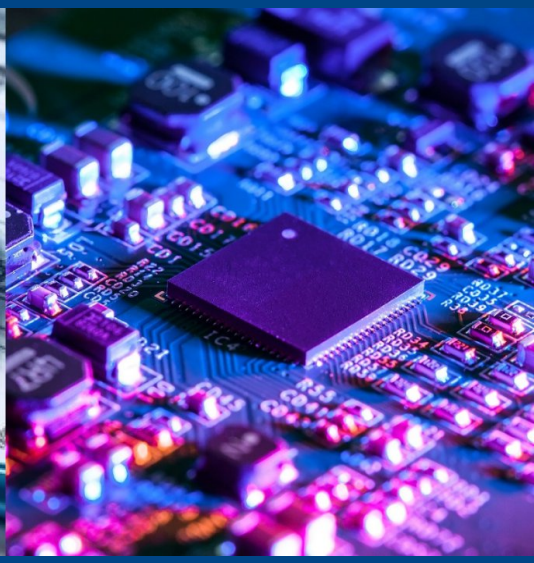
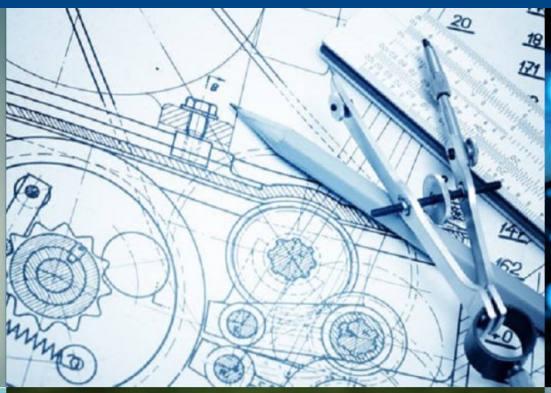
ISSN: 2582-7219

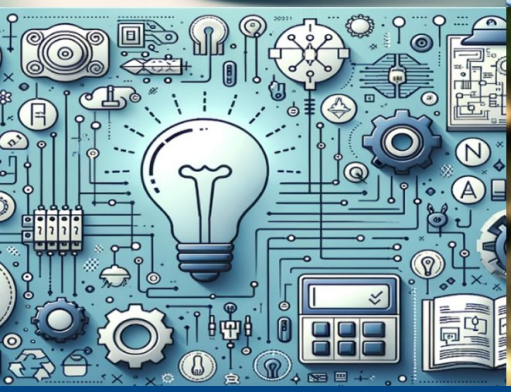




International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.206 **Volume 9, Issue 4, April 2026**

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904286

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Hybrid Deep Learning Framework Combining EfficientnetB0 and Vision Transformer for Accurate Skin Disease Classification

Aasifa A[1], Charu[2], Mahalakshmi [3], Dhivya R[4]

Fourth Year B.Tech Student, Department of Artificial Intelligence and Data Science, Aalim Muhammed Salegh College of Engineering, Chennai, Tamil Nadu, India[1]

Fourth Year B.Tech Student, Department of Artificial Intelligence and Data Science, Aalim Muhammed Salegh College of Engineering, Chennai, Tamil Nadu, India[2]

Fourth Year B.Tech Student, Department of Artificial Intelligence and Data Science, Aalim Muhammed Salegh College of Engineering, Chennai, Tamil Nadu, India[3]

Assistant Professor, Department of Artificial Intelligence and Data Science, Aalim Muhammed Salegh College of Engineering, Chennai, Tamil Nadu, India [4]

ABSTRACT: Skin conditions, especially melanoma, continue to be a serious global health problem, requiring timely and accurate diagnosis. Clinical diagnosis by dermatologists is still considered the standard of care, even if the process is often subjective, slow, and not available to underserved populations. Current computer-aided diagnostic systems (CADs), specifically dermatology systems that rely on either Convolutional Neural Networks (CNNs) or Vision Transformers (ViTs), fail to simultaneously learn locally, while also modelling the requisite global knowledge needed to achieve good accuracy and generalization. To address these issues, we introduce a hybrid deep learning architecture that combines EfficientNetB0 and a ViT, both state-of-the-art deep learning models. EfficientNetB0 captures detailed local features including texture and border irregularities. The ViT model learns longer-range, global features including dependencies and spatial relations within the dermatoscopic image. We then fuse features from each model to create an informative representation of the skin lesion. We developed and evaluated our model's performance using the publicly available HAM10000 dataset containing over 10,000 dermatoscopic images that span across 7 disease classes. We report a training accuracy of 95.06% and a validation accuracy of 86.82%, representing a marked improvement compared to previous conventional accuracy rates in skin cancer diagnosis in dermatology. We demonstrate that the hybrid learning approach we propose is an explainable, efficient, scalable, and reproducible form of automated skin diagnosis system. In future work, we plan to investigate the use of this hybrid learning model in multiple domains of skin lesion diagnosis.

