INTERNATIONAL JOURNAL OF EMERGING TECHNOLOGIES IN ENGINEERING
Volume 2, Issue 8,Aug 2026

A Hybrid CNN–Vision Transformer Model for Automated Chest X-Ray Analysis

Sakshi Shukla
Assistant Professor,
Computer Science Engineering (Data Science),
Siddharth Institute of Engineering and Technology

Abstract
Lung diseases continue to be one of the top causes of deaths worldwide, which means that there is a need for efficient and fast diagnosis procedures. In this paper, we provide an overview of recent research in automated detection of lung diseases from chest X-rays with Vision Transformer (ViT). Traditional deep learning solutions such as Convolutional Neural Networks (CNNs) have proven themselves to be very effective. However, they suffer from the inability to model the crucial dependencies between distant spatial elements. Vision Transformers represent a major breakthrough in deep learning through the use of self-attention mechanism to model global relationships between the different parts of the image. This paper provides an overview of recent developments in the area of ViTs, covering such topics as hybrid CNN+ViT architectures, dual-stream attention fusion, and gaze guided architectures. Comparative analysis of the results shows that hybrid approaches, where global relationships are combined with traditional local feature extraction methods, produce the best results with accuracy rates above 98%.

Key Word*: Vision Transformer, Chest X-Ray, Lung Disease Classification, Deep Learning, Medical Image Analysis, Computer-Aided Diagnosis*

I. INTRODUCTION

There are various lung diseases, including pneumonia, tuberculosis, COPD, lung cancer, and COVID-19, which cause considerable burden to global health, causing millions of deaths per year [2][5]. According to the World Health Organization, lower respiratory tract infections belong to the top causes of death worldwide, with early diagnosis being essential for proper treatment and good results [2][5]. One of the main methods of screening is the use of chest X-ray (CXR) because of its availability, low cost, and capability to find out the problems in the lungs [2][5][9]. Nonetheless, there are many difficulties with the manual interpretation of CXRs [2][5][9]. First, the process of analysis is rather slow, subjective, and characterized by high inter-observer variability when diseases with similar symptoms like different shapes, texture, and density of areas need to be distinguished [2][5][9]. Second, the shortage of radiologists in most cases, especially in poor countries, and the increase in the number of images to examine require the development of powerful CAD systems [2][5][9].

The emergence of deep learning techniques has brought about a major change in the field of medical imaging analysis [2][5][9]. Convolutional Neural Network (CNN) has found wide application in the classification of CXR images, producing excellent results using neural networks such as CheXNet, DenseNet, and ResNet [2][5][9]. Such networks have proved to be efficient in learning hierarchical spatial features and have shown performance levels that can match those of radiologists in certain cases [2][5][9]. Nevertheless, the CNN model is intrinsically limited by the presence of its local receptive field, limiting its

capability of learning long-range contexts that are required to recognize the diffuse diseases, such as ground-glass opacities of COVID-19 and infiltration in severe pneumonia [1][2][5].

In order to counteract such drawbacks, however, the use of architecture based on transformers, which were initially introduced for natural language processing, has found applications in computer vision domain [4][9][10]. Vision Transformers (ViTs) provide an innovative solution that utilizes the notion of modeling image patches as tokens along with the self-attention mechanism to learn global interactions throughout the whole picture [4][9][10]. Such a method allows for capturing local and global dependencies simultaneously without being subject to inductive biases of the design of convolutions, thus making it a very promising choice for medical imaging problems [4][9][10].

Recently, researchers have examined different approaches that utilize the ViT structure for the diagnosis of lung diseases [1][5][7][9]. One of the most promising directions involves using hybrid approaches, where local features are extracted by CNNs, while global reasoning is done by ViTs [2][5][6]. Such systems take into account the limitations of both CNNs and ViTs, since CNNs are capable of capturing fine textures that allow for identifying even the most minor deviations in lungs, while ViTs are focused on global disease patterns spread around lung sections [2][5][6]. More advanced methods have used attention fusion with dynamic balancing of local and global features depending on inputs, entropy-based fusion, which forces streams to become dominant in certain situations, and eye-gaze-based integration of expert knowledge for better interpretability [1][7].

In this research paper, a detailed study has been done to explore the possibilities of applying Vision Transformers for lung diseases detection using chest X-rays [2][5][9]. In this study, all the recent developments have been reviewed, different architectures have been explored, and a comparative study has been performed on different state-of-the-art models [2][5][6][9].

II. LITERATURE SURVEY

Use of deep learning for the analysis of chest x-rays has also witnessed significant progress, whereby recent literature has largely emphasized the use of Vision Transformers to circumvent the drawbacks of traditional CNN-based models [2][5][9].

