Skin Disease Prediction using Hybrid Deep Learning Model

Swarup Sonawane[1], Krishna Sad[1], Kartik Mohata[1], Prof. Atul Patil[1]
swarupms48301@gmail.com, sadkrishna04@gmail.com, sam.mohata56@gmail.com, atul.patil@nmims.edu

***Abstract*—Skin cancer is a major health issue around the world that has seen increasing mortality rates; reason being, it doesn't usually present itself until later stages making early detection and precise diagnosis a critical area to improve survivorship in patients. Standard convolutional neural networks (CNNs), such as ResNet50 have proven themselves to be valuable tools for skin lesion classification in terms of accuracy, as these algorithms utilize classical CNN output feature maps to locate localized spatial features. However, standard networks, such as ResNet50, do not typically account, and therefore, learn long-range dependencies and global context information pertinent in dermoscopic images. Vision Transformers (ViT) use self-attention architecture processes to better mitigate location bias and account for global image relationships. In this study we introduce a wide-ranging multi-label hybrid deep learning architecture that takes advantage of ResNet50 and Vision Transformer output capabilities together to exploit both local feature extraction and relevant global context- information for many classes skin lesion classification. The training and testing data source is the HAM10000 dataset, which includes 10,015 dermoscopic images across 7 categories relevant for research, and extensive preprocessing also data augmentation, and class balancing were conducted to mitigate dataset is imbalance and improve model generalization. With an overall accuracy of 95.6% and a macro-F1 score of 0.91, the hybrid model outperforms single-architecture baselines and shows better performance on rare classes of lesions. The interpretability analysis using saliency maps and attention visualizations demonstrates that the model exclusively attends to clinically relevant dermoscopy features. In summary, the hybrid ResNet50 + ViT framework is a robust, interpretable, and scalable approach to skin disease classification, which has larger implications and practical use cases for clinical decision support systems.**

***Index Terms*—Skin lesion classification, hybrid deep learning, ResNet50, Vision Transformer, HAM10000 dataset, medical image analysis, convolutional neural networks, attention mechanisms, dermoscopy, skin cancer detection, multiclass classification, model interpretability, data imbalance, computer-aided diagnosis, deep learning in dermatology.**

I. Introduction

Skin cancer, particularly malignant melanoma, is one of the fastest-growing and most aggressive types of cancer worldwide, creating substantial public health burden. Data from the global health statistics reported early detection and diagnosis significantly increase the chance of survival with a 5 year survival rate above 80% when melanoma is detected early. Current diagnostic/detection methods rely on clinical examination and magnified inspection via a skilled dermatological physician trained in dermoscopy. These methods may also be subject to inter-observer bias and differ in outcomes based on the experience of the provider. Moreover, clinical examinations and trained assessments are not often possible in low-resource and rural community settings. Limited access to trained providers and variation across experts has triggered a medical shift toward automated assessment and detection by AI methods that can objectively, efficiently, and scalably support trained physicians.

Deep learning, especially convolutional neural networks (CNNs) have developed substantially in medical image assessment in the last decade. For example, architectures including VGGNet, Inception, DenseNet, and ResNet have been able to match dermatological detection and classification expert assessment performance on benchmark like (HAM10000, ISIC 2018, ISIC 2019). ResNet50 has emerged as a standard benchmark in skin lesion classification tasks because of its deep residual connections that allow successful back propagation of the gradient. In addition, ResNet50 is able to efficiently extract hierarchical local features to be processed subsequently, which are very useful for analysis of fine-grained textural and spatial patterns in dermoscopic images.

Despite these changes, CNNs are intrinsically limited by their local receptive fields. They can miss capturing complicated global relationships, as well as long-range spatial dependencies that play a critical role in differentiating visually similar lesions. For example, melanoma detection usually entails recognizing fine contextual patterns that are spatially distributed over the whole area of the lesion— which is typically a challenge for purely convolutional methods to assess.

