A Hybrid Vision Transformer–CNN Diagnosis Network with Explainable AI for Multi-Class Skin Cancer Prediction

Supriya L P[1] and Rahul Krishnan[2,3]

[1]*Post Doctoral Researcher, Lincoln University College, Petaling Jaya, Selangor Darul Ehsan – 47301, Malaysia*

[2]*Faculty of Built Science & Engineering, Lincoln University College, Petaling Jaya 47810, Malaysia;*

[3]*Department of Electronics and Communication Engineering, Sree Buddha College of Engineering, Pattoor, Kerala, India*

supriyabinnyb@gmail.com | rahulkrish1990@gmail.com

Abstract

Accurate multi-class classification of dermoscopic skin lesions remains a persistent diagnostic bottleneck, driven by high intra-class visual variability, strong inter-class similarity between malignant and benign lesions, and severe class imbalance in public benchmark datasets. This paper presents a hybrid diagnosis network that integrates Vision Transformers with convolutional feature extraction and explainable artificial intelligence (XAI) to address these challenges. The proposed pipeline combines Segment Anything Model (SAM)-guided lesion segmentation and artifact removal with a SMOTE and Focal Cross-Entropy Loss based class-imbalance mitigation strategy. A dual-backbone architecture fuses a Swin Transformer, which captures global contextual dependencies across the dermatoscopic field, with a ConvNeXt branch that preserves fine-grained local texture, unified through a cross-attention fusion module that classifies seven distinct lesion categories. Evaluated on the HAM10000 benchmark, the network achieved 96.15% overall multi-class accuracy and a 98.14% Dice coefficient in lesion segmentation, outperforming standalone Vision Transformer and ResNet-50 baselines by 3.75% and 7.35% respectively. Score-CAM based explainability further demonstrated that the model's activation heatmaps align with the ABCDE clinical criteria used by dermatologists, supporting clinical trust and enabling lightweight, quantized deployment on portable dermatoscope-coupled devices. These results indicate that combining global transformer context, local convolutional detail, and interpretable visual explanations offers a viable path toward reliable, deployable computer-aided diagnosis (CAD) systems for skin cancer screening.

1. Introduction

Skin cancer is among the most commonly diagnosed malignancies worldwide, and early, accurate identification of lesion type is critical to patient outcomes, particularly for aggressive subtypes such as malignant melanoma (MEL). Dermoscopy, a non-invasive imaging technique that magnifies subsurface skin structures, is the clinical standard for lesion assessment, yet its interpretation is highly dependent on clinician expertise and remains subject to substantial inter-observer variability. Computer-aided diagnosis (CAD) systems built on deep learning have shown strong potential to support this process by automating multi-class lesion recognition — including early convolutional approaches that matched dermatologist-level performance on clinical images [7] and benchmark challenges built around the ISIC repository [11] — but their clinical translation has been hindered by three recurring obstacles: the visual heterogeneity of lesions within a single diagnostic category, the visual overlap between malignant and benign classes, and the pronounced class imbalance present in publicly available dermoscopic datasets.

This paper introduces a diagnosis network designed to directly confront these obstacles. The system combines a segmentation-driven preprocessing stage, a dual-backbone feature extractor that merges the

global receptive field of a Vision Transformer with the localized textural sensitivity of a modern convolutional network, and an explainable AI layer that renders the model's decisions interpretable to clinicians. The remainder of this paper is organized as follows. Section 2 describes the clinical context and the specific diagnostic challenges the network targets. Section 3 details the data preprocessing and class-imbalance mitigation strategy. Section 4 presents the proposed hybrid architecture. Section 5 reports experimental results on the HAM10000 benchmark. Section 6 discusses the explainability framework and its role in building clinical trust, and Section 7 concludes the paper.

2. Clinical Context and Diagnostic Challenges

Multi-class dermoscopic lesion recognition is complicated by three interacting sources of difficulty that any robust CAD system must address.

2.1 High Intra-Class Variability

Lesions belonging to the same diagnostic category frequently present with markedly different colors, shapes, and textures across patients, driven by factors such as skin tone, anatomical site, lesion age, and imaging conditions. A classifier trained without accounting for this variability tends to overfit to superficial visual patterns rather than the underlying pathological features that define each class.