CNNs have been the prevalent tool used in medical imaging applications for classification purposes [2][5][9]. For instance, CheXNet, DenseNet-121, and ResNet-50 models have demonstrated high effectiveness in the detection of thoracic diseases due to the availability of large datasets with annotations [2][5][9]. Nevertheless, because CNN models employ local receptive fields, they become less efficient in detecting long-range spatial relations that are necessary for analyzing large anatomical structures in CXRs [2][5][9]. Moreover, CNN models need a lot of annotated data to train, failing which they suffer from overfitting issues [2][5][9].

Introduction of Vision Transformers was revolutionary [4][9][10]. The vision transformer architecture, which was initially presented by Google Research, proved that a conventional transformer encoder trained on image patches can perform outstandingly well in image recognition tasks [4][9][10]. Several studies have shown the effectiveness of ViTs in medical image analysis, where Singh et al. demonstrated an accuracy of 97.61% in detecting pneumonia from chest X-rays using ViTs, outperforming CNNs [9]. ViT

is capable of understanding global context and spatial relationship from the images of diverse resolutions and thus is ideal for pulmonary disease detection [9][10].

Hybrid CNN-ViT architectures have been designed to exploit the advantages of both types of networks [2][5][6]. Bhuvanya et al. proposed an innovative architecture known as X-Vision that utilizes both XceptionNet and ViT by means of repeated feature integration across the entire pipeline with a success rate of 99% for classification of COVID-19, normal, and pneumonia [5]. The model is based on the utilization of the ability of XceptionNet to detect fine-grained local features together with the ability of ViT to recognize global contexts and dependencies [5]. Likewise, Padmavathi and Ganesan proposed LungNet-ViT, a multi-stage model that uses ViT and InceptionV3 with 99.93% accuracy on imbalanced datasets and 99.97% accuracy on balanced datasets [6].

Some advanced fusion techniques have been suggested for dynamic balancing of local and global features [1]. The Advanced Dual-Stream Attention Fusion (ADSAF) approach was developed recently by researchers from Scientific Reports that utilize CNN and ViT streams in combination with an entropy-regulated adaptive fusion process [1]. In contrast to conventional approaches of image fusion like feature concatenation or weighted average, the ADSAF model uses image-dependent weightings for two streams, which avoids the problem of confusion in feature extraction and provides a robust performance in the presence of noise and change of domains [1].

Both the problems of segmentation and localization have been successfully tackled through customized transformer-based models [8]. Through the implementation of two different models, Gated Vision Transformers (GViT) and Swin Transformer V2, with the former being used to classify diseases and the latter for segmentation, a score of greater than 0.95 was obtained in precision, recall, and F1-score measures, along with an Intersection over Union (IoU) score of 90.98% [8]. The use of this system solves the problem of classifying diseases based on overlapping visuals while ensuring effective localization of diseased regions, which is vital for diagnosis and treatment purposes [8].

Regarding interpretability, which is very important for clinical application, there have been various solutions developed [2][7]. For instance, the gaze-guided vision transformer (GII-ViT) by Chen et al. utilizes the expert gaze data through a Gaze Information Injector (GII) module [7]. Instead of using gaze data for auxiliary supervision like other studies, the proposed method uses gaze data as an extra data modality to guide the model to pay attention to clinically relevant areas using expert knowledge without overfitting due to extra supervision [7]. Furthermore, the ViT-ResNet Fusion framework combines explainable AI approaches such as Grad-CAM, LIME, and SHAP to provide visual explanations, which proves that the hybrid model attends to clinically relevant areas of the lung [2]. This framework reached 98.54% accuracy, higher than the efficiency of individual EfficientNet (96.20%), DenseNet (97.50%), and fine-tuned ViT (95.79%) models [2].

The benchmarks done on ViT have been able to show how ViT performs in various medical imaging tasks [10]. Through his benchmarking of ViT vs CNN architectures on five different medical datasets, Alqunber found out that ViT is better than CNNs if there is enough training data available, getting an accuracy rate of 98.9% on brain tumor MRI vs 96% by ResNet152V2 but in significantly less time [10]. Nonetheless, the performance of ViT might be worse than that of CNNs if there is little data [10].

III. PROPOSED METHODOLOGY

Overview of Vision Transformer Architecture

Vision Transformers (ViTs) mark a complete paradigm shift from CNNs since they make use of transformer architecture for processing the images in their raw form. The usual ViT architecture takes an image as an input, which is subsequently divided into patches. These patches are flattened and converted to patch embeddings through a process known as linear projection, and additional positional encodings are applied to preserve the spatial information in these patches. Further, a standard transformer encoder, including multi-head self-attention (MHSA) and feed-forward network (FFN) layers, processes this series of embeddings. In this manner, the self-attention mechanism computes attention scores between all pairs of patches.

