IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

TRANSPARENT AI FOR SKIN CANCER CLASSIFICATION USING EXPLAINABLE CNN–VIT MODELS

[1]Dhushyanth J, [2]Krishna K, [3]Mr. V. GokulaKrishnan

[1]Student, Department of Computer Science and Business Systems, [2]Student, Department of Computer Science and Business Systems, [3]Assistant Professor (Guide), Department of Computer Science and Business Systems

K. Ramakrishnan College of Engineering, Trichy, Tamil Nadu, India

Abstract: Skin cancer is one of the most prevalent and life-threatening forms of cancer worldwide, and early accurate detection is critical for improving patient outcomes. This paper presents a transparent and explainable artificial intelligence (AI) framework for skin cancer classification by integrating Convolutional Neural Networks (CNN) with Vision Transformer (ViT) models. While deep learning models have demonstrated remarkable accuracy in medical image analysis, their black-box nature limits clinical adoption. To address this, our approach leverages Gradient-weighted Class Activation Mapping (Grad-CAM) and attention rollout techniques to generate visual explanations of the model's decision-making process. The hybrid CNN–ViT architecture captures both local texture features and global contextual dependencies from dermoscopic images. Experiments conducted on the HAM10000 benchmark dataset demonstrate that the proposed model achieves a classification accuracy of 93.7%, with significant improvements in sensitivity and specificity compared to standalone CNN or ViT models. The explainability maps highlight clinically relevant lesion regions, thereby building trust for real-world dermatological diagnosis. This work contributes toward making AI-based skin cancer tools transparent, interpretable, and deployable in clinical environments.

Index Terms — *Skin Cancer Classification, Explainable AI (XAI), Convolutional Neural Networks (CNN), Vision Transformer (ViT), Grad-CAM, Dermoscopy, Deep Learning, Medical Imaging.*

I. INTRODUCTION

Skin cancer is one of the most commonly diagnosed cancers globally, with melanoma accounting for the majority of skin cancer-related deaths. Early and accurate diagnosis significantly improves the chances of successful treatment. Dermatologists traditionally rely on visual inspection and dermoscopy, but inter-observer variability and the sheer volume of cases pose considerable challenges. Deep learning, particularly Convolutional Neural Networks (CNNs), has emerged as a powerful tool for automated skin lesion classification.

However, despite achieving high classification accuracy, deep learning models are often perceived as opaque or black-box systems. In clinical settings, trust and transparency are paramount—medical professionals require interpretable reasoning behind any algorithmic decision before incorporating it into their diagnostic workflow. This limitation has motivated the growing field of Explainable Artificial Intelligence (XAI), which aims to make AI decisions understandable to human users.

Vision Transformers (ViT), originally proposed for natural language processing tasks, have demonstrated exceptional performance in image recognition by modeling long-range dependencies through self-attention mechanisms. When combined with CNNs, they offer complementary strengths:

CNNs effectively capture local texture patterns, while ViTs provide global contextual understanding of dermoscopic images.

This paper proposes a hybrid CNN–ViT architecture for skin cancer classification that is augmented with XAI methods including Grad-CAM and attention rollout. The goal is not only to achieve state-of-the-art accuracy but also to generate clinically meaningful visual explanations, enhancing transparency and supporting dermatologists in making informed diagnostic decisions.

II. LITERATURE SURVEY

Esteva et al. [1] demonstrated that a CNN trained on a large dataset of dermatological images could classify skin cancer at a level comparable to board-certified dermatologists. Their work established the foundation for deep learning applications in dermatology. Subsequently, numerous studies explored various CNN architectures including VGG, ResNet, and InceptionNet for lesion classification.

The introduction of Vision Transformers by Dosovitskiy et al. [2] marked a paradigm shift in computer vision. ViTs divide images into patches and apply transformer-based self-attention, enabling the model to capture long-range dependencies. Chen et al. [3] proposed TransMed, a hybrid model combining CNNs and transformers for medical image segmentation, demonstrating superior performance over standalone architectures.

On the explainability front, Selvaraju et al. [4] introduced Gradient-weighted Class Activation Mapping (Grad-CAM), which generates heatmaps highlighting regions most influential to the model's prediction. This technique has been widely applied in medical imaging to validate and interpret model decisions. Chefer et al. [5] proposed attention rollout as a method to propagate attention maps across transformer layers, enabling meaningful visualization of ViT decision boundaries.

More recent works such as those by Nauta et al. [6] have focused on combining CNN feature extraction with ViT attention for interpretable medical image classification. Despite progress, most existing approaches either sacrifice accuracy for interpretability or fail to provide explanations aligned with clinical expectations. Our work addresses this gap by integrating both accuracy and explainability in a unified framework.

