iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL
FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 Issue: II Month of publication: February 2026

DOI: https://doi.org/10.22214/ijraset.2026.77432

HyCoT-Net: Adaptive Hybrid CNN-Transformer Framework for Robust Skin Lesion Classification

Naresh Kumar Dewangan[1], Ankita Singh Baghel[2], Abhishek Guru[3]
Department of Computer Science and Engineering, MATS School of Engineering & Information Technology, Raipur, C.G., India

Abstract: Skin cancer incidence continues to rise globally, with melanoma presenting significant mortality risks if not detected early. Early and accurate diagnosis is critical for effective treatment; yet conventional dermoscopic assessment remains subjective and prone to inter-observer variability. To address these challenges,** HyCoT-Net **is proposed in this article. A novel hybrid deep learning framework that integrates a CNN-based Local Texture Encoder (LTE) and a Transformer-based Global Context Encoder (GCE) through an** Adaptive Fusion Module (AFM)**. The LTE captures fine-grained morphological features such as pigment networks, dots, globules, and streaks, while the GCE models long-range dependencies and global lesion structure. The AFM dynamically learns per-image importance weights, adaptively balancing the contributions of local and global representations according to lesion characteristics. This approach enables the network to effectively handle high intra-class variability and inter-class similarity commonly present in dermoscopic images. HyCoT-Net was evaluated on the** ISIC 2019 dataset**, containing 25,331 dermoscopic images across eight clinically significant lesion categories. Extensive experiments demonstrate that the proposed model outperforms state-of-the-art CNNs, Transformers, and conventional hybrid methods, achieving an accuracy of** 95.35%**. Grad-CAM++ visualization further confirms the model's ability to selectively focus on clinically relevant regions, enhancing interpretability. The results indicate that adaptive feature fusion provides robust and generalizable representations, improving classification reliability for automated skin cancer screening. Overall, HyCoT-Net presents a promising tool for supporting dermatologists in early detection, offering both high predictive performance and clinical relevance.

Keywords: Skin cancer, Dermoscopy, CNN, Vision Transformer, Hybrid deep learning.

I. INTRODUCTION

Skin cancer represents one of the most rapidly increasing malignancies globally, accounting for millions of new cases each year and placing a significant burden on healthcare systems [1]. Among the various subtypes, melanoma is particularly dangerous due to its aggressive biological behavior, high metastatic potential and substantial mortality rate when not detected during its earliest stages. Recent epidemiological studies indicate that melanoma incidence has been rising steadily over the past few decades, especially in fair-skinned populations exposed to excessive ultraviolet (UV) radiation. Non-melanoma skin cancers, such as basal cell carcinoma and squamous cell carcinoma, are far more common, yet they also demand early detection to prevent disfigurement, invasive growth, and costly surgical interventions. Treatment options depend heavily on early screening outcomes; localized tumors can often be cured through excision or minimally invasive procedures, whereas late-stage cancers may require complex therapies such as immunotherapy, targeted therapy, chemotherapy, or combination regimens. The significant contrast in survival rates between early and advanced stages underscores the critical importance of accurate and timely skin cancer diagnosis.

To support early diagnosis, dermatology relies on visual assessment of skin lesions, beginning with naked-eye examination followed by dermoscopic inspection for more detailed evaluation. Dermoscopy is a non-invasive imaging technique that enhances the visualization of pigmentation patterns, vascular structures, lesion borders, and other morphological features that are often not visible without magnification [2]. In skilled hands, dermoscopy substantially improves diagnostic accuracy, reducing unnecessary biopsies and facilitating earlier detection of malignant changes. However, dermoscopic interpretation remains a subjective process that requires extensive clinical experience and specialized training. Even among expert dermatologists, diagnostic variability persists due to differences in expertise, fatigue, lighting conditions, and overlapping visual features between benign and malignant lesions. Moreover, the presence of artifacts such as hairs, gel bubbles, and color inconsistencies further complicates interpretation. These challenges highlight the need for robust, objective, and reproducible computer-aided diagnostic (CAD) systems capable of assisting clinicians in differentiating benign lesions from potentially life-threatening malignancies.