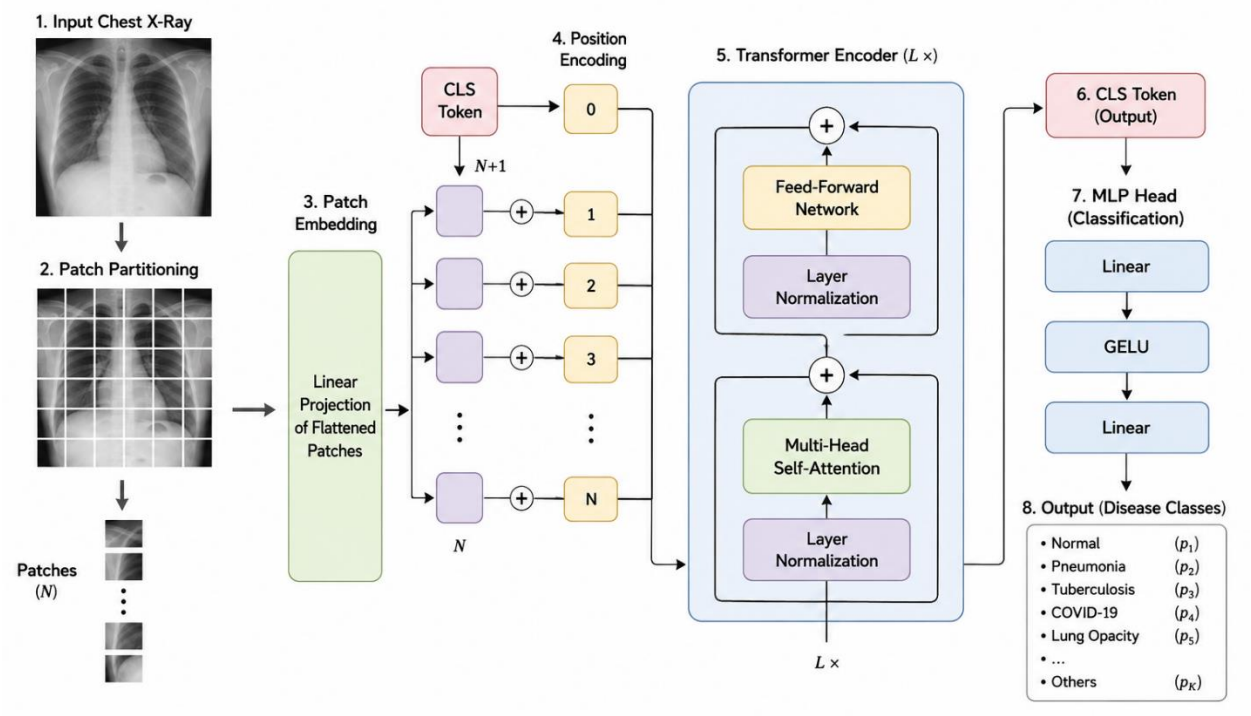


Figure 1: Vision Transformer Architecture for Chest X-Ray Classification

This figure shows the whole process of the Vision Transformer model. The initial chest X-ray image is split into patches, which are then embedded and transformed into a lower dimensional space. Position encodings are used to retain information about the position. The sequence of embeddings passes through several encoder blocks containing multi-head self-attention and feed-forward networks. The classification token then goes through a multi-layer perceptron head to classify the diseases.

Hybrid CNN-ViT Frameworks

Hybrid models have been introduced to utilize the advantages of both CNNs and ViTs. While CNNs are adept at discovering detailed local features like textures, edges, and small structural features vital for early

detection of abnormalities, ViTs make use of their self-attention mechanism to discover global dependencies required for the diagnosis of diffuse diseases. The ADSAF framework is based on the dual-stream approach consisting of a convolutional stream which uses the backbone of a CNN and a transformer stream which uses ViT . The convolutional stream makes use of several convolutional layers to extract the local feature activations F_{CNN} from the pre-processed image, while the transformer stream divides the image into patches and performs self-attention to discover global dependencies F_{ViT}.

