

Comparative Analysis of Single and Hybrid CNN-Transformer Models for Melanoma Detection Using the ISIC Dataset

Chenghan Wang

*Computer Science, Faculty of Engineering University of Bristol, Bristol, United Kingdom
ou23567@bristol.ac.uk*

Abstract: Early detection of skin cancer, particularly melanoma, is crucial for improving patient outcomes. This study presents a comprehensive comparative analysis of deep learning architectures for automated skin lesion classification. This study evaluated eight models across four architectural categories: traditional convolutional neural networks, residual networks, vision transformers and novel hybrid approaches combining CNNs with transformers. Using the International Skin Imaging Collaboration dataset with 3,267 dermatoscopic images, the study assessed both classification performance and computational efficiency. Hybrid architectures consistently outperformed individual models, with the ResNet+ViT hybrid achieving the highest accuracy of 85.1% and AUC of 0.932. However, these performance improvements came with significant computational costs, highlighting important trade-offs for practical deployment. MobileNetV2 offered the best efficiency while maintaining a reasonable accuracy of 79.1%, making it suitable for resource-constrained environments. The findings demonstrate that combining complementary feature extraction mechanisms from CNNs and transformers creates more robust representations for skin lesion classification, approaching the diagnostic accuracy of dermatologists reported in the literature. This research contributes valuable insights for the development of automated diagnostic support tools that could improve early detection rates and enhance patient outcomes in skin cancer management.

Keywords: Skin lesion classification, melanoma detection, convolutional neural networks, Hybrid Mode.

1. Introduction

Skin cancer represents one of the most prevalent forms of cancer globally, with melanoma being its most aggressive and potentially lethal variant. According to recent statistics, the incidence of melanoma has been rising at an alarming rate of 2-3% annually in many countries, with over 290,000 new cases and 60,000 deaths worldwide each year as reported by Siegel et al. [1]. Despite its lethality, melanoma exhibits one of the highest survival rates when detected early, with 5-year survival rates exceeding 95% for localized disease compared to below 25% for metastatic cases.

The critical importance of early detection in melanoma management cannot be overstated. However, accurate clinical diagnosis of skin lesions presents significant challenges even for experienced dermatologists. Visual examination of skin lesions is inherently subjective, with studies

reporting diagnostic accuracy rates among dermatologists ranging from 65-80% according to the literature [2, 3]. The challenge is further compounded by the global shortage of dermatologists, particularly in resource-limited and rural settings, creating barriers to timely and accurate diagnosis for many patients.

The integration of artificial intelligence into dermatological practice represents a significant opportunity to improve early detection rates and, consequently, patient outcomes in skin cancer management.

Deep learning approaches for skin lesion classification have evolved significantly over the past decade. Early efforts primarily utilized traditional convolutional neural networks with relatively simple architectures. Esteva et al. demonstrated a significant breakthrough in 2017 with a CNN-based system that achieved dermatologist-level classification of skin cancer using a dataset of 129,450 clinical images [4]. Their approach utilized transfer learning with the Inception v3 architecture pre-trained on ImageNet. The introduction of residual networks marked an important advancement in deep learning for medical image analysis. ResNet architectures pioneered by He et al. addressed the vanishing gradient problem through skip connections, which enabled the training of deeper networks [5]. Several studies have demonstrated the effectiveness of ResNets for skin lesion classification, with notably higher performance compared to traditional CNNs. Further architectural innovations such as MobileNetV2 [6] and EfficientNet [7] have focused on optimizing the efficiency-accuracy trade-off, potentially enabling deployment in resource-constrained environments. Recently, vision transformers have emerged as a promising alternative to CNNs for image classification tasks. Originally adapted from natural language processing by Dosovitskiy et al. transformers utilize self-attention mechanisms to capture global dependencies in images [8]. Several studies have explored their application to medical image analysis, including skin lesion classification, with competitive results compared to established CNN architectures.

A notable trend in recent literature is the development of hybrid approaches that combine CNNs and transformers to leverage their complementary strengths. However, the effectiveness of such hybrid approaches specifically for skin lesion classification remains relatively unexplored. Despite these advancements, significant challenges persist in the field. Many studies utilize different datasets, preprocessing techniques, and evaluation metrics, making direct comparisons difficult. Additionally, the computational requirements of various architectures are often not systematically assessed, limiting insights into practical deployment considerations. Furthermore, the extreme class imbalance present in realistic skin lesion datasets ongoing methodological challenges that require careful handling.

