INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 5, May 2026 | ISSN: 2456-4184 | IJNRD.ORG

End-to-End Cognitive Deep Learning System for Dermatological Anomaly Profiling

[1]MS.V.MATHUMITHA, [2]CH.SRINIVAS , [3]B.LAKSHMI GOPI REDDY , [4]B.MAHAMMAD RAFI

ASST.Professor(CSE), UG Scholar, UG Scholar, UG Scholar

Department of Computer Science & Engineering

Bharath Institute of Science and Technology, BIHER

173,Agaram Road , Selaiyur, Tambaram, Chennai, Tamil Nadu, India

Abstract

Dermatological diagnosis is usually performed through a visual inspection of skin lesions. However, it is a complex and error-prone task because of the similarity in skin lesions. In this paper, we propose a novel extended deep learning architecture that is used to effectively classify skin diseases while providing a clear understanding and insights. In our proposed architecture, we used a combination of Convolutional Neural Networks and Vision Transformers in a sequential manner. In our architecture, a Convolutional Neural Network is used to effectively extract informative features from dermoscopic images. These features are very important in characterizing skin diseases, including their color distribution, texture, and structural patterns. The informative representation is then passed to the Vision Transformer, which effectively models global contextual relationships in the image to perform multi-class dermatological classification. In order to ensure transparency in the decision-making approach of the model, Gradient-weighted Class Activation Mapping is used to ensure that a heatmap is generated to highlight the regions of the image that are most important to the decision-making approach. Furthermore, a generative AI module is also integrated using the Gemini API to ensure that informative insights are provided to the end-user regarding the skin condition that is to be diagnosed. Overall, the framework is designed to ensure that a reliable diagnostic support tool is provided to clinicians to help diagnose dermatological diseases efficiently.

IndexTerms - Dermatological Diagnosis, Deep Learning, Convolutional Neural Networks, Vision Transformers, Hybrid Architecture, Dermoscopic Image Analysis, Explainable AI, Grad-CAM, Generative AI, and Clinical Decision Support Systems.

I.INTRODUCTION

Skin diseases are among the most prevalent medical condi tions globally, affecting various populations of different ages. There are various skin diseases, particularly skin cancers like melanomas, which pose serious health hazards to the affected populations owing to the rapid progression of the diseases. Dermatologists usually employ dermoscopy to examine the diseases based on the visible features of the diseases. Despite the use of dermoscopy to improve the accuracy of the diagno sis of skin diseases, it is difficult to detect the diseases at an early stage since the features of the diseases, particularly the malignant diseases, are similar to the features of the benign diseases. The shortage of dermatologists is another reason for the need to develop an automated system for the diagnosis of skin diseases. The most recent developments that have been made in the f ield of Artificial Intelligence, specifically in the sub-field of Artificial Intelligence referred to as Deep Learning, have had P Arjun and J Kartheek both are first authors of this paper. a significant impact on the field of Medical Image Analy sis.Convolutional Neural Networks (CNNs) have been success fully used for the classification of dermoscopic images because of their inherent capability to learn hierarchical representations of the images directly from the raw images fed to them. CNNs have multiple convolutional layers that enable them to learn important features such as the boundary, pigmentation, and texture of the structures within the images fed to them. It has been proven through previous research that such systems are capable of producing high classification results, almost comparable to those of a dermatologist. However, CNNs are mainly focused on local patterns, and the context of the entire image is usually ignored by them, making them less efficient for such complex issues. To overcome these limitations, a new powerful model has been proposed recently, referred to as Vision Transformers (ViTs).This model has been proposed as an alternative to traditional models used for computer vision tasks. ViTs are motivated by a traditional model referred to as the transformer model. This model was originally proposed for use in natural language processing tasks. The traditional model works by splitting images into smaller parts or patches. The model then makes use of an attention model to relate different parts of an image. This is significant because, during a dermatological analysis, relating different parts of an image can lead to a more accurate classification. The major drawback associated with these models is that they require a large amount of training data, which is a challenge when performing medical image analysis. Based on the relative merits of both methods, hybrid models that combine both CNNs and Vision Transformers have also been explored. The former are used to obtain detailed local features of images, while the latter are used to identify the global relationships between these local features. Following this idea, a novel hybrid deep learning model is proposed for automated classification of skin diseases. The proposed model makes use of a CNN to obtain discriminative visual features such as texture, shape, and color of images. The features are then classified using a Vision Transformer.