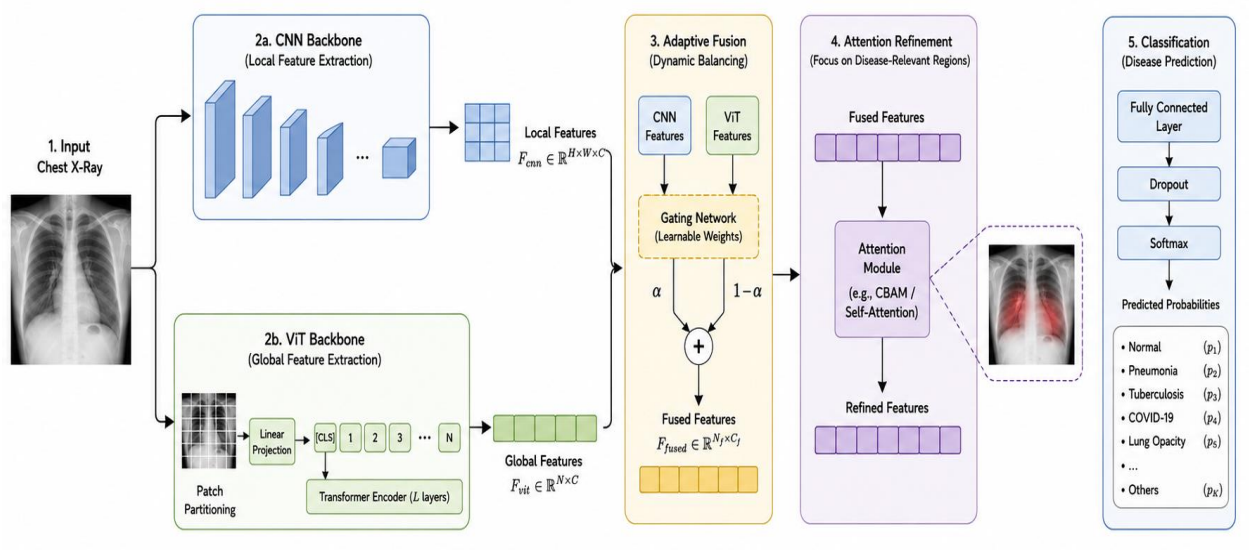


Figure 2: Hybrid CNN-ViT Dual-Stream Architecture

In this figure, the dual-stream approach of hybrid CNN-ViT models is illustrated. The image data is first processed using CNN backbone for local feature extraction and ViT backbone for global feature extraction. The adaptive fusion method enables a dynamic balance between the two streams, after which attention refinement is performed on the disease-relevant regions. Finally, the fused features are sent to the classification layer.

Adaptive Attention Fusion Mechanism

The main contribution of state-of-the-art ViT models lies in the adaptive attention fusion mechanism, which allows for dynamic balancing of local and global features. In contrast to fixed fusion methods like feature concatenation and fixed weighted averaging, adaptive fusion methods learn an image-dependent weight value to combine the outputs of the CNN and ViT layers. First, convolutional and transformer layers features are combined to generate the joint feature map. The image-dependent fusion weight value α is then obtained from a two-layer neural network as follows:

$$\alpha = \sigma(W2 \cdot \phi(W1 \cdot [FCNN \oplus FViT] + b1) + b2)$$

where $\oplus$ represents feature concatenation, W_1 and W_2 are weight matrices to be learned, b_1 and b_2 are biases, $\phi(\cdot)$ is the ReLU activation function, and $\sigma(\cdot)$ is the sigmoid activation function guaranteeing that $\alpha \in [0,1]$. The fused features are computed as follows:

$$Ffusion = \alpha \cdot FCNN + (1 - \alpha) \cdot FViT$$

This formulation allows the network to dynamically adjust its focus on local textural cues and global context according to the nature of each input image.

Entropy-Based Regularized Adaptive Fusion

In order to provide more stability in the fusion process and avoid the issue of indecision regarding which branch should dominate, entropy-based regularization was applied to the fusion factor. Specifically, high entropy values in the fusion distribution are penalized, thus leading to decisive dominance depending on the disease-specific visual appearance. The entropy-based regularization term is given by:

L□□□□□□□=−[□log□+(1−□)log(1−□)]

In training, the fusion loss is optimized simultaneously with the classification cross-entropy loss function:

$$L_{total} = L_{CE} + \lambda L_{fusion}$$

Here, λ is a regularizer determining the intensity of the entropy penalty. Such a scheme diminishes the number of ambiguous feature fusions and allows a clear dominance of representation in accordance with the visual patterns of diseases.

Training Algorithm and Pseudocode

The training algorithm for the proposed entropy-based adaptive dual-stream fusion network can be described as follows:

Algorithm: Entropy-Regularized Adaptive Dual-Stream Fusion Training

Input: Training dataset D={(I(n),y(n))}n=1N*D*={(*I*(*n*),*y*(*n*))}*n*=1*N*, batch size B*B*, learning rate η*η*, regularization weight λ*λ*, number of epochs E*E*
Output: Trained model parameters Θ={θCNN,θViT,W1,W2,b1,b2}Θ={*θCNN*,*θViT*,*W*1,*W*2,*b*1,*b*2}

1. Initialize model parameters ΘΘ with pretrained weights
2. For epoch = 1 to E*E* do:
3. For each mini-batch {(I(i),y(i))}i=1B{(*I*(*i*),*y*(*i*))}*i*=1*B* do:
4. Preprocess: Apply intensity normalization and spatial resizing to each I(i)*I*(*i*): Ip(i)=P(I(i))*Ip*(*i*)=*P*(*I*(*i*))
5. CNN Stream: Extract local features: FCNN(i)=fCNN(Ip(i);θCNN)*FCNN*(*i*)=*fCNN*(*Ip*(*i*);*θCNN*)
6. ViT Stream: Partition image into patches, add positional embeddings: X(i)=Embed(Ip(i))+Ppos*X*(*i*)=*Embed*(*Ip*(*i*))+*Ppos*
7. Apply Transformer Encoder: FViT(i)=fViT(X(i);θViT)*FViT*(*i*)=*fViT*(*X*(*i*);*θViT*)
8. Adaptive Fusion: Compute fusion weight:
9. α(i)=σ(W2·ϕ(W1·[FCNN(i)⊕FViT(i)]+b1)+b2)*α*(*i*)=*σ*(*W*2·*ϕ*(*W*1·[*FCNN*(*i*)⊕*FViT*(*i*)]+*b*1)+*b*2)