This study aims to address several key gaps in the current literature through a comprehensive comparative analysis of deep learning architectures for skin lesion classification. Specifically, the research objectives are to systematically evaluate and compare the performance of representative architectures across different deep learning paradigms using a common dataset and evaluation framework. And also to analyze the learning dynamics and convergence patterns of different architectures when applied to skin lesion classification, identifying potential optimization strategies. Through addressing these objectives, this study aims to provide a comprehensive understanding of the strengths and limitations of different deep learning approaches for skin lesion classification, contributing valuable insights for the development of more accurate and efficient automated diagnostic support systems.

2. Materials and methods

2.1. Dataset

The study utilized the International Skin Imaging Collaboration dataset from Kaggle, containing images of benign and malignant skin lesions. The dataset comprises a total of 3,267 images, which were divided between training and validation purposes. The training portion contains 2,637 images with 1,440 benign and 1,197 malignant cases, while the validation portion includes 630 images with 330 benign and 300 malignant cases.

All images in the dataset have a resolution of 224×224 pixels, with a balanced distribution of classes in both training and validation sets. It is important to note that all rights to the data are bound to the ISIC-Archive rights. The ISIC Archive contains over 23,000 dermoscopic images collected from clinical centers worldwide and serves as the largest publicly available collection of quality-controlled dermoscopic images [9]. The dataset was selected with the aim of developing machine learning models that might achieve better and more cost-effective predictions than dermatologists, potentially improving access to dermatological diagnostics in resource-limited settings.

2.2. Data preprocessing

The preprocessing pipeline focused primarily on data augmentation techniques, and all images were already standardized at 224×224 pixels, with the exception of AlexNet which required minor adjustment to 227×227 pixels. In order to enhance model generalization and robustness, the study applied multiple data augmentation techniques during the training phase. These included random horizontal flipping with a probability of 0.5, random vertical flipping also with a probability of 0.5, random rotation within ± 15 degrees, and color jittering with parameters of 0.1 for brightness, 0.1 for contrast, 0.1 for saturation, and 0.05 for hue adjustments.

All images were normalized using the mean values of 0.485, 0.456, and 0.406, along with standard deviation values of 0.229, 0.224, and 0.225 from the ImageNet dataset, which is standard practice when using pre-trained models. While the training images underwent these augmentation procedures, the validation set was processed with the same normalization steps but without any data augmentation to ensure accurate evaluation.

2.2.1. Model architectures

The study implemented and evaluated eight deep learning models across four architectural categories to provide a comprehensive comparison of approaches for skin lesion classification.

2.2.2. Traditional convolutional neural network

In the traditional CNN category, a pioneering architecture containing 5 convolutional layers followed by 3 fully connected layers, AlexNet, was implemented. For implementation, the study utilized ImageNet pre-trained weights while freezing the feature extraction layers to preserve generalized image feature recognition capabilities.

2.2.3. Residual networks

The study evaluated several residual network architectures. The first was ResNet50, a 50-layer deep residual network that employs skip connections to address the vanishing gradient problem [2]. The implementation froze the pre-trained network backbone and replaced the final fully connected layer with a custom classifier tailored to skin lesion detection.

The research also implemented MobileNetV2, which is a lightweight architecture specifically designed for mobile devices. This model utilizes depthwise separable convolutions and inverted residuals to maintain performance while reducing computational requirements [3]. The feature extraction layers were frozen during training, and the classifier was customized for the skin lesion binary classification task.

Additionally, the study evaluated EfficientNet-B0, a scalable architecture that employs compound scaling to balance network depth, width, and resolution [4]. Similar to other models, the pre-trained feature extraction layers were frozen, and a custom classification head optimized for skin lesion classification was added.

2.2.4. Transformer-based models

In the transformer category, the Vision Transformer (ViT-B/16) was implemented, an architecture that adapts the transformer model from natural language processing to image classification by treating image patches as tokens [5]. The pre-trained ViT backbone was frozen, the classification head was replaced with a custom one designed specifically for skin lesion classification.

2.2.5. Hybrid models

The research designed three hybrid architectures to investigate whether combining complementary feature extraction mechanisms could enhance classification performance. The first hybrid approach combined AlexNet with Vision Transformer. This architecture extracts features from both models, concatenates them, and passes them through a custom classifier. The AlexNet features (dimension 4096) and ViT features (dimension 768) were combined to create a rich representation of skin lesion characteristics. The second hybrid model combined EfficientNet-B0 with Vision Transformer. The EfficientNet features (dimension 1280) and ViT features (dimension 768) were concatenated and processed through a multi-layer classifier. The third hybrid architecture leveraged both the hierarchical feature extraction of ResNet50 and the attention mechanisms of Vision Transformer. The ResNet features (dimension 2048) and ViT features (dimension 768) were combined for classification.

