

Optimization of Skin Disease Classification with a Hybrid Convolutional Neural Network and Vision Transformer Approach

Joy Oyinye Orukwo, Constance Izuchukwu Amannah*

Department of Computer Science, Ignatius Ajuru University of Education, Port Harcourt, Nigeria

Email: *aftermymisc@gmail.com

How to cite this paper: Orukwo, J.O. and Amannah, C.I. (2026) Optimization of Skin Disease Classification with a Hybrid Convolutional Neural Network and Vision Transformer Approach. *International Journal of Communications, Network and System Sciences*, 19, 39-54.

<https://doi.org/10.4236/ijcns.2026.193003>

Received: February 26, 2026

Accepted: March 28, 2026

Published: March 31, 2026

Copyright © 2026 by author(s) and Scientific Research Publishing Inc. This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

Skin diseases remain a significant global health concern, with accurate diagnosis often challenged by visual similarities among lesion types and limited access to specialist dermatological services. This study presents an optimized hybrid Convolutional Neural Network-Vision Transformer (CNN-ViT) framework for automated skin disease classification using dermoscopic images from the HAM10000 dataset. The proposed architecture combines the local feature extraction capability of CNNs with the global contextual learning ability of Vision Transformers to improve classification performance. Prior to model training, the dataset was partitioned into training (70%), validation (15%), and testing (15%) subsets, with data augmentation applied exclusively to the training set to address class imbalance and improve generalization. The hybrid model was implemented using TensorFlow/Keras and evaluated using weighted accuracy, precision, recall, and F1-score metrics. Experimental results demonstrated a classification accuracy of 97.68%, precision of 89.80%, recall of 90.20%, and F1-score of 91.30%. Comparative analysis indicated that the hybrid architecture outperformed conventional CNN-based approaches due to its ability to simultaneously capture local texture patterns and global lesion structures. The findings demonstrate the effectiveness of integrating CNN and ViT architectures for intelligent dermatological image analysis and highlight the potential of hybrid deep learning models in supporting clinical decision-making and early skin disease diagnosis.

Keywords

Skin Disease Classification, Convolutional Neural Network, Vision Transformer, Deep Learning, HAM10000 Dataset, Medical Image Analysis

1. Introduction

The skin is the largest organ of the human body, comprising 16% of total body weight and covering an area of 1.5 to 2.0 m². It serves as a barrier against physical, chemical, and microbial threats, including UV radiation and pathogens. Additionally, it performs essential functions such as thermoregulation, immune defense, sensory reception, and vitamin D synthesis. Structurally, the skin is composed of three layers: the epidermis, dermis, and hypodermis, each contributing to the skin's protective and regulatory functions. The epidermis, rich in keratinocytes, forms a waterproof barrier and regenerates cells. The dermis, containing collagen and elastin, supports immune, vascular, and sensory functions. The hypodermis, consisting of adipose tissue, provides insulation and shock absorption [1].

Skin physiology involves continuous cell regeneration, with melanocytes producing melanin for UV protection and Langerhans cells functioning as immune sentinels [2]. The skin's interaction with the immune and endocrine systems makes it a critical organ for understanding systemic diseases and therapeutic interventions [3]. Exposure to environmental factors makes the skin susceptible to a wide range of conditions, significantly affecting individual health and quality of life [4].

Skin diseases are prevalent globally, affecting individuals of all ages and ethnicities. These conditions range from benign disorders to chronic and life-threatening diseases such as melanoma. They are commonly classified into infectious (e.g., cellulitis), inflammatory (e.g., eczema, psoriasis), neoplastic (e.g., melanoma), and autoimmune (e.g., lupus) categories [5]. Diagnosis traditionally relies on visual inspection, dermoscopy, and biopsy, but these methods are time-consuming, resource-intensive, and prone to inter-observer variability. Misdiagnosis due to the visual similarities between conditions often leads to delayed treatment and adverse outcomes [6].

In many low- and middle-income countries, access to dermatological care is limited, highlighting the need for automated diagnostic systems. AI, particularly in deep learning, has shown promise in real-time diagnostic support, enabling population screening and mobile health applications, thus improving the accessibility and equity of dermatology services.

Neural networks (NNs), a key component of deep learning, are computational models inspired by the human brain, capable of capturing complex patterns in large datasets. In medical diagnostics, NNs have demonstrated significant potential, particularly in image recognition tasks [7]. These models adjust weights through training to improve accuracy, making them valuable for detecting skin lesions and classifying conditions such as melanoma and keratinocyte carcinomas [8].

However, traditional NNs are computationally inefficient in visual tasks due to their fully connected architecture, which has led to the development of specialized models such as Convolutional Neural Networks (CNNs) designed to handle spatial data more efficiently.

