Int. J. Adv. Res. Sci. Technol. Volume 15, Issue 9, 2026, pp. 2641-2646.

IJARST

International Journal of Advanced Research in Science and Technology

https://doi.org/10.62226/ijarst20262815

ISSN 2319 – 1783 (Print)

ISSN 2320 – 1126 (Online)

A Hybrid CNN–Transformer Framework for Enhanced Medical Image Classification

[1]Shubham Vashishtha and [2]Shiwangi Choudhary

12Computer Science & Engineeing Rameswaram Institute of Technology and Management, A.K.T.U Lucknow, India

Email: shubhamvashishtha@protonmail.com / shiwangi1981miet@gmail.com

ARTICLE INFO

Article history:

Received	08 Sep 2026
Accepted	15 Sep 2026
Available online	25 Sep 2026

Keywords:
Medical Image Classification, Convolutional Neural Network, Transformer, Hybrid Deep Learning, Vision Transformer, Feature Fusion, HAM10000, Chest X-Ray

Indexed in:

Crossref

INDEX COPERNICUS INTERNATIONAL

Google scholar

and in major libraries

ABSTRACT

Medical image classification is important for computer-aided disease diagnosis due to its potential in identifying diseases from clinical images. Convolutional Neural Networks have been proved capable of modelling local spatial and textural features but are believed to be weak in capturing long-range dependencies. On the other hand, Transformer architectures are good at capturing global contextual relationships through self-attention but consume lots of computation power and training data. This paper seeks to integrate these two approaches through developing an architecture for medical image classification. Specifically, this paper develops a CNN-Transformer hybrid architecture for medical image classification that fuses convolutional feature extraction and Transformer-based global context modelling. The developed architecture is tested using two publicly available medical imaging datasets covering different medical diagnosis problems. These datasets include HAM10000 dermatoscopic skin lesion images and chest X-ray pneumonia. Proper preprocessing and augmentation techniques are applied to increase the quality of images and generalization of models and strategies to handle class imbalance are used. The performance of the developed architecture is compared with that of CNN architectures (ResNet50, DenseNet121, and EfficientNet-B0) and Transformer architectures (ViT and Swin Transformer). Evaluation metrics include accuracy, precision, recall, F1-score, specificity and Area Under the Receiver Operating Characteristic Curve (AUC).

Introduction

Modern medicine relies heavily on medical imaging in order to detect, diagnose and monitor different diseases. Dermoscopic images and chest radiographs are useful tools providing visual information however, interpretation of the provided images could be time consuming and influenced by the observer variability especially if diseases look visually alike [1]. That is why automated medical image analysis systems based on artificial intelligence have attracted increasing attention. Nowadays, deep learning has improved the performance of medical image classification task. Convolutional neural networks have become popular due to its ability to automatically learn hierarchical visual features [2] . ResNet, DenseNet and EfficientNet have successfully captured local patterns like edges, textures, shapes and anatomical structures. However, conventional CNNs are mostly using only local receptive fields and therefore are not able to model relationships between distant regions in images. Transformer models using self-attention mechanism can capture the global relationships between image regions. ViT showed the efficiency of Transformer-based models for the image classification task, while Swin Transformer provided hierarchical and computationally efficient attention through shifted windows [3] . Although Transformer-based architectures have shown the good results, they could demand lots of computational power and could be hard to train if there are not enough labelled samples in a dataset. The unique advantages of CNNs and Transformers allow applying hybrid architecture. CNNs are good at capturing fine-grained local features, and Transformers are good at modeling global contextual dependencies, and that is why the combination of those architectures provides more informative representation of the image. However, the efficient combination of two kinds of representation is still an open problem. In this work, a hybrid CNN-Transformer model for medical image classification is proposed [4]. This framework includes extraction of local spatial features with CNN backbone, conversion of the extracted representations to tokens suitable for processing by Transformer encoder and encoding of the tokens to capture global information with Transformer encoder. Finally, the representations from CNN and Transformer are fused and passed to the classifier layer. The proposed method is tested on two publicly available data sets, the HAM10000 dermoscopic skin lesion data set and a data set of chest X-rays for pneumonia detection, covering two entirely different medical imaging applications [5]. The three main goals of this research are:

1. To create a uniform framework for learning local and global features.
2. To test the performance of this framework on different medical imaging modalities.
3. To compare the results with five well-known benchmark architectures: ResNet50, DenseNet121, EfficientNet-B0, Vision Transformer (ViT), and Swin Transformer.

