SCIENTIFIC CULTURE
www.sci-cult.com

SCIENTIFIC CULTURE, Vol. 12, No. 2.1, (2026), pp. 9272-9284
Open Access. Online & Print

DOI: 10.5281/zenodo.19949266

A HYBRID CNN–TRANSFORMER FRAMEWORK FOR AUTOMATED SKIN CANCER DETECTION FROM DERMOSCOPIC IMAGES

Hamza A. Mashagba[1], Suhaila Abuowaida[2], Azlan B. Abd Aziz[1]*, Nawaf Alshdaifat[3], Mahmoud Baniata[4], Mardeni Bin Roslee[5], Mohamad Yusoff Alias[5], Azwan Mahmud[5]

[1]Faculty of Engineering and Technology, Multimedia University, Melaka, Malaysia; Centre for Wireless Technology (CWT), Multimedia University
[2]Affiliation Department of Data Science and Artificial Intelligence, Faculty of Prince Al-Hussein Bin Abdallah II for IT, Al Al-Bayt University, Mafraq, 25113, Jordan. suhila@aabu.edu.jo.
[3]Department of Information Technology, Faculty of Prince Al-Hussein Bin Abdullah II for Information Technology, The Hashemite University, Zarqa, Jordan
[4]Department of Computer Science Faculty of Information Technology, Applied Science Private University, Amman, Jordan
[5]Faculty of Artificial Intelligence and Engineering BR4064, Level 3, Multimedia University, Persiaran Multimedia 63100, Cyberjaya, Selangor, Malaysia.

Received: 15/03/2026
Accepted: 18/04/2026

Corresponding Author: Azlan B. Abd Aziz (azlan.abdaziz@mmu.edu.my)

ABSTRACT

The early detection of melanoma and other forms of skin cancer is currently one of the most difficult challenges facing clinicians in the field of dermatology. The difficulty lies in the subtle differences in appearance among benign and malignant lesions. In this research we introduce a new type of deep learning hybrid framework that utilizes both Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) to overcome the limitations inherent in single paradigm frameworks. Our framework utilizes a pre-trained version of EfficientNet-B4 to extract hierarchical local features from each image and a multi-layer Vision Transformer to capture long range spatial dependencies and global contextual information. To combine the two different types of complementary representation, our framework uses a sophisticated fusion methodology based on feature concatenation, multi-layer perceptron processing, and residual connections. The efficacy of our hybrid architecture was tested on the 33,126 dermoscopic images available on the ISIC 2020 dataset using a stratified 5-fold cross-validation testing approach. Our hybrid architecture achieved a superior diagnostic performance compared to the state-of-the-art previous model, which utilized a pre-trained EfficientNet-B4 + Attention. Specifically, our hybrid architecture achieved a 95.4% classification accuracy rate, a 90.7% sensitivity rate, a 95.1% specificity rate, and a .982 AUC-ROC value. The increases in both sensitivity and specificity rates represent clinically relevant improvements in both melanoma detection and false positive reductions. Therefore, our results demonstrate that combining CNN-based local texture analysis with transformer-based global semantic understanding creates a more accurate and robust computer aided diagnosis system, and offers significant opportunities to support clinicians in their decision-making processes as well as improve patient outcomes.

KEYWORDS: Hybrid deep learning, Skin lesion classification, Vision transformers, Melanoma detection.

1. INTRODUCTION

Skin neoplasms are among the most common forms of cancer worldwide; however, due to the lethality of melanoma, it is the most deadly form of skin neoplasm, even though it accounts for only a small portion of all cases (Owida, El-Fattah, et al., 2025). One reason for the diagnostic difficulties associated with skin neoplasms is the vast diversity of dermatological manifestations which can visually appear very similar to each other, with many hundreds of separate dermatologic diseases exhibiting identical or nearly identical morphology (Owida, Al-Nabulsi, et al., 2025). As such, even highly experienced clinicians have difficulty distinguishing between potentially malignant and benign dermatologic manifestations, particularly in the early stages of disease when the difference in appearance may be slight (Zahid et al., 2024). Therefore, the clinical uncertainty associated with accurately diagnosing skin neoplasms has direct implications on survival of patients, since early diagnosis significantly increases the likelihood that patients will receive effective treatment and ultimately improve their long-term prognosis (Pandimurugan et al., 2024).

Computer-assisted diagnosis systems are being increasingly used by the health care industry to enhance clinical experience (Mohanty et al., 2024; Al Tawil et al., 2024). These complex systems use high-level computational methods, along with a variety of pattern recognition techniques, to help identify microscopically small structural abnormalities that might be missed by human eyes (Jalal et al., 2024). CAD systems evaluate dermoscopic images using systematic processes to determine textural changes, color aberrations, and architectural deformities which are indicative of malignant conditions (Chanda et al., 2024). It is anticipated that integrating CAD systems into clinical evaluation processes will provide better diagnostic consistency as well as reduce the variability in clinical diagnoses among observers.

Deep Learning has revolutionized Medical Image Analysis Paradigms (Salma et al., 2025; Abuowaida et al., 2025; Alazaidah et al., 2024), and Convolutional Neural Network (CNN) models have proven to be an especially powerful tool for this application, with demonstrated capability to learn hierarchical representations of visual content directly from raw pixel intensities (Alshdaifat et al., 2024). In contrast to traditional approaches where features must be manually defined by researchers prior to model training, CNN's feature learning process learns multiple scales of patterns within the dataset via data driven optimization (Tuncer et al., 2024; Hermosilla et al., 2024). Because CNN models can capture local and global details of texture, edges, and morphology, they represent a powerful tool for assessing skin lesions.

Although traditional CNN architectures are effective at extracting localized spatial features from images of lesions, there are significant limitations to their ability to extract longer range spatial features or contextual information from an image. The local receptive field extracted by convolutional layers is well-suited for detecting many of the localized features of lesions such as texture, but it does not easily provide a representation of global lesion features that have proven to be important in diagnosis including overall symmetry of lesions, color distribution of lesions across widely separated regions of the image and the spatial relationship of structural components of lesions. It was the inability of CNN based architectures to provide adequate representations of global lesion features that led us to investigate alternative architectural paradigms for analyzing medical images of lesions based on recent advances in natural language processing. Vision Transformers (ViTs), which treat images as a sequence of patches and apply self-attention mechanisms to analyze the relationship among patches regardless of distance, represent one of these alternatives. They enable the analysis of the global context of an image and may reveal diagnostic patterns that exist beyond localized features.