CNNs are tailored for image analysis, particularly in medical imaging, where they excel in detecting spatial patterns, textures, and shapes in images. By using local receptive fields, CNNs recognize hierarchies of features such as edges, corners, and textures, making them ideal for tasks like skin lesion classification. CNNs have been shown to perform on par with dermatologists in diagnosing skin conditions, including melanoma and basal cell carcinoma [9].

However, CNNs face limitations in modeling long-range dependencies, such as the overall symmetry or shape of lesions, which are critical for accurate dermatological diagnosis. This has led to the exploration of Vision Transformers (ViTs) for capturing global context.

Vision Transformers (ViTs), introduced by [10], use self-attention mechanisms to model global relationships across image patches, offering a distinct advantage over CNNs for tasks requiring a holistic understanding of images. ViTs have demonstrated strong performance in image segmentation and melanoma classification, particularly in capturing irregular shapes and complex patterns [11] [12].

While ViTs outperform CNNs in tasks requiring global context, they are computationally expensive and require large datasets for training. These challenges have led to the development of hybrid models that combine CNNs' local feature extraction capabilities with ViTs' global context modeling.

The hybrid CNN-ViT model integrates the strengths of both CNNs and ViTs to address the limitations of each. CNNs excel at extracting fine-grained features such as edges and textures, while ViTs capture long-range dependencies and global context. Hybrid models like TransUNet and MC-ViTNet have shown superior performance in tasks such as skin lesion classification and tumor segmentation, offering improved robustness and interpretability [12] [13]. These models are particularly effective in dermatology, where lesions vary significantly in size, shape, and texture.

The proposed study aims to develop a hybrid CNN-ViT model for skin disease classification, leveraging both architectures to improve diagnostic accuracy, computational efficiency, and model interpretability. The model will be trained on the HAM10000 skin lesion dataset, incorporating both local texture features and global lesion patterns for more accurate classifications.

1.1. Theoretical Framework

This study is anchored on the theory of Feature Learning (Representation Learning) and the Transformer Attention Mechanism, which collectively explain the effectiveness of hybrid CNN-ViT architectures in medical image classification.

Feature learning enables deep neural networks to automatically discover discriminative representations directly from raw image data without manual feature engineering [14]. Through hierarchical learning, lower network layers capture simple visual patterns such as edges, textures, and color gradients, while deeper layers learn more complex and abstract lesion characteristics. In skin disease classification, feature learning facilitates the extraction of clinically relevant visual

cues necessary for differentiating among dermatological conditions.

The Transformer Attention Mechanism introduced by [10] provides the theoretical foundation for Vision Transformers. Unlike conventional CNNs that focus primarily on local receptive fields, self-attention enables the model to capture long-range spatial dependencies and contextual relationships across the entire image. This capability is particularly important in dermatology, where lesion symmetry, shape irregularity, and global structural characteristics are critical diagnostic indicators.

The integration of CNN and ViT architectures combines localized feature extraction with global contextual understanding, resulting in a more comprehensive representation of skin lesions and improved classification performance.

1.2. Empirical Research

The integration of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) represents a promising advancement in optimizing skin disease classification. Several studies have explored hybrid models combining these two architectures to enhance diagnostic accuracy by leveraging CNNs' local feature extraction capabilities and ViTs' ability to capture global dependencies. Below is a focused review of the most relevant works contributing to this optimization.

[15] developed a multi-task CNN for skin lesion classification, particularly melanoma detection. The model integrated both segmentation and classification tasks, achieving high performance in delineating lesion boundaries and classifying melanoma. However, the model was limited by the lack of diverse skin types in the training dataset, which restricted its generalizability. The study highlighted that hybrid models, incorporating both CNNs and ViTs, could improve both local feature extraction and global contextual understanding, which are crucial for accurate melanoma detection.

[15] developed a deep CNN for skin cancer detection, achieving dermatologist-level performance in melanoma and basal cell carcinoma diagnoses. While the CNN model demonstrated high accuracy, it was limited by the absence of clinical context such as patient history and lesion evolution. Esteva *et al.*'s findings suggest that a hybrid CNN-ViT model would address this limitation by combining CNNs' ability to extract local features with ViTs' capacity to model global relationships, enabling the integration of non-visual clinical data for improved diagnostic performance.

[3] evaluated the effectiveness of CNNs in detecting skin diseases like melanoma and psoriasis. The study found that while CNNs outperformed traditional methods, they struggled with modeling global dependencies and integrating clinical data. These limitations emphasize the potential of CNN-ViT hybrid models, which can incorporate both image features and clinical metadata, improving diagnostic accuracy and generalizability across various dermatological conditions.

