SKIN CANCER DETECTION USING HYBRID CNN–VISION TRANSFORMER ARCHITECTURE

1st Sanjinee
MS Artificial Intelligence
Fast, NUCES University
Karachi, Pakistan
sanjineehere@gmail.com

2nd Muhammad Abdullah Anwar
dept. name of organization (of Aff.)
Fast, NUCES University
Karachi, Pakistan
abdullah.anwar.muhammad@gmail.com

3rd Shiwam
BS Computer Science
Sukkur IBA University
Sukkur, Pakistan
gehanishiwam28@gmail.com

***Abstract*—Automated detection of skin cancer is hindered by the scarcity of annotated medical data and pronounced class imbalance. This paper introduces a hybrid CNN–Vision Transformer framework for multi-class skin lesion classification on the ISIC 2018 dataset, which includes seven diagnostic classes. Our method combines ResNet18 for extracting local image features with a domain-adapted Vision Transformer to model global contextual information. To counter class imbalance, we employ a dual strategy: Weighted Random Sampling and a weighted cross-entropy loss, with optimization targeted toward the Macro F1 Score. The proposed model attains a validation Macro F1 Score of 0.7751 and an accuracy of 80.98, surpassing conventional CNN-based baselines. For potential clinical use, we incorporate explainable AI methods such as Grad-CAM and SHAP-based patch attribution, showing that the model concentrates on clinically meaningful regions and resolves internal tensions between conflicting diagnostic cues. Overall, the findings indicate that hybrid architectures enhanced with domain adaptation and interpretability tools can perform reliably on small, imbalanced medical datasets while preserving the level of diagnostic transparency required for clinical practice.**

***Index Terms*—Skin cancer detection, Vision Transformer, hybrid CNN architecture, explainable AI, class imbalance, medical image classification, Grad-CAM, SHAP analysis**

I. Introduction

Skin cancer represents one of the most prevalent malignancies worldwide, with early detection being critical for patient survival. Traditional diagnostic approaches rely on visual examination by dermatologists, which is subjective and expertise-dependent. Automated diagnostic systems using deep learning offer promising solutions to augment clinical decision-making and improve diagnostic consistency.

Vision Transformers (ViTs) have demonstrated remarkable capabilities in capturing global contextual relationships within images. However, their application to medical imaging faces significant challenges. Medical datasets like the ISIC dataset are limited in size and exhibit severe class imbalances, where benign lesions vastly outnumber malignant cases. These constraints create critical obstacles for training models that must reliably detect rare but clinically significant conditions.

Traditional CNNs excel at extracting localized texture patterns but struggle to capture broader spatial relationships. Conversely, ViTs model global context effectively through self-attention mechanisms, yet their substantial parameter counts often lead to overfitting on small medical datasets. To address these complementary limitations, we propose a hybrid architecture combining ResNet18 for local feature extraction with a domain-adapted ViT for global attention mechanisms.

Our approach implements a dual-strategy framework for class imbalance: Weighted Random Sampling during training and weighted cross-entropy loss. We optimize for Macro F1 Score rather than overall accuracy, ensuring balanced performance across all diagnostic categories, including rare malignant lesions. Critically, we integrate Grad-CAM and SHAP-based attribution to provide clinical interpretability, offering visual evidence of features driving model predictions.

The primary contributions of this research are:

1. A hybrid CNN-ViT architecture adapted for small-scale, imbalanced medical datasets through domain-aware feature integration.

2. A dual-strategy imbalance mitigation approach ensuring equitable performance across all diagnostic classes.

3. Comprehensive explainability analysis through multiple XAI techniques supporting clinical interpretation.

4. Empirical validation demonstrating superior performance compared to pure-CNN baselines while maintaining diagnostic transparency.

II. Related Work

A. Deep Learning Approaches for Skin Cancer Detection

Automated skin cancer detection has evolved significantly with advances in deep learning architectures. Traditional CNN-based approaches have demonstrated strong performance in extracting localized features from dermatoscopic images. Musthafa et al. [5] proposed optimized CNN architectures with regularization techniques, achieving 92.9% accuracy on the HAM10000 dataset. While effective at capturing fine-grained texture patterns, pure CNN models face limitations in modeling long-range spatial dependencies critical for holistic lesion assessment.

