IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154282

Spatially-Gated CNN-Transformer Hybrids for Pneumonia Classification: A Unified Framework for Metric-Optimized Local-Global Explainability

Ch Mydhili[1*], B Madhav Rao[2]

M. Tech Student, Department of CSE, Sir C R Reddy College of Engineering, Eluru, India[1*]

Professor, Department of CSE, Sir C R Reddy College of Engineering, Eluru, India[2]

Abstract: The deep learning paradigm shift is completely transforming image classification especially in the medical scenario where proper classification helps establish the patients' treatment protocol. CNN has proven to be very successful in learning local spatial features such as textures and edges, whereas Swin-T excel in capturing global context dependencies through self-attention mechanism. However, it is evident that both methods have some shortcomings in case of considering only one approach since CNN fails to adequately capture the long-range dependency while Transformer models require huge amounts of data and powerful computations. Although hybrid CNN-Transformer architectures provide an answer to each architecture's limitation by combining their feature learning abilities into one, they remain black-box models which produce hard-to-explain predictions and hence cannot be trusted especially in the medical field where accurate classifications are required. This project aims at finding an answer to this problem by introducing an innovative model called Spatially-Gated CNN-Transformer Hybrid with Dual-Level Explainability. The model will consist of a ResNet-50 CNN backbone to learn local features, and a Swin Transformer to model global context through spatial attention gating mechanism. Besides, this research will introduce a dual-level explainability method involving both Grad-CAM and Attention Rollout.

Keywords: **CNN**–Transformer Hybrid Network, Dual-Level Explainability, Explanation Fusion, Quality Evaluation Metrics.

I. INTRODUCTION

Deep Learning has proved itself as a revolutionary innovation in the field of computer vision, allowing automated systems to develop more sophisticated abilities in analyzing images. The significance of Deep Learning in this area has become increasingly evident in the past few years in the domain of medical imaging, especially in terms of diagnosing diseases on time and correctly. While many medical imaging techniques are utilized today, chest X-rays are still considered one of the most frequently used diagnostics in the identification of lung ailments including pneumonia. Interpreting chest X-rays manually can be very tedious, not to mention that it leaves room for errors because of the inter-observer variability. This has inspired the creation of an automated system capable of assisting clinicians in accurately predicting.

Convolutional neural networks (CNNs) have been pioneers in navigating through this transition due to their ability to extract hierarchical representations of visual data. Specifically, thanks to the use of convolutional operators on visual images, CNNs are capable of representing spatial relationships in terms of edges, texture, patterns, etc., which makes it possible to recognize diseases from medical imaging. Moreover, some advanced CNN architectural design, including ResNet, DenseNet, and EfficientNet, have shown outstanding performance in various tasks, especially detection of diseases from chest x-ray images. Despite all of the above-mentioned benefits, however, the natural limitation of the CNN model is related to its bounded receptive field. This limitation is associated with the fact that CNN cannot learn long-term dependencies between image parts.

In order to resolve the above-discussed issues, another technique called Swin Transformers is being developed for image analysis tasks. Being motivated by the success of Transformers in natural language processing tasks, Swin-T rely on the concept of self-attention in order to learn interdependencies between different regions of an image. This is done by treating the image as a sequence of its patches, which then allows for global dependencies modeling. Due to this, Transformers are highly appropriate for dealing with complex spatial relationships in medical images. Moreover, such modifications of Transformer models as Swin Transformer allow learning hierarchies and using shifted window

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.154282

attention, which significantly increases efficiency in handling large images while keeping the computation affordable. Nevertheless, in contrast to CNNs, Transformer models need larger datasets for successful training and have more complicated computational structure.

Acknowledging the benefits and drawbacks of CNNs and Transformers, there has been considerable effort devoted to creating hybrid networks incorporating aspects from both techniques. Hybrid CNN-Transformer architectures are designed to leverage the ability of CNNs to extract features locally, along with the capacity of Transformers to model contexts globally. Although hybrid CNN-Transformers have exhibited success in many areas, including image classification, the primary focus remains on achieving high accuracy rather than understanding how the network makes its predictions, leading to a black-box model. For medical diagnoses, this is problematic because healthcare professionals need understandable reasoning behind the predictions.