[8] compared a CNN's diagnostic performance in melanoma recognition to that of 58 dermatologists, achieving 95% sensitivity. Despite this, the model's performance was limited by the lack of clinical metadata, underscoring the importance of integrating both image data and patient-specific information. The study advocates for hybrid CNN-ViT models, which combine local feature extraction from CNNs with global context modeling and clinical context integration from ViTs, to improve diagnostic consistency and applicability in real-world settings.

[12] proposed TransUNet, a hybrid model combining CNNs and ViTs for skin lesion segmentation. The model achieved superior performance in segmentation accuracy, capturing both local and global features. The study demonstrated that hybrid CNN-ViT models could significantly enhance skin disease classification by capturing fine-grained local features from CNNs and modeling long-range dependencies through ViTs. However, the computational complexity of the hybrid architecture remains a challenge, and future work needs to focus on optimizing this trade-off for clinical deployment.

[11] compared Vision Transformers (ViTs) with CNNs for skin lesion classification, finding that ViTs outperformed CNNs, especially in identifying malignant lesions. While ViTs showed promise, performance inconsistencies across different lesion types were noted. The study advocates for combining CNNs' localized feature learning with ViTs' global attention mechanisms to improve generalization and performance across diverse skin types and lesion presentations, supporting the development of hybrid CNN-ViT models for more robust skin disease classification.

[16] developed a CNN-based real-time skin disease diagnostic system that achieved high accuracy but was limited by its reliance solely on image data and the absence of clinical metadata. Their study suggests that hybrid models, integrating CNNs with other techniques such as ViTs, could improve contextual understanding and enhance system robustness in real-world clinical applications. The combination of CNN and ViT can overcome the limitations of image-only models by integrating global attention and contextual analysis.

[17] developed a multimodal CNN-based model that integrated clinical data with image analysis, improving skin cancer diagnosis. The hybrid model outperformed image-only CNNs, achieving higher sensitivity and diagnostic accuracy. The authors emphasize the need for hybrid models that combine visual and non-visual data to improve real-world clinical applicability. A CNN-ViT hybrid model could provide a better approach by capturing both local and global features and integrating clinical metadata for a more holistic and accurate diagnosis.

[15] developed a CNN model for psoriasis classification with 87% accuracy but noted limited generalizability due to dataset diversity. The study suggests that hybrid CNN-ViT models could improve generalizability by combining CNNs' efficiency in feature extraction with ViTs' ability to model global dependencies and contextual relationships, making the model more applicable to diverse patient

populations and various dermatological conditions. [6] proposed a hybrid CNN-SVM model to improve skin lesion classification. While the model achieved an accuracy of 85%, it was computationally complex and lacked clinical data integration. The study highlighted the need for models that balance performance with computational efficiency, suggesting that a CNN-ViT hybrid model could improve both performance and interpretability by combining CNNs' ability to extract local features with ViTs' capacity for global attention mechanisms. [18] compared a CNN-based AI system with experienced human dermatologists in detecting malignant and benign skin lesions. The CNN model achieved promising results, but its dependency on high-quality curated datasets limited its robustness in real-world scenarios. The study suggests that hybrid CNN-ViT models could enhance resilience to inconsistencies in visual data, improving accuracy across varied clinical conditions by incorporating global contextual understanding from ViTs.

2. Materials and Methods

This study adopts the Dynamic Systems Development Methodology (DSDM) framework for the development and optimization of the skin disease classification model using Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). DSDM provides an agile approach that facilitates structured development while emphasizing flexibility, stakeholder involvement, and iterative improvements throughout the process. The methodology was followed through the feasibility assessment, system design, and implementation phases, ensuring a systematic and user-centered approach to building the hybrid CNN-ViT model. **Figure 1** shows the framework of the Methodology.

2.1. Development Methodology

The development process adhered to the DSDM framework, structured into several key stages to guide the implementation of the CNN-ViT skin disease classification system.

This initial phase involved confirming the suitability of the dataset (HAM10000), verifying the computational resources required for training the model, and assessing the model's potential for deployment in clinical settings. A clear understanding of user needs, clinical relevance, and project constraints was established to guide the design and ensure the system meets healthcare requirements. The iterative process began with the implementation of the preprocessing pipeline, CNN-ViT hybrid modules, and the fusion of features from CNN and ViT for classification. These were refined across multiple experimental cycles to enhance performance. The final deployment involved integrating the model with a user-friendly interface, validating its performance through metrics such as accuracy, precision, and recall, and providing explainability outputs for clinical interpretation.