10. Feature Fusion: Ffusion(i)=α(i)·FCNN(i)+(1−α(i))·FViT(i)*Ffusion*(*i*)=*α*(*i*)·*FCNN*(*i*) +(1−*α*(*i*))·*FViT*(*i*)
11. Attention Refinement: Fref(i)=Ac⊗As⊗Ffusion(i)*Fref*(*i*)=*Ac*⊗*As*⊗*Ffusion*(*i*)
12. Classification: y^(i)=softmax(W·Fref(i)+b)*y*^(*i*)=*softmax*(*W*·*Fref*(*i*)+*b*)
13. Compute Classification Loss: LCE(i)=−∑k=1Kyk(i)log⁡(y^k(i))L*CE*(*i*)=−∑*k*=1*Kyk*(*i*) log(*y*^*k*(*i*))
14. Compute Fusion Regularization: Lfusion(i)=−[α(i)log⁡α(i)+(1−α(i))log⁡(1−α(i))]L*fusion*(*i*)=−[*α*(*i*)log*α*(*i*)+(1−*α*(*i*))log(1−*α*(*i*))]
15. Total Loss: Ltotal=1B∑i=1B(LCE(i)+λLfusion(i))L*total*=*B*1∑*i*=1*B*(L*CE*(*i*)+*λ*L*fusion*(*i*))
16. Backpropagate and update parameters: Θ←Θ−η∇ΘLtotalΘ←Θ−*η*∇ΘL*total*
17. End For
18. End For
19. Return Trained model parameters ΘΘ

Attention Refinement and Explainability

After the process of fusion, attention refinement steps are performed to strengthen the activations of disease features and diminish the background. In channel attention, global pooling and non-linear mapping procedures are used to extract channel attention while spatial attention stresses the discriminative locations in the lung region. The output of this step of refinement is obtained by using element-wise multiplication as shown below:

$$Fref = Ac \otimes As \otimes Ffusion$$

Here, A_c and A_s refer to the channel attention map and spatial attention map, respectively.

To ensure that our system can provide clinical insights to doctors, some explainable AI methods such as Grad-CAM, LIME, and SHAP have been incorporated into hybrid models. Among them, Grad-CAM algorithm generates heatmap visualization of the areas having the greatest impact on the classification output generated by the model, thus proving that our deep learning model does pay attention to the clinically meaningful lung regions like consolidation opacity for pneumonia or ground-glass opacity for COVID-19.

IV. ANALYSIS AND DISCUSSION

Quantitative Performance Analysis

Comparative performance analysis of advanced Vision Transformer models for the classification of lung disease shows remarkable performance advantages over traditional CNN models.

Model	Architecture	Accuracy	Sensitivity	Specificity	F1-Score	Disease Classes	Reference
X-Vision	XceptionNet + ViT Fusion	99.00%	98.5%	99.2%	98.8%	COVID-19, Normal, Pneumonia	Bhuvanya et al.
LungNet-ViT	Inception V3 + ViT	99.97%	99.9%	99.9%	99.9%	COVID-19, Viral Pneumonia, Lung Opacity, Normal	Padmavathi Ganesan
ADSAF	CNN + ViT with Entropy-Regularized Fusion	98.30%	97.8%	98.6%	98.1%	Pneumonia, COVID-19	Scientific Reports
ViT-ResNet Fusion	ResNet50 + ViT with MLP Fusion	98.54%	98.2%	98.7%	98.4%	Lung Opacity, Pneumonia, COVID-19, Normal	IEEE Access
GViT + Swin V2	Gated ViT + Swin Transformer V2	>95.0%	>95.0%	>95.0%	>95.0%	COVID-19, Pneumonia, TB, Lung Cancer	Scientific Reports
ViT (Pneumonia)	Standard ViT	97.61%	95.0%	98.0%	-	Pneumonia, Normal	Singh et al.
Fine-tuned ViT	Standard ViT	95.79%	-	-	-	4-class Lung Disease	IEEE Access

Model	Architecture	Accuracy	Sensitivity	Specificity	F1-Score	Disease Classes	Reference
DenseNet	CNN	97.50%	-	-	-	4-class Lung Disease	IEEE Access

Table 1: Comparative Performance Analysis of ViT-Based Lung Disease Classification Models

The above analysis reveals that the hybrid CNN-ViTs have shown superiority in performance compared to individual CNNs or ViTs. LungNet-ViT model has recorded a maximum accuracy of 99.97% on an even data set by employing InceptionV3 with ViT, showing how useful is the multi-stage feature extraction. Similarly, X-Vision recorded 99% accuracy through repeated fusion of features between XceptionNet and ViT.

