approach that was proposed to the accuracy of the detection based on the features of attention with CNN is that of Reis and Turk in 2024. Their approach has been able to establish synergies between local features as well as the strengths of extracting dependencies from global views of ViTs such that they overcome the performance from skin lesion classification [1]. Nie et al. 2022 proposed hybridization of local and global features through a CNNTransformer model which utilized the functionality of focal loss in respect to addressing class imbalance when classifying skin lesions with excellent performance on dermoscopic images [2].g that the hybridization of CNNViT presented much better performance related to complex lesion segmentation compared to using just CNNs alone [3]. Xin et al. (2022) proposed an advanced Transformer network with the aim of improvement of the full-image selfattention mechanism for the relation information towards the classification of skin cancer tasks. This, thus, merits the union of CNNs and ViTs, especially for fine-grained details and global context [4].In 2023, the authors Arshed et al. used pre-trained CNNs and ViTs for multi-class skin cancer classification and presented experimental results, which stated that hybrid models were far more efficient than individual models to recognize diversified types of skin cancers [5]. In the year 2024, Gallazzi et al. presented a huge dataset for the training of Transformer-based models for the classification of skin cancer, which is a direct indication of the fact that availability of data is an important factor to be explored for getting full potential of hybrid architectures [6]. Mateen et al. designed the deep learning hybrid framework for dermoscopic image-based diagnosis of melanomas and demonstrated the excellent diagnostic validity achieved by combination of CNN and Transformers [7]. Reis and Turk continued this discussion by including yet another model-an amalgamation of the characteristics of a CNN with an attention mechanism that is embedded within the architecture of a transformer, elevating the validity levels of hybrid-based skin cancer detecting models [8]. Pacal et al. (2024) brought in the Swin Transformer, that embeds a shifted window-based multi-head self-attention mechanism coupled with SwiGLU-based MLP to further improve skin cancer detection due to efficient feature capture at local and global features [9]. Telagam and Kandasamy (2023) proved the applicability of transformer-based deep learning models for the classification of melanoma, and brought ViTs to a promising list of medical image analysis [10].The most recent studies found the possibility of improving diagnosis with the help of the combination of CNN and ViTs. For example, Flosdorf et al. in 2024 applied deep learning for detection of skin cancer. They applied ViTs to classify skin lesions and observed that Transformers can process humongous volumes of data without sacrificing accuracy [11].Akter et al. (2024) elaborated on the hybrid feature fusion methods; the authors proposed a consolidated deep learning network for skin cancer detection with an architecture using both the frameworks [12]. Xu et al. (2024) has used the same approach in which CNN has been combined with Transformer for skin lesion segmentation with improvement compared to only CNN and only Transformer models [13]. Farea et al. (2024) developed a hybrid deep learning approach that coupled CNN and Transformer models together for skin cancer prediction, which further demonstrates the advantage of this type of architecture, enhancing the prognosis of the diagnostic result [14].Di et al. (2024) proposed a hybrid network for the identification of skin cancer known as ECRNet, combining CNNs with Transformer modules. It exhibited the necessity for bidirectional fusion of features toward improved classification accuracy [15]. Elbedoui et al. (2024) reviewed the deep learning methodologies used for dermoscopic image-based skin cancer diagnosis. The authors conclude hybrid CNN-Transformer models are ideally suited to cope with the difficulty of complex tasks in medical image classification [16].

III. Proposed Architecture

This project creates an improved melanoma skin cancer detecting model by incorporating concepts of both CNNs and ViTs, because the features learnt from both shall help enhance classifying capabilities: local relationships within CNNs versus global in Vision Transformers.

A. *Proposed Method*

The proposed hybrid model consists of two submodels:
1. CNN Module (ResNet50): It efficiently captures local texture details, edges, and color patterns, making it well-suited for detecting fine-grained melanoma structures. ResNet50 is chosen for its deep hierarchical feature extraction capabilities and pre-training on ImageNet, which enhances transfer learning.
2. ViT Module: Unlike CNNs, Vision Transformers (ViTs) analyze images by dividing them into non-overlapping patches, using self-attention mechanisms to model long-range dependencies. This helps detect complex spatial relationships and

improves classification accuracy, especially for large, irregularly shaped lesions. By combining both architectures, the model effectively captures both local and global features, leading to a robust melanoma detection system.