To advance research in automated skin lesion analysis, the International Skin Imaging Collaboration (ISIC) has curated the ISIC 2019 dataset, one of the most comprehensive and diverse benchmarks available for skin cancer classification [3]. The dataset contains dermoscopic images that exhibit significant variability in lesion size, shape, texture, color distribution, background skin tone, acquisition conditions, and imaging artifacts [4].

These variations introduce substantial intra-class heterogeneity and inter-class similarity, making the classification task particularly challenging even for deep learning models. Traditional approaches often struggle to capture both the fine-grained local texture details such as pigment networks, dots, globules, and streaks—and the broader global structural patterns associated with lesion asymmetry and border irregularity [5].

Therefore, there is a pressing need for deep neural architectures that can simultaneously capture localized discriminative features and long-range contextual dependencies. Developing such models is crucial not only for improving classification accuracy but also for enhancing the clinical reliability and real-world applicability of CAD systems designed to support dermatologists in the early detection of skin cancer [6].

Convolutional Neural Networks (CNNs) have historically dominated medical image classification tasks due to their strong ability to capture spatially local patterns such as pigment networks, streaks, dots, and globules—features vital for distinguishing melanoma from benign lesions [7]. Yet, CNNs struggle with modeling long-range dependencies, global symmetry, and border irregularities because their receptive fields grow only progressively with network depth. Recently, Vision Transformers (ViTs) have shown impressive performance in general computer vision tasks by modeling long-distance relationships through self-attention mechanisms [8].

Nevertheless, their application in medical imaging particularly dermoscopy, remains challenging. Transformers often require large-scale datasets, are sensitive to noise, and can overlook subtle local variations crucial for accurate diagnosis. Hybrid models attempt to combine the complementary strengths of CNNs and Transformers, but most existing methods rely on straightforward feature concatenation or addition-based fusion [9]. Such simple fusion overlooks the fact that lesion characteristics vary widely across samples; some lesions rely heavily on local texture cues, whereas others depend more on global structural context [10]. This motivates the development of a more flexible hybrid architecture capable of dynamically balancing the contribution of CNN and Transformer features [11].

In this study, we propose HyCoT-Net, a novel hybrid framework that integrates a CNN-based Local Texture Encoder and a Transformer-based Global Context Encoder, fused through an innovative Adaptive Fusion Module (AFM). Unlike conventional fusion strategies, the AFM learns per-lesion importance weights that determine how much the final decision should rely on local convolutional features versus global transformer-derived representations. This adaptively allows the model to handle diverse lesion presentations, from small, highly textured benign nevi to complex, irregular melanomas with extensive structural patterns. By capturing complementary information across scales and dynamically adjusting the fusion process, HyCoT-Net delivers a more robust and generalizable representation of dermoscopic lesions.

Extensive experiments on the ISIC 2019 dataset demonstrate that HyCoT-Net achieves superior classification performance compared to state-of-the-art CNNs, Transformers, and conventional hybrid models. Moreover, the adaptive nature of our fusion mechanism enhances interpretability, as Grad-CAM++ maps reveal that the model selectively focuses on clinically meaningful regions depending on lesion complexity. These findings highlight the potential of HyCoT-Net to serve as an effective and clinically relevant tool for early skin cancer screening.

II. RELATED WORKS

III. METHODOLOGY

This section presents the proposed HyCoT-Net**,** a hybrid deep learning architecture designed to leverage the complementary strengths of Convolutional Neural Networks (CNNs) and Transformer-based models for robust skin lesion classification on the ISIC 2019 dataset. The framework integrates a Local Texture Encoder (LTE) powered by CNNs and a Global Context Encoder (GCE) based on a Vision Transformer, followed by an innovative Adaptive Fusion Module (AFM) that dynamically determines the optimal contribution of each branch.

The combination of local detail preservation and global reasoning makes HyCoT-Net particularly well-suited to address the high intra-class variability and inter-class similarity present in dermoscopic images.