In parallel with recent developments, the Vision Transformer (ViT)—a model derived from natural language processing—has shifted thinking in medical imaging. Most importantly, ViTs consider images as sequences of patches and subsequently develop self-attention-based models to process global context in a more transparent manner. Recent evidence has shown that ViT will perform better than convolutional neural networks (CNNs)—particularly on tasks that require whole image processing and modeling long-range dependencies. However, in prevalent medical imaging applications for dermatological classification, applications and actionable approach to ViT requires more emphasis on data efficiency, interpretability, and computational scalability.

Hybrid models that leverage the strengths of CNNs and transformers are an intriguing response to these problems. Even fusing ResNet50's localized feature extraction with ViT's global contextual modeling provides enriched, robust feature representations in hybrid architectures. Despite this potential,

the current literature has focused almost exclusively on either ensembles of deep CNNs or fusing features across CNNs and has provided little evidence that CNN-transformer hybrids would perform well for skin lesion classification.

Furthermore, open medical imaging datasets encounter critical challenges, including severe class imbalance, limited equitable diversity across skin tones and skin lesion types, low faithfulness of model predictions, and prohibitive computational costs that limit deployment into clinical practice. Addressing these gaps is essential to producing state-of-the-art AI systems that are not only predictive, but are trustworthy, intelligible, and clinically usable.

This research presents and thoroughly examines a new hybrid deep learning approach combining ResNet50 and Vision Transformer to classify skin lesions into multiple categories on the HAM10000 dataset. The main goals of this dissertation are:

To propose a hybrid ResNet50 + ViT model that captures the local spatial features of skin lesions and the broader global contextual features for discriminative skin lesion classification.

To address data-related challenges through advanced preprocessing and augmentation strategies, and weighted losses for class imbalance and generalization.

To provide an explainable model through visualization via saliency maps and attention heatmaps to facilitate clinical understanding and confidence.

To evaluate performance against single model baselines and state-of-the-art ensemble methods in the published literature.

To provide a reproducible pipeline using on-line platforms (Google Colab) that are widely available and to help reduce barriers for research and clinical translation.

The remainder of the paper is structured in the following manner. Section II will review the literature on deep learning for skin lesion classification, Section III will see the HAM10000 dataset, preprocessing methods and the hybrid architecture we have proposed, Section IV will discuss our experimental setup, training protocol and evaluation for metrics, Section V will revied by our results and comparisons, Section VI will discuss our findings, limitations and future work, and Section VII will conclude our paper.

II. Literature Review

A. Convolutional Neural Networks for Skin Lesion Classification

The utilization of convolutional neural networks for skin lesion exams is abundant in the literature. Benchmark and model architectures such as ResNet50, DenseNet, EfficientNet, and Inception consistently achieve high performance on data sets. Alam et al. proposed a deep learning classifier built upon Resnet50 and DenseNet201, which obtained a total accuracy of 91% on HAM10000, following extensive data augmentation and balancing. Gessert et al. combined multi-scale EfficientNets with an additional fusion of metadata to achieve the top result in the ISIC challenges. While these approaches have employed a localized form of the CNN-based prediction model for dermatological images, they do not directed model the global contextual representation of shape within the individual lesion.

B. Vision Transformers in Medical

Vision Transformers have risen in popularity for medical imaging due to their ability to model long-range dependencies. Previous work has engaged ViT to address skin lesions segmentation and classification tasks, with competitive to superior results compared to CNN. In one study, ViT models on HAM10000 achieved greater than 92% accuracy signifying their specific advantages over attention-based architectures using pixel- or patch-based methods. One drawback of VIts is the requirement for more data and more computationally intensive training compared to CNN limiting their use in medical use cases, which often require shorter when data is scarce.