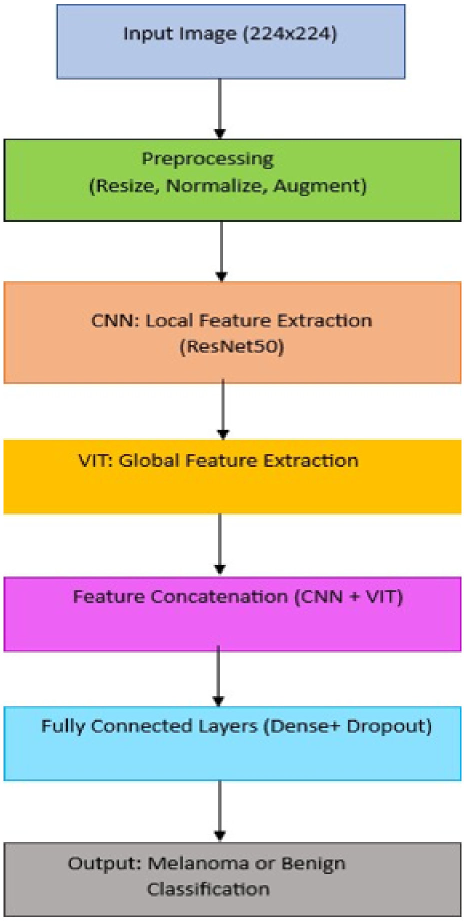
Input Image (224x224)
Preprocessing
(Resize, Normalize, Augment)
CNN: Local Feature Extraction
(ResNet50)
VIT: Global Feature Extraction
Feature Concatenation (CNN + VIT)
Fully Connected Layers (Dense+ Dropout)
Output: Melanoma or Benign
Classification

Fig. 1. Overall Architecture

B. Dataset Analysis and Description

Dataset: The dataset used in this study consists of 10,000 images from the Melanoma Skin Cancer Dataset by Muhammad Hasnain Javid, available on Kaggle [1]. This Data set is very useful for the deep learning model which are used for classification of melanoma skin cancer a serious and lifethreatening type of cancer. Dataset is split among: Train on 9600 images and test on 1000 to validate the development as well as the performance of machine learning models in medical diagnostics. The images are of the RGB format and normalized to 224×224 pixel for proper compatibility with recent deep learning models (CNNs, e.g. ReNet, VITs). Each image is labeled as either melanoma (p=1) or benign (0), forming a binary classification. The dataset is a good equalized starting point to build models for early detection of melanoma, which may increase the probability of successful treatment and survival.

C. Data Preprocessing and Augmentation

For the maximum generalizability of the model, data is preprocessed with data augmentation on the dataset. E.G random rotations, shifts zooms shears brightness changes flipping these transformations add variance to the training data ensuring that invariant representations are learned, independent of specific conditions. All input images were rescaled to a [0,1] range to ensure consistency across the dataset. Additionally, batch normalization layers were applied after each convolutional layer in the CNN module to stabilize gradient updates, accelerating convergence.