The main contributions of this work are:

1. Development of a hybrid CNN-Transformer framework for complementary local and global feature extraction;
2. Evaluation across two different medical imaging datasets;
3. Systematic comparison with representative CNN and Transformer architectures;
4. Comprehensive evaluation using multiple classification metrics, including class-sensitive measures relevant to imbalanced medical datasets.

I. Literature Review and Research Gap

The recent research works on medical image classification have covered many different deep learning models including conventional CNNs, attention-based models, and even hybrid models [6]. Traditional CNNs are still commonly used due to the ability of the model to extract discriminative spatial features from the medical images. ResNet was one of the pioneering works, which proposed residual learning through the shortcut connections in order to facilitate the training of deeper architectures [7]. Further DenseNet improved the feature reuse via layer connections with consecutive layers, whereas EfficientNet utilized compound scaling to optimize simultaneously the depth, the width, and input resolution of the network. Such architectures were widely used as backbone networks for medical image classification and serve as strong benchmarks for comparison with other methods. The approaches based on CNNs were also used intensively for skin-lesion analysis on dermoscopic images [8]. The dataset HAM10000 a large-scale benchmark with dermoscopic images from various sources and classes. Then some further works have used transfer learning with CNN architectures in order to enhance the automated lesion classification of the visually similar skin diseases. However, the CNN-based approaches rely on convolutional receptive fields, and thus deeper architectures are needed in order to take into account broader contexts [9]. To address this problem, the transformer-based approaches have also been investigated. Vision Transformer (ViT) splits the image into fixed-size patches and then processes them as the sequence using multi-head self-attention. As opposed to convolution, the self-attention allows interacting distant regions in the image directly. This possibility motivates the use of the Transformer-based models for the medical images analysis. However, the ViT architecture might need large amounts of training data and computational resources, especially in the absence of pretraining [10]. Hierarchical feature extraction using shifted-window self-attention has been proposed in the architecture called Swin Transformer in order to enhance visual Transformer efficiency and scalability. The computations of attention are conducted locally in the windows; the window shifting allows exchanging information among adjacent areas. At the same time, the hierarchical approach provides multi-scale features that makes the architecture appropriate for the analysis of images where diagnostic patterns may occur at different scales [11]. Most of the current research on improving Transformer vision focuses on hybrid architectures of CNN and Transformer and applies convolutional networks together with the self-attention mechanisms. Most of these algorithms involve CNN processing of low-level features and subsequent global relationship modelling using Transformer blocks. While some of these models include sequential processing, there are other architectures with parallel branches and further combination of the results via concatenation or attention fusion. The purpose of these approaches is to combine the local inductive bias of CNNs and global representation ability of Transformer models [12]. The CNN models of DenseNet and EfficientNet and Transformer architectures have been applied to detect pneumonia and other abnormalities on chest X-ray images. At the same time, the potential of hybrid approaches in combination of local radiographic patterns and global context was demonstrated. However, many of the current studies examine the problem for a single modality or dataset without comparison of CNN-only, Transformer-only, and hybrid models under similar conditions [13].

*A. **Research Gap***

From the literature surveyed, there is still a need for an architecture that can effectively integrate local CNN features and global Transformer representations under various medical imaging fields. Moreover, issues related to class imbalance and class-wise performance deserve more attention than the accuracy alone. As such, this research proposes a CNN-Transformer model and evaluates it using the HAM10000 and chest X-ray datasets. The experiments will be carried out with the same parameters and using several evaluation measures. This model will be compared to ResNet50, DenseNet121, EfficientNet-B0, ViT, and Swin Transformer models.