B. Vision Transformers in Medical Imaging

The emergence of Vision Transformers has introduced new paradigms for medical image analysis. Khan et al. [4] conducted a comprehensive scoping review identifying the potential of ViTs in skin cancer detection, highlighting their superior ability to capture global contextual information through self-attention mechanisms. Himel et al. [3] applied ViTs for both segmentation and classification tasks on HAM10000 and PH2 datasets, achieving 94.1% accuracy. However, the substantial parameter requirements of standard ViTs pose significant challenges when applied to limited medical datasets, often resulting in overfitting.

C. Hybrid CNN-ViT Architectures

Recent research has increasingly focused on hybrid approaches that leverage the complementary strengths of CNNs and Transformers. Reis et al. [7] developed a fusion framework combining transformer attention with CNN embeddings, achieving 94.4% accuracy on ISIC-2019. Yang et al. [9] introduced attention pooling mechanisms within transformer architectures, reaching 93.2% accuracy while addressing adversarial noise concerns. Xin et al. [8] implemented CutMix augmentation with improved transformer networks, demonstrating 94.7% accuracy on HAM10000. More sophisticated fusion strategies have emerged to optimize feature integration. Halawani [2] proposed EViT-DenseNet169, combining DenseNet-169 CNN features with ViT through hybrid attention mechanisms, achieving 95.8% accuracy on ISIC-2020. Pacal [6] developed a CNN-ViT hybrid specifically addressing data imbalance and small dataset constraints, reaching 95.1% accuracy across multiple datasets. Most notably, Zhang et al. [10] introduced AViT (Adapting Vision Transformers), achieving 96.3% accuracy through diagnosis-guided integration of CNN features with domain-adapted ViTs on HAM10000.

D. Addressing Class Imbalance

Class imbalance remains a critical challenge in medical datasets where benign lesions significantly outnumber malignant cases. While several studies acknowledge this issue, approaches vary in sophistication. Halawani [2] and Pacal [6] explicitly addressed class imbalance through architectural modifications and attention mechanisms. However, many existing works optimize primarily for overall accuracy, potentially masking poor performance on rare but clinically critical malignant classes. This limitation necessitates evaluation metrics such as Macro F1 Score that ensure balanced performance across all diagnostic categories.

E. Explainability and Clinical Interpretability

The clinical deployment of automated diagnostic systems demands transparency in decision-making processes. Dagnaw [1] pioneered the integration of explainable AI techniques, combining ViTs with Grad-CAM and SHAP analysis to achieve 93.8% precision while providing interpretable predictions. This work highlighted the critical gap between achieving high accuracy and providing clinicians with visual evidence supporting model decisions. Despite this advancement, comprehensive explainability analysis remains underexplored in hybrid architectures, particularly in quantifying internal model conflicts and patch-level attribution.

F. Research Gap and Motivation

While existing hybrid approaches demonstrate promising results, several limitations persist. First, most studies optimize for overall accuracy rather than balanced performance across imbalanced classes, risking poor detection of rare malignancies. Second, domain adaptation strategies that leverage acquisition metadata remain underutilized in hybrid architectures. Third, comprehensive explainability frameworks combining multiple XAI techniques are rarely integrated into hybrid models, limiting clinical trust and adoption. Our work addresses these gaps by proposing a hybrid CNN-ViT architecture with three key innovations:

1) dual-strategy imbalance mitigation through weighted sampling and loss functions, optimizing for Macro F1 Score
2) domain-aware feature integration leveraging acquisition metadata
3) comprehensive explainability analysis revealing both feature importance and internal model conflicts. This approach aims to achieve not only competitive classification performance but also the diagnostic transparency essential for clinical deployment.

III. Methodology

A. Dataset and Preprocessing

This study utilized the ISIC 2018: Skin Lesion Analysis Toward Melanoma Detection Challenge dataset, comprising 10,015 dermatoscopic images distributed across seven distinct skin lesion categories, including melanoma (MEL), melanocytic nevus (NV), and basal cell carcinoma (BCC). A critical challenge inherent to this dataset is the pronounced class imbalance, particularly the overrepresentation of melanocytic nevus relative to malignant and rare lesion types.

