

An Enhanced Skin Cancer Detection using Medical Transformers

Mohana Saranya Sellappan¹, Prasannndh Raaju M R², R.Mohammed Yousuf³, Nivethika S⁴, Pravin S⁵ and Geetha S⁶

¹Assistant Professor (SL.G), Department of Computer Science and Engineering, Kongu Engineering College, Erode, Tamil Nadu, India

Email: mohanasaranyaa@gmail.com

^{2,4-5}Department of Computer Science and Engineering, Kongu Engineering College, Erode, Tamil Nadu, India

Email: prasannndhraaju.25@gmail.com, nivethikas.22cse@kongu.edu, pravins.22cse@kongu.edu

³Assistant Professor, Department of Information Technology, Bannari Amman Institute of Technology, Erode, Tamil Nadu, India

Email: mohemmedyousuf@bitsathy.ac.in

⁶Assistant Professor, Department of Computer Science and Engineering, Nandha Engineering College, Erode, Tamil Nadu, India

Email: btechgeetha@gmail.com

Abstract—Skin cancer is among the most common and hazardous types of cancers, and the issue of early and accurate diagnosis is inherent in the context of raising the survival rates.

Improved models are proposed in this paper in order to classify skin cancer on dermoscopic skin using a better Vision Transformer (ViT). The proposed ViT model consists of patch-wise tokenization, multi-head self-attention, and positional embedding (in contrast to the existing convolutional neural networks (CNNs) that are limited by the capacity to reveal the long-range dependencies) to learn the informative global images of medical images.

Two benchmark datasets, i.e. ISIC-2019 Balanced Subset and ISIC9042 Full Dataset are used to train and test the model with a classification accuracy of 88.2 and 88.5 percent, respectively.

Grad-CAM is integrated to be capable of visualizing where the model concentrates to as a way of interpreting it and relying on its clinical applications to establish the regions of the injury that lead to the classification. The applicability of transformer based architecture to real world dermatological diagnostic can be explained by the fact that the architecture has been proven to achieve positive performance improvement in the accuracy of the diagnostic and the stability of the system relative to the latest state of art approaches and methods.

Index Terms— Vision Transformer (ViT), Deep Learning, Medical Image Analysis, ISIC Dataset, Attention Mechanism, Artificial Intelligence in Healthcare.

I. INTRODUCTION

Skin cancer is a significant health issue of the world whose rates of incidences continue to rise steadily in different parts of the world, and detection at early and accurate stages is paramount in enhancing patient outcomes [1].

Nevertheless, conventional dermoscopic imaging depends much on clinical judgment and human eye observation and may be associated with inconsistent diagnosis, whereas the initial CNN-based automatic systems were limited by localized receptive fields that limited their capability to resolve long-range feature interactions of skin lesions [2]. Transformer-based architectures, and especially the Vision Transformer (ViT), address these shortcomings by representing the global contextual relationships as patch-wise captured image representations, and multi-head self-attention, which provides a more flexible choice to the fixed spatial hierarchies of CNNs [3]. Recent findings further indicate that transformer-based and hybrid CNN-ViT models have a high performance in a melanoma and non-melanoma classification task, with a greater level of robustness, high interpretability, and increased recognition of fine-grained dermoscopic features, including asymmetry, pigment networks, and irregular lesion boundaries [4], [5], [6]. Continuing on these developments, the current study suggests an improved ViT-based skin cancer classification model that is tested on both the ISIC-2019 Balanced Subset and the ISIC9042 Full Dataset and offers a way to classify skin cancer diagnostic results better, with patch tokenization, positional embeddings, and transformer-driven attention as the improvements that are made to the existing model to offer better diagnostic accuracy and present clinically relevant explanations in terms of Grad-CAM visualizations highlighting areas of lesions as significant in model predictions [4], thus overcoming the performance limitations.

II. RELATED WORK

The first automated skin cancer detection was based on CNN-based models with decent segmentation and binary classification but with localized receptive fields and reduced generalization in complicated dermoscopic images. Later reviews also pointed to Vision Transformers (ViTs) as a global-context alternative that is based on attention mechanisms in NLP and achieves higher consistency on heterogeneous datasets [6]. The existing hybrid CNN-ViT models further improved the diagnostic accuracy of the models with local convolutional features and global self-attention.