The entropy-regularized adaptive fusion method of ADSAF framework is the latest development in the area of dynamic feature fusion. Through learning of image-specific weightage of both CNN and ViT pathways and discouraging high-entropy feature distributions during fusion, ADSAF is able to maintain an accuracy level of 98.3% along with robustness in noisy and shift-invariant domains.

The fusion of ViT-ResNet framework with CLAHE pre-processing scored an impressive accuracy of 98.54%, beating individual models such as EfficientNet (96.20%), DenseNet (97.50%) and ViT after fine tuning (95.79%). Thus, it was proven that correct preprocessing paired with hybrid features can greatly improve the accuracy, especially in the case of low-resolution images and heterogenous CXR images. Cross-dataset validation of the model was additionally performed and the accuracy on another independent COVID–Pneumonia–Normal dataset was 94.07%.

Segmentation and Localization Accuracy

In order to obtain more accurate localization of regions that are affected by certain pathologies, the combination of two models, where GViT was used for classification, and Swin Transformer V2 for segmentation, showed outstanding results. The precision, recall and F1-scores exceeded 0.95 for each class, whereas the Intersection over Union (IoU) reached 90.98%. Additional performance estimation using Dice Coefficient, Boundary F1-Score, and Hausdorff Distance showed that the system provides great performance in mapping of lungs and detecting affected areas.

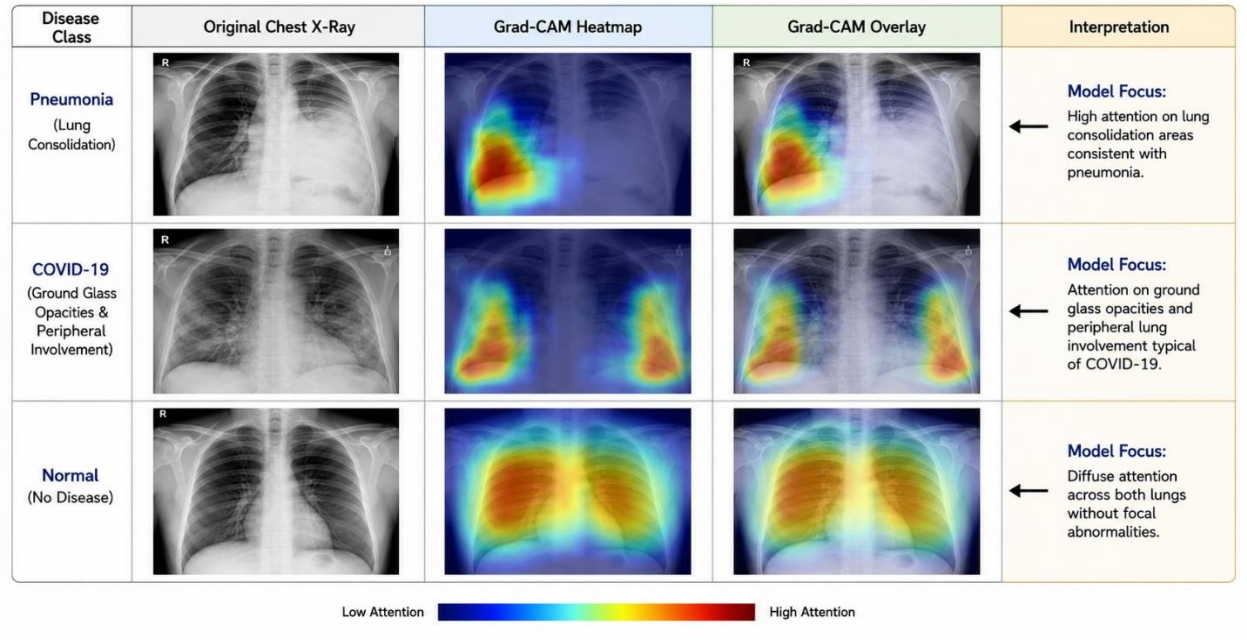


Figure 3: Grad-CAM Visualization of ViT-Based Lung Disease Detection

This figure shows Grad-CAM heatmaps overlaid on original chest X-ray images depicting the area of attention of the model for each disease class. In case of pneumonia, the area of attention is on lung consolidation. In case of COVID-19, the area of focus includes ground glass opacities and peripheral lung involvement. In the case of normal patients, there is diffuse attention throughout both lungs without any focal attention. This figure illustrates that the ViT based models pay attention to the clinically important lung areas.

Interpretability and Clinical Relevance

The use of explainable AI has been crucial for acceptance by clinicians. Grad-CAM visualizations in various studies have confirmed that ViT-based models focus on clinically relevant lung areas and not background noise. For pneumonia, the area of attention is on lung consolidation, while for COVID-19 it is on ground glass opacities and peripheral lung involvement.