The HAM10000 dataset was utilized for this study. The dataset contains 10,015 dermoscopic images representing seven diagnostic categories: Actinic Keratoses

(AKIEC), Basal Cell Carcinoma (BCC), Benign Keratosis-like Lesions (BKL), Dermatofibroma (DF), Melanoma (MEL), Melanocytic Nevi (NV), and Vascular Lesions (VASC). To prevent data leakage, dataset partitioning was performed prior to augmentation. Images associated with the same lesion were restricted to a single partition. The dataset was divided into training (70%), validation (15%), and testing (15%) subsets. To address class imbalance and improve model generalization, augmentation techniques including horizontal flipping, vertical flipping, random rotation, zooming, brightness adjustment, and image translation were applied exclusively to the training subset. The balanced class distributions presented in **Table 1** represent the post-augmentation dataset used during model training.

2.2. Hybrid CNN-ViT System Design

This study proposes a Hybrid CNN-ViT Model to enhance skin disease classification by leveraging both local feature extraction (via CNNs) and global context modeling (via ViTs). The model is a multi-stage pipeline consisting of: Input Layer, CNN Feature Extractor, Transformer Encoder (ViTs), Fusion Module, and Softmax Classifier.

2.3. Procedural Operation of the Hybrid System

The operational flow of the skin disease classification system is outlined as follows:

Input Module: Users upload images, which are then preprocessed for standardization (resizing, normalization, and augmentation).

Feature Extraction:

CNN Backbone: Extracts low-level features such as textures and edges.

ViT Encoder: Divides the image into patches, processes them with self-attention to capture global relationships, and outputs high-level semantic features.

Fusion Module: Combines the local CNN features and global ViT features to enhance feature representation.

Classification: The fused features are passed through a softmax classifier to predict the skin disease category.

2.4. Training, Validation, and Testing

Training: The model is trained on 70% of the dataset (HAM10000), with augmentation applied to improve generalization. Training proceeds through multiple epochs with backpropagation and gradient descent, and early stopping is applied to prevent overfitting.

Validation: A separate 15% of the dataset is used to validate the model, providing insights into generalization. Hyperparameters are adjusted based on validation loss to improve model performance.

Testing: The final 15% of the dataset is used to test the model's performance on unseen data. Performance metrics (accuracy, precision, recall, F1-score) are computed to evaluate the model's effectiveness.

2.5. Hybrid CNN-ViT System Components

Feature Extraction

CNN Module: Extracts fine-grained local features such as edges, textures, and color variations, essential for differentiating skin diseases.

ViT Module: Models global dependencies across the entire image, capturing contextual relationships that CNNs may miss.

Fusion Layer

The fusion layer merges the outputs of the CNN and ViT modules, creating a more comprehensive feature vector that incorporates both local and global features.

Softmax Classifier

The final decision-making stage uses the fused feature vector to predict the skin disease class, ensuring that the model outputs probabilities for each class and provides interpretable results.

Performance Evaluation

The model's performance is evaluated using standard classification metrics (accuracy, precision, recall, F1-score) on the test set. The system's ability to handle environmental variability, such as lighting changes and image quality, is also assessed, with results compared against traditional CNN models.

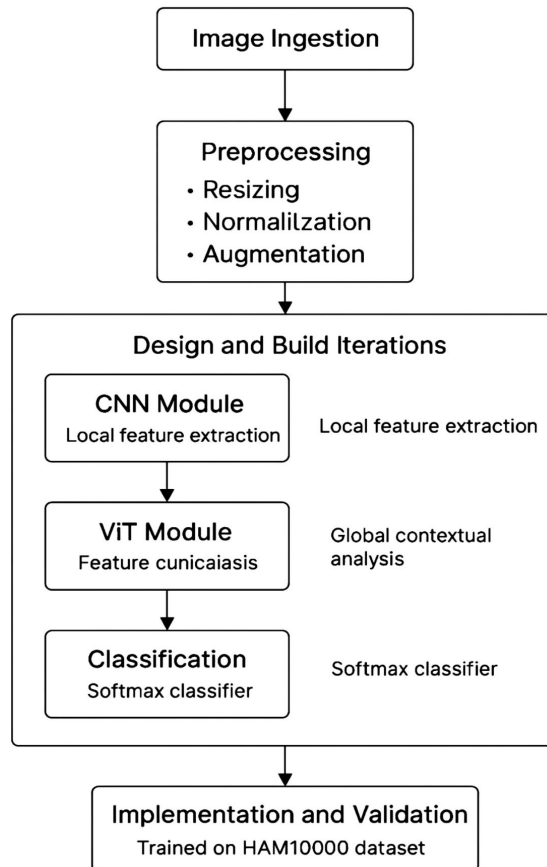


Figure 1. The framework in the methodology.

The proposed model combines EfficientNet-B0 and Vision Transformer (ViT-Base/16) architectures. The CNN branch employs EfficientNet-B0 for extracting local image features such as texture, color variation, lesion borders, and structural patterns. The ViT branch processes image patches through self-attention mechanisms to capture long-range spatial relationships and global contextual information.