For all models, the study employed a transfer learning approach, utilizing weights pre-trained on ImageNet. The feature extraction layers were frozen to prevent catastrophic forgetting of general image features, while only the classification heads were adapted for the specific skin lesion classification task.

2.2.6. Evaluation metrics

Model performance was assessed using several complementary metrics. For classification effectiveness, accuracy, weighted F1 score, Area Under the ROC Curve were measured to measure each model's ability to discriminate between classes across different classification thresholds, precision, recall, and loss values from both training and validation. To evaluate practical deployment considerations, efficiency metrics were collected including model size in megabytes, parameters count, inference time in milliseconds, throughput measured in images processed per second, peak memory usage in megabytes, and training time in seconds. These comprehensive metrics allowed for a holistic evaluation of each architecture's suitability for skin lesion classification, considering both performance and practical deployment aspects.

2.3. Experimental results

2.3.1. Hardware and software configuration

The experimental evaluation was conducted with NVIDIA RTX 4090 GPU, 64GB of DDR5 RAM, and an AMD Ryzen 7800X3D CPU. The software environment included PyTorch version 1.13.0 running on CUDA 11.7 with torchvision 0.14.0. Additional libraries utilized for the implementation and analysis included scikit-learn for metrics calculation, matplotlib for visualization, and thop for computing theoretical operations counts and model profiling. This configuration provided the necessary computational resources to efficiently train and evaluate the various deep learning architectures investigated in this study.

2.3.2. Training parameters

The training configuration for all models is summarized in Table 1.

Table 1: Training configuration

Parameter	Value
Image Size	224×224 (227×227 for AlexNet)
Batch Size	128
Number of Workers	16
Maximum Epochs	50
Learning Rate	Model-specific (5e-6 to 1e-4)
Weight Decay	1e-5
Optimizer	Adam
Scheduler	StepLR (step=5, gamma=0.1)
Loss Function	CrossEntropyLoss
Device	CUDA (NVIDIA RTX 4090)
Early Stopping Patience	8 epochs
Data Augmentation	Horizontal flips, vertical flips, rotation, color jitter

The learning rates were tuned for each model based on preliminary experiments: 5e-6 for AlexNet and ViT; 1e-5 for EfficientNet-B0, ResNet50, and the hybrid models; and 1e-4 for MobileNetV2.

2.3.3. Performance comparison

The comprehensive evaluation results across all models are presented in Table 2.

Table 2: Training history

Model	Accuracy	F1 Score (weighted)	AUC
AlexNet	78.9%	0.792	0.895
ResNet50	80.2%	0.801	0.886
MobileNetV2	79.1%	0.788	0.886
EfficientNet-B0	83.5%	0.835	0.883
Vision Transformer	81.7%	0.814	0.894
AlexNet+ViT Hybrid	81.5%	0.802	0.91
EfficientNet+ViT Hybrid	84.2%	0.847	0.927
ResNet+ViT Hybrid	85.1%	0.854	0.932

This reveals a progression in classification performance across architectural categories. Traditional CNN AlexNet demonstrated the lowest performance, while modern residual architectures, particularly EfficientNet-B0, showed substantial improvements. The hybrid models that combined CNN features with transformer attention mechanisms achieved the best overall performance, with the EfficientNet+ViT hybrid obtaining the highest metrics across all performance categories.

2.3.4. Learning dynamics

The training and validation learning curves provide valuable insights into model convergence patterns and generalization capabilities. Figures 1, 2, 3, 4, 5, 6, 7 and 8 illustrate the training dynamics of selected models.

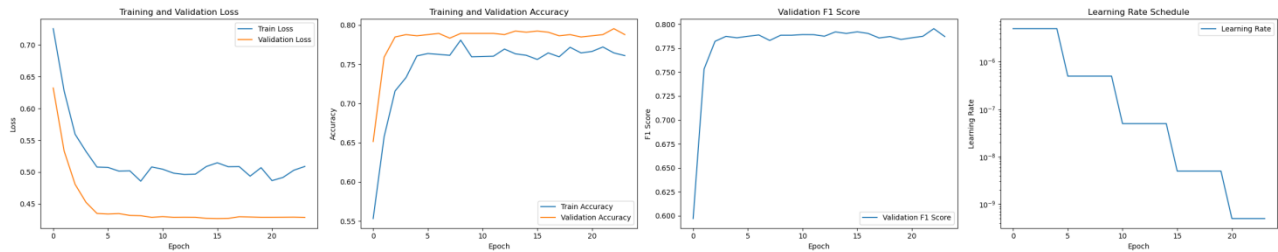


Figure 1: AlexNet training history (picture credit: original)