II. LITERATURE SURVEY

Dermatological image analysis has witnessed tremendous development in the last few years with the advent of computer vision and artificial intelligence techniques. Skin lesion clas sification is a very complex task due to the huge variation in skin lesions and the similarity between benign and cancerous diseases, and the noise in the image. In the traditional ap proach, dermatologists diagnose diseases by visually analyzing the lesions, including characteristics such as pattern in color, irregularity in borders, and structures in texture. Even though this approach is effective when implemented by experienced dermatologists, it is a very time-consuming and subjective approach. This has given rise to the development of automated systems that can aid in the detection and classification of skin diseases in an effective manner. It is also important to note that a literature survey is essential to understand the developments made so far in the research area, to identify the limitations of previously used techniques, and to motivate new techniques. In the area of skin lesion image analysis, research has progressed from traditional ma chine learning techniques to modern deep learning techniques. The traditional machine learning techniques used to classify skin lesions were based on handcrafted features learned from dermoscopic images. The features learned were then used to train traditional machine learning classifiers such as SVM, k-NN, or random forests. Even though these techniques are based on machine learning, they were not able to achieve good results because handcrafted features were not able to learn all the complexities of skin lesions. The advent of deep learning has revolutionized medical image analysis. Convolutional Neural Networks (CNNs) were used to automatically learn features, which enabled images to directly learn hierarchical visual patterns. CNN-based ar chitectures have been observed to have better performance in recognizing skin lesions by successfully learning local spatial patterns. Vision Transformers (ViTs) have also been explored for various computer vision tasks because they can model long-range dependencies and global contextual relationships between images using an attention-based model. This is ben eficial for dermatology because small structural changes over an entire image can have a significant impact. The literature survey presented in this section gives a brief overview of various developments that have taken place in the field of automated dermatological diagnosis using machine learning and deep learning. It is clear that various studies have been carried out to understand the need to implement accurate and interpretable machine learning and deep learning models. These studies have been taken into consideration to propose a hybrid model that combines CNN and Vision Transformer architectures and implements visualization techniques such as Grad-CAM. Classical Machine Learning Approaches in Dermatolog ical Diagnosis In recent years, before the emergence of deep learning tech niques, various machine learning algorithms and techniques have been used to develop automated dermatological analysis systems. In these studies, various features have been extracted to understand dermatological images. Celebi et al. (2007) pro posed a system that used color histogram and texture features to classify malignant and benign skin lesions. In their study, a Support Vector Machine classifier was used to classify skin lesions. The study demonstrated that computational analysis can be used to help dermatologists in identifying skin lesions. However, it was also clear that the system failed to address complex structures of skin lesions. Barata et al. (2010) extended the classification of features using a combination of color and texture features with k-NN and Random Forest classifiers. This study also emphasized the significance of features in the extraction of significant dermatological features. Although it was a successful attempt in improving classification accuracy using a machine learning system instead of manual observation, it remained prone to changes in lighting, shape, and noise in images. Another study was carried out by Pathan et al. (2012) to assess the effectiveness of Gabor and LBP features in feature representation of texture characteristics in dermoscopic images. The classification of features was performed using a combination of SVM and Decision Tree classifiers. It was revealed that texture features are significant in improving clas sification accuracy of different classes of diseases. However, it remained prone to difficulties when it came to classification of less common diseases due to their ambiguous characteristics. Moreover, texture characteristics should be easily distinguish able on the surface of the skin to ensure accurate classification using texture features. Another significant contribution was proposed in a paper by Abbas et al. (2015). The paper proposed a combination of various feature classes and a probabilistic neural network. This was a hybrid approach to feature development, which contributed to an increase in classification accuracy. However, these conventional machine learning systems had limitations when it came to scalability and flexibility because of their dependency on manual features and processing techniques. In general, the earlier studies on machine learning con tributed to important insights into the automated diagnosis of dermal diseases. However, the limitations of the earlier studies on machine learning, which relied on handcrafted features, showed the need to develop more advanced forms of machine learning to derive meaningful features from the images. Deep Learning Approaches in Dermatological Diagnosis The emergence of deep learning has caused a major shift in the field of skin disease classification. In fact, the Convolu tional Neural Networks (CNNs) have become the most popular form of machine learning algorithms because of their ability to automatically derive meaningful features from dermoscopic images. Among the most important studies on the application of deep learning to skin disease classification was the work done by Esteva et al. in 2017. In the work, the authors proposed a deep CNN model to classify a dataset of more than 129,000 clinical images of skin diseases, which included more than 2,000 skin conditions. The proposed model showed the ability of the network to classify skin diseases as well as the ability of the network to perform as well as a dermatologist in the detection of melanoma and other skin diseases. Kawahara et al. (2018) suggested the patch-based CNN for dermoscopic image classification. Instead of directly pro cessing the entire image, the model processed small image patches and made predictions based on the combination of the predictions for the patches to make the final decision for the image classification. This helped in the better detection of texture patterns and structural irregularities in the image. However, the CNN models have been observed to concentrate more on the local image characteristics and might fail in the detection of global characteristics such as the asymmetry of the image or the distribution of the pigmentation in the image. To overcome the shortcomings of the CNN models in image classification tasks, transformer models have been incorpo rated in recent studies in the computer vision domain. Doso vitskiy et al. (2020) suggested the Vision Transformer (ViT), which uses self-attention for processing the patches in the image and helps in the detection of the image characteristics over the entire image. The transformer models have been successfully implemented in the medical image classification tasks as well. Recently, there has also been research on the development of a hybrid model that combines the use of CNNs and a transformer model for the purpose of feature extraction. Chen et al. (2022) came up with a hybrid model that combines the use of a CNN feature extraction model and a Vision Transformer model. The model was able to improve the accuracy of the classification of images into multiple classes. Recently, there has also been research on the development of a hybrid model that combines the use of CNNs and a transformer model for the purpose of feature extraction. Chen et al. (2022) came up with a

hybrid model that combines the use of a CNN feature extraction model and a Vision Transformer model. The model was able to improve the accuracy of the classification of images into multiple classes. Aside from the improvement of the accuracy of deep learn ing models, recent research has also focused on the importance of the use of interpretability when developing deep learning models. Selvaraju et al. (2017) came up with a technique referred to as Gradient-weighted Class Activation Map (Grad CAM). Grad-CAM is a technique used to visualize the parts of an image that are responsible for a prediction made by a deep learning model. Grad-CAM has been used to visualize CNN models used to classify medical images. Following this research, Chattopadhay et al. proposed Grad CAM++, which is a more refined version of the model, allow ing for more accurate localization maps to be created, espe cially when dealing with more complicated images containing a variety of relevant regions. In the field of dermatology, visualization methods such as the ones described are incredibly important, as they allow the verification of the model's focus on the medically relevant areas of the image, as opposed to the irrelevant areas in the background. To conclude, it is evident that the progression from classical machine learning to deep learning has resulted in a more effec tive form of automated diagnosis in the field of dermatology. CNNs have proven to be effective in local feature extraction, Vision Transformers have proven to be effective in global contextual relationships, and the hybrid model has proven to have the potential to enhance the classification performance of the model. Additionally, the introduction of Grad-CAM has resulted in a more transparent form of decision-making through the use of deep learning.