With the growth of datasets, including HAM10000, PH2 and ISIC, researchers proposed hybrid pipelines with recurrent unit and GAN-based augmentation and channel-wise attention to deal with class imbalance and ambiguous lesions [7], then revisions to segmentation-classification integration were made to enhance clinical robustness [8], [9]. Transformer-based models showed great effectiveness in long-range dependencies and fine dermoscopic structures such as asymmetry and irregular borders [10], and domain-specific models such as SkinViT and xCViT showed better performance on multi-class accuracy and interpretability [11], [12]. Ensemble methods minimized variation in data sets [13], and new efficiency-based and hierarchical transformer architectures have solidified the movement of CNN-based methods to transformer-based and hybrid models to detect clinical skin cancer [14].

III. MOTIVATION

The growing complexity of the dermoscopic image and clinical heterogeneity is a major problem to automated diagnostic systems, especially when the lighting is inconsistent, the device not uniform or the populations are not homogeneous [15].

Although hybrid deep learning architectures, such as quantum-enhanced BiLSTM-MobileNetV2 models, have demonstrated the ability to perform better in melanoma and non-melanoma classification, due to the sensitivity of their generalization to the diversity of datasets [16]. Research on transformer-enhanced and multimodal learning to detect skin lesions, and detecting monkeypox in particular, shows that the increase in attention at the global scale leads to higher rates of recognition of finer structural elements, but problems with interpretability and class imbalance remain [17], [18]. Recent hierarchical transformer models with shifted-window attention have reinforced lesion representation in high-resolution images [19], and segmentation-classification hybrids like ESC-UNET have shown the advantages of combined pipelines when dealing with clinical variability [20]. Besides, optimization-based approaches emphasize the necessity of adaptable and robust solutions when it comes to rare or visually overlapping lesions [21], whereas ensemble transformer-based models such as SkinSight continue to support the incentive to have refined ViT-based models with enhanced reliability and interpretability [22].

IV. PROBLEM DOMAIN

Diagnosis of skin cancer is a complicated field of problem with high visual similarity between benign and malignancy lesions, high intra-class variation, and imaging artifacts like hair, shadow, and haphazard lighting

[16]. The conventional diagnostic routes are based on the skills of dermatologists, who might struggle to perceive the fine dermoscopic patterns, including pigment networks, streaks, and irregular edges, always across clinicians and clinical setting [9]. The automated deep learning systems have tried to solve these issues, but the traditional CNN-driven approaches fail to capture the long-range relationships and fine-structural information to be classified with high accuracy especially in case of large-scale heterogeneous datasets such as ISIC and HAM10000 [2], [8]. More recent methods that use attention models, CNN-Transformer hybrid and transformer-based segmentation pipelines have demonstrated better feature mapping and stability, but issues of generalizability, interpretability, and sensitivity to uncommon lesion types still exist [19]. With this topography, it is of crucial importance to devise a robust, explainable, and dataset-independent classification scheme that can be used to model both global and localized lesion phenotyping when it comes to further deploying clinically viable skin cancer detection informatics systems.

V. PROBLEM DEFINITION

The main issue that is discussed in this paper is the creation of an automated system of skin cancer classification that would be able to correctly identify malignant lesions and benign ones in case of considerable clinical and visual differences. Classical CNN-based systems are not very successful in capturing global contextual relations and are also not very effective in generalizing to heterogeneous datasets of dermoscopic observations, thus resulting in misclassification of cases with subtle or overlap lesion patterns [2], [8]. Transformer-based methods show better potential in capturing long-range interactions, but current models are still considered to have problems with data imbalance, interpretability, and ability to work across various imaging modalities [18]. Hence, the issue is to develop and deploy an improved Vision Transformer architecture that can be used effectively to extract discriminative features, provide steady performance on diverse datasets and make interpretable predictions that are applicable to clinical decision-making.

VI. STATEMENT

This study aims to develop a better Vision Transformer-based model that will provide reliable and high-quality skin cancer classification through efficient differentiation of malignant lesions and benign ones in a variety of dermoscopic appearance cases. Current deep learning methods, specifically CNN-based models, do not have the capability to model global contextual relationships and cannot be robust to changes in lesion phenotype, imaging equipment, and dataset bias.

This requires a model that does not only derive discriminative features using high-resolution dermoscopic images but also able to generalize well with heterogeneous datasets, and it should also produce interpretable results, which can be validated by clinical means. Consequently, this work aims at developing, deploying and testing a ViT-based solution to overcome these obstacles and to facilitate the creation of reliable and clinically significant computer-aid diagnosis systems.