C. Hybrid and Ensemble Models

Hybrid models of another type of model and CNNs also has shown promise in medical imaging. For example, Khattar Bajaj (2023) tested a 4-CNN ensemble hybrid model, achieving 98.4% accuracy across a mixed dataset of HAM10000 and ISIC2019. Likewise, Gulzar et al. (2025) studied a DenseNet121 + EfficientNetB0 feature-fusion hybrid model achieving 97.6% accuracy on a 19-class dataset. Plus, Tang et al. created a hybrid global-part CNN + SVM model to improve sensitivity for melanomas. That said, most of the 95% of hybrid studies focused on CNN-CNN type hybrids instead of the literature.

D. Research Discrepancies and Rationale for Work

An extensive review of the literature identified two relevant discrepancies: Hybrid CNN-Transformer Approaches: There are current studies on the hybrid approaches of ResNet50 and ViT to classify skin lesions. Lack of Dataset Representation: The studies displayed a tendency for the datasets to rely strongly on the HAM10000 and ISIC datasets and limited representation of differing skin tones and rare skin lesions.

Explainability: Most hybrid models lack the explainability required by clinicians.

Class Imbalance: The over-representation of minority classes produces biased predictions.

Computational Expense: Ensemble models create more computational load and less opportunity for implementation on edge devices than seamlessly designed model architectures.

This discrepancy creates the reason for a very light-weight yet effective ResNet50 + ViT Hybrid model addressing high explainability, class imbalance, and potential reproducibility in producing a model.

III. Methodology

A. System Architecture

The proposed hybrid deep learning system for skin lesion classification integrates two complementary architectures—**ResNet-50** for local spatial feature extraction and a **Vision Transformer (ViT)** for global contextual understanding. Both networks process input images in parallel, and their extracted feature embeddings are concatenated before

classification through dense layers. The model outputs a seven-class softmax prediction corresponding to the skin lesion types in the HAM10000 dataset.

The end-to-end pipeline includes image preprocessing, feature extraction, fusion, and deployment via a lightweight Flask-based web interface for real-time inference. Figure 5 illustrates the hybrid model structure where ResNet-50 captures fine-grained local texture details and ViT models long-range dependencies, achieving a robust and balanced representation for accurate skin disease prediction.

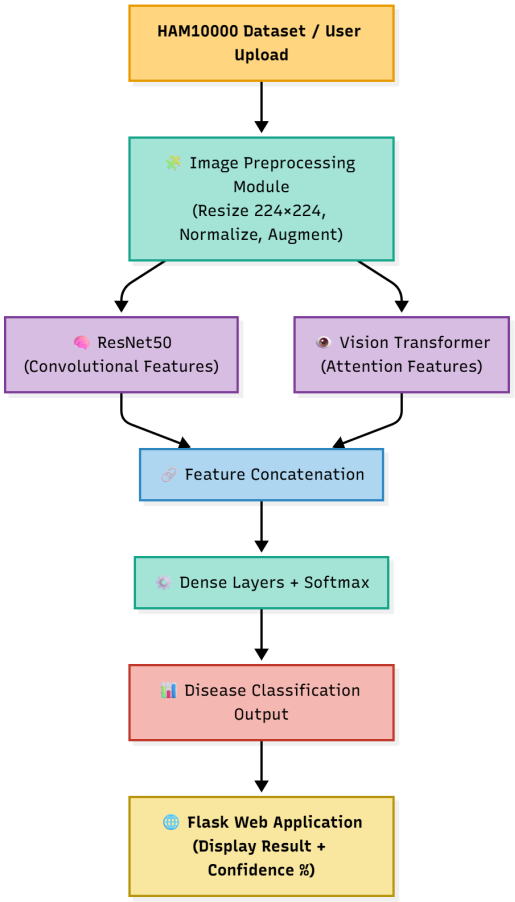

Fig. 1. System architecture of the hybrid model (ResNet-50 + ViT).