As a result, Explainable Artificial Intelligence (XAI) has become an indispensable field of study that addresses the problem of interpretability of deep learning models. Grad-CAM is one such method used for generating visual explanations by emphasizing crucial areas within the input image. On the other hand, SHAP and LIME are approaches used for explaining features. Attention mechanisms have also been devised for the purpose of explaining the Transformer model, with attention maps serving as visualization tools. While these techniques play a key role in promoting better transparency, most of them lack the ability to generate explanations that are both local and global at once. In addition, there is no standard measure to assess the accuracy of the generated explanations.

Another major problem in explainable AI research is the inadequacy of evaluation criteria that would measure how good these explanations are. Several measures, including faithfulness, fidelity, localization, stability, and robustness, have been defined; however, these measures remain standalone and are not incorporated into an overarching evaluation framework. Hence, it becomes hard to objectively compare one method of generating explanations to another. Consequently, there has emerged a clear need for frameworks that do not only explain how the machine works but also evaluate the efficacy of these explanations.

In light of the above concerns, this study proposes a CNN-Transformer hybrid model, which ensures good performance while being highly interpretable. In particular, the proposed model utilizes the capabilities of ResNet-50 CNN model for feature extraction and Swin Transformer for understanding global relations between features. To facilitate effective integration of the two types of features, spatial attention gating is used, which enables dynamic mixing of both features depending on their significance. Besides delivering superior performance, the proposed model uses a dual interpretation method, which entails conducting analysis locally and globally. The model uses Grad-CAM to generate heatmaps for disease detection, whereas Attention Rollout is used for global relation visualization.

Moreover, the framework also presents an explanation fusion approach to integrate both local and global explanations through an optimization approach to perform alpha searching. The framework guarantees that the generated explanation provides a balanced view of the local and global components. For evaluating the quality of the generated explanation, a thorough evaluation framework is adopted that uses various quantitative criteria to create a Q-score.

The contributions of this paper can thus be summarized in three ways. First, the authors provide an architecture which is based on hybrid CNN-Transformer networks, which allow the combination of both global and local learning of features. Secondly, it provides a framework of dual-level explainability and fused explainability. Finally, it provides a quantified metric for evaluation of the quality of explanations provided by models. Through their experiments, the authors were able to show that the model they propose not only works but is also highly accurate.

Finally, this study plays a significant role within the field of medical image processing from the perspective of gaining achievements in the correlation between performance and interpretability. The main contribution made by the authors of the paper can be explained through the combination of the methods used for deep learning and explainable artificial intelligence.

II. LITERATURE SURVEY

A. Overview of Deep Learning in Medical Image Analysis

The use of deep learning for the analysis of medical images has led to remarkable innovations in detecting diseases automatically and with high precision. As reported early on in a study by Daniel S. Kermany et al. [26], the efficacy of the application of deep learning in determining various diseases based on their images is well proven. In recent years,

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154282

improvements have been made in terms of accuracy and robustness, as noted by Ting Wang et al. [11], whose study explained the significance of explainable AI in various fields, including medicine.

B. CNN-Based Approaches for Medical Imaging

CNNs have extensively been applied to medical image classification, owing to their ability to extract local spatial features. This has been reported by Sheng Niu and Zhang [20] as well as Xiaobo Zhao et al. [25]. In terms of application, M. M. Auzine et al. [6] and M. I. Nazir et al. [7] proposed models that utilized CNNs with XAI to detect diseases. While CNNs can easily extract local features, they are not efficient at extracting global relationships and context.

C. Vision Transformer-Based Models

Vision Transformers overcome the drawbacks of CNNs through the use of self-attention models for capturing global dependencies. The effectiveness of Vision Transformers in capturing long-distance dependencies was shown by Laleh Barekatain and Ben Glocker [5] in medical imaging. However, despite their advantages, Vision Transformers require massive amounts of data and are resource-intensive. Moreover, their interpretability is less straightforward.

D. Hybrid CNN-Transformer Architectures

Hybrid methods incorporate both CNN and Transformer networks to gain advantages from both local and global feature extraction. Hybrid models have been introduced by Mohammed Alshomrani et al. [1] and K. Djoumessi et al. [2], which performed well in terms of classification. The same case applies to M. Alhumaid and A. G. Fayoumi [3] and D. Pantelaios et al. [4]. Nevertheless, the latter have not paid much attention to the problem of explainability.