III. PROPOSED METHODOLOGY

3.1 Dataset Description

The HAM10000 (Human Against Machine with 10000 training images) dataset was used for experimentation. It contains 10,015 dermoscopic images across seven skin lesion categories: Melanocytic nevi, Melanoma, Benign keratosis-like lesions, Basal cell carcinoma, Actinic keratoses, Vascular lesions, and Dermatofibroma. All images were resized to 224×224 pixels and normalized using ImageNet mean and standard deviation values.

3.2 CNN–ViT Hybrid Architecture

The proposed hybrid model employs a pre-trained ResNet-50 as the CNN backbone to extract local feature maps from dermoscopic images. These feature maps are then tokenized into patch embeddings and fed into a ViT encoder with 12 transformer blocks, each comprising multi-head self-attention layers and feed-forward networks. The final classification is performed by a multi-layer perceptron (MLP) head attached to the [CLS] token output.

The dual-stream design ensures that the model simultaneously leverages fine-grained local patterns (edges, textures, colors) captured by CNNs and global structural dependencies identified by the transformer. Skip connections are employed between the CNN and ViT modules to prevent information loss and facilitate gradient flow during training.

3.3 Explainability Methods

Two XAI techniques were integrated: (1) Grad-CAM was applied to the last convolutional layer of ResNet-50 to produce heatmaps indicating discriminative regions. (2) Attention Rollout was computed by multiplying attention weights across all ViT layers recursively, resulting in a global attention map across the input image. The final explainability visualization overlays both heatmaps on the original dermoscopic image, providing a comprehensive and interpretable output for clinical review.

3.4 Training Configuration

The model was trained for 50 epochs using the Adam optimizer with an initial learning rate of 1×10^{-4} and weight decay of 1×10^{-5}. Data augmentation techniques including random horizontal/vertical flips, rotation, and color jitter were applied to address class imbalance. A weighted cross-entropy loss function was used to handle the unequal distribution of lesion types in the dataset. The experiments were conducted on a system with NVIDIA RTX 3080 GPU and 32 GB RAM.

IV. RESULTS AND DISCUSSION

4.1 Classification Performance

Table 1 presents the performance comparison of baseline models and the proposed CNN–ViT hybrid on the HAM10000 dataset.

Table 1: Performance Comparison of Skin Cancer Classification Models

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)
ResNet-50 (CNN)	87.4	84.2	88.9	85.6
ViT-Base	89.1	86.7	90.3	87.8
CNN + ViT (Ours)	93.7	91.5	94.2	92.6

The proposed CNN–ViT hybrid model achieves the highest accuracy of 93.7%, outperforming the standalone ResNet-50 by 6.3% and ViT-Base by 4.6%. This demonstrates the complementary benefits of integrating local feature extraction with global attention modeling. The improved sensitivity of 91.5% is particularly noteworthy in a clinical context, as it minimizes false negatives—critical for cancer detection.

4.2 Explainability Analysis

The Grad-CAM heatmaps generated for melanoma cases consistently highlighted the lesion boundaries and asymmetrical color regions, which are clinically recognized hallmarks of malignancy. Attention rollout maps from the ViT module reinforced these regions, with high attention scores observed around lesion cores. Dermatologists who reviewed the explanation maps rated 89% of the highlighted regions as clinically relevant, confirming the practical utility of the XAI approach.

V. CONCLUSION

This paper presented a transparent AI framework for skin cancer classification using a hybrid CNN–ViT architecture augmented with Grad-CAM and attention rollout explainability methods. The proposed system achieves 93.7% classification accuracy on the HAM10000 dataset while providing interpretable visual explanations aligned with clinical understanding. By bridging the gap between model performance and human interpretability, this work advances the deployment of trustworthy AI in dermatological practice. Future work will explore multi-modal fusion incorporating patient metadata, as well as federated learning to enable privacy-preserving model training across multiple hospitals.

VI. ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the Department of Computer Science and Business Systems, K. Ramakrishnan College of Engineering, Trichy, for providing the necessary computational resources and academic support for this research. We also thank the creators of the HAM10000 dataset for making it publicly available to the research community.

VII. REFERENCES

[1] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115–118.

[2] Dosovitskiy, A., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. ICLR 2021.

[3] Chen, J., et al. (2021). TransMed: Transformers advance multi-modal medical image classification. Diagnostics, 11(8), 1384.

[4] Selvaraju, R. R., et al. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. ICCV 2017, 618–626.

[5] Chefer, H., Gur, S., & Wolf, L. (2021). Transformer interpretability beyond attention visualization. CVPR 2021, 782–791.

[6] Nauta, M., et al. (2023). Interpretable image classification with differentiable prototypes. ECML PKDD 2023.

[7] Tschandl, P., Rosendahl, C., & Kittler, H. (2018). The HAM10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data, 5(1), 1–9.

[8] Codella, N., et al. (2019). Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the International Skin Imaging Collaboration (ISIC). arXiv:1902.03368.

[9] Zhou, B., et al. (2016). Learning deep features for discriminative localization. CVPR 2016, 2921–2929.

[10] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. CVPR 2016, 770–778.