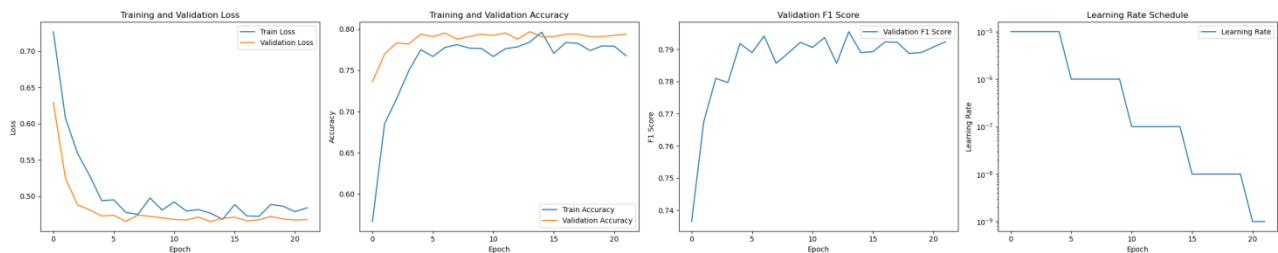


Figure 2: ResNet training history (picture credit: original)

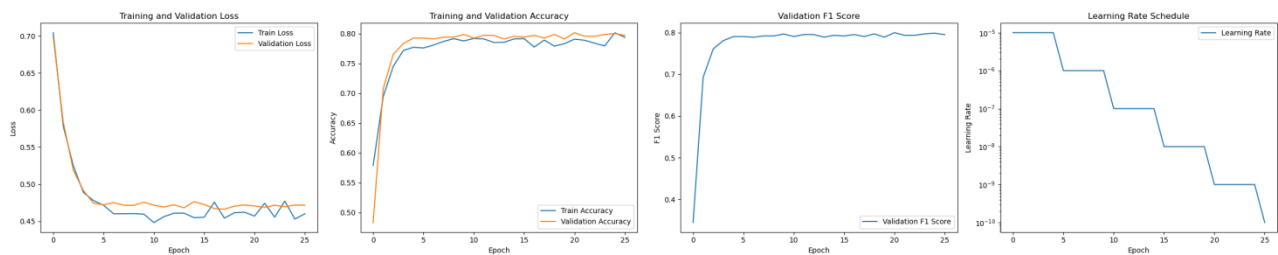


Figure 3: MobileNet training history (picture credit: original)

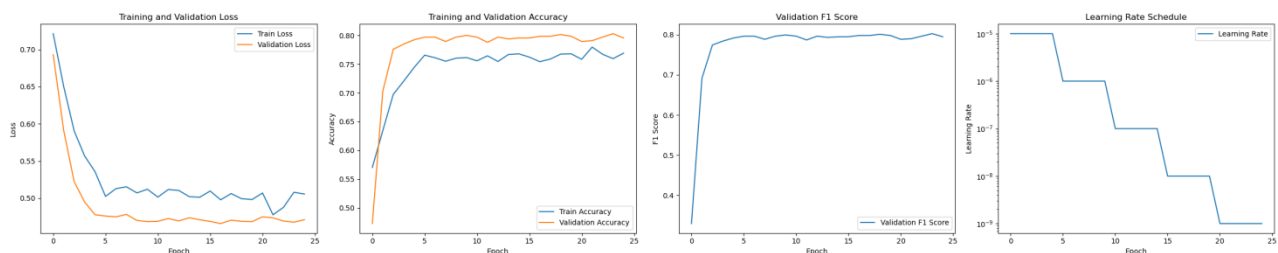


Figure 4: EfficientNet-B0 training history (picture credit: original)