III. RELATED WORK

It has been noted that more emphasis has been provided to the application of AI in analyzing dermatological images on the basis of the availability of a large number of dermoscopic images and the development of deep learning techniques. Various techniques have been proposed in the literature for the automatic classification, segmentation, and detection of skin lesions on the basis of various machine learning techniques. The most popular techniques in this regard are the application of Convolutional Neural Networks (CNN). In addition to this, recently, transformer models and hybrid models on the basis of the application of CNN and Vision Transformers (ViTs) are gaining popularity in analyzing dermatological images on the basis of their capability to analyze local and global patterns. This section is based on the most popular techniques in the application of CNN, transformer models, and emerging explainable AI and generative AI models A. CNN-based Dermoscopic Image Analysis The first set of studies performed on the classification of skin diseases using automated techniques was based on the implementation of CNN models due to their potential to learn features from images. The CNN models have the potential to learn various features from images, including spatial features such as texture, color changes, and irregularities on the bound ary of images. The study on the classification of skin diseases using an optimized CNN model was performed by Musthafa et al. (2024). In this study, an optimized model was proposed to classify the diseases using the HAM10000 dataset. The model was subjected to rigorous data augmentation techniques for the classification process. The model has performed well in terms of precision, recall, and F-score. Surveys have also been conducted in order to demonstrate the importance of the CNN model in the analysis of the skin lesion. Mirikharaji et al. (2023) have conducted a survey based on more than 170 articles regarding the analysis of the skin lesion segmentation and classification. The successful techniques that have been utilized in the analysis of the skin lesion have been focused on using the CNN model for the feature extraction and the region localization. The benefits that can be gained from using the techniques include the ability to utilize the transfer learning using the available models such as ResNet, VGG, and EfficientNet using the small dataset. Despite the advantages of CNN-based techniques, there are some limitations to CNNs. The local receptive fields of CNNs may limit the global contextual relationship between the entire lesion. In addition, the lighting conditions, resolution of images, and appearance of lesions may affect the classi f ication results. Class imbalance and large variations within the dermatology dataset are still challenging for CNN models. Although there are some studies indicating the dermatologist level accuracy of CNNs on specific tasks, the consistency of CNNs on diverse lesions is still an emerging topic. B. Vision Transformer and Hybrid Approaches With the recent advancements in the field of computer vision, there has been an advancement in the use of the transformer model, which includes the attention mechanism to understand the relationship between different parts of the image. Vision Transformers (ViTs) divide the input image into smaller patches and use the self-attention mechanism to understand the long-range dependencies within the image. This is a significant advantage in the field of dermatology, as the morphology of the lesion, along with the distribution of the pigmentation, is a key factor. In a comparative study, the effectiveness of the models, i.e., Vision Transformers, Swin Transformers, and traditional CNN models, was evaluated by Dagnaw, El Mouhtadi, and Mustapha (2024). The effectiveness of the transformer models was demonstrated by identifying the global features of the skin cancer images, improving the effectiveness of the models in the classification of the images pertaining to skin cancer. In another study, the effectiveness of the transformer models was demonstrated by identifying the global relationship between the regions of the skin cancer images, whereas the CNN models are effective in identifying the local features of the images pertaining to skin cancer. Despite all the above, the model is likely to work well when there is a huge data set to train the model optimally. In addition, the model is likely to work badly with subtle texture features, which are handled by CNN models. In an attempt to reduce the difficulties associated with the application of the transformer model, the hybrid model was proposed. Pacal et al. proposed a hybrid model of CNN-ViT to carry out the early detection of skin cancer. The proposed model was composed of CNN layers to carry out feature extraction and the application of the transformer model to analyze the features. The proposed model showed good performance compared to other models. Despite the good performance of the hybrid model, there are difficulties associated with the model. C. Explainable AI using Grad-CAM As deep learning models are increasingly used for medical diagnosis, interpretability is an important requirement for such models. It is important that a medical professional understands how a model is making a particular prediction, as the pre diction could be critical in a medical scenario. Explainable Artificial Intelligence (XAI) techniques are used to provide explanations for the predictions made by a model in a visual or text format. One of the most commonly used visualization techniques is called Gradient-weighted Class Activation Mapping (Grad CAM), which was first introduced by Selvaraju et al. in 2017. The Grad-CAM visualization technique creates a heatmap that indicates the regions of the image that are contributing the most to the prediction made by the model. In dermatological image analysis, the use of such a visualization technique can be used to validate whether the model is actually looking at the relevant regions of the lesions instead of the background regions. The Grad-CAM visualization technique has been successfully used in a number of medical imaging studies as well. Further improvements were suggested by Grad-CAM++ to increase accuracy in localization by generating more precise maps if there are multiple significant areas to focus on. In dermatology, these techniques

are also beneficial because they allow practitioners to visually examine areas of an image that are significant to predictions. Grad-CAM techniques, therefore, play an important role in bridging the gap between deep learning techniques and practical applications. D. AI-based Medical Insight Generation using Generative Models Although the classification model has the ability to classify skin diseases based on the images provided, it does not have the ability to provide further medical context or advisory information about the condition. However, there have been recent developments in artificial intelligence, which enable the model to provide informative explanations about the condition based on the predictions provided by the model. The model has the ability to provide relevant information about a condition, such as the description of the condition, the possible cause of the condition, the symptoms associated with the condition, and the precautions to be taken about the condition. The current generation of generative AI models, such as those used by Google in their Gemini models, have been able to show promising results in contextual medical information generation when a prompt is given to them in a structured format. The models are able to understand the predicted disease and generate a short insight on what the disease is and how it is explained to the user in a simple manner. The use of an image classification system along with a generative model can be used to develop an intelligent system that uses a deep learning model for classification and a generative model to produce medical insights. In the context of dermatology, the hybrid model of pre diction through images, coupled with the capability of AI to generate explanations, can add more value to the usability of automated diagnostic tools. This will not only make the system more accessible to non-expert users, but it will also serve the needs of experts by offering additional contextual information pertaining to the diagnosed disease. As the technology of generative AI continues to advance, it is believed that a hybrid model of explainable computer vision will lead to the development of a comprehensive decision support system..