E. Evaluation of Explainability Methods

These methods include Grad-CAM, SHAP, and LIME which have been widely used for the interpretations of predictions made by the model. The classification of these methods was done in the research conducted by D. Bhati et al. [16]. The issues related to inconsistency and reliability were identified by M. A. Awal and C. K. Roy [12], [15]. Concepts such as faithfulness and fidelity are among some criteria for evaluating that have emerged from the researches by X. Zheng et al. [13] and J. Zhou et al. [14].

F. Research Gap and Motivation

As seen in previous literature, even though CNNs and Transformers perform exceptionally well on their own, there exist some constraints to their effectiveness in identifying local and global features. Although hybrid models perform better compared to others, they fail in the aspect of being completely interpretable. Existing approaches for model explainability only offer partial results. In addition, these methods are not standardized for quantifiable assessment. This problem can be solved by our proposed method.

III.METHODOLOGY

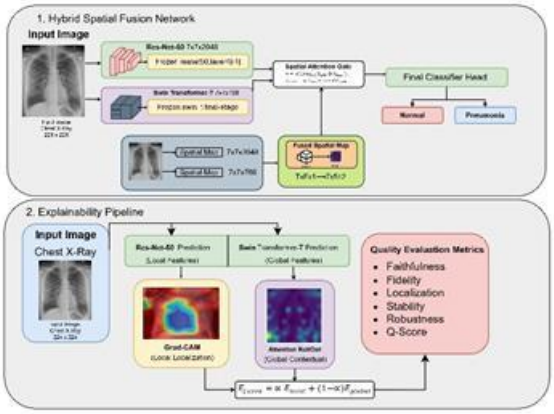

Fig: Proposed Methodology

A. Overview

In the suggested approach, a hybrid architecture that combines the advantages of convolutional neural networks and vision transformers is used to effectively extract both local and global representations from chest x-rays. The rationale for employing such an architecture is due to its complementary properties: convolutional neural networks are excellent

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154282

at capturing spatially localized features, while vision transformers can be more easily able to model relationships in space through the attention mechanism.

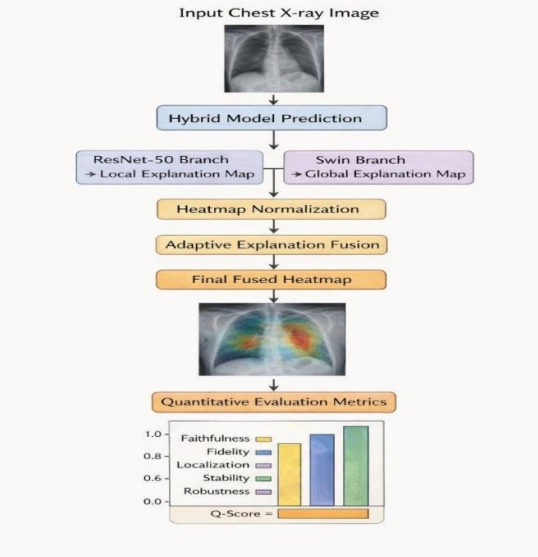

Fig: Model Workflow

B. Data Preprocessing and Input Representation

Input features include chest X-ray scans that go through different preprocessing procedures to stabilize the input process for consistent training of the model. The image scans are scaled to a uniform size of 224 × 224 pixels, which corresponds with the requirement for input dimensions by both the CNN and Transformer networks. The intensity values are normalized within a range to make sure there is numeric consistency to enhance faster convergence of the model. Random transformations such as rotation, flipping, and scaling are also employed to boost the variability of the input data and avoid overfitting of the model.