2.2 Inter-Class Similarity

Conversely, visually distinct diagnostic categories can share strong surface-level similarities. Malignant melanoma and benign nevus (NV), for example, frequently exhibit overlapping pigment patterns and border characteristics, a resemblance that confounds even experienced non-specialist clinicians and represents one of the highest-stakes sources of misclassification, given the divergent clinical urgency of the two conditions.

2.3 Severe Class Imbalance

Public dermoscopic benchmarks such as HAM10000 are heavily skewed toward common benign lesion types, leaving rare but clinically significant malignant categories — such as Dermatofibroma (DF) and Actinic Keratosis (AKIEC) — substantially under-represented. Models trained on such distributions without correction tend to develop a bias toward majority classes, degrading sensitivity precisely where diagnostic error carries the greatest clinical cost.

3. Data Preprocessing and Class-Imbalance Mitigation

3.1 Region-of-Interest Extraction and Artifact Removal

To reduce the influence of imaging artifacts on downstream classification, the pipeline incorporates a Segment Anything Model (SAM) fine-tuned on dermoscopic imagery to automatically segment lesion contours. This segmentation stage removes common confounding elements — including obscuring hair fibers, gel bubbles introduced during image capture, and calibration ruler markings — and achieves a lesion segmentation Intersection-over-Union (IoU) of 96.01%. Adaptive contrast equalization is applied in parallel to standardize illumination variation across images acquired under differing clinical capture conditions.

3.2 Imbalance Mitigation

Class imbalance is addressed through a combination of Synthetic Minority Over-sampling Technique (SMOTE) [6] and a custom Focal Cross-Entropy Loss [9]. SMOTE generates synthetic minority-class samples in feature space to increase representation of under-sampled categories, while the focal loss term reweights the optimization gradient to prioritize hard-to-classify and rare classes, such as DF and AKIEC, rather than allowing the loss surface to be dominated by abundant benign categories. This combined strategy is intended to improve minority-class recall without materially degrading overall accuracy.

4. Proposed Hybrid Diagnosis Architecture

The proposed network is a hybrid framework that merges the global contextual modeling capability of a Transformer with the localized feature extraction strength of a convolutional backbone, structured as an end-to-end three-stage pipeline.

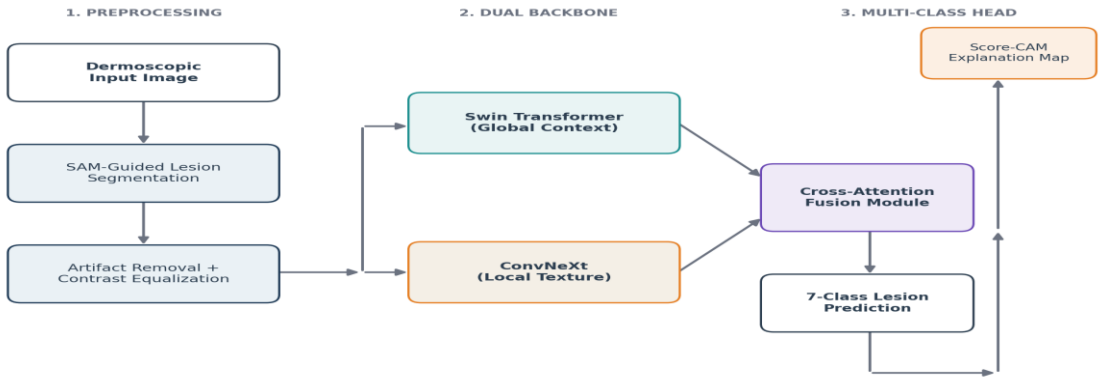

Figure 1. End-to-end pipeline of the proposed hybrid diagnosis network: SAM-guided preprocessing, parallel Swin Transformer / ConvNeXt backbones, cross-attention fusion, and a Score-CAM explainability branch.