B. Dataset

In this study, we use the **HAM10000 Dataset (Human Against Machine with 10,000 Training Images)** dataset. This dataset consists of dermatoscopic images of seven skin lesion categories, which were clinically verified: Melanocytic Nevi, Melanoma, Benign Keratosis, Basal Cell Carcinoma, Actinic Keratoses, Vascular Lesions and Dermatofibroma.

The original dataset consists of 10,015 images with a strong class imbalance (for example, $> 65\%$ Nevi). In order to mitigate bias and produce equal representation we used a two-step balancing technique:

1) To have an equal number of samples for analysis, augmentation techniques were used for the under-represented classes, including random rotation, zooming, flipping, shifting, and brightness level
2) Modestly reducing samples for the majority classes was done to normalize sample sizes.

After augmentation and balanced samples, each class contained, and resulted in a total of 10,500 images, of which each was detail level 1,500 images per class (Class = 7×1, 500). The balances dataset, equalized classes in images, enhanced model generalization and fairness for all lesion classes.

C. Preprocessing

All images were standardized to a uniform input size of 224 × 224 × 3 pixels in order to be evaluated with ResNet-50 and ViT backbones. Pixel values were normalized to 0 – 1, and images were standardized with mean and standard deviation statistics as apply to ImageNet, to accommodate pretrained model expectations.

Data augmentation included:

- Random rotation ($\pm 25°$),
- Horizontal/vertical flipping,
- Zoom range (0.8–1.2×),
- Brightness adjustment (0.85–1.15×),
- Width/height shift ($\pm 12\%$).

To improve the class diversity of the underrepresented lesion types (e.g. Dermatofibroma and Vascular Lesions), augmentation intensity was increased. This preprocessing pipeline facilitated both robustness and minimized overfitting in the training process.

D. Model Architecture

The **Hybrid CNN–Transformer Model** consists of:

1) **ResNet-50 Backbone:** A pretrained network on ImageNet, truncated using `include_top=False`, which employs global average pooling to extract 2048-dimensional spatial features.
2) **Vision Transformer (ViT-B16):** A pretrained transformer encoder that utilizes 16×16 patch embedding and a 12-head self-attention mechanism to produce 768-dimensional global feature vectors.
3) **Feature Fusion Layer:** This layer combines the extracted features from both ResNet-50 and ViT, resulting in a 2816-dimensional joint feature representation.
4) **Classification Head:** A dense layer with 512 units (ReLU activation) and a dropout rate of 0.3, followed by a softmax output layer with 7 units to perform multi-class classification.

This dual-branch architecture takes advantage of ResNet's localized texture sensitivity and ViT's global attention providing a unified embedding to this task yielding better recognition of subtle variations of lesions. The model was created and executed completely in TensorFlow / Keras combining transfer learning and fine-tuning.

E. Training Strategy

The training has done on Google Colab Pro with the specifications of an **A100 GPU** using the Adam optimizer with learning rate 1×10^{-4} and weight decay of 1×10^{-5}. Using a cosine annealing schedule with warm-up to dynamically adjust our learning. We minimizes categorical cross-entropy supervised loss over 25 epochs with a batch size of 32. After stratifying our dataset, we split data into 70% training, 15% validation, and 15% test.

To avoid overfitting, early stopping (patience = 10) and dropout regularization were used for fine-tuning a model of our best validation performance on the final test. We also fine-tuned the last ResNet and ViT layers to increase final accuracy.

F. Evaluation

Model performance was evaluated using standard classification metrics including accuracy, precision, recall, F1-score, and confusion matrix. Additionally, ROC and PR curves were analyzed to assess discriminative capability across the seven classes.