Model configuration parameters include

Parameter	Value
CNN Backbone	EfficientNet-B0
ViT Variant	ViT-Base/16
Input Resolution	224×224
Patch Size	16×16
Optimizer	Adam
Learning Rate	0.0001
Batch Size	32
Epochs	100
Early Stopping Patience	10
Random Seed	42
Framework	TensorFlow/Keras
Hardware	NVIDIA GPU

The extracted CNN and ViT features were concatenated through a fusion layer and subsequently passed through fully connected layers and a Softmax classifier for final disease prediction.

2.6. Mathematical Representation of the System

2.6.1. Convolutional Layer: Extracting Local Features

The CNN module starts with a convolutional layer, which applies a set of trainable filters (kernels) to detect patterns such as edges, textures, and shapes from the input image. The convolution operation between an input image $I(x, y)$ and a filter $K(x, y)$ (kernel) is mathematically expressed as:

$$F(i, j) = \sum_{m=0}^2 \sum_{n=0}^2 I(i+m, j+n) \cdot K(m, n) + C \quad (1)$$

- 1) $I(i+m, j+n)$ is the input image at position $(i+m, j+n)$.
- 2) $K(m, n)$ is the 3×3 Kernel (Filter).
- 3) C = bias.
- 4) The summation runs over the 3×3 kernel window.

Therefore, for multiple filters (F_m), the convolution is expressed thus;

$$F_k(i, j) = \sum_{m=0}^2 \sum_{n=0}^2 I(i+m, j+n) \cdot K_k(m, n) + C_k \quad (2)$$

where

K_k = kernel for filter k and.

C_k = bias term.

Therefore, the feature extraction within the CNN layer is handled using Equation (3)

$$F_{\text{CNN}} = \text{Flatten}(\text{CNN}_L(X)) \quad (3)$$

where

X = Input Image.

Flatten converts the spatial feature into a 1D vector using Equation (3).

2.6.2. Transformer Encoder Module: Global Feature Learning

Given an input image X (with height H , width C , and channels C), ViT splits the image into patches of size $P \times P$, giving N patches. This can be mathematically expressed thus;

$$N = \frac{H \times W}{P^2} \quad (4)$$

where;

N = total number of patches.

P = patch size.

H = Image Height.

W = Image Width.

Each patch is flattened into a vector and mapped to a fixed-size embedding using a linear transformation:

$$Z = W_e X + b_e \quad (5)$$

where;

X = Flattened Patch pixels.

W_e = Learnable weight matrix.

b_e = Bias term.

Z = Embedded feature representation of the patch.

Self-Attention Mechanism: Each transfer Encoder layer applies Multi-Head Self-Attention (MHSA)

$$\text{MHSA} = \text{Softmax}\left(\frac{(QK^T)}{\sqrt{F_d}}\right)V \quad (6)$$

where:

$Q = W_Q Z$, $K = W_K Z$, $V = W_V Z$ are the Query (Q), Key (K), and Value (V) Matrices.

F_d = feature dimension.

The softmax function ensures that the values sum to 1.

This step allows the Transformer Encoder module to learn global relationships across the image. Furthermore, the output from the self-attention layer is passed through a feedforward network. Therefore, the feature extraction within the Transformer Encoder is handled using Equation (7).

$$F_{ViT} = \text{MLP}\left(\text{MHSA}\left(X_{\text{patch}} * E\right)\right) \quad (7)$$

where:

X_{patch} = Image divided into patches

F_{ViT} = Transformer Encoder layer.

E = Patch embedding (Projection Matrix).

MHSA = Multi-Head Self-Attention, capturing global relationships.

MLP = Multi-Layer Perceptron, refining extracted features.

The Fusion Module of the model is responsible for integrating feature representations extracted from both the CNN and Vision Transformer (ViT) modules to enhance skin disease classification. Therefore, given an input image X , CNN processes the image and extract the local features and it passed into the transformer Encoder Module which processes the input patches.

Hence, the Skin Disease Classification Model (SDCM) is obtained by concatenating CNN features (Equation (1)) and ViT features (Equation (7))

$$\text{Hybridized Model}\left(F_{\text{Hybrid}}\right) = \text{Softmax}\left(W_{\text{FC}}\left([F_{\text{CNN}} \parallel F_{\text{ViT}}]\right) + b_{\text{FC}}\right) \quad (8)$$

where

W_{FC} and b_{FC} are the trainable weights and biases.

F_{CNN} = CNN Feature Layer.

F_{ViT} = Tranformer Encoder Layer.

$\parallel$ = Concatenation parameter.