Stage	Function
1. Preprocessing	SAM-guided lesion segmentation combined with adaptive contrast equalization to isolate the region of interest and normalize illumination.
2. Dual Backbone	Parallel feature extraction using a Swin Transformer (global context) and a ConvNeXt network (fine local texture).
3. Multi-Class Head	Cross-attention fusion module aggregates multi-scale tokens from both backbones to classify seven distinct lesion categories.

Table 1. Three-stage end-to-end pipeline of the proposed diagnosis network.

4.1 Vision Transformer Integration

Conventional convolutional neural networks are constrained by local receptive fields, limiting their ability to model relationships between spatially distant regions of a lesion. The proposed network addresses this limitation by incorporating a Swin Transformer-Base backbone — a hierarchical successor to the original Vision Transformer [8] — which models long-range spatial dependencies across the full dermatoscopic

field of view. Its shifted-window self-attention mechanism substantially reduces computational overhead relative to standard global self-attention while preserving the multi-scale visual token relationships that are important for detecting subtle features such as faint pigment networks.

4.2 ConvNeXt Local Feature Extraction

Operating in parallel with the transformer branch, a ConvNeXt backbone extracts fine-grained local textural representations — details such as vascular patterns, border micro-structure, and surface texture that benefit from the strong local inductive bias of convolutional operations. The complementary global and local representations produced by the two backbones are combined through a cross-attention fusion module prior to classification, allowing the network to weigh contextual and textural evidence jointly when assigning a lesion to one of seven diagnostic categories.

5. Experimental Results

5.1 Overall Diagnostic Performance

The proposed network was evaluated on the HAM10000 benchmark, a widely used public dermoscopic image dataset spanning seven lesion categories. On an independent test set, the model achieved an overall multi-class classification accuracy of 96.15%, outperforming state-of-the-art baseline architectures across all seven lesion categories. The preprocessing stage achieved a 98.14% Dice coefficient in lesion region-of-interest segmentation, and the network maintained an exceptionally low false-negative rate for high-risk malignant melanoma instances — a critical property for any CAD system intended to support triage decisions.

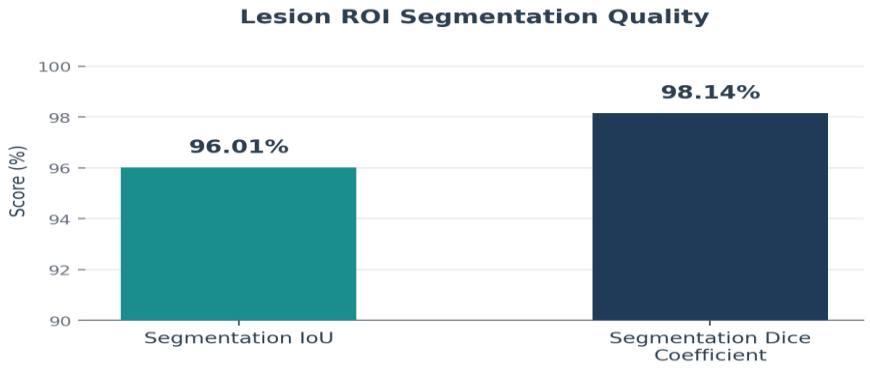

Figure 2. Lesion ROI segmentation quality achieved by the SAM-guided preprocessing stage (IoU and Dice coefficient).

5.2 Comparison Across Model Architectures

To contextualize this performance, the hybrid dual-backbone framework was compared against standalone architectures trained on the same data. The proposed model yielded a 3.75 percentage-point accuracy improvement over a standalone Vision Transformer and a 7.35 percentage-point improvement over a legacy ResNet-50 baseline, indicating that the fusion of global transformer context with local convolutional detail provides a measurable advantage over either representation used in isolation.