To address this imbalance, a dual-strategy approach was implemented. First, inverse frequency weighting was computed for each class based on sample distribution. These weights were subsequently employed in two complementary ways: through a weighted random sampler to increase minority class representation during training iterations, and via weighted cross-entropy loss to impose heavier penalties on misclassifications of underrepresented classes. Additionally, domain labels extracted from metadata—specifically center identification and acquisition source—were encoded and integrated into the model architecture to enhance generalization across diverse clinical settings. All images underwent standardization to 224×224 pixel resolution. The training dataset was augmented through random rotations, horizontal and vertical flips, color jittering, and normalization using ImageNet statistics to improve model robustness and prevent overfitting.

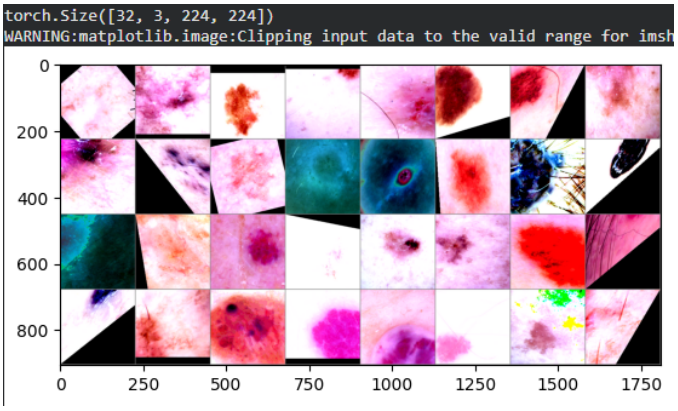

Fig. 1. Batch of 32 images

B. Hybrid CNN-ViT Architecture Design

The proposed architecture combines convolutional neural networks with vision transformers, drawing inspiration from adapted vision transformer principles tailored for small medical datasets. Rather than employing standard patch embeddings, the model utilizes a ResNet18 backbone pretrained on ImageNet as the initial feature extraction component. This convolutional frontend captures fine-grained textural and structural information from high-resolution dermoscopic images, generating feature maps that are subsequently flattened and linearly projected into token sequences for transformer processing.

To accommodate the limited dataset size and diverse acquisition conditions, a domain adaptation mechanism was integrated into the vision transformer component. This module accepts one-hot encoded domain labels representing image acquisition sources, which are incorporated into the self-attention computation. This design enables the model to learn source-specific biases while maintaining generalization capability across different clinical environments. The transformer architecture was configured with eight layers to balance representational capacity against overfitting risk given the constrained training data.

C. Training Protocol and Class Imbalance Management

Model training was conducted over 40 epochs using the Skorch framework, which provides a scikit-learn compatible interface for PyTorch models. The dual-strategy imbalance mitigation approach was implemented throughout training: the data loader incorporated a weighted random sampler using inverse class frequencies to ensure proportional representation of minority classes in each training batch, while the loss function applied class-specific weights to penalize minority class errors more severely than majority class mistakes.

Given the imbalanced nature of the dataset, the primary evaluation metric selected was the macro F1 score computed on the validation set. Unlike accuracy or weighted metrics, macro F1 calculates the F1 score independently for each of the seven classes and averages them without weighting, thereby ensuring that strong performance on abundant classes cannot obscure poor performance on clinically critical rare classes such as melanoma. Model checkpoints were saved based on validation macro F1 score improvements.