C. CNN-Based Local Feature Extraction (ResNet-50 Branch)

Once the preprocessing stage has been carried out, the image goes through a CNN branch using ResNet-50 as the pre-trained architecture. The idea behind having a CNN branch here is to capture local spatial features from the image that are likely to be indicative of any abnormalities in the lungs, such as textures and fine structures. With residual learning in ResNet-50, it becomes easy to train deep neural networks without degradation, thus enabling the extraction of very discriminative features.

$$F_{CNN} \in R^{7*7*2048}$$

Where:

- F_{CNN} is feature map extracted from the CNN (ResNet-50)
- R, represents real-valued feature space
- $7 * 7$, represents spatial dimensions of feature map
- 2048, represents no.of feature channels (deep features)

D. Transformer-Based Global Feature Extraction (Swin Transformer Branch)

The Swin Transformer processes the image parallel to the CNN branch, where it performs the job of capturing the global context of the relationship of the input image. This is because the Swin Transformer does not follow the principle used by CNN to model local connections; rather, it utilizes self-attention mechanisms, making it easier to capture the relationships between distant parts of an image, which is very essential in medical imaging due to the presence of diseases that could be located in different parts of the image.

$$F_{Swin} \in R^{7*7*768}$$

Where:

- F_{ViT}, feature map from Swin Transformer
- $7 * 7$, Spatial size of output feature map

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154282

- 768, no.of Transformer feature channels

E. Feature Projection and Dimensional Alignment

In order to ensure that there can be an efficient fusion of the features extracted from both the branches, it becomes imperative to project their outputs in the same embedding space. This task is performed using 1 × 1 convolution layers as they help in reducing dimensionalities without causing any loss of information. It plays a very important role in efficient feature fusion.

$$F' = Conv_{1*1}(F)$$

Where:

- F, Input feature map (CNN or transformer output)
- $Conv_{1*1}$, 1*1 convolution operation
- F', projected feature map (aligned dimensions)

F. Spatial Attention Gating Mechanism

Rather than combining these characteristics through a direct connection, the proposed scheme presents an attention-based spatial gating method, which enables the adjustment of contributions from each branch. It computes a weighting map for each point based on its relevance, thus determining the significance of both local and global characteristics. The attention gate was constructed with the use of convolutions and then applying a sigmoid function to create a normalized weight matrix.

$$\propto = \sigma(Conv_{3*3}(F_{CNN} \oplus F_{Swin}))$$

Where:

- $\propto$, Spatial attention weight map
- σ, Sigmoid activation function (range 0-1)
- $Conv_{3*3}$, convolution layer
- F_{CNN}, CNN map feature
- F_{ViT}, Transformer feature map
- $\oplus$, Concatenation operation

G. Adaptive Feature Fusion Strategy

The fused representation is obtained by performing a weighted sum of the outputs of CNN and Transformer. This technique makes sure that local and global information is preserved in the process of fusion. The use of complementary features makes the model learn a representation that can be used for both classification and localization of the image.

$$F_{fused} = \propto F_{CNN} + (1- \propto)F_{Swin}$$

Where:

- F_{fused}, final fused feature map
- $\propto$, Weight for CNN features
- $1-\propto$, weight for Transformer features
- F_{CNN}, CNN feature map
- F_{ViT}, Transformer feature map

H. Classification Head and Decision Layer

Then the merged features are sent to a classifier block to make the final prediction. A Global Average Pooling operation is first conducted to downsize the spatial dimensions, and then the features undergo fully-connected operations to reason at higher levels. The output layer applies a sigmoid activation function to compute the probability for binary classification. The use of dropout regularization ensures that the model does not overfit on training data.

$$y = \sigma(Wx + b)$$

Where:

- y, Output probability (prediction)
- W, weight matrix
- x, input feature vector
- b, bias term
- σ, sigmoid activation function

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154282

I. Dual-Level Explainability Framework

A significant contribution of this work is the integration of a dual-level explainability framework, which provides both local and global interpretations of the model's decisions.

$$E_{fused} = \propto E_{local} + (1-\propto)E_{global}$$

Where:

- E_{fused}, final explanation map
- E_{local}, Grad-CAM output
- E_{global}, Attention Rollout output
- $\propto$, weight controlling contribution

J. Local Interpretability using Grad-CAM

The Grad-CAM technique is used on the CNN side of the architecture to create heat maps that indicate which areas in an image play a significant role in influencing the prediction made by the model. Grad-CAM can accurately identify lung infection in cases using chest X-ray images.

K. Global Interpretability using Attention Rollout

In addition to the local explanation techniques, the method of Attention Rollout has been used on the Transformer branch. In order to get an overall explanation of the model, this technique uses attention weights from multiple layers for visualization of how various parts of the image are correlated with each other.