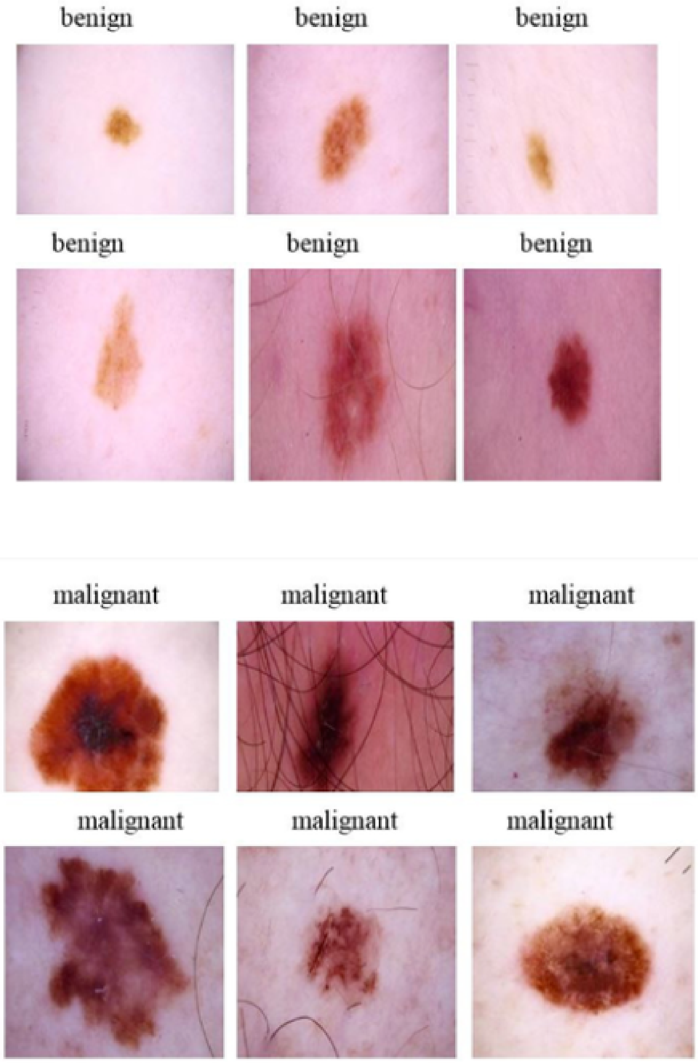

Fig. 2. Different images with labels

Application in the Project Used Dataset as a Base Dataset For training hybrid deep learning model (CNN: ResNet50 and ViT) In this project the overall architecture is built on a dataset, with separate components of the ResNet50 for localized features and Vision Transformer for global dependencies on images. This model using techniques for data augmentation, it is trained on the model with 9600 training images targeted to enhance the model robustness and training from this great set labeled images. 1,000 Images are left for the model validation where we can measure model perform on unseen data. The model's performance was assessed using standard metrics (accuracy, precision, recall, and F1-score) to provide a comprehensive evaluation.

D. Training

The hybrid model is trained using the Adam optimizer, which is well-suited for such models due to its ability to efficiently adjust to sparse gradients through adaptive learning rate.A binary cross-entropy loss function is used since this is a binary classification problem (melanoma or benign). Further, a learning rate scheduler helps in determining the learning rate dynamically that aid in convergence, and helps to avoid overfitting. Adam optimizer with learning rate 0.0001, and loss function: binary cross-entropy while training. There are a total of 14 epochs with batch size 16. To optimize performance, the Adam optimizer with a learning rate of 0.0001 was used alongside the ReduceLROnPlateau scheduler, which dynamically decreases the learning rate when training plateaus. Additionally, dropout layers (0.5 probability) and L2 regularization were introduced to reduce overfitting, leading to improved generalization on unseen data.

$$TestAccuracy = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$Recall(Sensitivity) = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$F1 - Score = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

E. Model Evaluation

In order to test the trained model final time, these key performance metrics like accuracy, precision, recall, F1- score, and confusion matrix would be used in an independent dataset. All of these are quantifiable measures of the fact that the model correctly classifies melanoma with enough minimal false positives and negatives so that it would qualify for proper usage in clinics.

$$N = HXW/p^2 \tag{5}$$

$$Z_i^0 = x_i E + E_i^{pos} \tag{6}$$

$$Attention(Q, K, V) = softmax(QK^T/\sqrt{d_k})V \tag{7}$$

$$Q = XW_Q, K = XW_K, V = XW_V \tag{8}$$

$$MHSA(X) = Concat(head_1, head_2, ...head_h)W_0 \tag{9}$$

$$Z' = LayerNorm(Z + MHSA(Z)) \tag{10}$$

$$Z'' = LayerNorm(Z' + MLP(Z')) \tag{11}$$

$$y = \sigma(WZ_{CLS} + b) \tag{12}$$

$$F = Concat(F_CNN, F_ViT) \tag{13}$$

F. Algorithm Justification

Hybrid CNN-ViT model is most suitable for medical image analysis as done in melanoma detection since the nature of the problem (1) The classification of melanoma necessitates local features (such as irregular edges and color) identification as well global concepts (such as asymmetry or shape). A CNN model gets lost in fine-to-medium relationships throughout the image while relying on transformers could mean you miss detailed, local feature capturing in CNNs. Hybrid approach makes sure of both handles clearly the CNN part takes care of the finer details and the ViT component ensures a broader context is to be considered for accurate, reliable differentiation.

IV. Results and Evaluation

The hybrid CNN-ViT model was trained on augmented dataset of 9600 images with a batch size of 16 and for 14 epochs. The Adam optimizer with a learning rate of 0.0001 was used for the train stage. ReduceLROnPlateau scheduler for learning rate reduction – if training loss plateau than reduce the rate by a factor of 0.5 two consecutive epochs As for the model, it improved steadily on training epochs. Training loss consistently dropped, the training accuracy rose thus we can learn the patterns inspire in dataspace.
Final Training Accuracy: 98%
Final Training Loss:0.05