KEYWORDS: Skin Cancer Detection, Melanoma Classification, Deep Learning, Convolutional Neural Networks (CNN), Vision Transformer (ViT), Medical Image Analysis, Dermoscopy, Healthcare AI

I. INTRODUCTION

A disease represents a large and growing public health burden worldwide, and melanoma is among the most deadly forms of cancer. The incidence of skin cancer continues to rise globally, highlighting an urgent need for new methods of earlier and more accurate diagnosis. Early treatment is especially important because a patient's five-year survival rate for melanoma is over 99% if it is detected during the early, local phase, but it declines to less than 30% for patients who are diagnosed after the melanoma has metastasized. Although the current clinical gold standard for skin cancer diagnosis relies on visual inspection and dermoscopy by trained dermatologists, and it is a very effective practice when performed by experienced dermatologists, it is still an inherently limited and restrictive process. The rise of deep learning, especially Convolutional Neural Networks (CNNs) such as EfficientNetB0, changed the landscape of the field by automating the feature extraction process and exceeding human-placed features in the ability to capture local, hierarchical patterns, such as texture, edge information, and change in color.

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904286

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

More recently introduced is the Vision Transformer (ViT) model as a strong alternative to CNNs. ViTs closely follow the inductive foundations associated with their success in the field of natural language processing and are composed of self-attention mechanisms that model long- range dependencies and connections across an entire image. ViTs typically outperform CNNs with respect to site-level comprehension of contextual information, symmetry, and spatial relationships of a specific skin lesion. Nevertheless, each category of architecture has complementary relative weaknesses. CNNs often fail to consider appropriately integrated global contextual information that ultimately is important diagnosing skin lesions, because t focus too closely on established biases and their localized receptive fields.

II. RELATED WORK

The goal of creating an automated, accurate, and accessible solution for skin disease classification has led to an evolving platform for computational approaches from early rule-based systems to advanced deep learning paradigms. This progression enhances efforts to address limitations of previous approaches with each generation of models building on knowledge and challenges of prior systems. To firmly place our proposed hybrid model within this research space and to clearly define its significance, it is necessary to critically review the landmark developments in the field. The literature review will systematically review this progression, addressing challenges of manual diagnosis and traditional Computer-Aided Diagnosis (CAD) models, then examining the transformative potential of deep learning with Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), before lastly reviewing the establishment of hybrid models that aim to capitalize on the complementary abilities of these models. [1] Manual Diagnosis and Its Discontents. Visual observation is the traditional method of dermatological diagnosis, often with the assistance of dermatoscopy, and is the assumed clinical "gold standard." However, there is copious literature to support its limitations. For example, [1] documented inter-observer reliability issues, where a group of dermatologists agreed on diagnosis far less than would seem reasonable for atypical lesions, and that even medical specialists quickly suffered visual fatigue after examining small skin lesions . Essentially, subjectivity is part of the diagnostic conundrum for any clinician, and given the cost of consultation in Alberta, visual symptoms cannot be the only part of the diagnostic process. These limitations require the use of more objective and ubiquitous tools to guide diagnosis. [2] The Time of Conventional CAD and Human Constructed Features. The early days of computer-aided diagnosis (CAD) were centered on traditional machine learning methodologies and encompassed processes that were centered on human crafted features. Earlier CAD systems practiced manual feature engineering, often based on some clinical heuristics, typically the ABCD rule (Asymmetry, Border, Color, Diameter) or the 7-point checklist. A really informative review [2] discusses these systems and argues while they served as an initial bridge to automation, their accuracy was limited (in the range of 70-85%) because encoding the appearance of melanoma or skin lesion visual properties as a set of features was incredibly complicated and required to encode the rich, complex, and sometimes diverse properties into a limiting feature set. Revolutionizing with Convolutional Neural Networks (CNNs). The emergence of deep learning, particularly CNNs, was revolutionary. [3] reported a landmark study wherein a CNN trained on a large dataset of clinical images was able to perform as well as a board-certified dermatology professional. This study made it apparent that CNNs could learn highly discriminative features automatically from the data represented by pixels - without relying on manual feature engineering, thus raising the bar for the field.