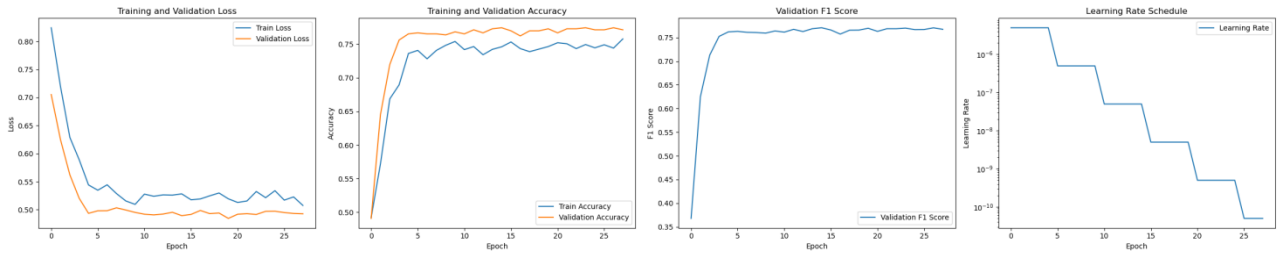


Figure 5: ViT training history (picture credit: original)

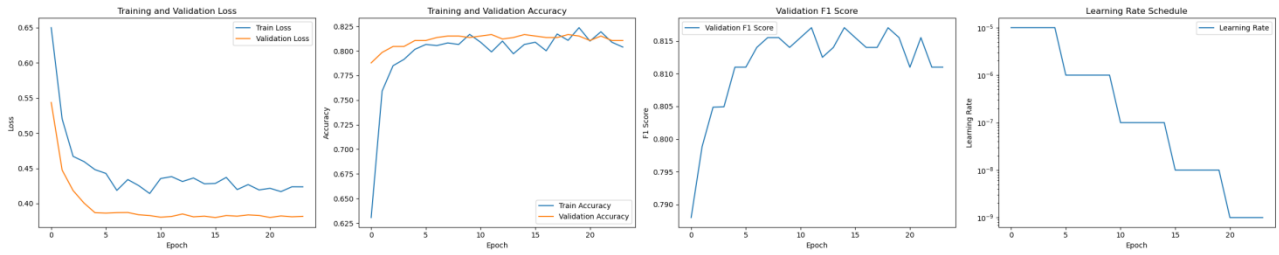


Figure 6: AlexNet&ViT training history (picture credit: original)

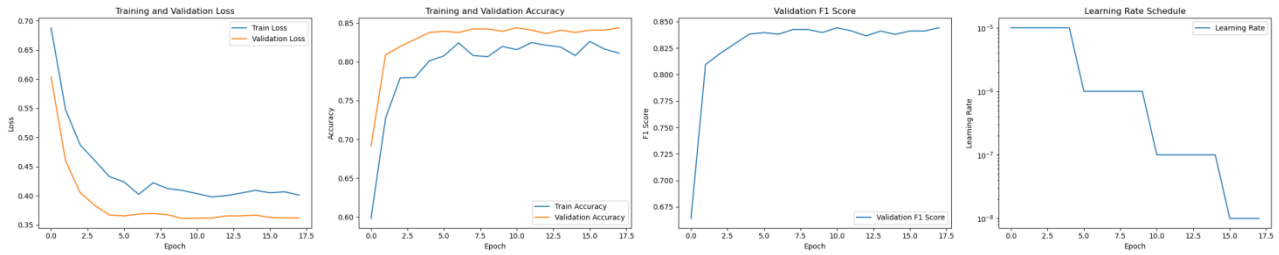


Figure 7: EfficientNet&ViT training history (picture credit: original)

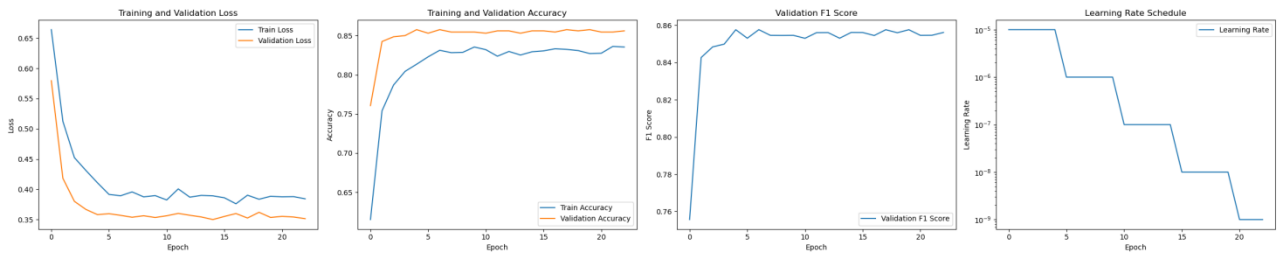


Figure 8: ResNet&ViT training history (picture credit: original)

All models exhibited a similar pattern of rapid improvement during the initial 5 epochs, followed by more gradual refinement. The ResNet+ViT hybrid model training history in Figure 8 demonstrated particularly stable convergence, with validation accuracy reaching approximately 86% and F1 score of 0.86 after about 10 epochs. The validation loss consistently decreased without significant oscillations, indicating good generalization properties.