VII. INNOVATIVE CONTENT

The novelty behind this work is the incorporation of an improved Vision Transformer architecture designed specifically to work with dermoscopic images where specific consideration should be given to the use of finer patch-centric tokenization, better positional embeddings and an attention-based feature extraction pipeline to detect the delicate lesion features that standard CNNs can be insensitive to. In contrast to the current hybrid or convolution-based frameworks, the suggested one is more focused on global contextual modeling, which allows the system to capture long-range association crucial to the ability to differentiate between visually similar non-malignant and malignant patterns. Also, the integration of an explainability module that is transformer-adapted, e.g., ViT-compatible Grad-CAM visualization, can give a better understanding of how the model works and facilitate clinical trust and transparency. The method also comprises of strong preprocessing plans and dataset conscious training improvements that encourage generalisation of heterogeneous dermoscopic datasets. Collectively, these works are a new step in creating clinically competent and interpretable AI-based diagnostic instruments related to skin cancer.

VIII. PROBLEM DESIGN

The proposed system is designed with the focus to build an effective and interpretable architecture based on Vision Transformer (ViT) designed to be used in dermoscopic skin cancer classification. The model (Fig. 1)

works on each dermoscopic image by dividing it into non-overlapping patches, converting them into token embeddings and adding positional encodings to ensure that the relationship between spaces is maintained. Such embeddings are subsequently subjected to several layers of multi-head self-attention and feed-forward networks, which allows the model to extract the global interaction of contextual features throughout the entire lesion area, instead of only using local features. The last representation is compiled and inputted into a classification head which produces benign or malign predictions. The design overcomes major issues related to the conventional CNN-based methods, including reduced receptive fields and vulnerability to variability in the datasets, due to the viability of the ViT to detect long-scale dependencies and subtle structural patterns. The use of explainability mechanisms also contributes to the use of transparency in decision-making whereby the model output is consistent with clinically relevant features of lesions.

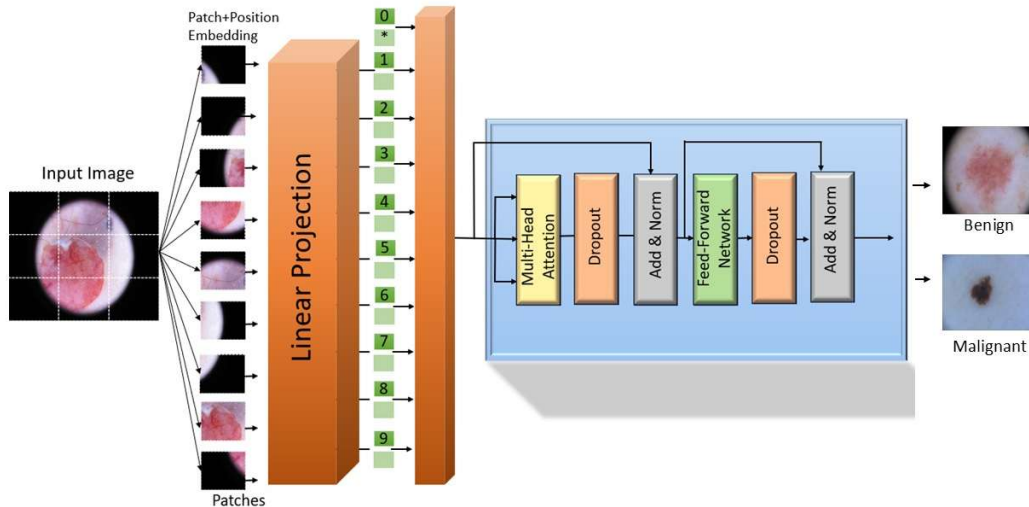


Figure 1. Architecture of the proposed Vision Transformer (ViT) model

IX. SOLUTION METHODOLOGIES

The offered framework of skin cancer detection follows a complex approach to solution involving dataset preparation, transformer-based image model, preprocessing tactics, and explainability mechanisms to provide correct and interpretable predictions by a clinician. All the methodological parts are aimed at making the approach more robust, reduce bias, and maximize the use of the Vision Transformer (ViT) model on various dermoscopic datasets.