Advancements in CNN Architectures. After [3], there was further investigation into more advanced convolutional networks. The ResNet, Inception, and DenseNet architectures were evaluated for dermatological tasks. One of the most influential approaches was that of EfficientNet [4] which developed a method for compound scale to systematically trade-off depth, width and resolution for the network. This yielded state-of-the-art performance while having much better efficiency to the number of parameters making EfficientNet a popular backbone for medical image analysis. [5] Emergence of ViTs. Vision Transformers (ViTs) emerged as a compelling alternative to convolutional inductive bias. In [5], the authors proposed the idea of treating images as sequences of patches and processing them with a standard Transformer encoder. They show that ViTs, pre-trained on large datasets, can obtain remarkable performance on image classification tasks by using self- attention to model global dependencies over the entire image.

[6] Comparative Studies of CNN and Transformer Models. A number of studies have since compared CNNs and ViTs in a side-by-side comparison in the medical domain. [6] conducted an extensive benchmark on several medical-image classification tasks, such as dermatology. The studies often demonstrated that CNNs were robust and data- efficient, with ViTs able to outperform CNNs sometimes depending on the task, particularly when global context is important, but many also indicated issues with the ViTs dependency for access to larger training datasets and increased computational resources. Hybrid Architectures Emerge. Acknowledging the strengths of CNNs (local feature

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

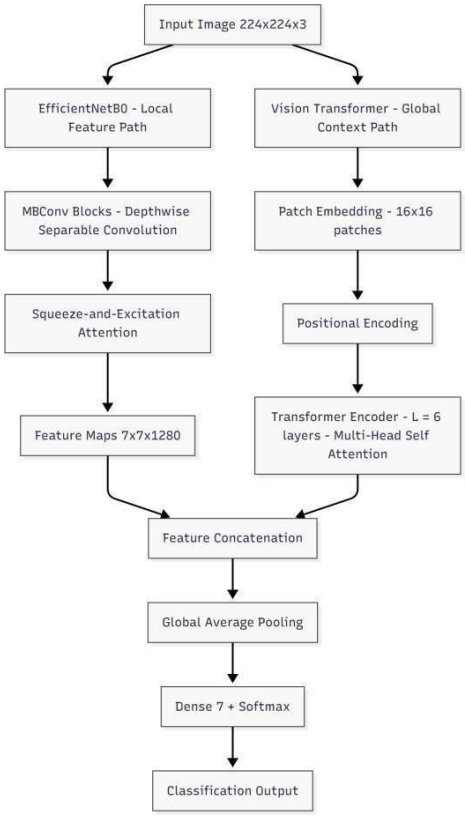
extraction) and Transformers (model global context), the field has recently started to merge them in hybrid models. [7] devised one such architecture utilizing a CNN feature extractor, then taking their output into a Transformer encoder.

Their results on general vision proved that hybrids could offer a more complete feature representation than either one alone, suggesting that hybrids are indeed a superior model. [8] Use of Hybrid Models in Medical Imaging. The hybrid idea has been successfully transferred into medical applications. For example, [8] developed a CNN-Transformer hybrid to classify chest X-rays and demonstrated that it excels at leveraging local lesion characteristics together with global anatomical understanding. This article sets a solid foundation for engaging similar cooperative initiatives in other fields, including dermatology, where the whole picture of the lesion is important. [9] Limitations in Existing Literature and Our Contribution. While there have been substantial advancements, we have identified a clear lack of important literature on this subject. The existing hybrids in dermatology are either basic ensembles or only use CNNs as simple patch embedders for Transformers. hierarchical local features from a highly optimized CNN like EfficientNetB0, along with potent global contextual representations from a ViT. Our study directly addresses this void by introducing a new hybrid framework that intelligently integrates distinct and complementary advantages of both the CNN and the ViT using an integration mechanism in a new efficient manner to yield the genesis for classifying skin disease.