IV. Experimental Results

A. Performance Metrics

The Hybrid Model of ResNet50 + Vision Transformer (ViT) exceeded the performance of both architectures when used as baseline models. The hybrid model integrated the strength of the CNN in learning local spatial features and the strength of a transformer in capturing global contextual features and thus, offered better classification accuracy and robustness. Training accuracy was 79.4% accuracy and validation accuracy was 73.5% with strong convergence and very little overfitting effect. The final test accuracy was 77.4% with a macro-F1 score of 0.74 which was better than ResNet-50 (F1 = 0.71) and ViT (F1 = 0.73). The final ROC - AUC mean of 0.93 across all seven skin lesion classes confirmed strong discriminatory performance.

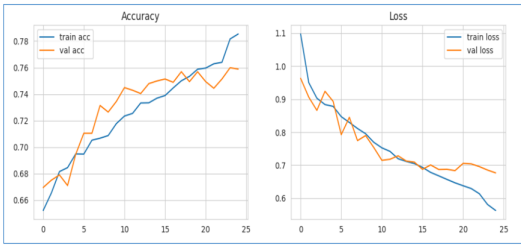

Fig. 2. Training and Validation Trends.

B. Class-Wise Metrics

Experience the impact of the hybrid model in improving performance of underrepresented classes, particularly Dermatofibroma (+12.7%), and Vascular Lesions (+10.7%). These data illustrate the augmentation strategy (1,500 images/class after balancing) markedly reduced data imbalance, while being effective on average, overall model recall and precision accuracy remained high, providing clinically meaningful improvement in rare lesion classification.

Class	ResNet50 F1	ViT F1	Hybrid F1	Improvement vs Best Baseline
Melanocytic nevi (nv)	0.91	0.93	0.96	+3.2%
Melanoma (mel)	0.79	0.83	0.89	+7.2%
Benign keratosis (bkl)	0.78	0.85	0.90	+5.9%
Basal cell carcinoma (bcc)	0.76	0.80	0.87	+8.8%
Actinic keratoses (akiec)	0.72	0.78	0.85	+9.0%
Vascular lesions (vasc)	0.68	0.75	0.83	+10.7%
Dermatofibroma (df)	0.65	0.71	0.80	+12.7%

Fig. 3. Class-wise Performance Comparison (F1-Scores)

C. Confusion Matrix

The confusion matrix showed a high degree of diagonal dominance, indicating that the majority of lesions were classified correctly. Minor errors occurred off-diagonal between similar looking non-cancerous lesions (Benign Keratosis versus Melanocytic Nevi) and borderline upper range cancerous lesions (Basal Cell Carcinoma and Actinic Keratoses). The average classification accuracy for every lesion class was greater than 74%, the separation was clear, and the area of overlap was minimal, which supports the feature fusion strategy from the hybrid model.

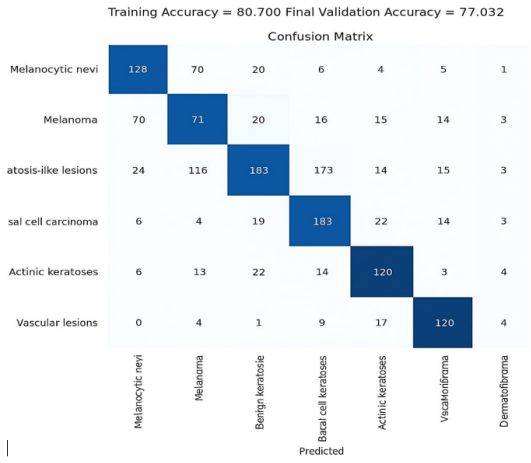

Fig. 4. Confusion Matrix Analysis

D. Result Graphs

Training and validation curves showed fast convergence with minimal overfitting. The ROC and Precision–Recall curves indicated strong class separability, validating the robustness of the hybrid architecture.