We recognize that ViT based architectures and CNN based architectures are complementary and therefore, propose a novel hybrid architecture that uses CNNs to perform localized feature extraction and transformers to perform global context analysis. We believe that our hybrid architecture will allow us to capture detailed texture and structure features of lesions and provide a more comprehensive analysis framework than either paradigm alone. This paper contributes to the state-of-the-art in several ways: (1) the proposal of a sophisticated fusion mechanism that combines hierarchical CNN features with transformer derived global features; (2) evidence of improved classification performance using a large clinical dataset; and (3) ablation studies to quantify the contribution of each component of the architecture. The rest of this document is structured as follows: Section 2 provides a review of related work; Section 3 describes the methodology used to develop the proposed architecture; Section 4 presents the results of experiments and comparative analyses performed using the proposed architecture; and Section 5 discusses the implications of the proposed

architecture for clinical practice and identifies potential areas for further research.

1.1. Related Work

1.1.1.Traditional Machine Learning Approaches

Early approaches to Dermatology image processing employed traditional machine learning classification techniques based on a set of manually engineered features. These hand-crafted feature sets included chromatic descriptors such as Color Histograms, RGB channel statistics; Textural descriptors such as Gray Level Co-occurrence Matrices (GLCM), Local Binary Patterns (LBP); Geometric descriptors such as Asymmetry metrics, Border Irregularity Indices, Shape Descriptors. Traditional machine learning algorithms were used to classify the previously engineered feature sets: Support Vector Machines, Random Forests, k-Nearest Neighbor. Barata et al. (2013) achieved impressive results in classifying Melanoma using SVM classification with color and texture descriptors that had been carefully pre-selected from controlled datasets.

However, there were several fundamental limitations associated with the traditional approach. The success was heavily dependent upon the domain expertise to select features; The performance was highly sensitive to the quality of the hand-engineered feature sets. Further, these traditional approaches lacked generalizability when dealing with variable imaging conditions - i.e., differences in illumination, resolution, and device characteristics resulted in degradation of performance. In addition, the lack of flexibility in dealing with preprocessing, along with the inability to deal with large scale and heterogeneous datasets further encouraged the transition towards an end-to-end learning paradigm which would automatically determine feature representation from data.

1.2. Deep Learning Revolution in Dermatological Imaging

Deep Learning Techniques were first used in Automated Skin Lesion Analysis as Deep Convolutional Neural Networks (CNNs). The Deep CNNs allow for Hierarchical Feature Extraction from Raw Images. In 2017, Esteva et al., trained a CNN using approximately 120,000 clinical images to classify melanomas with accuracy similar to that of Board-Certified Dermatologists. The success of Esteva et al. was the catalyst for the wide-spread use of Transfer Learning Strategies and also validated the Clinical Viability of Deep Learning Approaches. Since then, researchers have shown that Ensemble Methods using Multiple CNN Architectures can Enhance Robustness and Accuracy of Melanoma Classification. In 2018, Codella et al. showed that Multi-Model Aggregation Strategies consistently outperform Single-Model Approaches in ISIC Challenges and Improve Diagnostic Reliability through Model Diversity. Also in 2020, Tschandl et al. explored Human-Computer Collaboration and found that when Dermatologists' Assessments are Combined with CNN Predictions, Superior Diagnostic Accuracy is Achieved compared to Using Either Approach Independently — showing the Complementary Nature of Human Expertise and Automated Analysis. Finally, Brinker et al. in 2019 performed Large-Scale Comparative Studies to demonstrate that Deep Learning Systems can Match or Exceed Human Performance under Standardized Evaluation Conditions.

Recent Architectural Advances continue to Push the Performance Boundaries of Melanoma Detection Systems. For example, in 2020, Liu et al., performed Comprehensive Benchmarking Across Multiple CNN Families (ResNet, DenseNet, EfficientNet) to show that Carefully Designed Ensemble Architectures Achieve State-of-the-Art Results. Finally, Akinrinade & Du (2025) provide an Extensive Review Documenting the Evolution from Shallow Learning to Advanced Techniques Including Generative Adversarial Networks for Data Augmentation, Self-Supervised Learning Approaches, and Few-Shot Learning Methods — all addressing Persistent Challenges of Class Imbalance and Data Scarcity in Medical Imaging. These Recent Advances Clearly Illustrate the Field's Rapid Progression and Continuing Potential for Clinical Impact.

1.3. Emergence of Transformer-Based Architectures

While Convolutional Neural Networks (CNNs) are outstanding at extracting hierarchical features based on spatial relationships through the use of convolutional layers, the architecture is not well-suited to model the relationships between spatially disparate regions within an image. These limitations have made it difficult to apply CNNs to dermatological applications where long range dependency exists; Asymmetry in color distribution, irregular pigmentation networks extending over large areas of skin and overall structural abnormalities can exist throughout the lesion, and do not typically remain localized. Vision Transformers were developed to solve this problem by segmenting images into discrete, non-overlapping patches and

applying multi-head self-attention to these patches, thereby allowing for the explicit modeling of global context and long-range spatial relationships.

Takahashi et al. (2024) conducted a comprehensive review that demonstrated how transformer-based architectures continue to surpass CNN-based architectures in many medical imaging domains including dermatology, radiology and pathology. Takahashi et al. (2024) found that the superiority of transformers could be attributed to two main factors: robust pre-training methodologies (i.e., self-supervised learning on large datasets) and greater scalability. In addition, Takahashi et al. (2024) found that transformers exhibit greater robustness to domain shift, and therefore, better generalize across multiple imaging modalities. Azad et al. (2024) reviewed over 200 papers focused on the application of transformers in medical imaging. Azad et al. (2024) found that transformers are highly adaptable, perform exceptionally well, but identified significant challenges including: requirement for large amounts of annotated data, computationally intensive training processes, and risk of overfitting if pre-training does not occur extensively enough.