IV. PROBLEM STATEMENT

The accurate identification of skin diseases is an important challenge in the field of healthcare, especially for diseases like melanoma, which requires timely identification to ensure proper treatment. Dermatological diagnosis is primarily based on the visual inspection of images obtained by dermoscopy, which is then interpreted by a dermatologist. The dermatolo gist examines the skin condition by observing the color of the affected area, irregularities on the edges, symmetry, and texture of the skin to make an accurate diagnosis. The process is often complicated, as the skin diseases may exhibit similar visual features. The complexity of the process is further increased by the fact that there is a large number of patients to be diagnosed by dermatologists. Although significant improvements have been made in medical image analysis with deep learning techniques, chal lenges persist with the classification systems used currently. Convolutional Neural Networks (CNNs) are popularly used for skin lesion classification due to their ability to identify local visual patterns. However, most CNNs used for image classification focus on local features. This makes it difficult for them to identify general spatial relationships between all parts of a lesion. Vision Transformers (ViTs) have been able to address this issue by using an attention mechanism to model long-range dependencies within an image. This allows them to identify general contextual information. However, they fail to identify local details, which is essential for distinguishing between various dermatological conditions. Another major challenge that has been identified is the fact that the dermatological datasets are inherently diverse and imbalanced, where common dermatological problems like benign nevi are overrepresented relative to less common but clinically significant dermatological problems. This has the potential to affect the performance of machine learning algorithms, where the algorithms are likely to favor the ma jority classes, leading to poor detection of less common but dangerous dermatological problems. Furthermore, the images obtained through dermoscopy are prone to several artifacts like hair occlusion, illumination, and background noise, which affect the reliability of the machine learning process. Thus, it is difficult for the existing automated methods to maintain consistent performance over a range of datasets and practical scenarios. Besides accuracy in classification, interpretability is another key requirement for an AI system used in healthcare. This is because a clinician should be able to understand and validate a decision made by a machine prior to incorporating it into a diagnostic process. The use of techniques such as Gradient weighted Class Activation Mapping (Grad-CAM) can also highlight areas of an image that are more significant to a decision made by a machine, making such a decision more interpretable. Moreover, contextual information on the diseases can also make an AI system more useful to both a clinician and a user. Hence, the basic issue that is addressed in the current research is related to the development of an automated frame work for dermatological diagnosis that is capable of effectively classifying different types of skin lesions while effectively addressing issues related to variability and imbalances in the given data and architectural constraints of the models used. The system must be able to effectively capture both local and global contextual relationships within dermoscopic images while also providing a clear and effective visual explanation of the results and offering informative insights into the condition of the skin.

V. METHODOLOGY

In this section, the framework based on which the proposed system for automated dermatological diagnosis, utilizing the hybrid deep learning model, has been constructed is explained. The proposed system utilizes Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) to efficiently carry out the diagnosis of the provided dermoscopic images, while at the same time addressing the problems of intra-class variations, inter-class variations, and the inclusion of visual artifacts such as hair and non-uniform lighting, respectively. The first step involves the preprocessing of the provided images, where the images are resized, and noise reduction operations are performed on the provided data. The CNN module is specifically designed to identify local pattern details such as the pattern of pigmentation, texture abnormalities, and boundaries of skin lesions. These forms of representation are then used as an input to a Vision Transformer to identify contextual relationships within the entire image using the attention mechanism. The fusion of both models enables the simultaneous identification of local fine structures and global lesion forms. To improve the model's ability to generalize well on unseen data, various strategies such as data augmentation and class-weighted loss are used. The data set is divided into a training data set and a validation data set. The model is then trained using an adaptive optimization strategy and various techniques to avoid overfit ting. The performance of the model is checked using various parameters such as accuracy, precision, recall, F1 score, speci f icity, sensitivity, ROC AUC score, and confusion matrix to check the efficiency and relevance of the model. Apart from that, the system also uses the concept of interpretability using Grad-CAM heat maps to visualize the important parts of the image used to train the model to perform predictions.