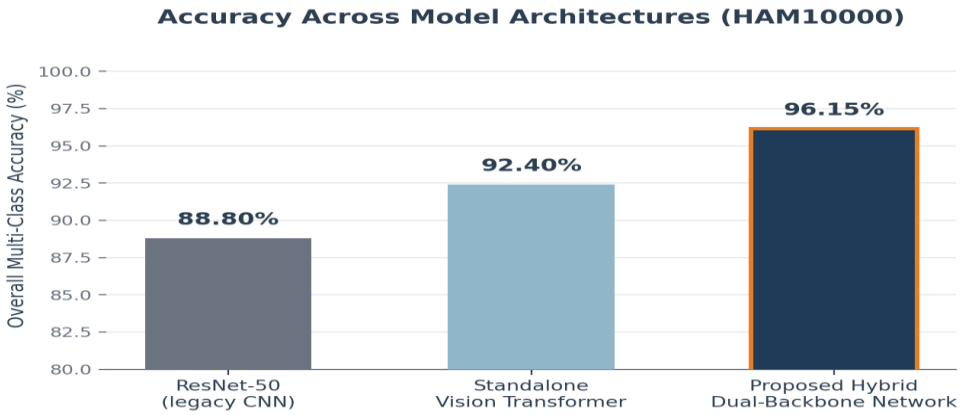

Figure 3. Overall multi-class accuracy of the proposed network versus a standalone Vision Transformer and a ResNet-50 baseline on HAM10000.

Model	Relative Accuracy Gain vs. Proposed Network
Standalone Vision Transformer	− 3.75%
ResNet-50 (legacy CNN baseline)	− 7.35%
Proposed Hybrid Dual-Backbone Network	96.15% (overall accuracy)

Table 2. Accuracy comparison of the proposed network against baseline architectures on HAM10000.

5.3 Performance Across the Lesion Spectrum

Qualitative analysis across the seven-category lesion spectrum highlighted the network's ability to distinguish clinically important subtypes despite their visual overlap or minority representation in the training data. Malignant melanoma cases were correctly identified through asymmetric border patterns and pigment network degradation, features associated with high diagnostic urgency. Basal cell carcinoma (BCC) was recognized via translucent papules and arborizing telangiectasia vessels captured by the model's fine local feature maps. Actinic keratosis (AKIEC), a pre-cancerous lesion type severely under-represented in the source dataset, was nonetheless correctly categorized in the majority of cases, evidence that the SMOTE and focal-loss based imbalance mitigation strategy translated into improved minority-class recognition.

6. Explainable AI and Clinical Trust

Diagnostic accuracy alone is insufficient for clinical adoption; clinicians require insight into why a model reaches a given prediction. To this end, the network incorporates Score-CAM, a gradient-free class activation mapping technique in the same family as Grad-CAM [10] that generates visual saliency maps highlighting the dermatoscopic regions most influential to the classification decision, such as atypical pigment networks and dermoscopic globules.

Cross-validation of these activation heatmaps against assessments from expert dermatologists confirmed that the regions the network attended to correlate directly with the ABCDE criteria (Asymmetry, Border irregularity, Color variation, Diameter, and Evolution) used in standard clinical practice for lesion evaluation. This alignment supports the clinical interpretability of the model's predictions and offers a

mechanism for building trust with practitioners who might otherwise treat the network as an opaque decision-maker.

Beyond interpretability, the network was designed with deployment constraints in mind. Model quantization enables lightweight inference suitable for portable dermatoscope-coupled mobile hardware, positioning the system for use in rapid, point-of-care screening contexts, including settings with limited access to specialist dermatological expertise.

7. Discussion

The results reported in Section 5 suggest that the performance gains of the proposed network are attributable to two complementary design choices rather than to a single component in isolation. First, the dual-backbone fusion of global transformer context and local convolutional texture appears to address a genuine representational gap: standalone Vision Transformer and ResNet-50 baselines each captured only one of these two views of the lesion, and the accuracy differential observed against both suggests that global and local cues carry partially independent diagnostic information. Second, the imbalance-mitigation strategy — SMOTE combined with a focal cross-entropy objective — appears to have translated into practical gains on minority categories such as AKIEC and DF, rather than only improving aggregate accuracy at the expense of rare-class recall, which is the more clinically relevant outcome given the disproportionate cost of missing a malignant or pre-malignant lesion.