L. Explanation Fusion using Alpha Optimization

The two explanations are then merged together through a weighting scheme using a parameter α. The parameter α will not be set manually but rather is determined automatically using the Alpha Search method, which helps determine the optimal value for α to obtain a balance between the two types of explanations. The system tests several α values to obtain the best explanation.

M. Qualitative Evaluation of Explainability

To ensure the reliability of explanations, the framework incorporates a comprehensive evaluation strategy based on multiple qualitative metrics.

Faithfulness Metric

Faithfulness measures how accurately the explanation reflects the model's decision-making process by analyzing the impact of removing important regions.

$$Faithfulness = f(x) - f(x \odot (1 - M))$$

Where:

- f(x), Model prediction for original input
- x, input image
- M, Explanation mask
- $\odot$, element wise multiplication
- $1 - M$, mask removing important regions

Fidelity Metric

Fidelity evaluates whether the model produces similar predictions when only the important regions are retained.

$$Fidelity = |f(x) - f(x \odot M)|$$

Where:

- f(x), original prediction
- M, important region mask
- $x \odot M$, image with only important regions
- $|\,.\,|$, absolute difference

Localization Accuracy

Localization measures how well the explanation highlights disease-relevant areas within the image.

$$Localization = \frac{\sum_{i,j} \quad M_{i,j} \,.\, G_{i,j}}{H * W}$$

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154282

Where:

- $M_{i,j}$, Explanation map pixel
- $G_{i,j}$, Ground truth pixel
- H, image height
- W, image width
- $\sum$, Summation over pixels

Stability Analysis

Stability assesses the consistency of explanations under small perturbations in the input.

$$Stability = 1 - \frac{1}{N}\sum_{i=1}^{N} ||E_i - E_{orig}||$$

N. Robustness Evaluation

Robustness measures the resilience of explanation maps to noise and variations.

$$Robustness = 1 - \frac{|E_{noise} - E_{orig}|}{|E_{orig}|}$$

Where:

- E_{noise}, Explanation with noise
- E_{orig}, original explanation
- $|\,.\,|$, norm function

O. Unified Q-Score Computation

All metrics are combined into a unified Quality Score (Q-Score), providing a single measure of explanation effectiveness.

$$Q = \frac{Faithfulness + (1 - Fidelity) + Localization + Stability + Robustness}{5}$$

Where:

- **Q**, overall explainability score (Quality score)
- **Faithfulness**, explanation accuracy
- **Fidelity**, prediction preservation measure
- **Localization,** region accuracy
- **Stability,** consistency measure
- **Robustness,** noise resistance

IV. RESULTS AND DISCUSSION

A. Experimental Setup and Evaluation Protocol

The experiment was performed on the basis of an openly accessible database of chest X-rays. The dataset was split into the training set, validation set, and test set in order to guarantee objective assessment. Several deep learning architectures were developed and compared.

B. Classification Performance Analysis

The proposed hybrid model achieved superior performance compared to existing architectures. The integration of local and global features significantly improved classification accuracy, demonstrating the effectiveness of the hybrid approach.

Accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Where:

- **TP,** True Positives
- **TN,** True Negatives
- **FP,** False Positives
- **FN,** False Negatives

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154282

Precision:

$$Precision = \frac{TP}{TP + FP}$$

Where:

- **TP,** True Positives
- **FN,** False Negatives

Recall:

$$Recall = \frac{TP}{TP + FN}$$

Where:

- **TP,** True Positives
- **FP,** False Positives

F1 Score:

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

Where:

- **Precision,** positive prediction accuracy
- **Recall,** Detection ability

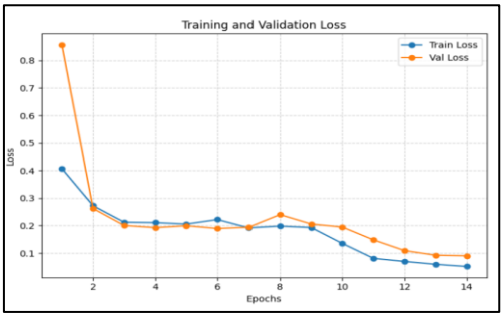

Fig: Training and Validation loss for ResNet50

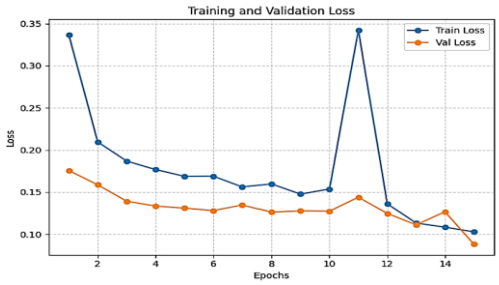

Fig: Training and Validation loss for Swin Transformer

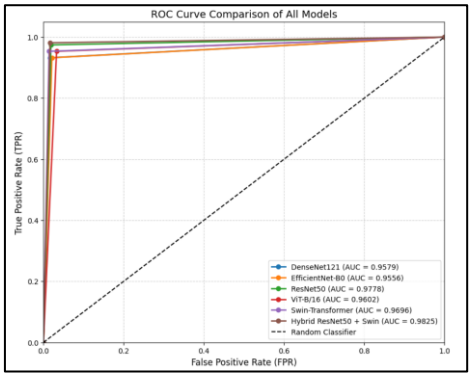

Fig: ROC Curve for all the models

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154282

Table 1: Confusion matrix of the hybrid model

Hybrid ResNet50 +Swin Transformer		
Actual Class	**Predicted Normal**	**Predicted Pneumonia**
Normal	**TN = 420**	**FP = 7**
Pneumonia	**FN = 8**	**TP = 420**

C. Comparative Analysis with Existing Models

Compared to ResNet50, DenseNet121, EfficientNet, and Vision Transformer, the hybrid model consistently outperformed all baselines. This highlights the advantage of combining CNN and Transformer architectures.

Table 2: Performance comparison of various models

Model	Accuracy (%)	Precision	Recall	F1 Score	AUC
DenseNet121	95.79	0.9828	0.9322	0.9568	0.9579
EfficientNet-B0	95.56	0.9779	0.9322	0.9545	0.9556
ResNet50	97.78	0.9812	0.9743	0.9777	0.9778
ViT-B/16	96.02	0.9668	0.9533	0.9600	0.9602
Swin-T	96.96	**0.9855**	0.9533	0.9691	0.9696
Hybrid ResNet50 + Swin T	**98.25**	0.9836	**0.9813**	**0.9825**	**0.9825**

D. Explainability Performance Evaluation

The explainability framework was evaluated using multiple metrics. The results indicate that the proposed fusion approach achieves a higher Q-score compared to individual methods, confirming its effectiveness.

E. Analysis of Local and Global Explanation Maps

Local explanations from Grad-CAM have been quite successful in identifying infections, whereas global explanations help us understand spatial relationships better. This combined approach gives us a holistic picture of the model.

Table 3: Comparison of explainability methods using evaluation Metrics

Method	Faithfulness	Fidelity	Localization	Stability	Robustness	Final Q-Score
DenseNet121						
Grad-CAM	0.0056	0.1991	**0.3673**	0.7871	0.7484	0.4215
SHAP	**0.9667**	**0.9988**	0.0062	0.9380	0.9348	**0.7689**
LIME	0.0376	0.6791	0.1876	0.7366	0.7292	0.4740
SmoothGrad	0.3575	0.9978	0.0004	**0.9624**	**0.9590**	0.6554
EfficientNet-B0						
Grad-CAM	**0.9231**	0.0852	**0.6939**	0.6213	0.6390	**0.5925**
SHAP	0.0052	0.9841	0.0014	0.9587	0.9573	0.5814
LIME	0.134	0.9387	0.3746	0.6636	0.6301	0.5482
SmoothGrad	0.0047	**0.9659**	0.0002	**0.9790**	**0.9842**	0.5868
ResNet50						
Grad-CAM	**0.3023**	**0.9998**	**0.3061**	0.8490	0.8046	**0.6523**
SHAP	0.0351	0.9998	0.0016	0.9535	0.959	0.5898
LIME	0.2534	0.9959	0.1668	0.7782	0.6589	0.5706
SmoothGrad	0	0.9996	0.0008	**0.9867**	**0.9867**	0.5948
Vision Transformer (ViT-B/16)						
Grad-CAM	0.0010	0.0015	**0.6837**	0.8623	0.8622	0.4821
SHAP	0.0012	0.0392	0.0022	0.9644	0.965	0.3944
LIME	**0.0055**	0.2712	0.3653	0.6688	0.7302	0.4082
SmoothGrad	0.0019	0.0010	0.0006	0.9865	0.9868	0.3950
Attention Rollout	0	**0.5575**	0.0008	**0.9978**	**0.9978**	**0.5108**
Swin Transformer						
Grad-CAM	**0.0505**	0.0417	**0.5631**	0.9085	0.9118	0.4951
SHAP	0.0352	**0.9337**	0.0005	0.9728	0.9654	**0.5815**