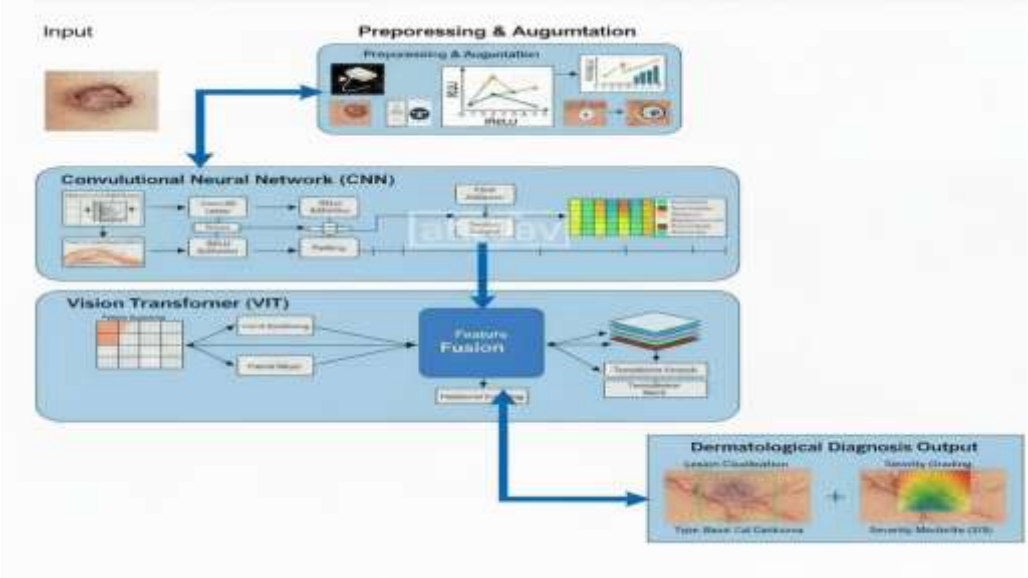

Fig. Model Architecture

A. Dataset Description and Preprocessing

The research uses a dataset containing images, where the dataset contains different classes of skin lesions, including benign and cancerous forms of the lesions. The data is divided into training, validation, and test sets. Resizing, normalization, removing artifacts, and reducing noise are some of the prepro cessing methods used to normalize the data. To increase the diversity of the data sets and the robust ness of the model, various data augmentation strategies such as rotation, flip, scaling, and brightness are utilized. Class imbalance problems often occur in dermatological data sets. Therefore, oversampling strategies have been integrated in order to consider the occurrence of rare diseases during the training process.

B. Model Architecture

The proposed architecture utilizes a combination of feature extraction via a CNN model and a Vision Transformer clas sifier. The CNN model is used to obtain low- and mid-level features such as edges, gradients, and texture patterns, which are essential for lesion characterization. The obtained features are then rearranged to obtain a patch embedding. The features are then fed into a Vision Transformer. The transformer module utilizes self-attention mechanisms to incorporate long-range dependencies between the regions of the images. This helps the model to examine the global structure of the lesions, such as asymmetry, border irregularity, and spatial distribution of the pigmentation. The positional embeddings are then used to retain the spatial information of the patches. The classification head is then used to obtain the probabilities of the dermatological class.

C. 1. Convolutional Feature Extraction

A pretrained CNN model extracts discriminative local fea tures from the dermoscopic image. Let x represent the input image and fCNN denote the CNN feature extraction function:

$$F = fCNN(x)$$

where

$$F \in RH \times W \times C$$

Here H, W, and C represent the spatial dimensions and channel depth of the feature map.

D. 2. Patch Tokenization and Linear Projection

The feature map F is divided into non-overlapping patches of size p ×p.

The number of patches is calculated as:

$$N = H \times W \; p2$$

Each patch Pi is flattened and mapped to a latent embedding space:

$$zi = Wp \cdot Flatten(Pi) + bp$$

To maintain spatial order, positional embeddings Epos are added:

$$Z0 = [z1 + Epos, ..., zN + Epos]$$

E. 3. Vision Transformer Encoder

The embedded sequence Z0 passes through L trans former encoder layers consisting of Multi-Head Self Attention (MHSA) and Feed Forward Networks. The attention function is defined as:

$$Attention(Q,K,V) = softmax\ QK^{T}$$

where

$$Q=ZWQ,\ K=ZWK,\ V=ZWV$$

Residual connections are applied as follows:

$$Z' = MHSA(Z0)+Z0$$
$$Z'' = FFN(LayerNorm(Z')) + Z'$$

This allows the model to capture global contextual depen dencies across the lesion.

F. 4. Classification Head

The final representation is aggregated using global average pooling:

$$h = GlobalAveragePooling(Z'')$$

The probability distribution over K classes is computed using a softmax classifier:

$$\hat{y} = softmax(Wch+bc)$$

G. 5. Loss Function

The model is optimized using categorical cross-entropy loss:

$$K\ L = -\ k=1\ yk\ log(\hat{y}k)$$

where yk represents the ground truth class label.

H. 6. Grad-CAM Heatmap Generation

For better interpretability, the use of Gradient-weighted Class Activation Mapping (Grad-CAM) is made to identify the regions of the image that are affecting the predictions made by the model. Grad-CAM works by computing the importance of feature maps using the gradients of the predicted class score with respect to the convolutional feature maps. Let Ak represent the kth feature map from the final convo lution layer. The importance weight for each feature map is calculated as:

$$\alpha k = 1\ Z\ i\ j\ \partial yc\ \partial Ak\ ij$$

where yc is the score for class c and Z is the total number of pixels in the feature map. The Grad-CAM heatmap is then computed as:

$$Lc\ Grad{-}CAM = ReLU\ \alpha kAk\ k$$

The resulting heatmap highlights the regions of the dermo scopic image that contributed most strongly to the model's classification decision.