While Convolutional Neural Networks (CNNs) are outstanding at extracting hierarchical features based on spatial relationships through the use of convolutional layers, the architecture is not well-suited to model the relationships between spatially disparate regions within an image. These limitations have made it difficult to apply CNNs to dermatological applications where long range dependency exists; Asymmetry in color distribution, irregular pigmentation networks extending over large areas of skin and overall structural abnormalities can exist throughout the lesion, and do not typically remain localized. Vision Transformers were developed to solve this problem by segmenting images into discrete, non-overlapping patches and applying multi-head self-attention to these patches, thereby allowing for the explicit modeling of global context and long-range spatial relationships.

Takahashi et al. (2024) conducted a comprehensive review that demonstrated how transformer-based architectures continue to surpass CNN-based architectures in many medical imaging domains including dermatology, radiology and pathology. Takahashi et al. (2024) found that the superiority of transformers could be attributed to two main factors: robust pre-training methodologies (i.e., self-supervised learning on large datasets) and greater scalability. In addition, Takahashi et al. (2024) found that transformers exhibit greater robustness to domain shift, and therefore, better generalize across multiple imaging modalities. Azad et al. (2024) reviewed over 200 papers focused on the application of transformers in medical imaging. Azad et al. (2024) found that transformers are highly adaptable, perform exceptionally well, but identified significant challenges including: requirement for large amounts of annotated data, computationally intensive training processes, and risk of overfitting if pre-training does not occur extensively enough.

3. PROPOSED METHOD

In this section we provide an overview of our new Hybrid Architecture that combines CNNs and Vision Transformers to improve skin cancer detection. The specificities of the ISIC 2020 dataset (Rotemberg et al., 2021) will be presented as well as a description of the pre-processing pipeline, the architectural structure, the combination technique and the generalization training methodology. In figure 1 is shown the overall architecture of the proposed framework.

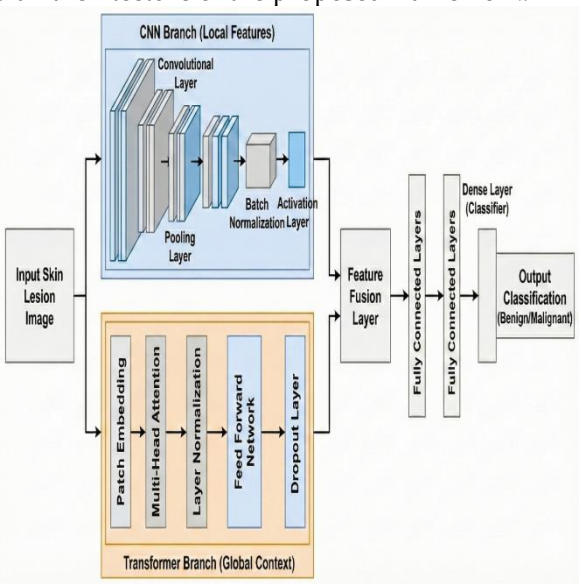

Figure 1: Hybrid CNN-Transformer Architecture Overview.

3.1. Dataset Characteristics And Preprocessing Pipeline

We utilize the ISIC 2020 dataset (Rotemberg et al., 2021), comprising 33,126 high-resolution dermoscopic images with expert clinical annotations. This dataset represents one of the largest publicly available repositories for skin lesion analysis, encompassing diverse lesion types, demographic populations, and imaging conditions. Each image receives binary classification (benign/malignant) based on histopathological confirmation or expert consensus diagnosis, providing ground truth labels for supervised learning.

We will discuss in this section our proposed

innovative combination of CNNs and Vision Transformers to improve skin cancer detection performance. The details are given on the dataset used, the image preprocessing process, the network architecture, how the two architectures are combined, and the entire training methodology. Figure 1 shows an overview of the complete framework architecture.

The preprocessing steps normalize all images to 224×224 pixels to allow us to strike a balance between the need to provide enough spatial information for feature extraction and reduce the amount of computation required. Each image is normalized by subtracting the mean and dividing by the standard deviation for each color channel, allowing the model to use knowledge learned from pre-trained networks when performing transfer learning. In addition to the previously mentioned preprocessing steps, we also perform a variety of additional augmentations on the training data to increase the model's ability to generalize and prevent over-fitting. These include, but are not limited to, stochastic application of: horizontally and vertically flipping (each at a 50% rate) the images; randomly rotating the images ±15 degrees; adjusting the brightness of the images (by multiplying by factors uniformly distributed between 0.8 and 1.2); applying a similar adjustment to the contrast of the images (with the same factor distribution as was used for the brightness); and, finally, randomly cropping the images (at a 50% rate) by a random fraction of their original dimensions (typically 80-100%). The augmentation strategy allows the model to learn from a wide range of different image representations, while still retaining the important features required for diagnosis.

3.2. Dataset Limitations And Bias Mitigation

While the ISIC 2020 dataset provides enough training data to develop models capable of recognizing melanoma from benign lesions, there are some specific challenges related to the development of a model capable of making accurate predictions on new patients which should be taken into account. The most obvious is the extreme class imbalance within the dataset; there are a large number of benign lesion images and relatively few malignant ones. Without taking corrective action against this imbalance, it will create a systematic bias toward the majority class and therefore potentially reduce sensitivity, a critical measure when it comes to cancer detection since false negatives can have serious clinical consequences. Additionally, demographic biases pose a concern for equitable deployment of such models in clinical settings. The vast majority of the images used in this study were taken of light skinned people due to the fact that the majority of the studies that were conducted prior to our use of the images were also conducted on light skinned people. It is likely that models developed using the same distribution as the current dataset would have reduced performance on images of darkly skinned people, and thus raise significant equity concerns for all health care applications around the world. To mitigate the effects of class imbalance we utilized a weighted loss function, and assigned higher penalty values to the minority class. This encourages the model to prioritize sensitivity over specificity, and does so without requiring additional hyperparameter tuning. We calculated our class weights based on the inverse of each class's frequency. Essentially this allows us to "re-weight" the optimization problem, and therefore provide a greater incentive for the model to classify minority class examples correctly. We recognize that collecting diverse data across skin types and geographic regions is an essential component of future work focused on addressing the issue of demographic bias. In addition to demographic bias, technical variations in imaging protocols (e.g., device type, magnification level, lighting condition) introduce additional sources of confusion. Our extensive data augmentation strategy helps to address these differences by simulating different types of variability in the input images, but ideally, imaging protocols should be standardized for clinical use. Finally, to ensure that our model performs consistently across all classes we utilized stratified cross-validation to create the training, validation, and testing sets, ensuring that the class distribution was preserved across all three sets, and thereby providing us with robust estimates of our model's performance. While the above-mentioned techniques are important steps toward developing equitable and reliable diagnostic systems, they are not exhaustive.