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154282

LIME	0.035	0.6494	0.2009	0.6890	0.6405	0.4430
SmoothGrad	0	0.0012	0.0007	**0.9862**	**0.9869**	0.3950
Attention Rollout	**0.0418**	0.8808	0.0017	0.9660	0.9627	**0.5706**

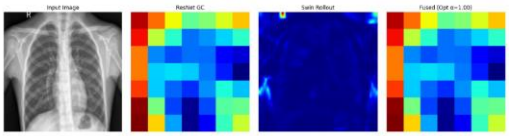

Fig: Analysis of Local and Global Explanation Maps

Table 4: Comparison of evaluation Metrics for same sample image

Model & Method	Faithfulness	Fidelity	Localization	Stability	Robust-ness	Final Q-Score
ResNet50 Grad-CAM	0.3023	0.9998	0.3061	0.8287	0.781	0.6436
Swin Attention Rollout	0.0029	0.9139	0.0017	0.9939	0.9916	0.5808
Hybrid Adaptive Fused Map	0.3023	0.9998	0.3061	**1**	**1**	**0.7216**

F. Impact of Explanation Fusion Strategy

The fusion of local and global explanations significantly improves interpretability. The Alpha optimization ensures that the best balance is achieved, enhancing the overall quality of explanations.

G. Clinical Significance and Practical Implications

From a clinical perspective, the proposed model provides reliable predictions along with interpretable explanations. This enhances trust among healthcare professionals and supports its application in real-world diagnostic systems.

V. CONCLUSION

The proposed network in this paper uses a combination of CNN and Transformer models called ResNet-50 and Swin Transformer, which can help identify both the spatial features and global dependencies of the input image. Our proposed architecture also includes the spatial attention gated fusion module that can be used to benefit from the advantages of the suggested architectures and their outputs according to some aspects of the input image. In our paper, the unique aspect of our contribution lies in the application of a dual-level explainer, which utilizes the features of Grad-CAM and Attention Rollout approaches. Moreover, in this paper, we developed an explanation fusion technique based on Alpha Search that can enhance the explanation of the suggested approach. Apart from these innovations, the application of quantitative criteria to explainability allows us to compare our model with others. Our experiments revealed that our model shows high performance and provides reliable explanations, indicating the necessity of balancing performance and explainability in health care projects. The next steps for further research are related to implementing our model for multi-class classification and computation optimization.

REFERENCES

[1] M. Alshomrani, A. Albeshri, A. A. Alsulami, and B. Alturki, “An explainable hybrid CNN–Transformer architecture for visual malware classification,” Sensors, vol. 25, no. 15, p. 4581, 2025.

[2] K. Djoumessi, S. O. Mensah, and P. Berens, “A hybrid fully convolutional CNN–Transformer model for inherently interpretable medical image classification,” Hertie Institute for AI in Brain Health, University of Tübingen, 2025.

[3] M. Alhumaid and A. G. Fayoumi, “Hybrid CNN–Swin Transformer model to advance the diagnosis of maxillary sinus abnormalities on CT images using explainable AI,” Computers, vol. 14, no. 12, p. 419, 2025.

[4] D. Pantelaios, P.-A. Theofilou, P. Tzouveli, and S. Kollias, “Hybrid CNN–ViT models for medical image classification,” in Proc. IEEE Int. Symp. Biomedical Imaging (ISBI), 2024, pp. 1–4.