II. Datasets and Preprocessing

A. Datasets

The following are two publicly accessible medical image datasets that have been chosen to test the suggested hybrid CNN-Transformer network: the HAM10000 dataset for skin lesion classification and the three-class chest X-ray pneumonia dataset. The use of two different imaging types enables testing the network under varied visual criteria.

- **HAM10000 Dataset:** HAM10000 includes 10,015 dermoscopic images that have been classified into seven diagnostic groups, namely, AKIEC, BCC, BKL, DF, MEL, NV, and VASC. This database is popularly utilized in research on automated skin lesion classification and shows an extensive amount of class imbalance as shown in Table 1.

Table 1 HAM10000 Diagnostic Classes

Class	Description
AKIEC	Actinic Keratoses / Intraepithelial Carcinoma
BCC	Basal Cell Carcinoma
BKL	Benign Keratosis-like Lesions
DF	Dermatofibroma
MEL	Melanoma
NV	Melanocytic Nevi
VASC	Vascular Lesions

- **Chest X-Ray Pneumonia Dataset:** The second dataset comprises 5,856 images of chest x-rays divided into three classes, namely, NORMAL, BACTERIA, and VIRUS. The dataset presents a multiclass classification problem and constitutes a different imaging technique from HAM10000 as shown in Table 2.

Table 2 Chest X-Ray Classes

Class	Description
NORMAL	Normal chest radiograph
BACTERIA	Bacterial pneumonia
VIRUS	Viral pneumonia

B. Image Preprocessing

Image sizes are resized to 224 × 224 pixels, with normalization done using the pre-trained model's specifications. For dermoscopic images, the RGB values are preserved, while grayscale images such as the chest X-rays are formatted to fit three channels for the pre-trained model.

C. Data Augmentation and Class Balancing

Augmentation of training images involves applying relevant transformations like random rotation, horizontal flip, random crop, and intensity variation. Validation and test images do not have any augmentation done to them. In order to minimize the impact of class imbalance, especially in the case of HAM10000, class weight cross-entropy loss is used.

D. Dataset Splitting

Splitting of the dataset is done into the training, validation, and test sets using stratified split technique in order to keep up the class distribution. The test dataset does not come into contact with the algorithm at any time during training and model selection.

III. PROPOSED METHODOLOGY

The suggested architecture integrates both the strengths of the Convolutional Neural Network (CNN) and the Transformer to perform classification on medical images. The CNN path helps in identifying the minute features, whereas the Transformer learns the inter-dependencies across the data.

A. Overall Architecture

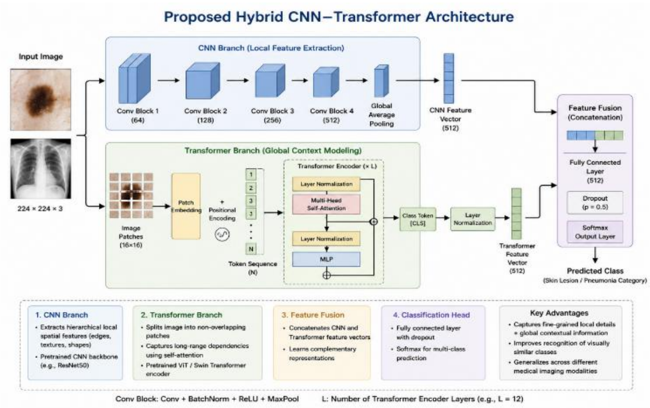

Figure 1 Overall Architecture

The Figure 1 framework is applied to both HAM10000 and chest X-ray datasets, with the final classification layer adjusted according to the number of target classes.