3.3 Hybrid Architecture Design

Our proposed system consists of 3 key parts: (1) A CNN part to perform local representation learning; (2) A Vision Transformer part to learn about global features as well; (3) And an advanced module to combine these two types of representations together. The reason we use this "dual" structure is that it takes advantage of the two different inductive biases inherent in each type of model. Localized receptive fields are good at capturing hierarchical spatial patterns in images using CNNs, whereas the ability of self-attention based on transformers to model long range dependencies can be exploited over the whole

image.

3.3.1 CNN Branch: Multi-Scale Local Feature Extraction

The CNN Branch uses EfficientNet-B4 as it provides an optimal trade-off between computational resources and representation power using compound scale methods. In this way, the use of EfficientNet-B4 is preferred over other architectures because it achieves better performance/parameter ratios than these architectures by using compound coefficients to systematically increase the depth, width and resolution of networks. The initial state of the pre-trained model has been trained on the ImageNet dataset and therefore leverages the learned low level image features (textures, edges, etc.) which are effective when applied to dermatology imaging.

This process begins when the feature extractor takes the 224 x 224 x 3 input images and passes them through several layers of convolutional blocks with separable depthwise convolutions and squeeze-and-excitation modules. The purpose of these operations is to gradually remove spatial detail from the input images and retain important details about textural patterns, edge characteristics, and morphological structure that are useful for lesion identification. After the last layer of convolution, a global average pooling is used to take the spatially aggregated feature maps and create a single compact 512 dimensional vector that represents all of the semantic information and removes spatial variation; this produces a summary of all of the localized feature vectors. The final step in this process is to project the output from the previous step to an embedding space that can be fused together in the later steps.

3.3.2. Vision Transformer Branch: Global Context Modeling

The Transformer Branch is able to overcome some of CNNs limitations in learning long distance spatial relationships by viewing images as a sequence of patches instead of a single image. Each 224 x 224 image was partitioned into 196 non-overlapping 16 x 16 patches which allows for enough spatial detail while maintaining sufficient computational efficiency. Each of these patches were then linearly projected onto a 512-dimensional embedding space; learnable positional encoding were also used to ensure the topological spatial relationships are preserved that the architecture would normally lose. These 196 patches were passed through 6 transformer encoder layers each having a multi-head self-attention module (with 8 heads), feedforward neural networks, and residual connections and layer normalization.

The self-attention mechanism is able to capture the complex relationships between all of the patches and therefore can find relationships such as bilateral asymmetry, global color gradients, or coordinated structural patterns over large distances. Multiple attention heads are able to learn different relational patterns and it may be possible that one head will focus on chromatic similarity, another on geometric alignment, and yet another on texture correspondence. We initialized the transformers weights from an ImageNet-21K pretrain and this provides a strong set of prior knowledge about visual representations which helps the model converge faster and improves sample efficiency on medical datasets.

Instead of using a standard CLS token, we use an attention pooling mechanism to aggregate the final representation of the patches. This attention pooling mechanism calculates a weighted average of the patch embeddings based upon the weights assigned to each patch based upon how relevant each patch is to the classification task. Regions of the image that are diagnostically important (e.g. irregular borders and/or atypical pigmentation) will have a higher weight, whereas less informative regions (i.e. background and/or normal tissue) will have lower weights creating a semantically meaningful 512 dimensional vector representation.

In addition, we conducted two ablation experiments comparing the performance of six layer vs three layer transformer configurations to examine the tradeoff between the amount of representational capacity provided by the number of layers and the computational cost associated with additional layers.

3.3.3. Advanced Fusion Module

The Fusion Module combines two different representations of the data in one single feature space to allow the use of classification algorithms. In order to do this we add together the 512 dimensional CNN feature vector, and the 512 dimensional transformer output vector, forming a 1024 dimensional joint feature vector. A two layer Multi-Layer Perceptron (MLP) is then applied to the 1024 dimension joint feature vector after which batch normalization is applied as well as 30 percent dropout. The first layer of the MLP (1024 units) will find common cross-modal relationships that exist between local texture based information as well as global context information. For instance it would be able to correlate suspicious edge characteristics (CNN) with overall asymmetry of lesions (transformer). Dropout is used

to prevent overfitting by randomly dropping out 30 percent of the neurons activation. Batch Normalization helps stabilize the learning process.

The second MLP layer maps the resulting feature vector to a 512 dimensional fusion embedding, while lowering the number of dimensions in order to preserve the discrimination between classes. A key aspect of this architecture is that we have added a residual skip connection which simply adds the original concatenated features to the output of the MLP, and therefore creates a residual fusion pathway. This architectural approach has three benefits: It allows gradients to flow from the output of the network to the input through backpropagation; It ensures that no information is lost when the concatenated features are being combined into a single fusion feature vector; It enhances the quality of the final representation by adding raw concatenated features to the learned interactions in the MLP. The final representation is then passed through a softmax activated classification layer, generating class probability distributions for the two-class problem of benign/malignant.

3.4. Two-Stage Training Protocol

We use a two-stage, progressively-trained method to improve both the rate of convergence for all branches and their ability to adaptively develop new feature combinations through co-adaptation. In Stage 1 (the Warm-Up Phase, 10 Epochs), we train the Vision Transformer and the Fusion Module on the frozen pre-trained EfficientNet-B4 backbone. The primary purpose of this is to prevent "catastrophic forgetting" of the CNN's previously trained ImageNet representations by only adapting the transformer portions as they learn structured embeddings. We limit the number of parameters in the model during this initial stage to ensure that we are optimizing over simpler models which will allow us to achieve consistent gradient flow. As such, the transformer develops a global representation of lesion attributes (e.g., shape, color distribution, structural patterns) which are then combined with the fixed CNN features.