3. Results

The hybrid CNN-ViT model, was implemented on Python (v3.8+) given its readability, simplicity, and extensive support for scientific computing and machine learning. The project leverages TensorFlow, Keras, and OpenCV for deep learning, image processing, and numerical computations, respectively. The following results were achieved.

Result 1:

A hybrid CNN-ViT system was successfully developed, integrating CNN and ViT models to enhance skin disease classification. The CNN module extracts local features such as textures and edges, while the ViT module captures global dependencies and contextual relationships. These features are then fused to create a comprehensive feature vector, improving the system's classification performance.

Result 2:

The system was trained, tested, and validated using the HAM10000 dataset with a split ratio of 70:15:15 (training: validation: testing). The data instances and relevant features for skin disease classification are provided in **Table 1**, with each feature contributing to the differentiation of diseases like melanoma, actinic keratosis, and basal cell carcinoma.

Training and Validation: The model achieved strong performance during training, with 97.68% accuracy, 90.20% recall, and 91.30% F1-score. These results in-

dicate that the hybrid CNN-ViT model effectively learned to classify skin diseases, demonstrating robustness in performance across various datasets.

Table 1. Data split for SDCM.

Disease	Total	Train (70%)	Validation (15%)	Test (15%)
Benign Keratosis-Like Lesions	1300	910	195	195
Dermatofibroma	1500	1050	225	225
Melanoma	1400	980	210	210
Vascular Lesions	1400	980	210	210
Basal Cell Carcinoma	1400	980	210	210
Melanocytic Nevi	1500	1050	225	225
Actinic Keratoses	1500	1050	225	225

Performance Evaluation Metrics

Performance evaluation was conducted using Accuracy, Precision, Recall, and F1-score. For the seven-class classification problem, Precision, Recall, and F1-score were computed using weighted averaging to account for class imbalance.

Accuracy measures the proportion of correctly classified samples, while Precision evaluates the proportion of correctly predicted positive samples. Recall measures the ability of the model to identify actual positive samples, and F1-score provides a harmonic balance between Precision and Recall. **Table 2** aptly highlights the performance of the SDCS.

Table 2. Performance evaluation of SDCS.

Metric	Value
Accuracy	97.68%
Precision	89.80%
Recall	90.20%
F1-Score	91.30%

Confusion Matrix: The confusion matrix for SDCS (**Figure 2**) shows strong classification performance for most diseases, with a few instances of misclassification, particularly between diseases with similar visual patterns (e.g., melanoma and melanocytic nevi). This indicates the need for further refinement in differentiating these diseases, which may require additional data augmentation or model tuning.

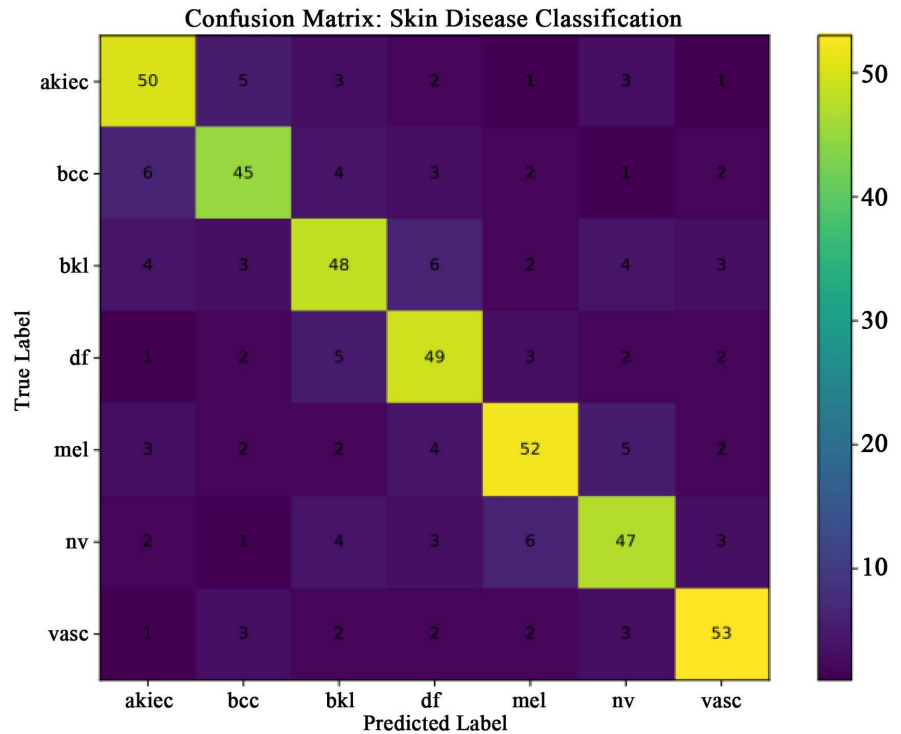


Figure 2. Confusion matrix of SDCS.