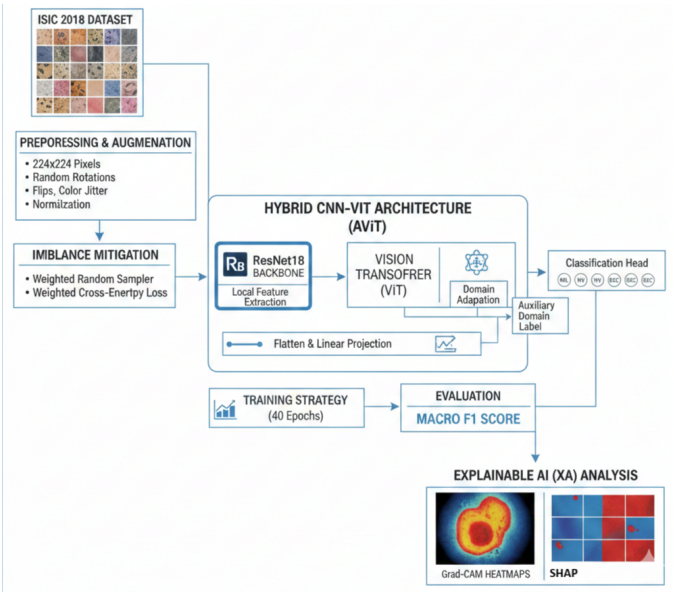

Fig. 2. Research Methodology

D. Explainability Analysis Framework

To ensure clinical interpretability and trustworthiness, two complementary post-hoc explanation techniques were applied to the trained model. Gradient-weighted Class Activation Mapping (Grad-CAM) was implemented by targeting the final attention block output, generating spatial heatmaps that highlight image regions most influential to classification decisions. These visualizations reveal whether the model focuses on clinically relevant features such as lesion borders, asymmetry, or pigmentation patterns.

Additionally, a patch-level attribution analysis based on Shapley value concepts was performed to quantify the contribution of individual image patches to the predicted class probability. This approach decomposes the model's decision into token-level contributions, identifying patches that provide strong positive evidence for the predicted class as well as patches that generate conflicting signals by suggesting alternative diagnoses. This granular analysis reveals the internal reasoning process and demonstrates the model's ability to reconcile competing visual features when making diagnostic predictions.

RESULTS

The Hybrid CNN–Vision Transformer (ViT) model, integrating the ResNet18 backbone with a Domain Adaptation Block, was evaluated on the seven-class ISIC 2018 dataset. The primary objective of the training strategy was to achieve balanced classification performance across all lesion types, prioritizing the detection of rare malignancies over mere overall accuracy.

E. Quantitative Performance Analysis

Due to the profound class imbalance, performance assessment was critically anchored on the Validation Macro F_1 Score, which ensures that high performance on majority classes (e.g., Nevus) does not obscure poor recall on crucial minority classes (e.g., Melanoma).

The training concluded after 40 epochs, with the model achieving optimal generalization performance on the validation set at Epoch 32. Table **??** provides a comparison of the Hybrid CNN-ViT architecture against a state-of-the-art pure CNN baseline, DenseNet121, both trained under identical conditions, including the dual imbalance mitigation strategy (Weighted Random Sampler and Weighted Cross-Entropy Loss).

The Hybrid CNN-ViT model achieved a best Validation Macro F_1 Score of 0.7751, confirming its robust ability to generalize across all seven classes. This strong F_1 performance is further complemented by a Validation Accuracy of 0.8098, which represents a substantial improvement over the DenseNet121 baseline, demonstrating the efficacy of combining local feature extraction (CNN) with global attention modeling (ViT) for small, imbalanced medical datasets.

Model Architecture	Val Macro	Val Accuracy	Train Loss
Hybrid CNN-ViT	**0.7751**	**0.8098**	0.0319
DenseNet121	N/A	0.5785	N/A

TABLE I

NOTE: THE MACRO F_1 SCORE PRIORITIZES BALANCED PERFORMANCE ACROSS ALL SEVEN CLASSES, MAKING IT THE DEFINITIVE METRIC FOR THIS IMBALANCED DATASET.

F. Qualitative Explainability Analysis (XAI)

To ensure the model's diagnostic decisions are clinically transparent and reliable, two post-hoc Explainable AI techniques were applied: Grad-CAM and SHAP-based attribution.