The Vision Transformer model showed slightly slower initial convergence compared to CNN-based architectures but ultimately achieved competitive performance. This behavior aligns with findings in the literature that suggest transformers require more data to effectively learn patterns compared to CNNs with built-in inductive biases.

Notably, the gap between training and validation accuracy remained relatively small across models, suggesting that the implemented regularization strategies effectively mitigated overfitting despite the limited dataset size.

2.3.5. ROC curve analysis

The Receiver Operating Characteristic curves for representative models are presented in Figures 9 and Figure 10. These curves illustrate each model's ability to discriminate between benign and malignant lesions across different classification thresholds.

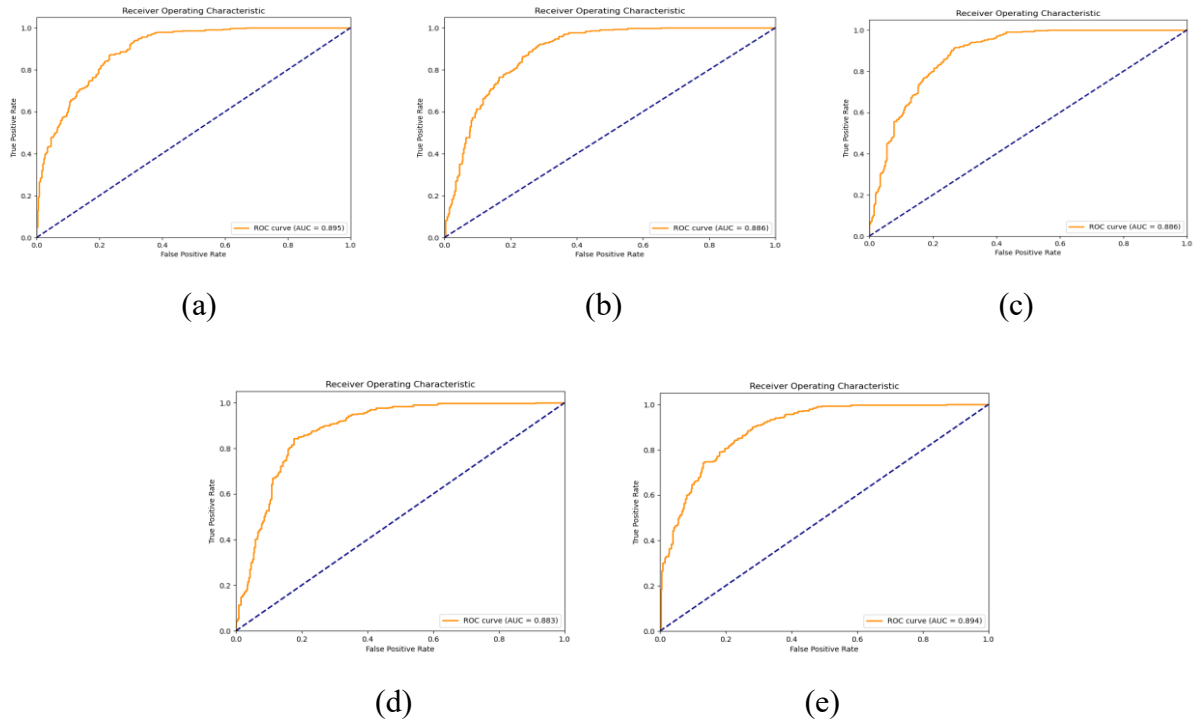


Figure 9: Roc curve (a) is the AlexNet ROC, (b) is the ResNet ROC, (c) is the MobileNet ROC, (d) is the EfficientNet ROC, and (e) is the ViT ROC