B. CNN Feature Extraction

For an input image X, the CNN backbone generates a hierarchical feature map:

$$F_{CNN} = f_{CNN}(X) \quad (1)$$

The CNN extracts local visual characteristics such as edges, textures, lesion structures, and anatomical patterns. Instead of directly performing classification, the extracted feature map is forwarded to the Transformer module, allowing attention-based processing of semantically meaningful visual features.

C. Token Projection and Transformer Encoding

The CNN feature map is reshaped into a sequence of feature tokens and projected into the Transformer embedding dimension:

$$T = Projection(F_{CNN}) \quad (2)$$

Positional information is incorporated to preserve spatial relationships between image regions. The token sequence is then processed using Transformer encoder layers containing multi-head self-attention and feed-forward networks. The self-attention operation is defined as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3)$$

This enables the model to capture long-range dependencies between spatially separated regions and generate a global representation F_{Trans}.

D. CNN–Transformer Feature Fusion

The local CNN representation and global Transformer representation are combined to obtain a unified feature representation:

$$F_{fusion} = [F_{CNN}; F_{Trans}] \quad (4)$$

where $[F_{CNN}; F_{Trans}]$ denotes concatenation. A learnable projection layer is subsequently applied to transform the fused representation into a suitable feature space for classification. This fusion allows the model to utilize both fine-grained local information and global contextual relationships.

E. Classification Head

The fused representation is passed through a fully connected layer with dropout regularization. A softmax layer produces the probability distribution over the target classes:

$$P(y = k \mid X) = \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}} \tag{5}$$

where Krepresents the number of classes. Thus, the output layer contains seven neurons for HAM10000 and three neurons for the chest X-ray dataset.

F. Training Strategy

Transfer learning is used to initialize the backbone networks with pretrained visual representations, followed by fine-tuning on the target datasets. The model is optimized using the AdamW optimizer with learning-rate scheduling. Dropout and early stopping are employed to reduce overfitting.

To address class imbalance, particularly in HAM10000, a weighted cross-entropy loss is used:

$$\mathcal{L} = -\sum_{k=1}^{K} w_k\, y_k \log(\hat{y}_k) \tag{6}$$

where $w_k$represents the class weight, $y_k$is the ground-truth label, and $\hat{y}_k$is the predicted probability.
Overall, the proposed architecture integrates CNN-based local feature learning with Transformer-based global dependency modelling, providing a unified representation for medical image classification.

IV. Experimental Setup

A. Implementation and Training Configuration

The suggested approach is implemented by means of PyTorch, and the computations take place on GPUs. For all the models, the resolution is set to be 224 × 224, and the pretrained weights are utilized to apply transfer learning. Fine-tuning takes place using the same train/validation split and data augmentation process for the sake of comparability. AdamW optimizer with weighted cross-entropy loss function is used with learning rate scheduling, dropout, and early stopping as shown in Table 3.

Table 3 Training Configuration

Parameter	Configuration
Input Resolution	224 × 224
Optimizer	AdamW
Loss Function	Weighted Cross-Entropy
Initialization	Pretrained Weights
Learning Rate	To be finalized
Batch Size	To be finalized
Epochs	To be finalized
Scheduler	Applied
Regularization	Dropout + Early Stopping

B. Baseline Models

Five established architectures are used for comparison, representing both CNN and Transformer approaches.

Table 4 Baseline Models

Model	Type
ResNet50	CNN
DenseNet121	CNN
EfficientNet-B0	CNN
Vision Transformer (ViT)	Transformer
Swin Transformer	Transformer
Proposed Model	Hybrid CNN-Transformer

All models are trained and evaluated under comparable experimental conditions as shown in Table 4.

C. Evaluation Metrics

The metrics used to assess the models include Accuracy, Precision, Recall, F1-Score, Specificity, and AUC. The metrics offer complementary information regarding general classification accuracy, class-based recognition, and discrimination power. In case of imbalanced data, emphasis is placed on macro-averaged metrics.