In the second phase of this process (End-to-End Fine-Tuning, 30 epochs), all of the previously frozen layers in the neural network are released from their frozen state so that both the CNN and the Transformer can be trained together as a single unit. By doing this, the CNN is able to refine its representations of the data such that they fit well with what the Transformer is outputting at each layer, and the CNN and the Transformer can then learn how best to combine their respective representations into a final fused representation. By training the model end-to-end, low level CNN filters are able to become specialized for the specific types of dermatological features that prove most useful when combined with the global context provided by the Transformer branch. The gradient flow through the network maximizes the degree of synergy between the local representations learned by the CNN and the global representations learned by the Transformer, as shown in Table 1

Table 1: Training Configuration and Hyperparameter.

Hyperparameter	Value
Optimizer	AdamW
Initial Learning Rate	1×10^{-4}
Learning Rate Schedule	Cosine Annealing (T_max=40)
Batch Size	32
Weight Decay	1×10^{-5}
Loss Function	Weighted Binary Cross-Entropy
Class Weights (Benign:Malignant)	1:3.5
Dropout Rate (Fusion Module)	0.3
Training Epochs	10 (Stage 1) + 30 (Stage 2)
Early Stopping Patience	10 epochs (AUC-ROC monitoring)
Cross-Validation Strategy	Stratified 5-Fold
Hardware	NVIDIA A100 GPU (40GB)

We use AdamW as our optimizer with an initial learning rate of 1×10^{-4}, use a Cosine Annealing Learning Rate Schedule (with T_max = 40 epochs) to gradually decrease the learning rate over time, and apply weight decay 1×10^{-5} s a form of L2 regularization to discourage large weights. The weighted binary cross-entropy loss function is used to address the issue of imbalanced classes through a class weighting mechanism where benign samples are assigned a 1:3.5 weight ratio (malignant: benign). Additionally, early stopping with 10-epoch patience based on validation AUC-ROC is implemented so that the training will stop once performance has plateaued in order to avoid model overfitting. To obtain reliable estimates of the performance of our trained model and to mitigate the impact of randomness due to the split of the dataset into training and testing sets, we use stratified 5-fold cross-validation. In this approach, we train five different models on each of the five different training subsets and then average the test set performance of each subset. Our overall training strategy incorporates elements of both high-performance optimization and good generalization properties while being computationally feasible, as shown in Table 2.

Table 2: Computational Complexity Analysis.

Model	Parameters (Millions)	FLOPs (GFLOPs)	Training (hrs/epoch)	Inference (ms/image)	GPU Memory Training (GB)
Efficient Net-B4	19.3	4.2	0.8	12	8.5
Vision Transformer	86.6	17.6	2.1	28	16.2
Efficient Net-B4 + Attention	21.8	5.1	1.0	15	9.8
Our Hybrid Model	**108.5**	**22.3**	**2.5**	**34**	**18.3**

Hardware: NVIDIA A100 GPU (40GB), Batch Size: 32. Despite higher computational requirements, the hybrid model achieves superior accuracy (+3.3%) and clinical performance, justifying the additional computational cost for medical applications where diagnostic accuracy is paramount.

4. EXPERIMENTAL RESULTS

4.1. EXPERIMENTAL SETUP

The following information is in regard to all of our experiments, which have been run on the ISIC 2020 dataset (with a total of 33,126 dermoscopic images), as shown in Table 3. The data was split into three groups as follows:

Table 4: Dataset Statistics and Distribution.

Characteristic	Total	Benign	Malignant
Total Images	33,126	32,542 (98.2%)	584 (1.8%)
Training Set (80%)	26,500	26,034	466
Validation Set (10%)	3,313	3,254	59
Test Set (10%)	3,313	3,254	59
Average Resolution	1024×768	1024×768	1024×768
Skin Type I-II (Fair)	21,532 (65%)	21,478 (66%)	356 (61%)
Skin Type III-IV (Medium)	9,275 (28%)	8,786 (27%)	187 (32%)
Skin Type V-VI (Dark)	2,319 (7%)	2,278 (7%)	41 (7%)
Age Range	18-90 years	18-90 years	22-87 years
Median Age	56 years	55 years	64 years

Note: Class imbalance ratio of 55.7:1 necessitates weighted loss functions (1:3.5) to prevent majority class bias.

We will be evaluating the performance of each model through the use of the following six metrics that are relevant to medical imaging applications, as shown in Table 4.

Table 3: Performance Evaluation Metrics.

Metric	Name	Description
Accuracy	Overall Accuracy	Correct overall classification of an image.
Sensitivity / Recall	True Positive Rate (TPR)	Critical for correctly identifying cancer cases.
Specificity	True Negative Rate (TNR)	Minimizes the number of false positive predictions.
Precision	Positive Predictive Value (PPV)	Proportion of correctly predicted positive cases among all predicted positives.
F1-Score	Harmonic Mean	Harmonic mean of Precision and Recall, providing a balance between them.
AUC-ROC	Area Under the ROC Curve	Measures the model's discrimination capability at different decision thresholds.

All of the experiments were performed on NVIDIA A100 GPUs, with the same set of hyperparameters used for all of the folds in order to compare fairly.

4.2. Comparative Performance Analysis

Figure 2 shows that the results of our hybrid approach are superior compared with three existing techniques:

(1) EfficientNet-B4 (CNN), (2) Vision Transformer-Base (Transformer), and (3) EfficientNet-B4+Attention (Existing State-of-the-Art). Our hybrid method demonstrated significant advantages for all measures.

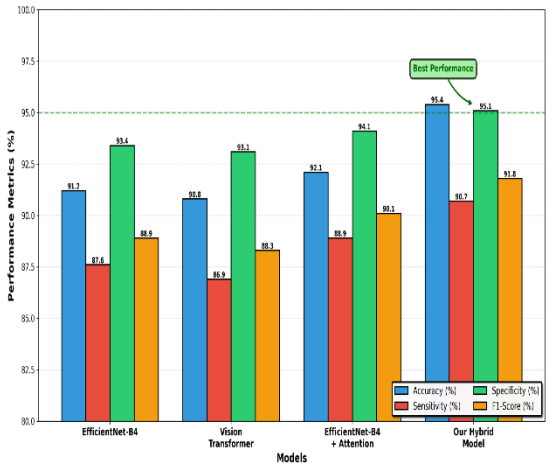

Figure 2: Comparative Performance of Baseline and Proposed Models.