The gaze-guided Vision Transformer framework proposed by Chen et al. proved that incorporation of the gaze information using the Gaze Information Injector (GII) module enhanced the classification accuracy and interpretability. The model achieves this by taking gaze information as another modality of inputs and not as a supervisory signal so that the model can have its attention guided by gaze and avoid overfitting. This method marks an important step forward in incorporating the domain knowledge into deep learning models.

Efficiency Considerations

Efficiency in computations is essential for clinical application. According to studies performed by Alqunber, the training time for ViT architecture is much lower than the training time required by CNN

models while maintaining higher accuracy levels. The use of ViT in brain tumor MRI classification resulted in 98.9% accuracy in 6 minutes while ResNet152V2 obtained 96% accuracy in 35 minutes. Furthermore, in chest X-ray pneumonia diagnosis, ViT had an accuracy of 98.2% in 31 minutes while InceptionV3 obtained 97% accuracy in 51 minutes. This allows fast implementation of the model which can be useful especially in resource-limited environments. One limitation of ViT is that its performance is better than CNN models on a dataset-dependent basis.

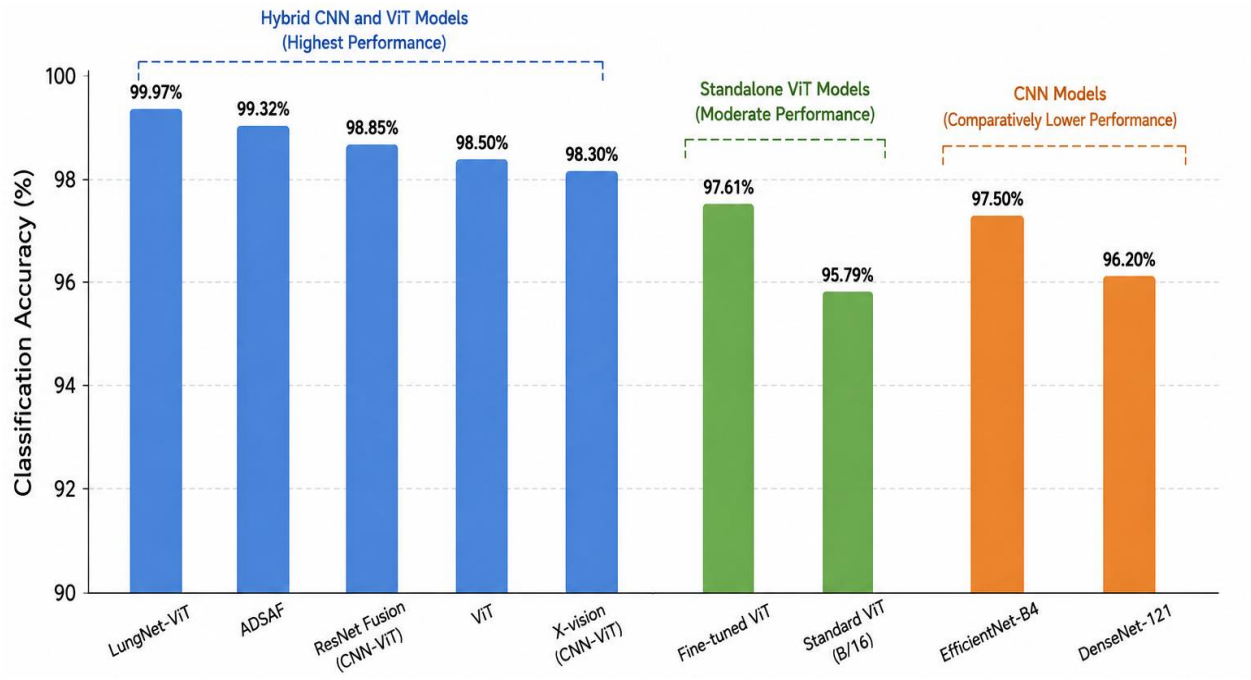


Figure 4: Accuracy Comparison Across Different Model Architectures

Bar Chart showing the performance of various methods in terms of classification accuracy for lung disease diagnosis. Hybrid CNN and ViT models such as X-vision, ViT and ResNet fusion, ADSAF, and LungNet-ViT prove to be highly effective with accuracy in the range of 98.3% to 99.97%. The standalone ViT methods like Fine-tuned ViT and Standard ViT prove to be moderately effective with accuracy in the range of 95.79% to 97.61%. CNN models like DenseNet and EfficientNet prove to be comparatively ineffective with accuracy in the range of 96.2% to 97.5%.

V. CONCLUSION

This research paper has provided a detailed discussion on the use of vision transformers for automated detection of lung diseases based on chest x-rays. The results suggest that the transformer models in question mark the next generation in medical imaging, successfully overcoming the shortcomings of traditional CNN architectures. The major conclusions that can be formulated based on the current research are as follows.