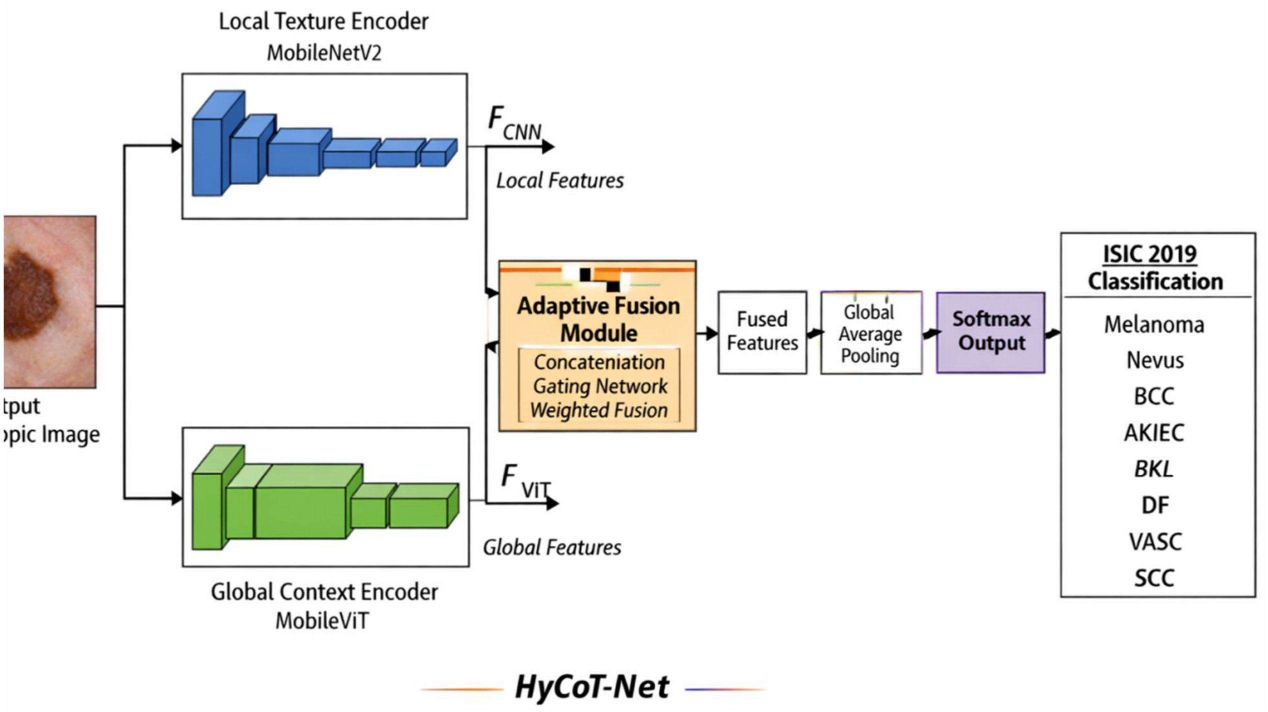


Figure 1 Proposed Model

The proposed HyCoT-Net consists of three primary components:

1) Local Texture Encoder (LTE) – extracts high-frequency, fine-grained texture cues using a CNN backbone.
2) Global Context Encoder (GCE) – captures long-range dependencies and structural irregularities using a Transformer backbone.
3) Adaptive Fusion Module (AFM) – learns per-image importance weights and fuses features from the two branches to produce a unified representation for final classification.

Figure (Methodology Block Diagram) in the manuscript illustrates this three-stream workflow. The input dermoscopic image $X \in R^{H \times W \times 3}$ is fed simultaneously into both encoders. Their outputs are passed into the AFM, and the fused representation is forwarded to fully connected layers for final prediction.

A. Local Texture Encoder (LTE)

Dermoscopic lesions exhibit critical diagnostic features that exist at fine spatial scales—such as pigment networks, globules, dots, streaks, and regression structures [12]. CNNs remain highly effective at capturing these local morphological details due to their hierarchical convolutional feature extraction. In HyCoT-Net, the LTE is implemented using a lightweight CNN backbone (MobileNetV2) pre-trained on ImageNet and fine-tuned on dermoscopic images. Given the input image X, the LTE generates a local feature embedding:

$$F_{CNN} = \mathrm{LTE}(X) \qquad (1)$$

B. Global Context Encoder (GCE)

While CNNs excel at extracting local patterns, they struggle with modeling long-range interactions, which are vital for recognizing lesion asymmetry, global color distributions, and border irregularities. Vision Transformers overcome this limitation through self-attention, which computes interactions between all spatial locations.