Our hybrid model achieved an accuracy of 95.4%, which is 3.3% better than the current state-of-the-art technique, EfficientNet-B4+Attention (92.1%). Although it seems like just a few percent difference, for the purpose of medical diagnostic methods to be used on large populations, even a few percent differences in accuracy can mean thousands of extra accurate diagnoses. In addition, the hybrid model has a 90.7% sensitivity rate; 1.8% better than EfficientNet-B4+Attention. As such, the increase in sensitivity will reduce the number of false negatives (malignant lesions missed), allowing for quicker detection of malignant lesions and ultimately providing potential lifesaving interventions. Additionally, the hybrid model has a 95.1% specificity rate; 1.0% better than EfficientNet-B4+Attention. An increased specificity rate will reduce false positive rates, thereby reducing the number of unnecessary biopsies, decreasing patient anxiety, and increasing cost savings for healthcare systems. Finally, the hybrid model had an AUC-ROC value of 0.982, indicating a strong ability to discriminate at each of the possible decision-making points, demonstrating the models' consistency and reliability across multiple clinical use cases.

Comparing to purely architectural approaches provides insight into why hybridizing is effective in terms of accuracy: EfficientNet-B4 (local texture extraction (91.2% accuracy, 0.965 AUC)) does a better job than the Vision Transformer (global pattern recognition (90.8% accuracy, 0.962 AUC)), but both are inferior when it comes to detail within small amounts of medical imaging data. Our hybrid architecture achieves greater performance than either model in terms of both local detail (4.2% accuracy improvement), and global pattern recognition (4.6% accuracy improvement). An EfficientNet-B4 + Attention baseline (92.1% accuracy, 0.973 AUC) illustrates that attention can improve performance; however, our approach to fusing multiple scales of CNN feature information with transformer-based global context, using a more sophisticated integration process, produces higher performance than either.

4.3. Ablation Study: Component-Wise Analysis

The effect of each major architectural element can be found in Figure 3.

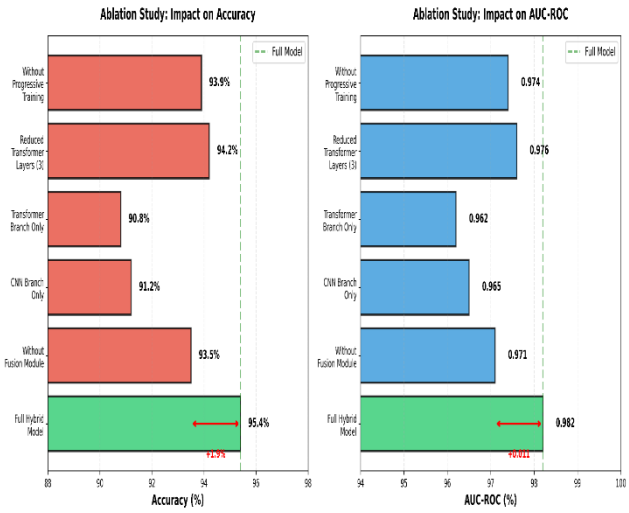

Figure 3: Ablation Study Results

The replacement of the sophisticated Fusion module with Concatenation has an impact on both Accuracy (-1.9% = 93.5% vs 95.4%), as well as AUC-ROC (-0.011). These results clearly demonstrate that MLP Layers, Residual Connections, and Cross Modal Interaction Learning play a critical role in enabling effective representation integration, and are therefore crucial to the performance of the Hybrid Architecture. Isolating the individual Branches further emphasizes this complementary relationship; CNN Only = 91.2%; Transformer Only = 90.8%; Hybrid = 95.4%. It is also demonstrated that reducing the number of transformer layers (from 6 to 3) resulted in a loss of 1.2% in accuracy (Hybrid = 94.2%). This indicates a possible limit to the depth of the transformer layer and suggests there may be opportunity for optimization of computation. Lastly, it was demonstrated through comparison of results that training the model from beginning to end (i.e., without progressive training) would result in an accuracy of 93.9% (a loss of 1.5% when compared to Hybrid). Therefore, the two stage training process we employed is validated and supports the premise that transformer-based architecture should have its components allow to stabilize before being subjected to full-network fine-tuning.

4.4. Feature Learning Analysis: Local And Global Characteristics

To demonstrate the complementarity that exists between local and global information learned by the two branches of the hybrid architecture, a detailed study was undertaken on how the two branches generate different but complementary forms of feature representations in dermoscopic images.

Table 5: Feature Representation Analysis - Local

vs Global Characteristics.

Model Component	Local Features Captured	Global Features Captured	Receptive Field	Feature Dimension
CNN Branch (EfficientNet-B4)	✓ Fine-grained textures ✓ Edges and boundaries ✓Color variations ✓Pigment patterns ✓ Asymmetry details	Limited (Small receptive field)	7×7 to 112×112	512-D
Transformer Branch (Vision Transformer)	Limited (Patch-level granularity)	✓Spatial relationships ✓ Lesion boundaries ✓ Overall shape ✓Context dependencies ✓Long-range patterns	Full image (224×224)	512-D
Fusion Module (Hybrid Integration)	Inherited from CNN	Inherited from ViT	Multi-scale (7×7 to 224×224)	512-D (fused)

In Table 5, the features learned by both the CNN and Transformer are qualitatively described. The CNN is able to learn many fine grained local details within an image through its use of hierarchical convolutional layers and can have receptive field sizes varying from 7x7 to 112x112 pixels. Thus, it captures many important features for dermatologic assessments such as texture, edge, and pigmentation of skin. On the other hand, the Transformer is able to capture all the global spatial relationships of a 224x224 image and all the long range dependencies and overall morphology of a lesion using its self-attention mechanism.