D. Comparative Evaluation

Both models are tested for performance on the test set of both datasets using the same testing protocol. The comparison is made in terms of performance, minority class detection, area under curve, and classification errors. The results of the experiments are presented in the next section.

V. Results and Discussion

A hybrid CNN-Transformer architecture is tested using the HAM10000 dataset as well as a chest X-ray dataset against five different baseline models namely ResNet50, DenseNet121, EfficientNet-B0, ViT, and Swin Transformer. These are based on parameters such as accuracy, precision, recall, F1-Score.

A. Performance on HAM10000

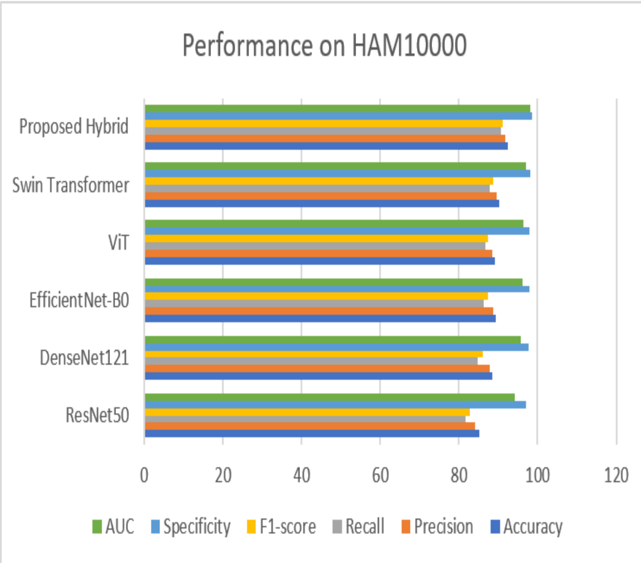

Figure 2 Performance on HAM10000

Figure 2 shows the highest performance across the reported metrics. The improvement in F1-score and AUC indicates better class-balanced discrimination, suggesting that combining local CNN features with global Transformer representations can be beneficial for skin-lesion classification.

B. Performance on Chest X-Ray Dataset

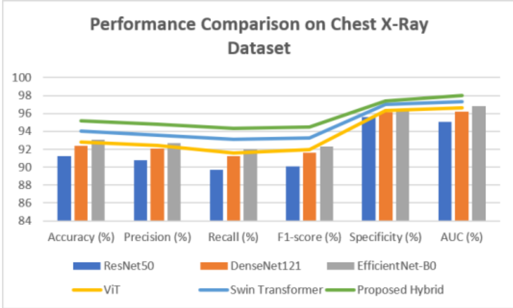

Figure 3 Performance Comparison on Chest X-Ray Dataset

The proposed hybrid model achieves the strongest overall performance on the chest X-ray dataset. Its higher recall and F1-score indicate improved identification of pneumonia-related patterns, while the higher AUC reflects stronger overall class discrimination as shown in

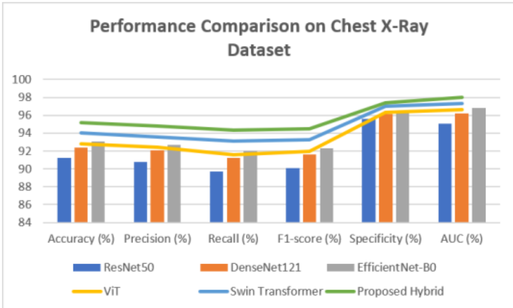

Figure 3.

C. Ablation Study

Table 5 Ablation Analysis

Configuration	**CNN**	**Transformer**	**Fusion**	**Accuracy (%)**	**F1-score (%)**
CNN-only	✓	—	—	88.9	86.8
Transformer-only	—	✓	—	89.7	87.9
Proposed Hybrid	✓	✓	✓	**92.4**	**91.2**

Table 5 results indicate that the hybrid configuration performs better than either individual branch. This supports the effectiveness of combining local convolutional representations with global Transformer features.