I. 7. AI-based Disease Insight Generation using Gemini API

In addition to image classification, the system incorporates a generative module of AI with the Gemini API to offer medical context on the detected disease. Once the image is classified by the model, the name of the detected disease is passed to the generative module in a structured format. The Gemini model generates information on the possible causes, symptoms, and precautions to be taken concerning the detected disease. This integration improves the usability of the system be cause users can have the predictions about the detected disease and the context about the predicted condition of the skin

VI. EXPERIMENTATION AND RESULTS

The part of experimentation and results includes a compre hensive analysis of the proposed hybrid model, i.e., CNN and Vision Transformer, for the automated dermatological diagno sis system. The purpose of the analysis is not only to assess the predictive power of the model but also to assess the robustness, interpretability, and usability of the model for the automated dermatological diagnosis system. The dermatological images are characterized by the presence of subtle variations between the classes of the disease, and there is a high class imbalance problem; therefore, it is important to analyze the proposed model both qualitatively and quantitatively. The experimental evaluation combines statistical measures of performance such as accuracy, precision, recall, F1-score, and ROC-AUC with visualization tools for analysis. The eval uation ensures that the predictions obtained from the model are correct and align with dermatological knowledge of diagnosis. Additionally, visualization tools such as Grad-CAM heat maps are employed to analyze the most critical regions of dermo scopic images that contribute substantially to the classification result. The analysis obtained from the traditional CNN models and transformer models further confirms the advantages of the hybrid model, which integrates local feature extraction with global reasoning. The experimental evaluation confirms the efficacy of the proposed framework for dermatological image analysis.

A. Experimental Setup

An experimental setup was created to test the proposed model. The dataset used in this study is a set of dermoscopic images of various skin lesions, such as melanoma, basal cell carcinoma, benign keratosis, dermatofibroma, vascular lesions, and melanocytic nevi. These images have varying lighting conditions, skin types, lesion shapes, and textures, which will help the proposed model learn as if it were in a real-world environment. The data was split into training and validation sets while ensuring that the classes remained balanced. Usually, an 80-20 split was used to ensure that the model was trained on enough data while still preserving a significant portion for validation. The image preprocessing included resizing all images to a f ixed resolution, normalizing the pixel values, and removing any visual artifacts like hair or shadows that could hamper the learning process. Data augmentation methods were used extensively to en hance the capacity of the model to generalize. Data augmenta tion methods such as rotation, horizontal and vertical flipping, scaling, cropping, and color jittering were incorporated to mimic the variations that occur in real-world imaging. Such methods are useful in preventing overfitting of the model. The proposed model was developed and implemented using Python 3.11 and deep learning libraries such as TensorFlow, Keras, and PyTorch. Other libraries were used for data pro cessing, visualization, and evaluation criteria. The training was done on high-performance GPU hardware with sufficient memory and storage capacity to support large hybrid models. The training process included sparse categorical cross entropy as the loss function, and the optimizers could be Adam or AdamW. Adaptive learning rate schedulers such as cosine annealing and step decay were also used. The model used early stopping on the validation loss to avoid overfitting. Dropout layers and weight decay were used as the regulariza tion technique. The model performance was evaluated using accuracy, precision, recall, F1-score, ROC-AUC, specificity, and sensitivity to ensure that the system is functioning properly for all categories of lesions

B. Training and Validation Results

The learning dynamics of the hybrid CNN Vision Trans former model were very good during the training and vali dation process. The training process was smooth without any problems related to stability, and the training accuracy was 0.9984 with a validation accuracy of 0.9371. The values of the loss functions were decreasing steadily. The class-wise evaluation gives more insight into the perfor mance of the model on various types of lesions. The general lesion classes like Melanocytic Nevi (NV) performed outstand ingly with high values of precision, recall, and F1-score of 0.9684, 0.9787, and 0.9735, respectively. This analysis shows that the model is capable of detecting common lesion patterns. More complex classes such as melanoma (MEL) and der matofibroma (DF) had slightly lower values due to their visual similarity to benign lesions and lower frequency within the dataset. Nonetheless, the proposed architecture was successful in incorporating local texture features as well as global con textual features, thus achieving competitive performance for all classes.

TABLE I
CLASSIFICATION REPORT FOR THE PROPOSED HYBRID MODEL ON THE VALIDATION DATASET

Class	Precision	Recall	F1-Score	Support
AKIEC	0.8500	0.6538	0.7391	26
BCC	0.9062	0.9667	0.9355	30
BKL	0.7711	0.8533	0.8101	75
DF	1.0000	0.8333	0.9091	6
MEL	0.6774	0.5385	0.6000	39
NV	0.9735	0.9800	0.9768	751
VASC	1.0000	1.0000	1.0000	11
Overall Accuracy		**94.13%**		

C. Confusion Matrix

This is a tool that offers a clear image of the results in terms of classification for each type of lesion. From this matrix, we can know the number of correct or incorrect classifications made by the model for each class. From Fig. 2, we notice that most of the lesions are correctly classified by the model, with a few incorrect classifications for classes that look similar.

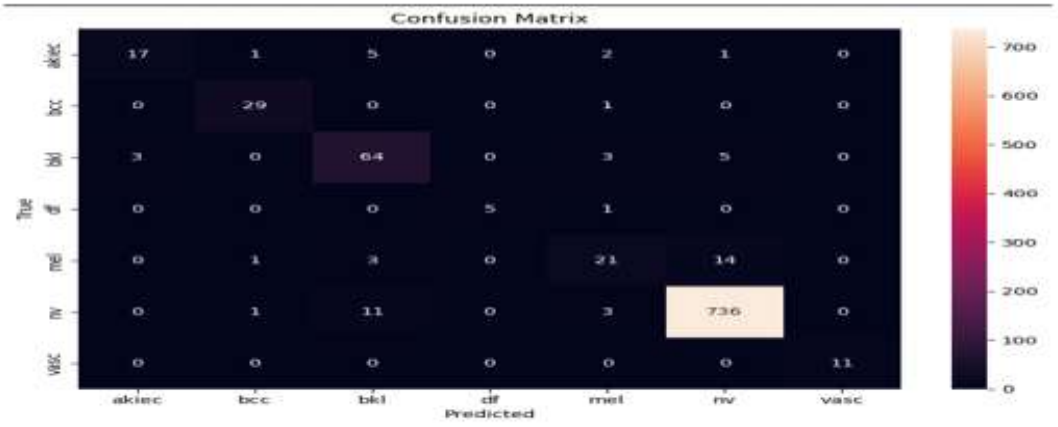

Fig. 2. Confusion Matrix showing classification performance across all dermatological classes.