A. Dataset Preparation and Standardization

The methodology uses the ISIC-2019 Balanced Subset and ISIC-9042 Full Dataset to deal with the problem of class imbalance and dataset bias. Any dermoscopic image is converted to 224 x 224, which is compatible with the input requirements of the Vision Transformer (ViT) and improves its consistency levels between training and evaluation stages [4].

B. Image Preprocessing and Patch Extraction

Preprocessing tasks like scaling and normalization are used to increase the stability of learning. Individual images are partitioned into 16 x 16 non-overlapping patches, transforming the critical lesion properties into 16 x 16 patch-level tokens, and allowing long-range dependencies to be modeled in ways that are beyond the traditional CNN abilities [3].

C. Feature Encoding by Transformers

Patch embeddings are used and enhanced with positional encodings that are processed with stacked layers of self-attention, residual connections, and feed-forward networks. This architecture supports global contextual modeling of lesion areas and the sensitivity of the subtle patterns of malignancy thus, as shown in recent dermatological studies [6].

D. Training Strategy and Optimization

The loss used to optimize an MLP classification is Adam and binary cross-entropy loss. Embedding dimension, attention heads, learning rate, and dropout regularization are used to avoid overfitting but achieve high discriminative performance due to best practices in transformer-based medical imaging [5].

E. Performance Evaluation and Explainability

Transformer embeddings with grad-CAM are visual explanations of the regions that influence the decision that allow to increase transparency and clinical trust [4]. The use of accuracy, precision, recall, specificity, MCC, AUC, and Cohen kappa as performance evaluation measures are compared with the state-of-the-art hybrid CNN-transformer models, which confirms the usefulness of pure ViT approaches [12], [19].

X. RESULTS AND SENSITIVITY ANALYSIS

A. AI Model Performance Evaluation

The Vision Transformer (ViT) achieved 88.2% accuracy on the ISIC-2019 Balanced Subset and 82.5% on ISIC9042, demonstrating strong performance across varied imaging conditions. Its global self-attention enabled effective recognition of subtle lesion patterns often missed by CNN-based systems [3], with stable interpretability supported by attention-based mechanisms [5].

B. Classification Stability and Scalability

The model was also found to be constant in its accuracy with various dataset sizes and clinical variations. With a larger and more heterogeneous ISIC9042 dataset, the performance was also reliable, as previous results indicated that transformers perform well with a wide range of medical images [4]. There were also preprocessing strategies that enhanced management of the imbalance of classes and variability of data [11].

C. User-Oriented Responsiveness and Explainability

Grad-CAM module allows generating easily interpretable heatmaps of key areas that affect prediction, enhancing clinical trust and interpretability without introducing obvious processing time. This type of explainability is consistent with the suggestions regarding AI-based medical diagnostics [4].

D. Sensitivity Analysis

There were slight performance problems in low-light conditions, image artifacts, unusual lesion patterns - problems that are often encountered during dermoscopic imaging [16]. The global awareness of ViT however maintained stability by maintaining long range spatial relations, as well as making reliable predictions even in cases when local features were ambiguous [6].

XI. DATA MODEL

The system has an organized data model that arranges the dermoscopic images, metadata, and diagnostic products. The images are further stored with corresponding labels of their classes, preprocessing information, and patch level embeddings so that the Vision Transformer can remember both global and local lesions [3]. Confidence scores and Grad-CAM heatmaps are also elements of metadata, which aids clinical traceability and enhance interpretability as per the current practices of explainable AI [4]. The folders of datasets are organized in the same way as the ISIC archive, so that they can be used with standard deep-learning pipelines and recreate experiments [18]. The predictions, activation maps and evaluation results are stored in the user-facing records, which allow one to easily review previous analyses. It is a lightweight data model that stores only important diagnostic data and also has minimal storage overhead.

XII. COMPARISON OF RESULTS

The proposed Vision transformer (ViT) model has an accuracy of 88.2 on ISIC-2019 and 82.5 on ISIC9042, which is very close to some of the state-of-the-art models in Table I. Although other hybrid and multi-modal models, including xCViT and RemixFormer++, do attain in the high-90s range of accuracy, these are based on more complex models, that is, CNNs, attention blocks, or multimodal inputs. Conversely, the suggested pure ViT methodology can be compared with the same level of competitiveness and has the simpler design, demonstrating that transformer-based global attention is able to identify key diagnostic features without further fusion networks. This shows that there is a trade-off between the accuracy, simplicity of the model, and

interpretability, which proves the feasibility of the use of transformer-only solutions to the reliable classification of skin cancer.