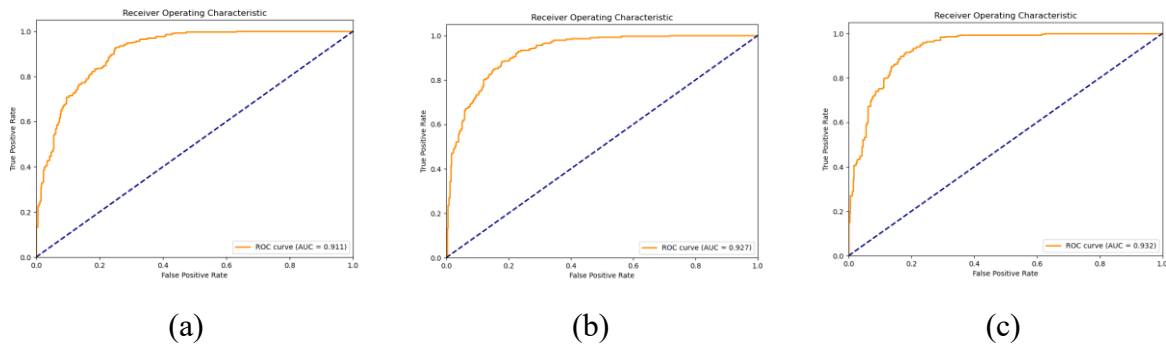


Figure 10: Hybrid model Roc curve (a) is AlexNet&ViT ROC, (b) is EfficientNet&ViT ROC, (c) is ResNet&ViT ROC

The ResNet+ViT hybrid achieved the highest Area Under the Curve of 0.932, and was followed by the EfficientNet+ViT hybrid and traditional AlexNet. The EfficientNet architecture demonstrated

the lowest AUC 0.883, though still showing reasonable discriminative ability. The ROC curves reveal that hybrid models consistently outperformed their individual component architectures, suggesting that the combination of CNN's hierarchical feature extraction with the transformer's attention mechanism creates a more robust feature representation for skin lesion classification.

2.3.6. Computational efficiency

Since classification performance is critical, analysis of computational efficiency metrics in practical deployment should be considered. Table 3 summarizes key efficiency metrics across architectures.

Table 3: Efficiency metrics

Model	Parameters (M)	Inference Time (ms)	Throughput (img/s)	Model Size (MB)	Training Time (s)
MobileNetV2	4.0	1.97	515.5	15.7	2328
EfficientNet-B0	5.8	2.29	458.7	22.7	2610
AlexNet	61.1	2.54	408.2	235.0	2540
ResNet50	27.6	3.26	306.7	100.0	2360
Vision Transformer	86.6	5.32	188.0	332.3	2959
EfficientNet+ViT Hybrid	91.9	7.26	137.7	353.1	3365
AlexNet+ViT Hybrid	147.7	7.64	130.9	565.7	3385
ResNet+ViT Hybrid	112.2	8.51	117.5	430.4	3547

MobileNetV2 demonstrated exceptional efficiency with the lowest parameter count, fastest inference time, and smallest model size, which makes it suitable for deployment on resource-constrained devices. EfficientNet-B0 provided an excellent balance between efficiency and performance, with relatively low resource requirements while achieving near-top classification performance.

The hybrid models, while offering superior classification performance, incurred significant computational costs with inference times approximately 3-4 times slower than MobileNetV2. This performance-efficiency tradeoff is an important consideration for real-world deployment scenarios, particularly in resource-limited healthcare settings.

2.3.7. Analysis and discussion

The experimental results reveal several important insights regarding the application of deep learning architectures to skin lesion classification.

Traditional AlexNet provided reasonable baseline performance but was outperformed by more modern architectures. The hierarchical feature extraction paradigm of CNN appears well-suited to capturing the visual patterns of skin lesions, but the limited depth and absence of skip connections in these early architectures constrain their effectiveness.

Residual networks especially EfficientNet-B0 demonstrated substantial performance improvements over traditional CNN. The residual connections and architectural optimizations in these models enable more effective training of deeper networks, allowing for more complex feature representation. EfficientNet-B0's compound scaling approach, which balances network depth, width, and resolution, proved especially effective for skin lesion classification, achieving 83.5% accuracy and 0.883 AUC while maintaining reasonable computational efficiency.

The Vision Transformer model performed competitively gives 81.7% accuracy, 0.894 AUC despite lacking the inductive biases of CNNs. The attention mechanism enables the model to focus on the most relevant regions of skin lesions while capturing global contextual information. However, the transformer's slower inference time and higher computational requirements present challenges for deployment in resource-constrained environments.

2.3.8. Hybrid architecture effectiveness

The most significant finding from this study is the superior performance of hybrid architectures that combine CNN and transformer features. The ResNet+ViT hybrid achieved the highest performance across all metrics. This suggests that the complementary nature of hierarchical feature extraction from CNN and attention mechanisms from transformer creates a more robust representation for skin lesion classification.

The performance improvement of hybrid models over their individual components indicates that skin lesions contain both local textural patterns best captured by CNNs and global contextual relationships best captured by attention mechanisms. The fusion of these complementary features enables a more accurate distinction between benign and malignant lesions, a finding consistent with recent research on hybrid architectures in medical image analysis [10].