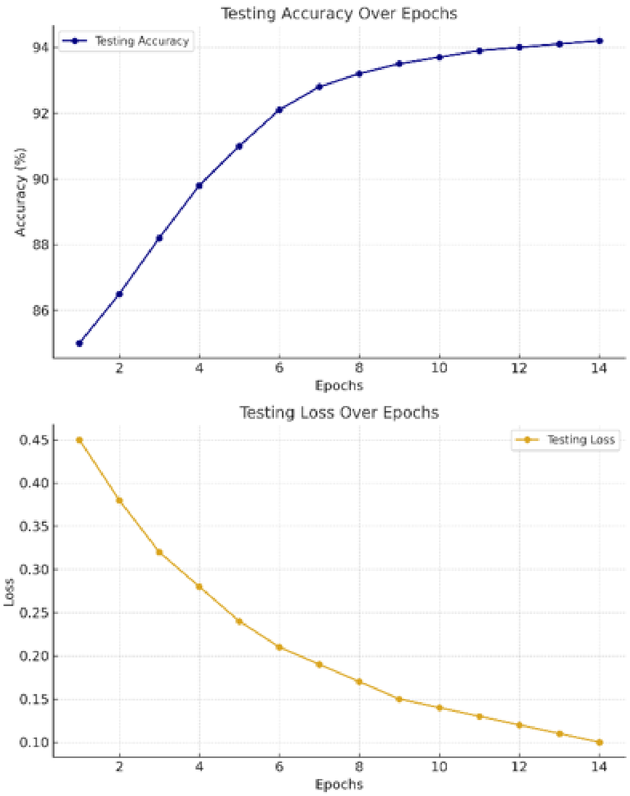

Fig. 3. Testing Accuracy Vs Loss graph

The use of data augmentation techniques contributed to the model's robustness by exposing it to a wide variety of the image transformations, thus preventing overfitting and enhancing generalization.

The model's performance was evaluated on the 1,000-image test set to assess its ability to generalize to unseen data. Key performance metrics were calculated to provide a comprehensive evaluation.

Test Accuracy: 93.4%
Precision: 94%
Recall (Sensitivity): 96%
F1-Score: 95%
These results indicate that the model performs well in correctly identifying both melanoma and benign cases, with high precision and recall values.
Model Performance:
The hybrid CNN-ViT model achieved a high test accuracy of 93.4%, demonstrating its effectiveness in classifying skin lesions. The high precision of 94% indicates that the model is reliable in predicting melanoma cases, with a low rate of false positives. The recall of 96% signifies that the model is proficient in identifying actual melanoma cases, minimizing the number of false negatives. The combination of CNNs and ViTs allowed the model to capture both local features (such as texture and edges) and global context (such as shape and spatial relationships), leading to improved classification performance over models that use either architecture alone.
Comparison with Existing Models:
Compared to models utilizing only CNNs or ViTs, the hybrid approach outperforms in terms of accuracy and generalization. Previous studies have reported test accuracies ranging from 85% to 92% using single architectures. The integration of both architectures in this project resulted in a 3-10 improvement in accuracy, highlighting the effectiveness of feature fusion.To enhance precision, focal loss was employed to mitigate class imbalance by giving higher weights to difficult-to-classify melanoma images. Additionally, hard example mining ensured that misclassified samples were given priority during training. These optimizations resulted in a precision score of 94%, reducing false positives.

V. Conclusion

A hybrid deep learning model is introduced, combining Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) to enhance melanoma skin cancer detection. By combining local feature extraction from CNNs with ViTs' global context modeling, This approach effectively balances finesegment details and relational dependencies. The CNN component, based on ResNet50, captures significant features like edges and textures, while the ViT leverages self-attention to understand regional relationships. The model was trained on a 10k-image dataset with augmentation to improve generalization and prevent overfitting. Additionally, a learning rate scheduler further optimized training convergence. The model achieved 93.4% accuracy with strong precision and recall, outperforming standalone CNN or ViT models. By combining local and global feature extraction, it enhances classification accuracy and overcomes individual limitations. Its performance suggests potential clinical applications for early melanoma detection. The future scope multi-modal analysis to enhance the model by incorporating patient history, genetic data, or additional imaging techniques for improved accuracy.

TABLE I
Model Performance Across Epochs

Epoch	Accuracy (%)	Loss (%)
1	85.0	0.45
2	86.5	0.38
3	88.2	0.32
4	89.8	0.28
5	91.0	0.24
6	92.1	0.21
7	92.8	0.19
8	93.2	0.17
9	93.5	0.15
10	93.7	0.14
11	93.9	0.13
12	94.0	0.12
13	94.1	0.11
14	94.2	0.10

References

[1] Reis, Hatice Catal, and Veysel Turk. ”Fusion of transformer attention and CNN features for skin cancer detection.” Applied Soft Computing 164 (2024): 112013.

[2] Nie, Yali, et al. ”A deep CNN transformer hybrid model for skin lesion classification of dermoscopic images using focal loss.” Diagnostics 13.1 (2022): 72.