TABLE I. COMPARISON OF PROPOSED ViT-BASED CLASSIFIER WITH STATE-OF-THE-ART (SOA) ALGORITHMS FOR SKIN CANCER DETECTION

Algorithm Used	Dataset Used	Accuracy (%)	Citation No.
Personalized Multimodal Fed. Learning	Private/ISIC	89.8	[1]
CNN–Transformer–GRU Fusion	HAM10000/PH2	92.1	[2]
Hybrid CNN & ViT	ISIC-2018	87.6	[5]
Multi-Model Attentional Fusion	ISIC 2019	90.7	[13]
Hybrid QDL BiLSTM–MobileNetV2	ISIC 2018	85.4	[15]
Hybrid LSTM–CNN	DermIS/DermQuest	86.3	[7]
xCViT (CNN–Xception–ViT Fusion)	Private/ISIC	91.2	[12]
RemixFormer++ Multi-modal Transformer	ISIC archive	93.5	[14]
Swin Transformer (Hybrid)	ISIC 2018/PH2	92.5	[18]
YoTransViT (CNN + Transformer)	Derm7pt/HAM10000	88.8	[17]
ViT-based classifier (Ours)	ISIC 2019	88.2	-
ViT-based classifier	ISIC9042	82.5	-

XIII. JUSTIFICATION OF RESULTS

The results obtained are in line with the anticipated behaviour of skin cancer classifiers which is based on transformers. The high performance on both datasets was enabled by the fact that the Vision Transformer (ViT) is able to capture global contextual relationship in dermoscopic images, which has been shown to be an effective method in previous dermatology research [3], [6]. The increased accuracy with the ISIC-2019 Balanced Subset is explained by the controlled distribution of the dataset that minimizes the imbalance between the classes and variability of the imaging. The reduced yet still constant precision on the more varied dataset of ISIC9042 is due to having a wider variety of acquisition conditions and lesion heterogeneity, which is a difficulty of medical imaging models. Grad-CAM visualizations also confirm the accuracy of predictions by underlining areas of lesions which dermatologists would assess, which proves the model interpretable. The proposed method provides competitive results compared to state-of-the-art methods in Table I even though the pure transformer design is used in the proposed method. This verifies that despite the absence of necessary complex hybrid pipelines, ViT structures can reach a high-quality diagnostic accuracy, and the relevance and effectiveness of the suggested system should be justified.

XIV. CONCLUSION

This paper involved an improved Version Transformer (ViT) model in detecting skin cancer with automated detection using the ISIC-2019 Balanced Subset and ISIC9042 datasets. It has reached an accuracy of 88.2 and 82.5 respectively, which proves that the system can extract global features and discover malignancy patterns with high accuracy. Transparency was also enhanced because the Grad-CAM was easily integrated, which allowed areas of critical lesions to be visualized, and such areas were supported by clinical interpretability. The findings on the whole indicate the effectiveness of transformer-based architectures in medical imaging tasks as it gives an efficient alternative to conventional CNN or hybrid designs. The suggested model offers a balanced approach that is accurate, efficient, and explainable hence it can be used in dermatology-related diagnostic tasks.

FUTURE WORK

The model can be improved in a number of ways in the future research. More generalization would be achieved by increasing the diversity of datasets, particularly, the number of skin tones, rare lesion types and a variety of imaging environments. There is also the potential to enhance noise and lighting sensitivity and acquisition artifact resistance using sophisticated augmentation methods and domain adaptation techniques. In addition, the incorporation of the hybrid architectures, where CNNs are used to extract the features, and transformer-level global attention is introduced, might potentially enhance the current accuracy to an even higher level. The application of federated learning can potentially resolve the issue of privacy since training can be conducted across numerous medical institutions without transferring any confidential patient information. Lastly,

implementation of the model within the clinical workflow and testing it with the help of dermatologists will be significant milestones on the way to creating reliable, clinically deployable skin cancer detection systems.