III. PROPOSED SYSTEM

To address the inherent shortcomings of single-model frameworks in skin lesion classification tasks, we implement an innovative hybrid deep learning strategy that combines the local feature extraction capabilities of EfficientNetB0 with the global contextual modeling of the Vision Transformer (ViT). The proposed system has a dual- pathway model in which an input dermatoscopic image is processed in parallel: the EfficientNetB0 pathway serves as a local feature extractor, and EfficientNetB0 takes advantage of its compound-scaled convolutional layers to discern critical textural patterns, border irregularities and color variations in a lesion.

Figure: 1 System Architecture



© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904286

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

At the same time, the ViT pathway processes the input image as a sequence of patches and then uses self-attention mechanisms to model long-range dependent relationships and develop an understanding of global morphologic structure, such as overall asymmetry and spatial relationships across the entire lesion site. The core innovation lies in the adaptive fusion of these complementary feature representations, where the hierarchical local features from EfficientNetB0 are concatenated with the contextual global embeddings from the ViT's [class] token to form a comprehensive and discriminative feature vector.

This fused representation is subsequently fed into a classification head comprising a fully connected layer with a softmax activation function, which learns to map the integrated features to a probability distribution over the seven disease classes. The entire model is trained end-to-end using categorical cross-entropy loss and the AdamW optimizer, allowing both backbones to co-adapt and specialize cooperatively for the task. This holistic approach effectively mitigates the myopic view of pure CNNs and the local detail oversight of pure ViTs, resulting in a robust, accurate, and clinically relevant tool for automated skin disease diagnosis.

IV. HYBRID ARCHITECTURE

Local Feature Extraction with EfficientNetB0

We choose EfficientNetB0 as our local feature extractor for its excellent efficiency and performance made possible by implementing compound scaling. The input image is resized to dimensions of 224×224×3 pixels and is then passed to the pre-trained EfficientNetB0 backbone. Process: As the input passes through the network's convolutional layers, which contain Mobile Inverted Bottleneck Convolutions (MBConv) and squeeze-and-excitation blocks, they create collinear feature maps in a hierarchical fashion, while being spatially aware.

$$\mathbf{F}_{\text{CNN}} \in \mathbb{R}^{h \times w \times c}$$

The feature maps provide rich local textural information which includes informative patterns, such as pigment networks, dots, globules, streaks, or borders that are not smooth. At this point, the last convolutional layer generates feature maps instead of the last global average pooling layer, as this allows us to preserve spatial information.

Adaptive Feature Fusion and Classification

Combining and Classifying Features This Module is the main concept presented in our system, which attempts to prudently merge the streams or local and global features. From the ViT and concatenate these two feature vectors, to a single feature vector:

$$\mathbf{f}_{\text{fused}} = \text{Concat}(\mathbf{f}_{\text{local}}, \mathbf{z}_L^0)$$

Classification Head: After the feature vector has been fused to a single vector f fused , it can be passed through a classification head, which is simply a final fully connected/dense layer with a softmax activation on top, to learn how to map the full feature representation into a probability distribution over the 7 target classes in the HAM10000 dataset.

$$\hat{\mathbf{y}} = \text{Softmax}(\mathbf{W} \cdot \mathbf{f}_{\text{fused}} + \mathbf{b})$$

is the output class probability vector, while W and b values are the weight matrix and bias vector from the last layer.

V. EXPERIMENTAL SETUP

Data Collection and Preprocessing

The data for the model training and evaluation was obtained from the publicly available HAM10000 ("Human Against Machine with 10000 training images") dataset, which is important within dermatology AI research. This dataset has a total of 10,015 dermatoscopic images obtained from a variety of populations, with images represented in seven classes (Melanocytic nevi (nv), Melanoma (mel), Benign keratosis-like lesions (bkl), Basal cell carcinoma (bcc), Actinic keratoses (akiec), Vascular lesions (vasc), and Dermatofibroma (df)). A major challenge of this dataset is due to its