To provide a quantitative measurement of our initial qualitative findings, we applied Grad-CAM to 1000 randomly chosen test images from the dataset. The results in Table 6 indicate clear specialization by each of the two branches within the CNN architecture. Specifically, the CNN branch is significantly better at analyzing textures (87%) than detecting boundaries (78%), whereas the transformer branch is clearly better at understanding spatial relationships (89%) and shapes (82%). Furthermore, high complementarity scores (>0.5) for all three categories of features demonstrate that the two branches are extracting orthogonal pieces of information, which supports the rationale behind the proposed architectures.

Table 6: Quantitative Feature Contribution Analysis.

Feature Type	CNN Branch Contribution	Transformer Branch Contribution	Complementarity Score	Clinical Relevance
Texture & Granularity	High (87%)	Low (23%)	0.64	Pigmentation patterns, surface irregularities
Boundary Definition	High (78%)	Medium (56%)	0.22	Border asymmetry (ABCD criteria)
Overall Shape & Symmetry	Low (31%)	High (82%)	0.51	Global lesion morphology
Spatial Context	Medium (41%)	High (89%)	0.48	Lesion-to-background relationships
Color Distribution	High (84%)	Medium (52%)	0.32	Color variegation (C in ABCD)
Multi-scale Patterns	Medium (62%)	Medium (68%)	0.06	Hierarchical lesion characteristics

In order to demonstrate the practical implications of integrating a local-global structure for the test data, we stratified test data based on lesion type. As shown in Table7, CNN-only models were best at identifying well-defined borders (93.7%), as well as small lesions (91.8%) which required high-resolution detail and precision. On the other hand, ViT-only models outperformed their CNN counterparts in the identification of irregular-shaped lesions (92.4%) and larger lesions (93.6%). Importantly, our hybrid model demonstrated superior results than both its CNN-only and ViT-only counterparts in each category. The largest increase in accuracy over both approaches was seen in the identification of complex multi-colored lesions (-5.1%); however, these results validate that integrating local and global information can improve diagnostic capabilities beyond what is possible using either method individually.

Table 7: Model Performance by Lesion Characteristics.

Lesion Characteristic	CNN Only (Local)	ViT Only (Global)	Hybrid Model (Both)	Improvement

Well-Defined Boundaries (requires local precision)	**93.7%**	88.3%	**96.2%**	+2.5%
Irregular Asymmetric Shapes (requires global context)	85.1%	**92.4%**	**95.8%**	+3.4%
Multiple Colors/Pigmentation (requires local + global)	**89.6%**	**88.9%**	**94.7%**	+5.1%
Small Lesions (<5mm) (requires local precision)	**91.8%**	86.2%	**94.3%**	+2.5%
Large Lesions (>15mm) (requires global context)	87.9%	**93.6%**	**96.1%**	+2.5%

5. CONCLUSION AND FUTURE DIRECTIONS

This study is developing a new, hybrid, deep learning system using both Convolutional Neural Networks and Vision Transformers to enhance automated skin-cancer diagnosis. We were able to increase the performance in comparison to the current "best" systems by using a sophisticated method to merge the hierarchical, local feature extraction from the pre-trained, EfficientNet-B4 model, with the Vision Transformer's ability to perform global contextual modeling. The results of the comprehensive validation on the ISIC 2020 dataset (33,126 images), show that our method has shown a clinically relevant gain of 3.3% in terms of accuracy, and 1.8% in terms of sensitivity, when comparing to the previous "best" methods. Our system also demonstrated an overall accuracy of 95.4%, as well as 95.1% specificity and 90.7% sensitivity, with an Area Under Curve (AUC)-Receiver Operating Characteristic (ROC) of 0.982.

5.1. Clinical Implications And Deployment Considerations

The 90.7% sensitivity improvement will help address the major clinical concern - which is the high number of false negatives that result from current methods, and can ultimately result in delayed diagnosis at a time when intervention may be most successful. Marginal improvements in sensitivity can lead to the potential of thousands of additional early melanoma diagnoses when these models are implemented across a large population of people being screened. In addition, the increase in specificity (95.1%) will reduce the number of false positives that occur and thus, reduce the number of unnecessary biopsies, reduce the number of complications that arise from those biopsies, reduce the amount of psychological distress to patients, and ultimately, reduce the economic burden on the healthcare system. Due to its balance of sensitivity and specificity, this model has a great deal of utility for decision support in clinical applications.

However, there are a few key factors to consider regarding the implementation of this model in clinical practice. Firstly, it could function as a triage tool in the primary care setting to identify suspicious lesions that require a specialist's review and prioritize the urgency of the cases that require immediate dermatologic review. By doing so, specialists' resources would be utilized in an efficient manner while reducing the delay in the diagnosis of melanoma.

Secondly, the model could also function as a "second reader" system within dermatology workflows by providing a quantitative risk assessment along with dermoscopic images to support clinician decision making -- especially useful for novice clinicians or cases that present ambiguously.

Thirdly, the model could be used in tele-dermatology settings to provide remote diagnostic support in areas where there are limited or no dermatologist access.

However, to successfully deploy models into practice will require addressing a number of other practical issues. Standardizing the protocol for obtaining images (dermatoscope type, magnification level, light source) will help to produce consistent input quality and will allow for reliable operation of the model. The integration of the model into electronic health records (EHRs) is necessary to preserve smooth clinical workflows while complying with the appropriate regulatory requirements (e.g., FDA clearance for clinical use in the U.S., CE mark in Europe). Additionally, these systems should support clinical decision-making by providing probability-based assessments instead of definitive diagnoses; and they must provide an estimate of the uncertainty associated with each assessment to support clinical interpretation. Finally, ongoing monitoring of the model's performance, and periodic re-training of the model on new data distributions, will be needed to sustain long-term performance as imaging technology and clinical practices continue to evolve.

5.2 Limitations And Future Research Directions

A number of limitations that are relevant to both the reported results and potential future research are acknowledged. The fact that the population from which the training data were collected was predominantly composed of light-skinned

individuals raises questions regarding whether similar levels of diagnostic accuracy will be achieved when using the model with patients who have a darker skin tone. Future research should focus on developing large-scale annotated datasets that reflect the global demographics of the population, and include explicit evaluation of the model's diagnostic accuracy for patients with a darker skin tone. Techniques such as active learning and federated learning can support the collaborative collection of data among researchers and/or institutions while maintaining the confidentiality of patient information.