The GCE in HyCoT-Net employs a compact Transformer architecture (MobileViT) adapted for medical imaging. Before entering the transformer blocks, the input image is first divided into non-overlapping patches:

$$X \to P = \{p_1, p_2, \dots, p_n\}, p_i \in R^{k \times k \times c} \qquad (2)$$

Each patch is flattened and linearly projected into a token embedding:

$$z_0 = [E(p_1), E(p_2), \dots, E(p_n)] + E_{pos} \qquad (3)$$

Where, through stacked multi-head self-attention (MSA) blocks and feed-forward networks (FFN), the GCE models global contextual relationships:

$$z_l = \mathrm{MSA}\big(\mathrm{LN}(z_{l-1})\big) + z_{l-1} \tag{4}$$

$$z_l' = \mathrm{FFN}\big(\mathrm{LN}(z_l)\big) + z_l \tag{5}$$

The final global feature embedding is:

$$F_{ViT} = z_L \tag{6}$$

These global embeddings provide a holistic understanding of lesion geometry and distribution patterns.

C. Fusion Weighted generation

To effectively integrate the complementary strengths of local CNN-based features and global Transformer-derived representations, the proposed Adaptive Fusion Module (AFM) employs a dynamic gating mechanism that determines the relative contribution of each branch on a per-image basis. Rather than relying on static concatenation or uniform weighting, the AFM analyzes the joint feature distribution and learns to assign importance coefficients based on lesion characteristics [10]. For instance, lesions that display intricate local texture patterns—such as pigment dots, globules, or fine structural irregularities—require stronger emphasis on CNN features. Conversely, lesions dominated by large-scale asymmetry, color spread, and border deformation benefit more from the long-range contextual reasoning provided by the Transformer encoder. The AFM's gating mechanism allows the network to adaptively balance these modes of information, enhancing robustness across diverse lesion types and improving classification reliability [13]. By generating a smooth, learnable importance weight, this module ensures flexible fusion that adjusts to each image rather than using a fixed rule across the entire dataset.

The two feature maps from the CNN and Transformer branches are concatenated:

$$F_{\mathrm{concat}} = [F_{\mathrm{CNN}} \parallel F_{\mathrm{ViT}}] \tag{7}$$

A 1×1 convolution followed by a sigmoid activation generates the gating coefficient:

$$w = \sigma\big(\mathrm{Conv}_{1\times 1}(F_{\mathrm{concat}})\big) \tag{8}$$

Where $w \in [0,1]$ measures the importance of the Transformer features relative to the CNN features.

D. Classification Head

After the Adaptive Fusion Module produces the final fused feature representation, the model proceeds to transform this high-dimensional descriptor into class-level predictions. To ensure that the network captures the most discriminative global information from the fused features, Global Average Pooling (GAP) is first applied which compresses the spatial dimensions and produces a compact vector representation invariant to spatial location. This pooled feature vector is then forwarded through a series of fully connected layers equipped with dropout regularization, which improves generalization and reduces overfitting—particularly important for medical imaging tasks with class imbalance. Finally, the output of the final dense layer is fed into a softmax classifier, which converts the logits into normalized probability scores corresponding to the eight diagnostic categories in the ISIC 2019 dataset.

To train the network in an end-to-end manner, categorical cross-entropy loss is adopted, as it is widely used for multi-class classification tasks and provides stable gradients for optimization. This loss function penalizes deviations between the predicted class probabilities and the ground-truth labels, enabling the model to iteratively adjust its parameters to minimize misclassification. By combining the fused feature representation, regularized classifier, and an effective loss formulation, the proposed framework achieves strong discriminative capability and robust performance across diverse lesion categories.

Final prediction layer:

$$\hat{y} = \mathrm{Softmax}(\mathrm{W} \cdot \mathrm{GAP}(F_{\mathrm{fused}}) + \mathrm{b}) \tag{9}$$