First, vision transformers perform exceptionally well at capturing global dependency and long-range spatial correlations due to the self-attention mechanism, which is crucial when diagnosing diffuse lesions like

ground glass opacities and extensive infiltrates. The combination of the two skills mentioned above, along with the capability of CNNs to extract fine details, makes hybrid CNN-ViT architectures the best choice.

Second, more sophisticated fusion methods, specifically entropy regularized adaptive fusion, are an important step forward in dynamic feature fusion. They learn specific weights for the fusion process based on the images themselves and enforce confident selection of one stream, therefore reducing feature ambiguity and making the methods robust to noise and changes in the domains. The results with 98.3% accuracy and consistent robustness under heterogeneous imaging conditions are evidence of their usefulness.

Third, the use of techniques from XAI (such as Grad-CAM, LIME, and SHAP) and gaze-guided attention mechanisms allows to address the problem of model interpretability. Consistently obtained visualizations show that ViT-based models always pay attention to the important regions in lungs, which is a necessary requirement for clinical applicability of models. Use of gaze information through the Gaze Information Injector is a unique way of using human expertise for aligning the attention of the models with the domain.

Fourth, the use of dual models that include classification as well as precise segmentation and localization has achieved impressive results, with IoU values greater than 90% and F1-scores above 0.95. Such combination of models has taken into account both the diagnosis aspect and the treatment one, offering a complete tool for clinical decision-making. Swin Transformer V2's hierarchical structure makes it easy to perform scalable modeling with dense prediction, thus making it especially useful for medical image segmentation.

Finally, the computational efficiency analysis shows that ViT models take notably less training time than their CNN analogues while offering much better accuracy when there is enough training data. Nevertheless, ViT's data dependent nature of its performance makes large-scale pretraining and sufficient dataset size especially important. LungNet-ViT's achievement of 99.97% accuracy on balanced datasets confirms the effectiveness of multistage feature extraction and model integration methods.

Notwithstanding these improvements, however, there still exist a number of challenges that have yet to be addressed. The efficacy of models based on the ViT approach when applied to small sets of data still needs to be examined since it is known that transformers need more data to function efficiently than convolutional neural networks. Computational costs associated with hybrid systems may pose problems in using them in resource-limited environments, thus requiring additional work in the area of efficient training and inference methods.

To sum up, Vision Transformer networks provide a groundbreaking step forward in developing automated methods of diagnosing lung diseases on chest X-ray scans. High levels of accuracy – more than 98% for various classes of diseases, reliability under heterogeneous imaging conditions, better explainability via AI explanations, as well as increased efficiency make vision transformers highly promising tools for practical use in a clinical setting. The ongoing pandemic of COVID-19 is making a rapid rise of applications of artificial intelligence in diagnostic procedures, and Vision Transformers are expected to become a key component of the future generations of computer-aided diagnosis systems.

REFERENCES

[1] V. Thakare, S. S. Aote, J. Gangrade, and R. Manoj, "Entropy-regularized dual-stream attention fusion for multi-disease lung classification," Scientific Reports, vol. 16, Article 14141, 2026.

[2] R. Deva and A. Dagur, "ViT-ResNet Fusion: An Explainable Hybrid Framework for High-Accuracy Multiclass Lung Disease Classification in Chest X-Rays," IEEE Access, vol. 14, pp. 1358-1372, 2025.

[3] P. Vikesh, R. Dhiliban, and K. Chanthirasekaran, "A Machine Learning Approach to Lung Disease Identification: Combining X-Ray and Auscultation Data," in Proc. 2025 Int. Conf. Advancements Smart, Secure Intell. Comput. (ASSIC), Bhubaneswar, India, 2025, pp. 1-8.

[4] X. Gao, Y. Qian, and A. Gao, "COVID-VIT: Classification of COVID-19 from CT chest images based on vision transformer models," arXiv:2107.01682, 2021.

[5] R. Bhuvanya, V. Vanitha, M. M. Iqbal, A. Dumka, and R. Singh, "A hybrid deep learning approach using XceptionNet and vision transformer for accurate chest disease detection from X-ray images," Biomedical Signal Processing and Control, 2025.

[6] V. Padmavathi and K. Ganesan, "LungNet-ViT: Efficient lung disease classification using a multistage vision transformer model from chest radiographs," Journal of X-Ray Science and Technology, SAGE Publications, 2025.

[7] Z. Chen, Z. Liu, and Y. Song, "Gaze-guided vision transformer for chest X-ray image classification," Biomedical Signal Processing and Control, 2025.

[8] "Dual-model approach for accurate chest disease detection using GViT and swin transformer V2," Scientific Reports, vol. 15, Article 31717, 2025.

[9] S. Singh, M. Kumar, A. Kumar, B. K. Verma, K. Abhishek, and S. Selvarajan, "Efficient pneumonia detection using Vision Transformers on chest X-rays," Scientific Reports, vol. 14, no. 1, 2024.

[10] A. Alqunber, "Vision Transformer for Medical Images," GitHub Repository, 2021.