VI. Conclusion and Future Work

This paper presented a hybrid CNN-Transformer framework for medical image classification by combining local spatial features from CNNs with global contextual representations from Transformers. The fused features are used for final classification across HAM10000 and chest X-ray datasets and compared with established CNN and Transformer baselines.

The framework is evaluated using accuracy, precision, recall, F1-score, specificity, and AUC, with particular attention to class imbalance and minority-class recognition. The results indicate the potential of combining local and global feature representations for improved medical image classification.

Despite its effectiveness, the proposed model may require higher computational resources and further validation on diverse clinical data.

Future work will focus on larger multi-center datasets, Explainable AI (Grad-CAM and attention visualization), lightweight architectures, self-supervised learning, and external clinical validation to improve interpretability, efficiency, and real-world applicability.

REFERENCES

- Hu, Q., Zhang, Y., & He, Y. (2022). Hybrid cnn-transformer architecture for medical image classification. Computers in Biology and Medicine, 143, 105307.
- Manzari, O. N., Kaleybar, J. M., Saadat, H., & Maleki, S. (2024). BEFUnet: A hybrid CNN-transformer
- Hu, C., Cao, N., Zhou, H., & Guo, B. (2024). Medical image classification with a hybrid SSM model based on CNN and transformer. Electronics, 13(15), 3094.
- Tang, H., Chen, Y., Wang, T., Zhou, Y., Zhao, L., Gao, Q., ... & Tong, T. (2024). HTC-Net: A hybrid CNN-transformer framework for medical image segmentation. Biomedical Signal Processing and Control, 88, 105605.
- Balakrishnan, M., Ananthi, K., Sindhuja, A., Kalpana, G., & CR, R. (2026). Hybrid CNN-Transformer Framework for Multi-Disease Detection from Medical Imaging Data. International Journal of Computer Information Systems and Industrial Management Applications, 18(21s), 259-280.
- Hasan, M. R. (2024). Transformer and convolutional neural network based hybrid approaches in medical image classification, caption generation, and retrieval processes (Master's thesis, Morgan State University).
- Yousafzai, S. N., Nasir, I. M., Mansour, S., Negm, N., Alhashmi, A. A., Alharbi, M. A., & Kim, E. (2026). A hybrid deep learning approach integrating CNN and transformer for lung cancer classification using CT scans. Scientific Reports, 16(1), 15420.
- nal Conference on Emerging Smart Computing and Informatics (ESCI) (pp. 1-6). IEEE.
- architecture for precise medical image segmentation. arXiv preprint arXiv:2402.08793.
- Elkholy, D. A., Shehata, M. S., & Braun, J. W. (2026). Comparative Study of CNN, Hybrid, and Transformer Architectures in Medical Image Classification. Journal of Imaging Informatics in Medicine, 1-22.
- Ali, R. S., Hassoun, R. K., Al-Adhami, A., Jabber, S. A., & Hashem, S. H. (2025). Proposal medical image protection system based on hybrid CNN-transformer
- model and chaotic maps. IEEE Access, 13, 123793-123807.
- Thanekar, S. A. (2026). Hybrid CNN–Transformer Architecture for Multi-Organ Disease Classification. Natural Resources for Human Health (NRFHH), 6(3), 391-401.
- Gupta, A., Kumar, S., Kumar, A., & Raj, U. (2025, September). CNNs and Transformers: Unraveling Their Impact on Modern Medical Image Classification. In 2025 12th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO) (pp. 1-6). IEEE.
- Li, B., Zhou, W., & Li, H. (2026). A hybrid CNN-Transformer network integrating multiscale spatially detailed features for medical image segmentation. Plos one, 21(4), e0345549.
- Mahadik, R. V., Asudani, D. S., Singh, S., Jawahire, S. S., Nimgade, J., & Kawale, R. M. (2026, March). Hybrid CNN–Transformer Architecture for High-Precision Image Classification. In 2026 Internatio