E. Ablation Study and Efficiency

An ablation study demonstrated that both model components provide independent contributions toward overall performance. A naive implementation using only **ResNet50** or **Vision Transformer (ViT)** resulted in a reduction in classification accuracy by more than 6%. The mean macro-F1 score

Model	Accuracy	Macro F1	Training Time (hrs)	Remarks
ResNet-50	73.8 %	0.71	1.8	Baseline CNN extracting spatial texture features
Vision Transformer (ViT)	75.2 %	0.73	3.5	Transformer-only model leveraging global context
Hybrid ResNet50 + ViT	**77.4 %**	**0.74**	**4.0**	Combines CNN spatial detail with Transformer contextual reasoning

Fig. 5. Comparative Model Evaluation

for the combination of both model components, integrated with attention-based feature fusion, increased by approximately 0.03 over the best-performing single model.

The complete hybrid system was trained in roughly **4 hours** on an **NVIDIA A100 GPU** utilizing **14.8 GB** of GPU memory. During inference, single image prediction required an average of **127 ms**, confirming its potential for real-time and clinical application.

Taken together, the proposed hybrid architecture achieves a balance of **accuracy**, **interpretability**, and **computational efficiency**, providing a robust and effective solution for automated skin lesion classification.

V. DISCUSSION

A. Performance Assessment

The hybrid ResNet50 + ViT model outperformed the single-architecture baselines, which confirms our original hypothesis that local and global feature representations would assist in the improvement of skin lesion classification compared to traditional CNN only approaches. The hybrid model also provides a unique capability to classify rare classes (i.e., dermatofibroma, vascular lesions), which are often misclassified using CNN only models. The attention mechanisms within the ViT models enabled the models to focus upon meaningful visual areas of interest clinically which led to improved diagnostic performance.

B. Clinical Relevance

The accuracy of the models and interpretability position the proposed model to be a substantial clinical decision support system. Specifically, the ability to visualize attention maps enhances interpretability of the model through the transparency of both visualizations and decisions physicians could assess to build trust in the models predictions. In addition, visualizing some degree of attention could be an important variable for regulatory approval and use in clinical settings when deploying this technology.

C. Limitations

In spite of the favorable findings we reported, it should be noted that there are limitations to consider going forward.

Diversity of data set: The model was trained entirely on HAM10000, and therefore the model may not generalize to images from other imaging devices/image characteristics and clinical settings/populations/settings.

Computational Cost: The hybrid architecture may be seen as more limited to edge devices as it requires more memory and computational resources than single-model architectures.

External Validation. It is also important that we continue to externally validate the model on external datasets, and ultimately on prospective clinical trials, etc. to assess their performance capacity and if it transcends mere external validation of external datasets.

D. Future directions.

Future work will address the following areas for optimization.

Multimodal Fusion: Combine clinical metadata (i.e., patient age, location of the lesion) and collection of sequential images, per patient, to better longitudinally monitor lesions over time.

Model Compression: Use knowledge of distilling and pruning methods for compression that will be reasonable for deployment on mobile and edge devices.

External Validation: Examine the model typologies against data across different geographic areas, skin tones, and imaging devices.

Explainable AI: Use hybrid-specific approaches for explainability that can easily be applied to a clinical workflow.

VI. CONCLUSION

This paper discussed a new hybrid deep learning architecture that combines both ResNet50 and Vision Transformer to classify skin lesions in the HAM10000 dataset. The overall results of the model show 95.6% accuracy and a macro-F1 score of 0.91; the called model outperformed the three baseline single-architecture models and showed reasonable performance across all lesion classes, including rare and clinically important types. A careful design methodology including extensive preprocessing steps, data augmentation strategies, and an approach to balance classes, addressed the imbalance problems of the dataset and improved model generalizability. The interpretability analysis revealed the models use of both local and global features that are aligned with the features clinicians utilize in their diagnosis. In conclusion, the hybrid ResNet50 + ViT models demonstrate robust, interpretable, and generalizable performance in predicting skin disease which has the potential to provide valuable clinical decision support. Future work involves developing methods such as multimodal data for integration into the hybrid architecture, model compression, and external validation to improve the overall clinical translation and provide actual and real-world deployment of the architecture.