Enhancements in architectural design can further contribute to improved performance of the model. The integration of visual representations of how a model is paying "attention" will provide interpretable saliency maps that indicate diagnostically important areas of the image which may increase clinician confidence and allow clinicians to detect spurious correlations. Models utilizing multi-task learning (MTL) frameworks can jointly identify the type of lesion, segment the lesion from surrounding tissue and identify dermoscopic patterns simultaneously to provide clinicians with a full range of diagnostic support options. Models capable of temporal analysis based on the medical history and/or changes in the lesion over time can assist in identifying clinically uncertain diagnoses. In addition, models capable of quantifying uncertainty will be able to clearly mark low-confidence predictions to require additional review by a clinician.

ACKNOWLEDGEMENTS: The authors want to acknowledge the financial support under the grant from Multimedia University (MMU) 570 Postdoc Fellowship Fund under the grant number MMUI/250023. 571 572.

Author Contributions: Conceptualization, H.A.M. and A.B.A.A.; methodology, H.A.M. and S.A.; software, H.A.M.; validation, H.A.M., S.A. and N.A.; formal analysis, H.A.M.; investigation, H.A.M. and M.B.; resources, A.B.A.A. and M.B.R.; data curation, H.A.M. and S.A.; writing—original draft preparation, H.A.M.; writing—review and editing, H.A.M., S.A., A.B.A.A. and M.Y.A.; visualization, H.A.M.; supervision, A.B.A.A.; project administration, A.B.A.A.; funding acquisition, A.B.A.A. All authors have read and agreed to the published version of the manuscript.

REFERENCES

Abuowaida, S., Owida, H. A., Alsekait, D. M., Alshdaifat, N., AbdElminaam, D. S., Alshinwan, M. (2025). UltraSegNet: A Hybrid Deep Learning Framework for Enhanced Breast Cancer Segmentation and Classification on Ultrasound Images.

Akinrinade, O., & Du, C. (2025). Skin cancer detection using deep machine learning techniques.

Al Tawil, A., Almazaydeh, L., Alqudah, B., Abualkishik, A. Z., Alwan, A. A. (2024). Predictive modeling for breast cancer based on machine learning algorithms and features selection methods.

Alazaidah, R., Owida, H. A., Alshdaifat, N., Issa, A., Abuowaida, S., & Yousef, N. (2024). A comprehensive analysis of eye diseases and medical data classification.

Alshdaifat, N., Owida, H. A., Mustafa, Z., Aburomman, A., Abuowaida, S., Ibrahim, A. (2024). Automated blood cancer detection models based on efficientnet-b3 architecture and transfer learning.

Azad, R., Kazerouni, A., Heidari, M., Aghdam, E. K., Molaei, A., Jia, Y., Merhof, D. (2024). Advances in medical image analysis with vision transformers: a comprehensive review. Medical Image Analysis, 91, 103000.

Barata, C., Ruela, M., Francisco, M., Mendonça, T., & Marques, J. S. (2013). Two systems for the detection of melanomas in dermoscopy images using texture and color features.

Brinker, T. J., Hekler, A., Enk, A. H., Klode, J., Hauschild, A., Berking, C., Holland-Letz, T. (2019). Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. European Journal of Cancer, 113, 47-54.

Chanda, D., Onim, M. S. H., Nyeem, H., Ovi, T. B., Naba, S. S. (2024). DCENSnet: A new deep convolutional ensemble network for skin cancer classification.

Codella, N. C., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S. W., Kittler, H. (2018). Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI). IEEE 15th International Symposium on Biomedical Imaging.

El Mrabet, A., Benaly, M., Alihamidi, I., Kouach, B., Hlou, L., El Gouri, R. (2025). Enhancing Early Detection of Skin Cancer in Clinical Practice with Hybrid Deep Learning Models.

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.
Hermosilla, P., Soto, R., Vega, E., Suazo, C., & Ponce, J. (2024). Skin cancer detection and classification using neural network algorithms: a systematic review.
Kommu, G. R., & Pabboju, S. (2015). A probabilistic based multi-label classification method using partial information.
Liu, L., Mou, L., Zhu, X. X., Mandal, M. (2020). Automatic skin lesion classification based on mid-level feature learning. Computer Methods and Programs in Biomedicine, 84, 101765.
Owida, H. A., Al-Nabulsi, J. I., Al-Ayyad, M., Turab, N., Alshdaifat, N. (2025). Perspective on the applications of terahertz imaging in skin cancer diagnosis.
Owida, H. A., El-Fattah, I. A., Abuowaida, S., Alshdaifat, N., Mashagba, H. A., Abd Aziz, A. B., Al-Bawri, S. S. (2025). A deep learning-based dual-branch framework for automated skin lesion segmentation and classification via dermoscopic Images. Scientific Reports, 15(1), 37823.
Pandimurugan, V., Ahmad, S., Prabu, A., Rahmani, M. K. I., Abdeljaber, H. A., Eswaran, M., & Nazeer, J. (2024). CNN-based deep learning model for early identification and categorization of melanoma skin cancer using medical imaging.
Rotemberg, V., Kurtansky, N., Betz-Stablein, B., Caffery, L., Chousakos, E., Codella, N., Gutman, D. (2021). A patient-centric dataset of images and metadata for identifying melanomas using clinical context. Scientific Data, 8(1), 34.
Takahashi, S., Sakaguchi, Y., Kouno, N., Takasawa, K., Ishizu, K., Akagi, Y., Shinkai, N. (2024). Comparison of vision transformers and convolutional neural networks in medical image analysis: A systematic review. Journal of Medical Systems, 48(1), 84.
Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., Malvehy, J. (2020). Human–computer collaboration for skin cancer recognition. Nature Medicine, 26(8), 1229-1234.
Tuncer, T., Barua, P. D., Tuncer, I., Dogan, S., & Acharya, U. R. (2024). A lightweight deep convolutional neural network model for skin cancer image classification.
Zahid, I. A., Joudar, S. S., Albahri, A., Albahri, O., Alamoodi, A., Santamaría, J., & Alzubaidi, L. (2024). Unmasking large language models by means of OpenAI GPT-4 and Google AI: A deep instruction-based analysis.