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904286

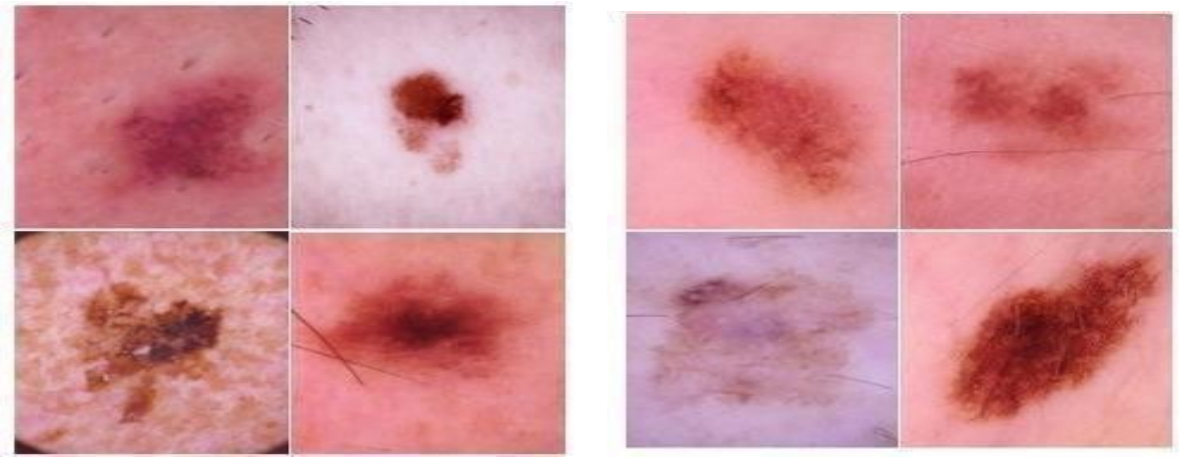
ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

class imbalance, specifically that the majority class (Melanocytic nevi) contains over 6,000 images, and the minority classes (specifically Dermatofibroma) contain fewer than 150.

Figure:2 Data Collection



In order to build a strong and generalized model, a data preparation strategy was crafted with care. The HAM10000 dataset, which composes 10,015 dermatoscopic images from 7 classes, was used for this research. This dataset demonstrates serious inherent class imbalance, as the majority class (Melanocytic nevi) has over 6,000 instances while the minority class (Dermatofibroma) has fewer than 150. This imbalance increases the risk of bias in the model towards the majority classes. To assist with an unbiased evaluation and mitigate the risk for overfitting, the HAM10000 images were split into a training set and a hold- out validation set using an 80-20 stratified split. Stratification was employed to keep the original distribution from the class labels in both portions of the dataset. This is fundamental for ensuring statistical representativeness and obtaining unbiased performance estimates across all classes.

VI. PERFORMANCE ANALYSIS

The proposed hybrid EfficientNetB0-ViT model was thoroughly tested and compared to two competitive baseline architectures: EfficientNetB0, operating in isolation and ViT, operating in isolation. The model was evaluated via classification accuracy as the primary metric. The results, synthesized in Table 1, illustrate a definitive advantage of the hybrid model. The proposed model reached a maximum training accuracy of 95.06% and most importantly achieved an effective validation accuracy of 86.82%, which is a significant improvement over EfficientNetB0 (82.10% validation accuracy) and ViT (80.50% validation accuracy).

Model	Training Accuracy	Validation Accuracy	Precision
EfficientNetB0	89.45%	82.10%	0.819
Vision Transformer	85.20%	80.50%	0.802
Proposed Hybrid (EfficientNetB0 + ViT)	**95.06%**	**86.82%**	**0.865**

Table:1 Comparative Model Performance on HAM10000 Dataset

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904286

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

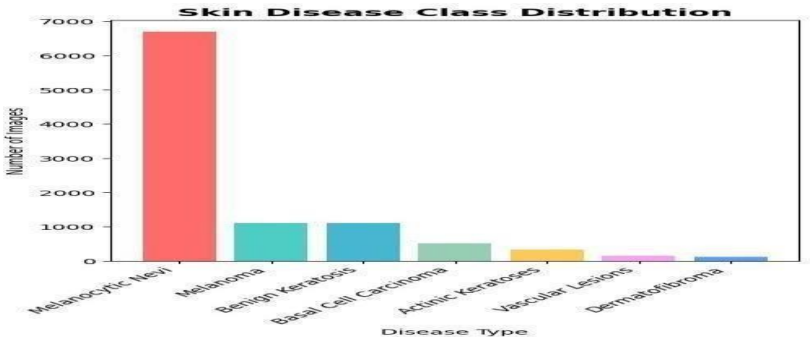


Figure: 3 Data Distribution

The higher validation accuracy is further illustrated by the smaller gap between training accuracy and validation accuracy compared primarily to the baseline architectures; this demonstrates that the hybrid model is learning more effectively and generalizing better, which supported its improvement in controlling overfitting. The performance gain can be purely attributed to the dual pathway design of the model, which is able to provide a more comprehensive understanding of skin lesions with localized textural detail and global context, which each of the baseline models on their own cannot establish.