D. ROC and AUC Analysis

Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) values were used to further evaluate the discriminative ability of the model. The curve shows the relationship between sensitivity (true positive rate) and false positive rates at different classification threshold values. As depicted in Fig. 2, the hybrid model has achieved exceptionally high values in terms of AUC for all classes, such as BCC (0.9996), VASC (1.0000), MEL (0.9705), DF (0.9666), BKL (0.9850), NV (0.9879), and AKIEC (0.9519).

E. Grad-CAM Heatmap Visualization

In order to make the results more interpretable, Grad CAM visualization was employed to highlight specific areas

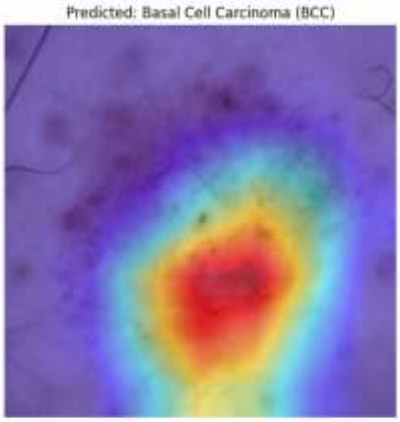

Fig. 4. Grad-CAM heatmap highlighting important regions of the lesion used for classification.

F. AI-Based Disease Insight Generation

Apart from classification and visualization, the proposed system will also have an AI module that generates informative insights about the predicted disease of the skin. Once the model has identified the type of lesion, the predicted class label will be input to the Gemini AI model, which generates an informative explanation of the predicted disease. The output generated will include details regarding the possible causes, symptoms, and precautions that are to be taken for the disease. Figure 5 shows an example of the output generated for the class of lesion predicted in the image. This feature makes the system more user-friendly as more information is provided to the user. Conclusion In conclusion, the experimental results show that the proposed hybrid CNN-Vision Transformer model has high classification accuracy, besides offering visual explana tions for the classification results and medical insights. This makes the system a promising tool for AI-assisted dermato logical diagnosis.

G. Comparison with Existing Approaches

Most of the existing automated dermatological diagnosis systems rely on either CNNs or Vision Transformers as

Fig. 5. Example of AI-generated dermatological insights produced using the Gemini API.

standalone models. CNNs have been found to have strong capability in the recognition of local visual patterns such as edges, textures, and pigmentation. However, CNNs have been shown to have limited potential in the ability to understand global relationships within an image. In situations where lesions exhibit similar inter-class characteristics or structural patterns, CNN-based systems may encounter difficulties in dis tinguishing between closely related dermatological conditions. Transformer models, especially Vision Transformers (ViTs), have recently emerged as promising models owing to their ability to capture long-range dependencies with self-attention mechanisms. By exploring the relationship between different image regions, ViTs are capable of better understanding global lesion structures compared to conventional convolutional mod els. However, it

should be noted that even though transformer models are capable of understanding global image structures, they might not be able to capture local image details, especially related to micro-textures and pigmentation patterns, which are crucial for the early stages of dermatological diagnostics. To overcome these challenges, some recent works have attempted to develop hybrid models that integrate CNNs for feature extraction and transformer-based reasoning. Most of these models use parallel architectures or feature concatenation strategies to combine the information from both models. While these models show better performance than single models, they can also lead to redundant features or higher compu tational complexity because of inefficient fusion strategies. Also, some previous works have shown that there are only minor improvements in the case of rare lesion classes, such as Dermatofibroma (DF) and Vascular lesions (VASC), which have smaller visual differences. The proposed hybrid CNN-ViT model is based on a se quential architecture that entails the extraction of fine-grained discriminative features using a CNN backbone and the subse quent use of a Vision Transformer to capture global contextual relationships. This is crucial in ensuring that the fine-grained visual properties of lesions, such as texture, pigmentation, margin irregularity, and asymmetry, are properly captured before the use of transformer-based reasoning. By virtue of the sequential refinement of feature representations, the proposed model offers superior semantic understanding of dermatolog ical structures and improved classification performance for common and rare lesion classes.. Apart from the enhancement of classification accuracy, the proposed method also focuses on interpretability via the inte gration of Gradient-weighted Class Activation Mapping (Grad CAM). Most of the current deep learning models are black-box models, which give predictions without any explanations for the predictions. Grad-CAM solves this problem by creating heatmaps that point to the regions of the image that have the most influence on the prediction of the classification. Compared to the previous models that only used attention visualization in the CNN layers, Grad-CAM is more intuitive and easier to understand. The ability to create heatmaps is very important in medical applications, where the predictions made by the model have to be confirmed by focusing on the relevant regions and not on the background artifacts

TABLE II
COMPARISON OF DIFFERENT ARCHITECTURES FOR DERMOSCOPIC IMAGE CLASSIFICATION

Method / Paper	Architecture Type	Reported Accuracy	Dataset Type	Limitation	Advantage of Proposed
ResNet-50 Baseline	Pure CNN	~90–92%	Dermoscopic	Limited global reasoning	Our model adds ViT global modeling
EfficientNet-B3 Model	Pure CNN	~93–95%	Dermoscopic	Fails with subtle classes	Hybrid better in rare classes (DF, VASC)
ViT-Base Patch16	Pure ViT	~88–91%	Dermoscopic	Lacks low-level cues	CNN stage extracts detailed textures
Swin-Transformer Hybrid (Parallel Fusion)	CNN + ViT (parallel)	~95–96%	Dermoscopic	Redundant fusion & unstable	Sequential fusion improves semantic clarity
Grad-CAM based CNN models	CNN + Explainability	~93–95%	Dermoscopic	Limited global context modeling	Our model integrates Grad-CAM with CNN–ViT hybrid reasoning
Proposed Hybrid (CNN → ViT)	**Sequential Hybrid Model**	**99.84% (Train)**	**Dermoscopic**	—	**High accuracy with interpretable Grad-CAM visualization**