Result 3:

A comparative performance evaluation between the hybrid SDCS model and the CNN-only model developed by Agarwal and Godavarthi (2023) showed that the SDCS outperforms the CNN model across all metrics. The SDCS achieved a +14.35% improvement in precision, +12.68% improvement in accuracy, and +17.16% improvement in recall compared to the CNN-only model.

This comparison highlights the effectiveness of integrating CNN and ViT models, combining the strengths of both local feature extraction (CNN) and global context modeling (ViT) for improved classification of skin diseases. **Table 3** shows the performance comparison of the study's model (SDCS) against the Agarwal and Godavarthi model. Graphical highlights in **Figure 3**.

Table 3. Proposed SDCS performance comparison against the existing model.

Metric	Proposed SDCS	Agarwal and Godavarthi (2023) Model
Algorithm	CNN and ViT Algorithms	CNN algorithm
Input Type	Image	Image
Precision	89.80%	75.45%
Accuracy	97.68%	85.00%
Recall	90.20%	73.04%
F1-Score	91.30%	Not Applicable

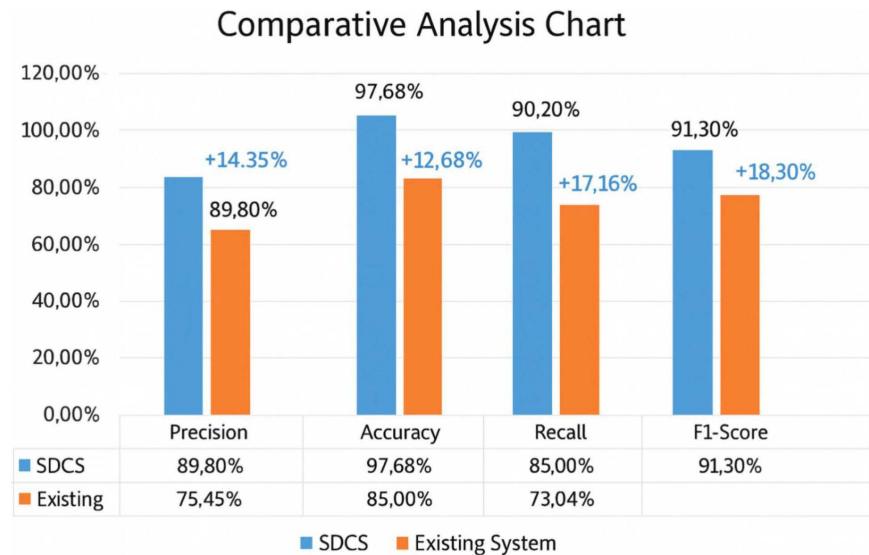


Figure 3. Comparative analysis chart of the CNN model with SDCS.

4. Discussion

The results demonstrate the effectiveness of integrating Convolutional Neural Networks and Vision Transformers for skin disease classification. The hybrid architecture achieved a classification accuracy of 97.68%, indicating a substantial improvement in the model's ability to discriminate among diverse dermatological conditions.

The CNN component effectively extracted localized lesion characteristics, including color variation, border irregularities, pigmentation patterns, and texture information. These features are particularly important in differentiating visually similar dermatological conditions. However, CNN architectures are inherently constrained by local receptive fields and may struggle to capture broader lesion structures and contextual information.

The Vision Transformer component complemented the CNN by capturing long-range spatial relationships through self-attention mechanisms. This enabled the model to analyze global lesion characteristics such as symmetry, shape distribution, and contextual dependencies across the entire image. The fusion of local and global representations significantly enhanced feature discriminability and classification robustness.

The confusion matrix demonstrated strong classification performance across all disease categories. Nevertheless, a limited number of misclassifications were observed between melanoma and melanocytic nevi. This finding is consistent with previous studies [11] [18] and can be attributed to the considerable visual similarities that exist between these lesion types.

The performance achieved in this study is comparable to recent hybrid CNN-ViT frameworks reported in the literature [11]-[13]. The results support emerging evidence that transformer-based architectures can significantly enhance medical image analysis when combined with convolutional feature extraction mechanisms.

Despite these encouraging findings, several limitations exist. First, model evaluation was restricted to the HAM10000 dataset. Consequently, the generalizability of the proposed framework to external clinical datasets remains unverified. Second, patient metadata such as age, gender, lesion location, and clinical history were not incorporated into the classification process. Third, Vision Transformer architectures require substantial computational resources, which may present deployment challenges in resource-constrained healthcare environments.

Future studies should investigate multimodal approaches that integrate image data with clinical information, validate the framework using independent datasets, and explore lightweight transformer architectures suitable for real-time deployment in clinical practice.