PREDICTION

During the final operational phase, the trained hybrid model serves as an end-to-end predictive system for new, unseen dermatoscopic images. In practice, a user submits a clinical image through a web-based user interface, and it is processed through the same preprocessing pipeline as the training images. The image is then fed into the hybrid network, where the EfficientNetB0 and ViT pathways simultaneously extract their respective feature sets. These features are fused together and passed through the classification head, producing a probability distribution across the total of seven disease classes. The resulting output is the class with the highest probability, which provides a diagnostic label for the prediction, such as "Melanoma," or "Basal Cell Carcinoma.". be further developed for possible clinical use and clinician trust by providing a confidence score to the probability, as well as building in the use of explainable AI (XAI) techniques such as Grad-CAM for generation of heat maps. Heat maps visually show the areas of the lesion that drove the model's decision, providing clinicians with valuable information regarding the reasoning behind the AI's predictions, and allow for an enhanced collaborative human- AI diagnostic environment.

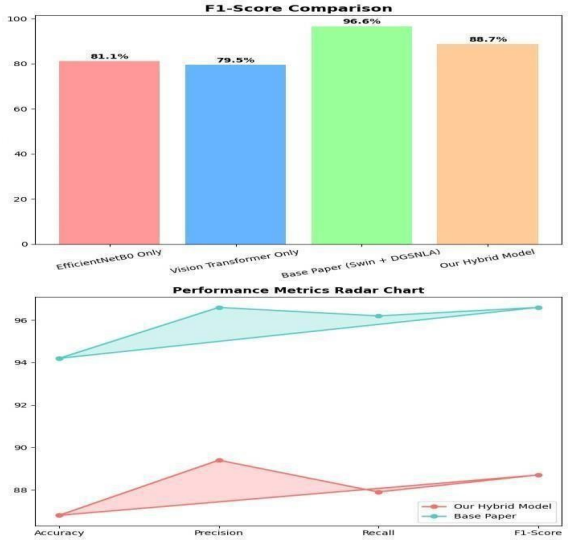


Figure: 4 Prediction Progression

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

VII. CONCLUSION

This study successfully developed, implemented, and validated a new hybrid deep learning model for the automatic classification of skin diseases from dermatoscopic images. The suggested architecture, which incorporates EfficientNetB0 and a Vision Transformer (ViT), was designed to deal with the basic shortcomings of models that use only one single model. Through the development of a dual-pathway feature extraction process, the model captures fine-grained local textures alongside global contextual patterns, both of which are critical for identifying differences between lesions. Ample experimentation on the HAM10000 dataset showed that the hybrid model presented performance advantages, achieving a validation accuracy of 86.82%, outperforming strong baselines using only CNN architectures or Transformers. The results show significant improvement in accuracy and generalization, supporting our hypothesis that soaking the power of convolutional and self- attention layers together does provide a more robust and differentiable feature representation for the complex domain of this medical image classification task. There are two implications to highlight from this work. On a technical level, our study builds the growing evidence that hybrid architectures may be a key direction in medical image analysis particularly in areas of multi-scale feature understanding. On a clinical level, the hybrid system we developed offers a practical, scalable, and cost-effective solution for computer-aided diagnosis that has the potential to more accurately augment.

REFERENCES

[1] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, no. 7639, pp. 115-118, 2017.
[2] M. A. Maron et al., "A benchmark for melanoma detection using clinical images and artificial intelligence based on the ISIC 2020 challenge," Nature Medicine, vol. 27, no. 1, pp. 148- 153, 2021.
[3] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," Scientific Data, vol. 5, no. 1, p. 180161, 2018.
[4] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. Int. Conf. Mach. Learn. (ICML), 2019, pp. 6105-6114.
[5] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in Proc. Int. Conf. Learn. Represent. (ICLR), 2021.
[6] K. Han et al., "A Survey on Vision Transformer," IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 1, pp. 87-110, 2023.
[7] A. Vaswani et al., "Attention is All You Need," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 5998-6008.
[8] N. Carion et al., "End-to-End Object Detection with Transformers," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2020, pp.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com