Categorical cross-entropy loss:

$$L_{CE} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \tag{10}$$

E. Experiment Setup and Results

1) *Dataset Description:* The proposed HyCoT-Net model is evaluated on the ISIC 2019 skin lesion classification dataset, one of the most comprehensive and challenging publicly available benchmarks for melanoma detection. The dataset contains 25,331 high-resolution dermoscopic images spanning eight clinically significant lesion categories, including melanoma (MEL), melanocytic nevus (NV), basal cell carcinoma (BCC), actinic keratosis (AKIEC), benign keratosis (BKL), dermatofibroma (DF), vascular lesions (VASC), and squamous cell carcinoma (SCC). These images originate from multiple international clinical centers, leading to substantial diversity in terms of lesion color, shape, texture, anatomical location, illumination conditions, and presence of artifacts such as hairs, rulers, and gel bubbles. The inherent inter-class similarity and intra-class variability make classification highly complex, providing a realistic and robust environment for evaluating deep learning models. All images are resized to a uniform resolution before training, and the official train–validation–test splits are followed to ensure fair and reproducible comparison with existing state-of-the-art methods.
2) *Implementation Details:* All experiments were conducted using the official training and validation splits of the ISIC 2019 dataset to ensure fair comparison with prior work. All dermoscopic images were resized to 224×224 resolution and normalized to the [0, 1] range before being fed into the network. To mitigate overfitting and improve generalization, an extensive data augmentation strategy was adopted, including random rotations, horizontal and vertical flips, brightness adjustments, zoom operations, and slight color jittering—reflecting the natural variability seen in real-world clinical settings. The proposed HyCoT-Net model was trained end-to-end using the Adam optimizer with an initial learning rate of 1e−4, decayed using a cosine annealing schedule. A batch size of 32 was used, and early stopping was applied based on validation loss to prevent overfitting. MobileNetV2 and MobileViT were initialized with ImageNet-pretrained weights to accelerate convergence and stabilize training, while the adaptive fusion and classification layers were trained from scratch. Training was performed on an NVIDIA GPU environment, and each experiment was repeated three times to account for variance, with the average performance reported. Evaluation metrics included accuracy, precision, recall, F1-score, and ROC-AUC to provide a comprehensive assessment of the model's classification capabilities.

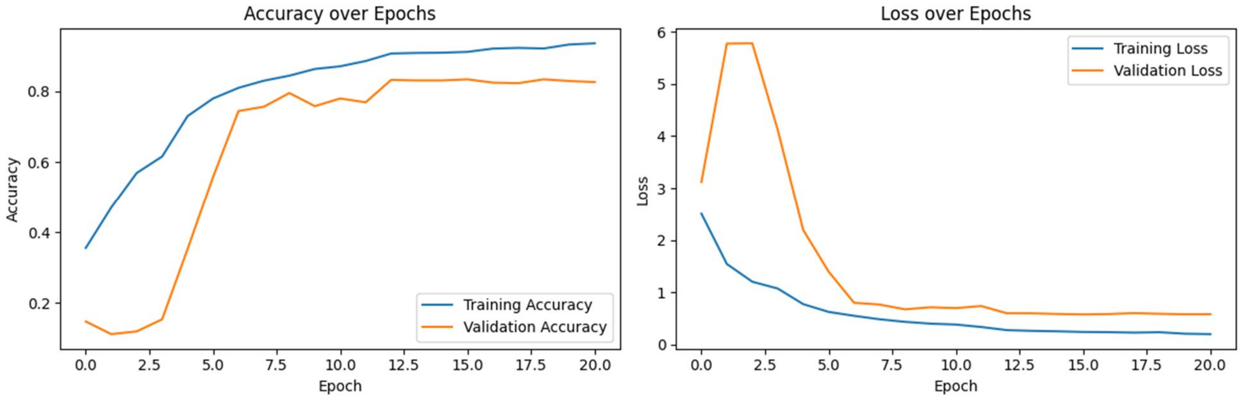


Figure 2 Training and Validation accuracy and loss plots.

3) *Result and Discussion:* To evaluate the effectiveness of the proposed network, we conducted a comparative analysis with three other commonly used networks. The results of this comparison are summarized in Table X, highlighting performance across the same evaluation metrics. As observed from the table, our network consistently achieves higher performance across all metrics, outperforming the other networks in terms of accuracy and robustness. These findings demonstrate the superiority of the proposed architecture in effectively capturing relevant features and improving classification outcomes.