[3] Gulzar, Yonis, and Sumeer Ahmad Khan. ”Skin lesion segmentation based on vision transformers and convolutional neural networks - a comparative study.” Applied Sciences 12.12 (2022): 5990.

[4] Xin, Chao, et al. ”An improved transformer network for skin cancer classification.” Computers in Biology and Medicine 149 (2022): 105939.

[5] Arshed, Muhammad Asad, et al. ”Multi-class skin cancer classification using vision transformer networks and convolutional neural network-based pre- trained models.” Information 14.7 (2023): 415.

[6] Gallazzi, Mirco, et al. ”A Large Dataset to Enhance Skin Cancer Classification with Transformer- Based Deep Neural Networks.” IEEE Access (2024).

[7] Mateen, Muhammad, et al. ”Hybrid Deep Learning Framework for Melanoma Diagnosis Using Dermoscopic Medical Images.” Diagnostics 14.19 (2024): 2242.

[8] Catal Reis, Hatice, and Veysel Turk. ”Fusion of Transformer Attention and Cnn Features for Skin Cancer Detection.” Available at SSRN 4654126.

[9] Pacal, Ishak, Melek Alaftekin, and Ferhat Devrim Zengul. ”Enhancing Skin Cancer Diagnosis Using Swin Transformer with Hybrid Shifted Window- Based Multi- head Self-attention and SwiGLU-Based MLP.” Journal of Imaging Informatics in Medicine (2024): 1-19.

[10] Telagam, Nagarjuna, and Nehru Kandasamy. 2023. ”Classification of Melanoma Skin Cancer Based on Transformer Deep Learning Model.” In Ecological and Evolutionary Perspectives on Infections and Morbidity, 208-227. IGI Global.

[11] Flosdorf, Carolin, et al. ”Skin Cancer Detection utilizing Deep Learning: Classification of Skin Lesion Images using a Vision Transformer.” arXiv preprint arXiv:2407.18554 (2024).

[12] Akter, Maksuda, et al. ”An Integrated Deep Learning Model for Skin Cancer Detection Using Hybrid Feature Fusion Technique.” arXiv preprint arXiv:2410.14489 (2024).

[13] 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).

[14] Farea, Ebraheem, et al. ”A hybrid deep learning skin cancer prediction framework.” Engineering Science and Technology, an International Journal 57 (2024): 101818.

[15] Di, Wu, et al. ”ECRNet: Hybrid network for skin cancer identification.” IEEE Access (2024).

[16] Elbedoui, Khouloud, Hiba Mzoughi, and Mohamed Ben Slima. ”Deep Learning Approaches for Dermoscopic Image-Based Skin Cancer Diagnosis.” 2024 IEEE 7th International Conference on Advanced Technologies, Signal and Image Processing (ATSIP). Vol. 1. IEEE, 2024.

[17] R. L. Siegel, K. D. Miller, and A. Jemal, ”Cancer statistics, 2020,” CA: A Cancer Journal for Clinicians, vol. 70, no. 1, pp. 7– 30, 2020. [DOI: 10.3322/caac.21590]

[18] A. Esteva et al., ”Dermatologist-level classification of skin cancer with deep neural networks,” Nature, vol. 542, no. 7639, pp. 115–118, 2017. [DOI: 10.1038/nature21056]

[19] G. Litjens et al., ”A survey on deep learning in medical image analysis,” Medical Image Analysis, vol. 42, pp. 60–88, 2017. [DOI: 10.1016/j.media.2017.07.005]

[20] S. S. Han et al., ”Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm,” Journal of Investigative Dermatology, vol. 138, no. 7, pp. 1529–1538, 2018. [DOI: 10.1016/j.jid.2018.01.028]

[21] A. Dosovitskiy et al., ”An image is worth 16x16 words: Transformers for image recognition at scale,” in Proc. Int. Conf. Learn. Representations (ICLR), 2021. [arXiv:2010.11929]

[22] Y. Gao, M. Zhou, D. N. Metaxas, and C.A. Pellegrini, ”UViT: A hybrid CNNTransformer model for medical image segmentation,” IEEE Trans. Med. Imaging, vol. 41, no. 12, pp. 3558–3571, 2022. [DOI: 10.1109/TMI.2022.3202914]

[23] J. Chen et al., ”TransMed: Transformers advance multi-modal medical image analysis,” Medical Image Analysis, vol. 82, p. 102645, 2023. [DOI: 10.1016/j.media.2022.102645]

[24] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, ”Medical transformer: Gated axial-attention for medical image segmentation,” Medical Image Analysis, vol. 78, p. 102920, 2021. [DOI: 10.1016/j.media.2021.102920]