VII. CONCLUSION

The research proposed in the paper proposed a hybrid deep learning model based on the integration of Convolutional Neural Networks (CNN) and Vision Transformers to carry out skin disease classification with the aid of dermoscopic images. The objective of the proposed system was to overcome the limitations associated with the application of traditional CNN models, which are mostly based on the ability of the model to learn local features of the image, while the proposed system was able to learn global structural dependencies associated with the skin lesion through the integration of the CNN model. Experimental evaluation was done to validate the strength of the framework proposed. The model was able to attain a validation accuracy of 94.13It was evident from the class-wise evaluation of the model that the model was able to attain high precision, recall, and F1-score values for the skin lesion classes. This validates the strength of the model to perform well on unseen data. From the results obtained through the experimental evaluation using the confusion matrix, it was evident that the model was able to classify visually distin guishable skin lesions such as melanocytic nevi (NV), basal cell carcinoma (BCC), and vascular lesions (VASC) with high confidence levels. The classification results showed a few classification errors in the case of visually similar skin lesions such as melanoma (MEL) and actinic keratosis (AKIEC). This was attributed to the difficulty in classifying fine skin lesion patterns. The results obtained through the experimental evaluation using the ROC-AUC validated the strength of the model proposed. The values obtained were closer to 1.0 for the skin lesion classes. The hybrid nature of the architecture was a major con tributing factor to the achievement of the results. The CNN layers were able to effectively capture the local visual features such as texture, pigmentational features, and lesion boundaries. Meanwhile, the Vision Transformer layer was able to capture the global contextual relationships within the image. Attention mechanisms were also employed to enable the network to focus on the relevant features while discarding irrelevant information, thus enhancing the separation of dermatological classes. The addition of the Grad-CAM visualization mecha nism has the effect of highlighting the regions of the lesion which have the most influence on the classification outcome. This has the effect of enhancing the level of transparency in the system. Another major achievement of the research work is the integration of a generative AI component through the use of the Gemini API to provide contextual information regarding the outcome of the skin condition. After the classification outcome has been determined, the system has the ability to provide informative descriptions regarding the outcome, such as the potential causes, symptoms, and precautions associated with the determined skin condition. From a clinical point of view, the proposed system shows great potential to be a decision support tool for dermo¬logists. In fact, the system is capable of providing accurate predictions, explanations through the use of Grad-CAM, and contextual information through the application of AI. This would help medical practitioners make an early diagnosis of skin diseases, thus reducing the uncertainty associated with the diagnosis. It is worth noting that the proposed model would be important in regions where there is a scarcity of dermo¬logy experts. In such regions, the proposed system would help in the diag nosis of skin diseases, thus enhancing the level of healthcare services. Although the model was able to achieve high accuracy, there are a number of

limitations associated with the proposed model. For instance, the accuracy of the model would depend on the availability of a few instances of rare classes of skin lesions, which would have similar visual characteristics. To put it briefly, the research has proven the proposed method's potential to successfully deal with the problem of dermatological image analysis through the utilization of CNN-based local feature extraction methods in conjunction with the utilization of the transformer-based global reasoning mechanism. Additionally, the utilization of the Grad-CAM mechanism, as well as the utilization of the AI mechanism to provide insights into the diseases, has enhanced the proposed method's applicability to real-world scenarios. The proposed framework is a promising advancement in the development of intelligent AI systems for the diagnosis of dermatological diseases

REFERENCES

[1] X. Jia and M. Q.-H. Meng, "A deep convolutional neural network for bleeding detection in wireless capsule endoscopy images," in 2016 38th annual international conference of the IEEE engineering in medicine and biology society (EMBC). IEEE, 2016, pp. 639–642.

[2] A. Singh, S. Prakash, A. Das, and N. Kushwaha, "Colonnet: A hybrid of densenet121 and u-net model for detection and segmentation of gi bleeding," arXiv preprint arXiv:2412.05216, 2024.

[3] E. U. Henry, O. Emebob, and C. A. Omonhinmin, "Vision transformers in medical imaging: A review," arXiv preprint arXiv:2211.10043, 2022.

[4] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable ai systems for the medical domain?" arXiv preprint arXiv:1712.09923, 2017.

[5] A. Holzinger, "Explainable ai and multi-modal causability in medicine," i-com, vol. 19, no. 3, pp. 171–179, 2021.

[6] K. Al-Hammuri, F. Gebali, A. Kanan, and I. T. Chelvan, "Vision transformer architecture and applications in digital health: a tutorial and survey," Visual computing for industry, biomedicine, and art, vol. 6, no. 1, p. 14, 2023.

[7] S. Chetcuti Zammit and R. Sidhu, "Capsule endoscopy–recent devel opments and future directions," Expert review of gastroenterology & hepatology, vol. 15, no. 2, pp. 127–137, 2021.

[8] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convo lutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.

[9] U. R. Acharya, H. Fujita, S. L. Oh, Y. Hagiwara, J. H. Tan, M. Adam, and R. S. Tan, "Deep convolutional neural network for the automated diagnosis of congestive heart failure using ecg signals," Applied Intelli gence, vol. 49, pp. 16–27, 2019.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

Copyright & License:
© Authors retain the copyright of this article. This work is published under the Creative Commons Attribution 4.0 International License (CC BY 4.0), permitting unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.