[5] L. Barekatain and B. Glocker, “Evaluating the explainability of vision transformers in medical imaging,” Imperial College London, 2025.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.154282

[6] M. M. Auzine, M. Heenaye-Mamode Khan, S. Baichoo, N. G. Sahib, P. Bissoonauth-Daiboo, X. Gao, and Z. Heetun, "Development of an ensemble CNN model with explainable AI for the classification of gastrointestinal cancer," PLoS ONE, vol. 19, no. 6, e0305628, 2024.
[7] M. I. Nazir, A. Akter, M. A. H. Wadud, and M. A. Uddin, "Utilizing customized CNN for brain tumor prediction with explainable AI," Heliyon, vol. 10, no. 20, e38997, 2024.
[8] S. S. Shuvo, S. R. Refat, F. F. Preotee, and T. Muhammad, "Advanced CNN and explainable AI-based architecture for interpretable brain MRI analysis," in Proc. Int. Conf. Computing Advancements (ICCA), 2024.
[9] S. Iftikhar, N. Anjum, A. B. Siddiqui, M. U. Rehman, and N. Ramzan, "Explainable CNN for brain tumor detection and classification through XAI-based key features identification," Brain Informatics, vol. 12, art. no. 10, 2025.
[10] M. A. Lago, G. Zamzmi, B. Eich, and J. G. Delfino, "Evaluating explainability: A framework for systematic assessment and reporting of explainable AI features," arXiv preprint arXiv:2506.13917, 2025.
[11] T. Wang et al., "Explainable AI across domains: Techniques, domain-specific applications, and future directions," 2024.
[12] M. A. Awal and C. K. Roy, "EvaluateXAI: A framework to evaluate the reliability and consistency of rule-based XAI techniques for software analytics tasks," Journal of Systems and Software, 2024.
[13] X. Zheng et al., "F-Fidelity: A robust framework for faithfulness evaluation of explainable AI," in Proc. Int. Conf. Learning Representations (ICLR), 2025.
[14] J. Zhou, A. H. Gandomi, F. Chen, and A. Holzinger, "Evaluating the quality of machine learning explanations: A survey on methods and metrics," Electronics, vol. 10, no. 5, p. 593, 2021.
[15] M. A. Awal and C. K. Roy, "A meta-survey of quality evaluation in explanation methods," Journal of Systems and Software, vol. 217, p. 112159, 2024.
[16] D. Bhati et al., "A survey on post-hoc explanation methods for XAI visualization," TechRxiv preprint, 2025.
[17] A Batool and Y.-C. Byun, "A lightweight multi-path convolutional neural network architecture using optimal feature selection for multiclass classification of brain tumor using MRI," Results in Engineering, vol. 25, p. 104327, 2025.
[18] Appasami and N. Savarimuthu, "A novel lightweight CNN design for MRI brain tumor image classification with performance-driven optimization," Discover Computing, vol. 28, p. 206, 2025.
[19] C. Zhang et al., "A fine-grained car recognition method based on a lightweight attention network and regularized fine-tuning," Electronics, vol. 14, no. 1, p. 211, 2025.
[20] S. Niu and J. Zhang, "Image processing based on convolution neural network," Electronics, vol. 14, p. 4649, 2025.
[21] A Salih et al., "A review of evaluation approaches for explainable AI with applications in cardiology," TechRxiv preprint, 2023.
[22] Kim, H. Maathuis, and D. Sent, "Human-centered evaluation of explainable AI applications: A systematic review," Frontiers in Artificial Intelligence, vol. 7, p. 1456486, 2024.
[23] A Wijekoon et al., "XEQ scale for evaluating XAI experience quality," 2025.
[24] Y. Yu and J. Wang, "Hybrid granularities transformer for fine-grained image recognition," Entropy, vol. 25, no. 4, p. 601, 2023.
[25] X. Zhao et al., "A review of convolutional neural networks in computer vision," Artificial Intelligence Review, vol. 57, p. 99, 2024.
[26] D. S. Kermany, M. Goldbaum, W. Cai, C. C. S. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, J. Dong, M. Prasadha, J. Pei, M. Ting, J. Zhu, C. Li, S. Hewett, J. Dong, K. Ziyar, A. Shi, R. Zhang, L. Zheng, R. R. L. Chan, K. Vajzovic, H. P. Sadda, D. Huang, and M. F. Chiang, "Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning," Cell, vol. 172, no. 5, pp. 1122–1131.e9, Feb. 2018, doi: 10.1016/j.cell.2018.02.010.