REFERENCES

- [1] S. Fan, A. Ahmed, X. Zeng, R. Xi, and M. Hou, "A personalized multimodal federated learning framework for skin cancer diagnosis," *Electronics*, vol. 14, no. 14, p. 2880, 2025.
- [2] A. Varghese, A. Jain, M. I. Rahman, M. Khan, A. K. Dubey, I. Ahmed, Y. P. Narayan, A. Panwar, A. Choubey, and S. Mallik, "Robust skin cancer detection through CNN–Transformer–GRU fusion and generative adversarial network-based data augmentation," *Comput. Model. Eng. Sci. (CMES)*, vol. 144, no. 2, 2025.
- [3] S. Khan, H. Ali, and Z. Shah, "Identifying the role of vision transformer for skin cancer—A scoping review," *Front. Artif. Intell.*, vol. 6, p. 1202990, 2023.
- [4] 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," *IEEE Access*, vol. 12, pp. 109544–109559, 2024.
- [5] G. D. George and T. Y. Satheesha, "A novel deep learning approach for automated melanoma classification using hybrid CNN and vision transformer model," *Fusion Pract. Appl.*, vol. 19, no. 2, 2024.
- [6] G. Yang, S. Luo, and P. Greer, "A novel vision transformer model for skin cancer classification," *Neural Process. Lett.*, vol. 55, no. 7, pp. 9335–9351, 2023.
- [7] S. M. M. Abohashish, H. H. Amin, and E. I. Elsedimy, "Enhanced melanoma and non-melanoma skin cancer classification using a hybrid LSTM-CNN model," *Sci. Rep.*, vol. 15, no. 1, p. 24994, 2025.
- [8] A. Ahmed, G. Sun, A. Bilal, Y. Li, and S. A. Ebad, "A hybrid deep learning approach for skin lesion segmentation with dual encoders and channel-wise attention," *IEEE Access*, vol. 13, pp. 42608–42621, 2025.
- [9] G. Cirrincione et al., "Transformer-based approach to melanoma detection," *Sensors*, vol. 23, no. 12, p. 5677, 2023.
- [10] A. K. M., "Hybrid approach for skin lesion analysis integrating modified U-Net segmentation with vision transformers for multi-class skin cancer detection," 2024.
- [11] S. Khan and A. Khan, "SkinViT: A transformer-based method for melanoma and nonmelanoma classification," *PLOS ONE*, vol. 18, no. 12, p. e0295151, 2023.
- [12] A. Ali, H. Shahbaz, and R. Damaeviius, "xCViT: Improved vision transformer network with fusion of CNN and Xception for skin disease recognition with explainable AI," *Comput. Mater. Continua*, vol. 83, no. 1, 2025.
- [13] I. Ahmed et al., "Multi-model attentional fusion ensemble for accurate skin cancer classification," *IEEE Access*, 2024.
- [14] J. Xu et al., "RemixFormer: A multi-modal transformer model for precision skin tumor differential diagnosis with memory-efficient attention," *IEEE Trans. Med. Imag.*, 2024.
- [15] A. A. Hussein, A. M. Montaser, and H. A. Elsayed, "Skin cancer image classification using hybrid quantum deep learning model with BiLSTM and MobileNetV2," *Quantum Mach. Intell.*, vol. 7, no. 2, p. 66, 2025.
- [16] S. Mushtaq and O. Singh, "Implementing image processing and deep learning techniques to analyze skin cancer images," *Int. J. Comput. Digit. Syst.*, vol. 15, no. 1, pp. 1243–1257, 2024.
- [17] G. Y. Oztel, "Vision transformer and CNN-based skin lesion analysis: Classification of monkeypox," *Multimedia Tools Appl.*, 2024.
- [18] D. K. Saha, A. M. Joy, and A. Majumder, "YoTransViT: A transformer and CNN method for predicting and classifying skin diseases using segmentation techniques," *Inform. Med. Unlocked*, vol. 47, 2024.
- [19] I. Pacal, M. Alaftekin, and F. D. Zengul, "Enhancing skin cancer diagnosis using swin transformer with hybrid shifted window-based multi-head self-attention and SwiGLU-based MLP," *J. Imaging Inform. Med.*, vol. 37, no. 6, pp. 3174–3192, 2024.
- [20] A. Jimi et al., "ESC-UNET: A hybrid CNN and swin transformers for skin lesion segmentation," *Intell.-Based Med.*, p. 100257, 2025.
- [21] N. J. Jaber Faisal and A. Akbas, "Melanoma skin cancer detection based on deep learning methods and binary Harris Hawk optimization," *Multimedia Tools Appl.*, vol. 84, no. 22, pp. 25709–25722, 2025.
- [22] N. Ashfaq et al., "SkinSight: Advancing deep learning for skin cancer diagnosis and classification," *Discover Comput.*, vol. 28, no. 1, p. 63, 2025.