1) Grad-CAM Visualization: Gradient-weighted Class Activation Mapping (Grad-CAM) was applied to the final attention block of the Vision Transformer to localize the regions contributing most significantly to the final classification. As illustrated in Figure **??**, the resulting heatmaps consistently localized high-attention regions on clinically relevant structures, such as the lesion borders, areas of asymmetry, and complex pigment networks. This qualitative evidence validates that the model effectively synthesizes the local feature extraction (CNN) with the global contextual understanding (ViT) to focus on dermatologically important features.

2) SHAP-based Patch Attribution and Internal Conflict: A quantitative patch-level attribution analysis, conceptually related to SHAP (SHapley Additive exPlanations), was used to decompose the model's output and measure the contribution of each image token.

As shown in Figure **??**, patches with high positive contribution (e.g., deep blue/green scores) represent features that strongly confirm the predicted class (e.g., uniform texture for a benign lesion). Conversely, a significant and noteworthy finding was the presence of **internal model conflict** in complex or visually ambiguous cases. These cases were characterized by patches exhibiting high negative contribution (deep red scores). These negative scores indicate that the features within those specific image patches strongly pushed the model away from its final prediction, often signaling characteristics of other, potentially malignant, classes (such as irregular borders or specific vascular patterns).

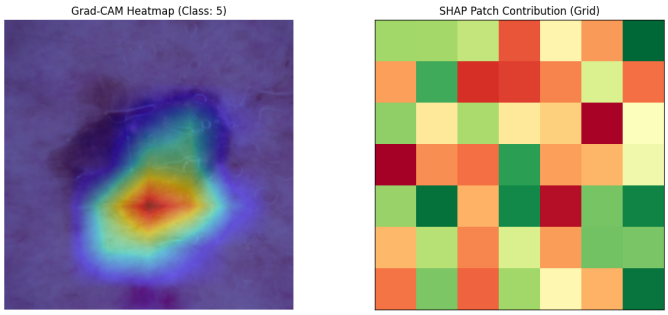

Fig. 3. GradCam: Visualizing model focus on lesion borders and features for different classes. SHAP:Visualizing positive (confirming) and negative (conflicting) feature contributions.

The model's final, correct classification in these complex instances was therefore a result of successfully resolving this internal conflict by assigning a higher overall weight to the features confirming the predicted class. This finding provides crucial evidence of the model's robustness and complexity, confirming that its decision-making is non-trivial and based on a detailed weighing of conflicting visual evidence, mirroring the diagnostic process of expert dermatologists.

REFERENCES

[1] G. H. Dagnaw, "Skin cancer classification using Vision Transformers and explainable artificial intelligence," Journal of Medical Artificial Intelligence, 2024.

[2] H. T. Halawani, "Enhanced early skin cancer detection through fusion of vision transformer and CNN features using hybrid attention of EViT-DenseNet169," Scientific Reports, 2025.

[3] G. M.-A. Himel, "Skin cancer segmentation and classification using Vision Transformer," International Journal of Biomedical Imaging, vol. 2024, 2024.

[4] M. A. Khan, "Identifying the role of Vision Transformer for skin cancer detection: A scoping review," Frontiers in Artificial Intelligence, vol. 6, p. 102345, 2023.

[5] M. M. Musthafa, "Enhanced skin cancer diagnosis using optimized CNN architecture and checkpoints for automated dermatological lesion classification," BMC Medical Imaging, vol. 24, no. 1, pp. 1–20, 2024.

[6] I. O. Pacal, "A novel CNN-ViT-based deep learning model for early skin cancer diagnosis," Biomedical Signal Processing and Control, vol. 104, p. 107627, 2025.

[7] H. C. Reis, "Fusion of transformer attention and CNN features for skin cancer detection," Applied Soft Computing, vol. 164, p. 112013, 2024.

[8] L. H. Xin, "An improved transformer network for skin cancer classification," Computers in Biology and Medicine, vol. 146, p. 105725, 2022.

[9] L. X. Yang, "A novel Vision Transformer model for skin cancer detection," Neural Processing Letters, 2023.

[10] X. L. Zhang, "AViT: Adapting Vision Transformers for Small Skin Lesion Segmentation Datasets," Bioengineering, vol. 12, no. 4, p. 421, 2025.