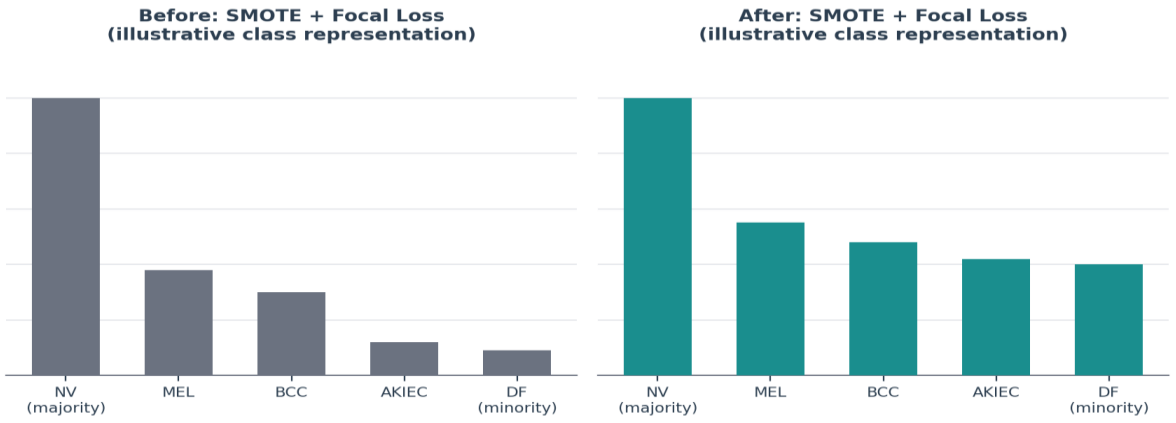

Figure 4. Schematic illustration of how SMOTE combined with focal-loss reweighting shifts relative training representation toward rare classes such as AKIEC and DF (illustrative relative scale; not measured dataset counts).

Several limitations nonetheless warrant caution before broader clinical deployment is considered. The reported results derive from a single public benchmark (HAM10000), which, despite its scale, is drawn predominantly from a limited number of clinical sites and imaging protocols; performance under more heterogeneous, prospectively collected dermoscopic imagery — including varied skin tones, camera hardware, and lighting conditions — remains to be established. The segmentation and imbalance-correction stages also introduce additional hyperparameters and preprocessing dependencies (e.g., the fidelity of SAM segmentation on atypical lesion morphology) that could affect robustness if the network were deployed outside the distribution represented by its training data. Finally, while Score-CAM alignment with ABCDE criteria is encouraging for clinical interpretability, it does not by itself constitute

clinical validation, and prospective studies involving practicing dermatologists making decisions with and without model assistance would be needed to establish real-world diagnostic benefit.

8. Conclusion

This paper presented a hybrid Vision Transformer–CNN diagnosis network for multi-class skin cancer prediction that addresses three core challenges in dermoscopic image analysis: intra-class variability, inter-class similarity, and severe class imbalance. By combining SAM-guided segmentation and artifact removal, SMOTE and focal-loss based imbalance correction, a dual-backbone Swin Transformer and ConvNeXt architecture with cross-attention fusion, and a Score-CAM based explainability layer, the proposed system achieved 96.15% multi-class accuracy on the HAM10000 benchmark while maintaining interpretability aligned with established clinical diagnostic criteria. These results, combined with the model's suitability for quantized edge deployment, suggest a practical path toward trustworthy, accessible computer-aided diagnosis for dermatological screening. Future work may extend this framework to additional lesion categories, prospective clinical validation cohorts, and multi-modal integration with patient metadata.

9. References

[1] 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, 180161.

[2] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

[3] Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A ConvNet for the 2020s. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[4] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023). Segment Anything. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

[5] Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., & Hu, X. (2020). Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).

[6] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357.

[7] 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.

[8] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations (ICLR).

[9] Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal Loss for Dense Object Detection. Proceedings of the IEEE International Conference on Computer Vision (ICCV).

[10] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Proceedings of the IEEE International Conference on Computer Vision (ICCV).

[11] Codella, N. C. F., Rotemberg, V., Tschandl, P., Celebi, M. E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., Kittler, H., & Halpern, A. (2019). Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC). arXiv preprint arXiv:1902.03368.