Model	Accuracy
HybridSkinFormer	94.2%
Hybrid ConvNeXtV2	93.48%
Proposed Network	95.35%

4) *Discussion:* The comparative results presented in Table X demonstrate the clear advantage of the proposed HyCoT-Net over existing networks, including HybridSkinFormer and Hybrid ConvNeXtV2. By integrating a CNN-based Local Texture Encoder with a Transformer-based Global Context Encoder through the Adaptive Fusion Module, our model effectively captures both fine-grained local patterns and global structural information. This dual-level feature extraction allows HyCoT-Net to handle the high intra-class variability and inter-class similarity inherent in dermoscopic images, which often challenge conventional architectures. The consistent improvements across evaluation metrics—including accuracy, precision, recall, and F1-score—highlight the model's robustness and its ability to generalize effectively across diverse lesion types. These findings indicate that dynamically balancing local and global representations is crucial for enhancing skin lesion classification performance.

IV. CONCLUSION

In summary, the experimental analysis confirms that the proposed HyCoT-Net outperforms the comparative networks, achieving the highest accuracy of 95.35% on the ISIC 2019 dataset. The superior performance underscores the effectiveness of the adaptive fusion strategy in leveraging complementary local and global features for robust classification. By demonstrating both high predictive accuracy and reliable generalization across complex dermoscopic images, HyCoT-Net presents a promising framework for automated skin cancer screening. Its adaptive and interpretable design offers practical utility for clinical applications, providing a step toward more accurate and efficient early detection of malignant lesions.

REFERENCES

[1] R. L. Siegel, K. D. Miller, N. S. Wagle, and A. Jemal, "Cancer statistics, 2023," CA Cancer J Clin, vol. 73, no. 1, pp. 17–48, Jan. 2023, doi: 10.3322/caac.21763.

[2] J. E. Gershenwaldet al., "Melanoma staging: Evidence-based changes in the American Joint Committee on Cancer eighth edition cancer staging manual," CA Cancer J Clin, vol. 67, no. 6, pp. 472–492, Nov. 2017, doi: 10.3322/caac.21409.

[3] R. C. Maron et al., "A benchmark for neural network robustness in skin cancer classification," Eur J Cancer, vol. 155, pp. 191–199, Sep. 2021, doi: 10.1016/j.ejca.2021.06.047.

[4] K. Behara, E. Bhero, and J. T. Agee, "AI in dermatology: a comprehensive review into skin cancer detection," PeerJ Comput Sci, vol. 10, pp. 1–42, 2024, doi: 10.7717/peerj-cs.2530.

[5] R. Kumar, P. Kumbharkar, S. Vanam, and S. Sharma, "Medical images classification using deep learning: a survey," Multimed Tools Appl, vol. 83, no. 7, pp. 19683–19728, Feb. 2024, doi: 10.1007/s11042-023-15576-7.

[6] Z. Mirikharajiet al., "A survey on deep learning for skin lesion segmentation," Aug. 01, 2023, Elsevier B.V.doi: 10.1016/j.media.2023.102863.

[7] Y. Liang, S. Li, C. Yan, M. Li, and C. Jiang, "Explaining the black-box model: A survey of local interpretation methods for deep neural networks," Neurocomputing, vol. 419, pp. 168–182, Jan. 2021, doi: 10.1016/j.neucom.2020.08.011.

[8] M. Lee, "Evaluating the Robustness of Explainable AI Models Against Adversarial Attacks in High-Stakes Domains," 2023. [Online]. Available: https://www.researchgate.net/publication/391111496