5. Conclusions

This study developed and evaluated a hybrid CNN-ViT framework for automated skin disease classification using dermoscopic images from the HAM10000 dataset. By integrating CNN-based local feature extraction with Vision Transformer-based global contextual modeling, the proposed system achieved an accuracy of 97.68%, precision of 89.80%, recall of 90.20%, and F1-score of 91.30%.

The findings demonstrate that hybrid deep learning architectures can significantly enhance dermatological image classification by combining complementary feature representations. The proposed model offers a promising foundation for intelligent computer-aided diagnostic systems capable of supporting dermatologists in clinical decision-making and early disease detection.

Future research should focus on external validation, multimodal learning, model explainability, and computational optimization to facilitate real-world clinical deployment.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

- [1] Del Rosso, J.Q. and Levin, J. (2011) The Clinical Relevance of Maintaining the Functional Integrity of the Stratum Corneum in Both Healthy and Disease-Affected Skin. *Journal of Clinical and Aesthetic Dermatology*, **4**, 22-42.
- [2] Apalla, Z., Nashed, D., Weller, R.B. and Castellsagué, X. (2017) Skin Cancer: Epidemiology, Disease Burden, Pathophysiology, Diagnosis, and Therapeutic Approaches. *Dermatology and Therapy*, **7**, 5-19. <https://doi.org/10.1007/s13555-016-0165-y>
- [3] Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., *et al.* (2017) A Survey on Deep Learning in Medical Image Analysis. *Medical Image Analysis*, **42**, 60-88. <https://doi.org/10.1016/j.media.2017.07.005>
- [4] Tschandl, P., Rosendahl, C. and Kittler, H. (2018) The HAM10000 Dataset, a Large Collection of Multi-Source Dermatoscopic Images of Common Pigmented Skin Lesions. *Scientific Data*, **5**, Article No. 180161. <https://doi.org/10.1038/sdata.2018.161>
- [5] Aydin, Y. (2023) A Comparative Analysis of Skin Cancer Detection Applications Us-

- ing Histogram-Based Local Descriptors. *Diagnostics*, **13**, Article 3142. <https://doi.org/10.3390/diagnostics13193142>
- [6] Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., *et al.* (2018) Skin Lesion Analysis toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC). arXiv: 1902.03368.
- [7] Bengio, Y., Courville, A. and Vincent, P. (2013) Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35**, 1798-1828. <https://doi.org/10.1109/tpami.2013.50>
- [8] Brinker, T.J., Hekler, A., Enk, A.H., Berking, C., Haferkamp, S., Hauschild, A., *et al.* (2019) Deep Neural Networks Are Superior to Dermatologists in Melanoma Image Classification. *European Journal of Cancer*, **119**, 11-17. <https://doi.org/10.1016/j.ejca.2019.05.023>
- [9] LeCun, Y., Bengio, Y. and Hinton, G. (2015) Deep Learning. *Nature*, **521**, 436-444. <https://doi.org/10.1038/nature14539>
- [10] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017) Attention Is All You Need. *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS 2017)*, Long Beach, 4-9 December 2017, 5998-6008.
- [11] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., *et al.* (2021) An Image Is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations (ICLR 2021)*, Vienna, 4 May 2021.
- [12] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L. and Zhou, Y. (2021) TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. arXiv: 2102.04306.
- [13] Adebiyi, A., Abdalnabi, N., Simoes, E.J., Becevic, M., Smith, E.H. and Rao, P. (2024) Transformers in Skin Lesion Classification and Diagnosis: A Systematic Review. medRxiv. <https://doi.org/10.1101/2024.09.19.24314004>
- [14] Goodfellow, I., Bengio, Y. and Courville, A. (2016) Deep Learning. MIT Press.
- [15] Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., *et al.* (2017) Dermatologist-level Classification of Skin Cancer with Deep Neural Networks. *Nature*, **542**, 115-118. <https://doi.org/10.1038/nature21056>
- [16] Agarwal, R. and Godavarthi, D. (2023) Skin Disease Classification Using CNN Algorithms. *EAI Endorsed Transactions on Pervasive Health and Technology*, **9**. <https://doi.org/10.4108/eetpht.9.4039>
- [17] Alshammari, H., Gasmi, K., Ben Ltaifa, I., Krichen, M., Ben Ammar, L. and Mahmood, M.A. (2022) Olive Disease Classification Based on Vision Transformer and CNN Models. *Computational Intelligence and Neuroscience*, **2022**, Article ID: 3998193. <https://doi.org/10.1155/2022/3998193>
- [18] Tan, M. and Le, Q.V. (2019) EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *Proceedings of the 36th International Conference on Machine Learning*, Long Beach, 9-15 June 2019, 6105-6114.