2.3.9. Performance-efficiency tradeoffs

The experimental results highlight the important tradeoff between classification performance and computational efficiency. While hybrid models achieved the highest classification metrics, they also incurred the highest computational costs with inference times 3-4 times slower than MobileNetV2. This tradeoff necessitates careful consideration of deployment constraints when selecting a model architecture.

2.3.10. Clinical implications

From a clinical perspective, the ROC curve analysis provides particularly relevant insights. The highest AUC values achieved by the best models ResNet+ViT hybrid indicate excellent discrimination between benign and malignant lesions across different classification thresholds. This flexibility allows for threshold adjustment based on clinical priorities, whether maximizing sensitivity to minimize missed melanomas or optimizing specificity to reduce unnecessary biopsies.

The performance levels achieved by these models, particularly the hybrid architectures, approach the reported diagnostic accuracy of dermatologists (75-84% according to literature, [10]), suggesting potential clinical utility as diagnostic support tools. However, further validation of external datasets and prospective clinical studies would be necessary before deployment in real-world clinical settings.

3. Conclusion

This study presented a comprehensive comparative analysis of deep learning architectures for skin cancer classification, encompassing traditional CNNs, modern residual networks, vision transformers, and novel hybrid approaches. The systematic evaluation of a dataset of 3,267 dermatoscopic images yielded several important insights for automated skin lesion classification. The experimental results demonstrated a clear performance progression across architectural categories, with hybrid models combining CNN and transformer features achieving the highest classification metrics. The ResNet+ViT hybrid model attained 85.2% accuracy and 0.932 AUC, outperforming all individual architectures. This finding suggests that the complementary nature of hierarchical feature extraction in CNNs and attention mechanisms in transformers creates a more robust representation for

distinguishing between benign and malignant skin lesions. The performance-efficiency analysis revealed important tradeoffs that must be considered for practical deployment. While hybrid models achieved superior classification performance, they incurred significantly higher computational costs. MobileNetV2 offered the best efficiency with reasonable performance that makes it suitable for resource-constrained environments. EfficientNet-B0 provided an excellent balance, achieving near-top performance with moderate computational requirements. Several limitations of this study should be acknowledged. The dataset size, though substantial, could benefit from further expansion to improve generalization across diverse patient populations and imaging conditions. The binary classification approach represents a simplification of the clinical reality, where multiple skin conditions with varying degrees of severity must be distinguished. Additionally, the models were evaluated on a validation set from the same distribution as the training data, potentially overestimating real-world performance where domain shifts may occur.

Future research should explore several promising directions. Integration of clinical metadata could provide valuable context to improve classification accuracy. Extension to multi-class classification would better reflect real-world diagnostic scenarios. Investigation of model interpretability techniques would enhance clinician trust and facilitate clinical adoption by providing explanations for model predictions. Finally, prospective clinical validation studies would be essential to establish the true clinical utility of these models.

The performance levels achieved in this study, particularly by hybrid architectures, approach the reported diagnostic accuracy of dermatologists according to the literature. This suggests significant potential for these models to serve as valuable diagnostic support tools, potentially improving early detection rates and enhancing patient outcomes in skin cancer management. As deep learning technologies continue to mature and integrate with clinical workflows, they hold promising potential to augment dermatological practice and increase access to high-quality skin cancer screening in diverse healthcare settings.

References

- [1] Siegel, R. L., Miller, K. D., Fuchs, H. E., & Jemal, A. (2021). *Cancer statistics, 2021*. CA: A Cancer Journal for Clinicians, 71(1), 7–33.
- [2] Yap, J., Yolland, W., & Tschandl, P. (2018). *Multimodal skin lesion classification using deep learning*. Experimental Dermatology. Advance online publication.
- [3] Brinker, T. J., Hekler, A., Enk, A. H., Berking, C., Haferkamp, S., Hauschild, A., ... & von Kalle, C. (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.
- [4] 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.
- [5] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778).
- [6] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4510–4520).
- [7] Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning* (pp. 6105–6114).
- [8] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houshy, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations*.
- [9] Tschandl, P., Rosendahl, C., & Kittler, H. (2018). The HAM10000 dataset, a large collection of multi-source dermoscopic images of common pigmented skin lesions. *Scientific Data*, 5, Article 180161.
- [10] Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., ... & Reader Study Level-I and Level-II Groups. (2018). Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology*, 29(8), 1836–1842.