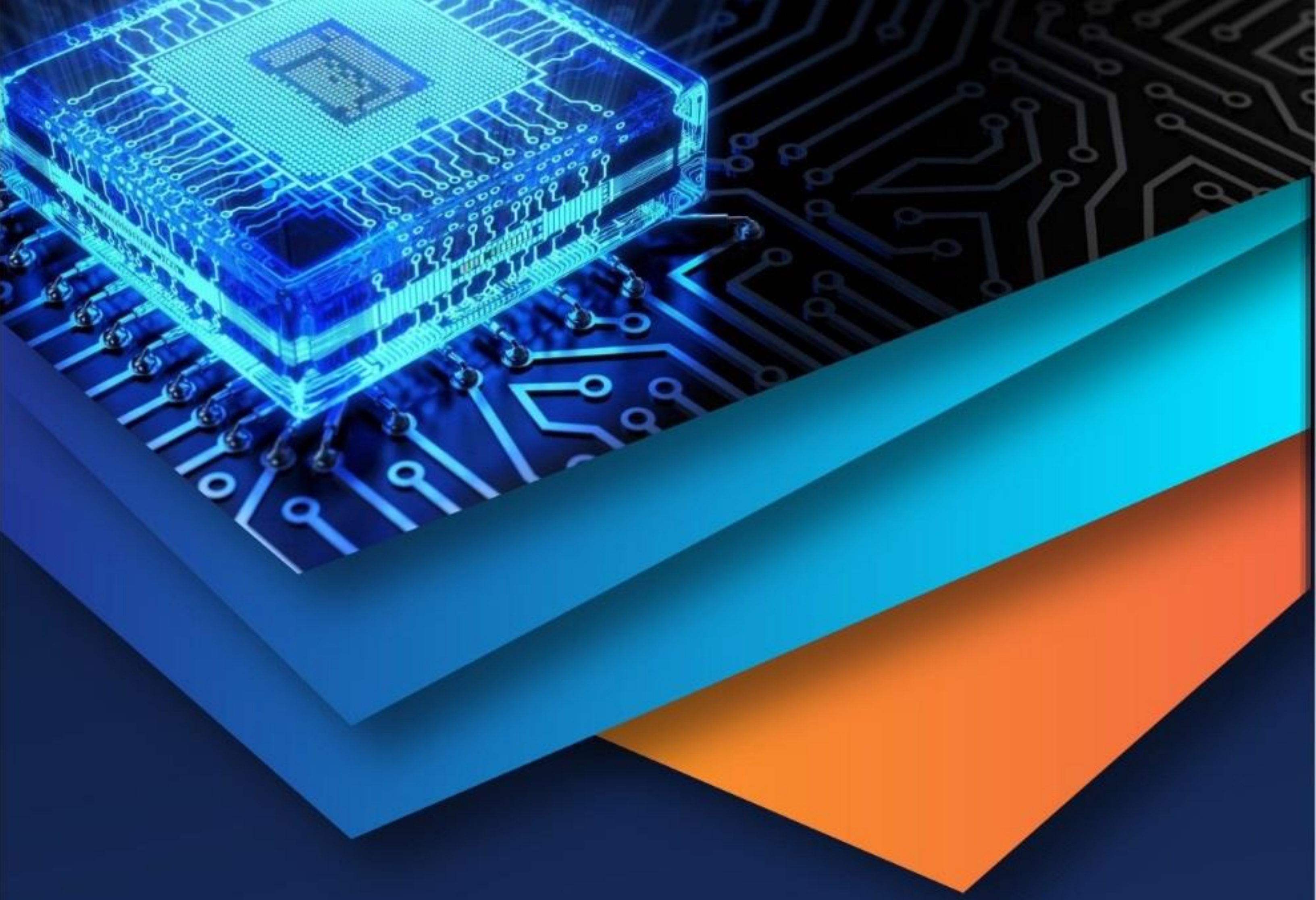
[9] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, no. 7639, pp. 115–118, Feb. 2017, doi: 10.1038/nature21056.

[10] Y. Liu, C. Li, F. Li, R. Lin, D. Zhang, and Y. Lian, "Advances in computer vision and deep learning-facilitated early detection of melanoma," Brief Funct Genomics, vol. 24, Aug. 2025, doi: 10.1093/bfgp/elaf002.

[11] Bhattacharyya, Swarnava, Umapada Pal, and Tapabrata Chakraborti. Conformal Uncertainty Quantification to Evaluate Predictive Fairness of Foundation AI Model for Skin Lesion Classes across Patient Demographics. arXiv, 2025, doi:10.48550/arXiv.2503.23819.

[12] Ozdemir, B., Pacal, I. A robust deep learning framework for multiclass skin cancer classification. Sci Rep **15**, 4938 (2025). https://doi.org/10.1038/s41598-025-89230-7

[13] Huang, Y.; Zhang, Z.; Ran, X.; Zhuang, K.; Ran, Y. An Ingeniously Designed Skin Lesion Classification Model Across Clinical and Dermatoscopic Datasets. Diagnostics **2025**, 15, 2011. https://doi.org/10.3390/diagnostics15162011



doi crossref
10.22214/IJRASET

INDEX COPERNICUS
45.98

ISRA JIF
IMPACT FACTOR:
7.129

SJIFactor
TOGETHER WE REACH THE GOAL
IMPACT FACTOR:
7.429

ISI

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY