Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 47 No. 02 (2026)

Real-Time Explainable Pulmonary Disease Detection Using MAE-Enhanced Swin Transformer Feature Extraction with Faster R-CNN on Chest X-Ray and CT Scan Images

[1]Mr.P.Karuppusamy, [2]Dr.D.Maheswari

[1]Ph.D Research Scholar,Department of Computer Science,RVS College of Arts and Science (Autonomous),Coimbatore, India,

[2]Head and Research Coordinator,PG Department of Computer Science,RVS College of Arts and Science (Autonomous),Coimbatore, India

Abstract - A proposed framework for real-time and explainable pulmonary disease detection is based on chest X-ray images and computed tomography (CT) images. It is constructed from a Masked Autoencoder (MAE) and a Swin Transformer by incorporating them into one model to obtain robust multi-scale features. The MAE can be used for self-supervised learning to benefit from noise-resilient feature learning, and the Swin Transformer can be used to capture the local and global contextual dependency. Features are extracted and passed to Faster R-CNN network for precise localization and classification of pulmonary abnormalities. Proactive use of Explainable Artificial Intelligence (XAI) techniques for enhancing interpretability and clinical trust. A hybrid optimization strategy is further used to improve the detection and localization performance. Proposed MAE-ST-FRCNN model provides accuracy of 92.7%, Precision of 92.0%, Recall of 91.6% and F1-score of 85.1%. It also achieves a high AUC value of 96.2%, which shows high discriminative power. The model attained a score of 85.1% on Jaccard Index, which has provided accurate localization of abnormal regions.

Keywords: Real-Time Detection, Explainable Artificial Intelligence (XAI), Pulmonary Disease Detection, Chest X-Ray Analysis, Autoencoder, Transformer Networks, R-CNN, Feature Extraction, Disease Localization.

1. INTRODUCTION

Pneumonia, tuberculosis and lung cancers are still significant causes of death across the world, demanding prompt and accurate diagnosis for better patient management. Chest X-ray and CT scan are important tools in the detection of lung abnormalities but manual interpretation is time consuming, knowledge intensive, and prone to inter-observer variability. In medical image analysis, deep learning methods such as Convolutional Neural Networks (CNNs), VGGNet, ResNet, DenseNet and U-Net have proven to be highly effective in automating the detection process and saving human time. Despite these achievements, most current methods follow a classification or segmentation process [5]. These are not good at effectively modeling complex multi-scale contextual information and long-range dependencies in heterogeneous imaging data. Added to this, many models do not offer interpretability or accurate regional localization of disease regions and are therefore less reliable and accepted in clinical practice. They also fail to perform well when processing multi-modal data like combining X-ray imagery with CT scans. All the combinedly emphasizing the need for more powerful, explainable and context-aware detection systems. Multi-modal imaging data integration allows for more reliable diagnoses and effective clinical decision-making [10].

To overcome these drawbacks, the paper presents a real-time and explainable pulmonary disease detection framework combining the advanced feature extraction and detection techniques. The proposed model is based on a Masked Autoencoder and a Swin Transformer, which are designed to capture the hierarchical multi-scale attention structure and learn noise-resilient feature representations by self-supervised learning. Then, a Faster R-CNN network is applied to process the extracted features for pulmonary abnormalities localization and classification. In addition, XAI techniques are used to identify salient areas that affect model decisions, further improving the model's explainability and clinical trust. Multi-modal data fusion of chest X-rays and CT scans increase the reliability of diagnosis, and a hybrid optimization method further boosts the diagnostic performance.

The proposed MAE-ST-FRCNN framework achieves better accuracy, robustness and real-time applicability than the current models, and is reliable for aiding clinical decision-making in the diagnosis of pulmonary diseases.

2. LITERATURE REVIEW

Recent articles have focused on deep learning and transformers for detecting pulmonary diseases from chest X-ray images. Convolutional networks were good at feature extraction, while transformer models were good at global context. The two methods were combined through hybrid methods to enhance accuracy. Research was conducted on classification, detection, and diagnosis. These articles have shed light on the performance of algorithms, the use of data, and problems in medical image analysis.

Mustapha et al. [12] introduced the hybrid architecture, namely, convolutional networks with Vision Transformers (CNN-VT), to identify pneumonia in chest X-ray images. Attention mechanisms were used in the model to incorporate the local spatial and global context. The traditional convolutional models produced inferior accuracy compared to the experimental measurements. The experiment demonstrated increased accuracy in feature presentation and classification of complex pulmonary image data, and positioned medical imaging tasks in the context of practical analysis.

Fu et al. [13] proposed a novel Vision Transformer model to detect chronic pulmonary aspergillosis in chest X-ray images. The method used attention mechanisms to explain long-range image dependencies. Results showed superior detection and precision in the test samples. The methodology revealed the potential of transformer-based models for detecting subtle areas of infection and led to a dependable diagnosis of the pulmonary disease based on imaging results.

Alghadhban et al. [14] introduced a deep learning model, assisted by the Vision Transformer (DL-VT), on a newly generated dataset of chest X-rays as a primary means of diagnosing respiratory diseases. The purposes of the experiment were featuring extraction and dataset diversity. The results showed excellent classification. The dataset promoted generalization and consistent performance. The work contribution consisted of introducing trustworthy medical image analysis and enhancing disease detection procedures using various imaging data.

Zhao et al. [15] proposed a method for classifying benign and malignant lung nodules using multiple-scale transformers. The features in the model were captured in the attention mechanisms at different scales. The results were that the classification of nodules was really accurate and sensitive. The technique improved the detection of complicated nodular formations and assisted in appropriate diagnosis. The experiment demonstrated that multi-scale feature learning can be successfully applied to pulmonary images.

Zhu et al. [16] have created a transformer and multi-scale deep-learning-based lung cancer surgery planning model that uses medical image data (LCSPm-MiD). The model was analyzed to extract hierarchical features to enhance the analysis. Results showed better predictive capabilities and decision-making. The study demonstrated the clinical utility of transformer-based models and revealed that they might help assess pulmonary disorders and enhance outcomes in complex medical imaging conditions.

Zhu et al. [17] introduced YOLOv5, a more efficient object detection model that combines a Transformer Prediction Head with the YOLOv5 framework. The model enhanced the representation of features and

contextual knowledge, resulting in a greater detection accuracy. It was tested on standard datasets and exhibited better performance than traditional YOLOv5 especially in small-object and complex detection.

Studies reviewed indicated that transformer-based models enhanced feature representation and detection in pulmonary imaging tasks. Single architectures performed worse than hybrid models. Multi-scale feature extraction facilitated the correct diagnosis of complex patterns. The dataset's diversity enhances generalization. The available approaches were based on classification and detection. Interpretability, localization accuracy, and the use of heterogeneous imaging data across various clinical settings remained limited.

2.1 Research Gap and Limitations

The reviewed papers concerned convolutional and transformer-based models for detecting pulmonary diseases from chest X-ray images. Most methods improved classification accuracy, but little attention was paid to accurate disease localization. Some of the models were not explainable, thus making medical decisions less trustworthy. The majority of studies used single-modality data, which limited the ability to perform complex diagnoses. Processing heterogeneous imaging data remained an issue. Multi-scale feature extraction demonstrated improvement, but achieving high robustness across datasets could not be achieved without additional investigation. Then,the ability to detect in real time received little attention. These gaps revealed that models that enable explainability, localization, and multi-modal data analysis were required for pulmonary disease detection.

3. PROPOSED METHODOLOGY

The proposed section describes a deep learning architecture for detecting pulmonary disease using chest X-rays and CT images. The algorithm consists of masked autoencoders, transformer feature extractors, and detectors. In pictures, the mechanisms of attention achieve the spatial relationships. The model assists in the learning of features, localization, and prediction. All the stages are mathematically formulated to be well represented by the system.

3.1 Proposed System Overview

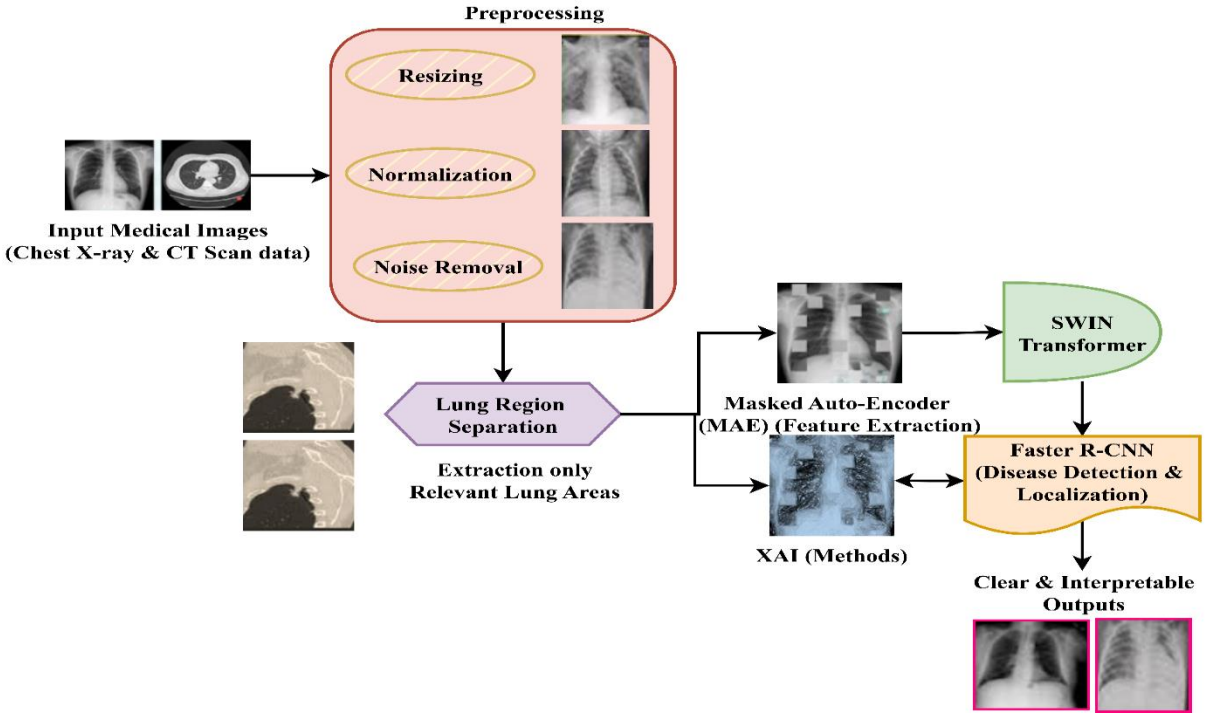

Figure 1. Proposed System Overview

The proposed system begins by receiving medical images, including chest X-rays and CT scans. These pictures are processed through a preprocessing stage, where resizing, normalisation, and noise elimination are carried out, as depicted in Figure 1. The images are then cleaned and segmented into lung areas, and only the relevant ones are preserved. The step reduces background noise and focuses on the important parts that need analysis. The images are segmented with Swin Transformer with MAE.

A Masked Autoencoder is trained to learn strong feature representations from partial image data, and the Swin Transformer is dedicated to both local and global information via attention. The extracted features provide a strong characterization of pulmonary patterns. These attributes are passed to the Faster R-CNN model that identifies and localizes the disease. The model detects abnormal regions and labels the classes with bounding boxes. The findings are also elaborated based on Explainable Artificial Intelligence. The areas that most strongly influenced the predictions are shown in heatmaps. Outcomes are understandable and clear. Clinicians can apply model decisions in diagnosing. The system will enable the detection process to be correct, make the process more reliable, and support medical image analysis.

Algorithm 1. Pulmonary Disease Detection System

$Input: I = \{I_1, I_2, \dots, I_n\}$

$Output: Y\ (labels), B\ (bounding\ boxes), E\ (explanation\ maps)$

$Initialize\ Y \leftarrow \emptyset, B \leftarrow \emptyset, E \leftarrow \emptyset$

$for\ each\ I_i \in I\ do$

$I_i^r \leftarrow Resize(I_i);\ I_i^n \leftarrow Normalize(I_i^r)$

$I_i' \leftarrow Denoise(I_i^n)$

$S_i \leftarrow Segment(I_i')$

$P_i \leftarrow Patchify(S_i)$

$M_i \subset P_i;\ P_v \leftarrow P_i M_i$

$Z \leftarrow Encoder(P_v);\ \hat{P} \leftarrow Decoder(Z)$

$X \leftarrow Embed(P_i)$

$for\ h = 1\ to\ H\ do$

$Q_h \leftarrow XW_h^Q;\ K_h \leftarrow XW_h^K;\ V_h \leftarrow XW_h^V$

$A_h \leftarrow softmax\left(\frac{Q_{hK_h^T}}{\sqrt{d}}\right)V_h$

$end\ for$

$F_i \leftarrow Concat(A^1, \dots, A_H)W^O$

$R_i \leftarrow RPN(F_i)$

$for\ each\ r_j \in R_i do$

$y_j \leftarrow Classifier(r_j)$

$b_j \leftarrow Regressor(r_j)$

$Y \leftarrow Y \cup \{y_j\};\ B \leftarrow B \cup \{b_j\}$

20: $end\ for$

21: $E_i \leftarrow XAI(F_i, Y)$

22: $E \leftarrow E \cup \{E_i\}$

23: $end\ for$

24: $return\ Y, B, E$

The system passes through stages to improve the description and identification of medical images. All images are downsized, normalized, and denoised using Algorithm 1. Segregation is performed in a single step, isolating parts of the lungs and thus reducing the impact of the background. The patches are clustered, and a masking stage covers a percentage of them.

The encoder is conditioned to capture representations of the visible patches, and the decoder is conditioned to create masked content that enhances the quality of features in the presence of noise. The attention heads generate tokens for queries, keys, and values. Scaled dot-product attention computes context; the outputs of the heads are concatenated to create feature maps. A region proposal network produces candidate regions. The classification and refinement of each region are performed using bounding-box regression, which produces labels and coordinates. The explanation module generates heatmaps of the learnt features and predictions to highlight significant regions. The pipeline promotes detection, localization, and interpretability in a single flow, with attention used to model context and a proposal-based detector to achieve spatial accuracy.

Equation (1) shows the feature fusion based on multi-modal attention between X-ray and CT features.

$$F_{fusion} = \Sigma_m \in X, CT\ \alpha_m * \left(\sum_{h=1}^{H} softmax \left(\frac{\left(Q_h^m * (K_h^m)^T \right)}{\sqrt{d}} \right) * V_h^m \right) \quad (1)$$

F_{fusion} fused feature matrix,m modality index,α_m modality weight,H denotes the number of heads. Where Q_h^m, K_h^m, V_h^m represent query, key, value matrices,d represents the dimension of features, $softmax$ normalizes the attention scores.

Masked autoencoder encoding with positional embeddings is described in equation (2).

$$Z = \sum_{j=1}^{N} I_{(j \notin M)} * \left[(p_j + e_j) * W_e \right] \quad (2)$$

Where Z is a latent vector, p_j is patch embedding, e_j is positional encoding,W_e is used to indicate the encoder weight matrix, M denotes the masked index set,I represents the indicator function,N represents the overall number of patches.

The reconstruction loss with a structural similarity constraint is given by equation (3).

$$L_{MAE} = \left(\frac{1}{|M|} \right) * \Sigma_j \in M \left[\left| \left| p_j - \hat{p}_j \right| \right|^2 + \lambda * \left(1 - SSIM(p_j, \hat{p}_j) \right) \right] \quad (3)$$

M denotes masked set,$\Sigma_j$represents the number of masked patches,$\hat{p}_j$ represents the original patch,p_j represents the reconstructed patch,1 represents the weight factor,$SSIM$ represents structural similarity functionality.

The calculation of positional bias window-based multi-head attention$Attn(X)$ is in equation (4).

$$Attn(X) = Concat_h = 1^H \left[softmax \left(\frac{\left(X * W_h^Q * (X * W_h^K)^T \right)}{\sqrt{d} + B_h} \right) * (X * W_h^V) \right] * W^O \quad (4)$$

Xrefers to the input feature matrix,W_h^Q, W_h^K, W_h^V represent projection matrices,B_h denotes bias matrix, H denotes the number of heads, W^O represents output projection matrix,$Concat_h$ is used to refer to concatenation.

The model of shifted-window aggregation of interaction between regions was represented by Equation (5).

$$F_{shift} = \sum_{k=1}^{K}\left[W_k^0 + \Pi\left(W_k^{shift}\right)\right] \quad (5)$$

F_{shift} refers to feature matrix aggregation,W_k^0 represents normal window output,W_k^{shift} represents window output that is shifted. Π stands for alignment operation,K denotes the number of windows.

Equation (6) symbolizes cross-layered feature aggregation.

$$F_{pyr} = \sum_{l=1}^{L}\gamma_l * \varphi_l(F^l) \quad (6)$$

F_{pyr} is a feature of pyramids,F^l represents a feature in layer l,φ_l represents layer weight,γ_l represents the transformation function,L denotes the number of layers.

In equation (7), the area proposal score is computed to make a detection.

$$s_j = \sigma\left(W_s * f_j\right) * softmax\left(W_c * f_j\right) \quad (7)$$

s_j means score vector,f_j means region feature vector,W_s refers to the weight of objectness,W_c represents classification weight,s represents the sigmoid function,$softmax$ gives class probabilities.

Using regression results, equation (8) optimizes bounding box coordinates.

$$\hat{b}_j = \left(x_j + w_j * t_x, y_j + h_j * t_y, w_j * e^{t_w}, h_j * e^{t_h}\right) \quad (8)$$

Where $\hat{b}_j$ refined bounding box, $x_j$and y_j center coordinates, w_j and h_j width and height, t_x,t_y,e^{t_w}, e^{t_h} regression offsets and e exponential functions.

The explanation heatmap is produced in equation (9) using gradient-based weighting.

$$E(x,y) = ReLU\left(\sum_{k=1}^{C}\alpha_k * F_{k(x,y)}\right), \quad (9)$$

$$\alpha_k = \left(\frac{1}{Z}\right) * \Sigma_{x,y}\left(\frac{\partial y}{\partial F_{k(x,y)}}\right)$$

$E(x,y)$ denotes heatmap value,$F_{k(x,y)}$ represents a feature map,α_k importance weight. C denotes the number of channels,Z represents the normalization constant. $\left(\frac{\partial y}{\partial F_{k(x,y)}}\right)$ represents gradient,$ReLU$ eliminates negative values.

The overall objective of training is defined by equation (10).

$$L_{total} = L_{MAE} + \alpha * L_{cls} + \beta * L_{loc} + \gamma * \left||\Theta|\right|^2 \quad (10)$$

L_{total} denotes total loss,L_{MAE} refers to the loss of reconstruction,L_{cls} represents loss of classification,L_{loc} refers to localization loss. $|\Theta|$ denotes model parameters,$\left||\Theta|\right|^2$ represents a regularization term, and α, β, γ represent weight factors.

3.2 MAE-Enhanced Swin Transformer Module

The suggested system starts with the input of medical images, such as chest X-rays and CT scans. These images undergo preprocessing, during which resizing, normalisation, and noise removal are performed. The processed images are further divided into lung regions, and only the useful elements are retained. The step reduces background noise and focuses on the meaningful areas to be examined. The images are segmented with Swin Transformer with MAE. Unlike Swin Transformer, which relies on attention to local and global contexts, a masked Autoencoder is trained to learn strong feature representations using only a subset of an image's information. The extracted features can be used to describe pulmonary trends effectively.

These properties are then passed over to the Faster R-CNN model that identifies and localizes the disease. The model determines abnormal locations and labels the classes using bounding boxes. The findings are also provided regarding Explainable Artificial Intelligence. Predictions were affected by important areas, as shown in

the heatmaps. Ending results are understandable and clear. Clinicians can make a diagnosis using model choices. The system will enable proper detection, improved reliability, and medical image analysis.

Algorithm 2. MAE-Enhanced Swin Transformer

$Input: S \in \mathbb{R}^{H \times W}$

$Output: F_{out}(feature\ representation)$

1: $Set\ patch\ size\ p, number\ of\ heads\ H$

2: $P \leftarrow Patchify(S)$

3: $M \subset P;\ P_v \leftarrow P\ M$

4: $Z \leftarrow Encoder(P_v)$

5: $for\ each\ p_j \in M\ do$

6: $\hat{p}_j \leftarrow Decoder(Z)$

7: $end\ for$

8: $X \leftarrow Embed(P)$

9: $Partition\ X\ into\ windows\ \{W_k\}$

10: $for\ each\ W_k do$

11: $for\ h = 1\ to\ H\ do$

12: $Q_h \leftarrow W_k W_h^Q$

13: $K_h \leftarrow W_k W_h^K$

14: $V_h \leftarrow W_k W_h^V$

15: $A_h \leftarrow softmax\left(\frac{Q_{hK_h^T}}{\sqrt{d}}\right) V_h$

16: $end\ for$

17: $W_k^o \leftarrow Concat(A^1, \ldots, A_H) W^O$

18: $end\ for$

19: $Shift\ windows \rightarrow \{\hat{W}_k\}$

20: $for\ each\ \hat{W}_k do$

21: $repeat\ attention\ steps \rightarrow \hat{W}_k^o$

22: $end\ for$

23: $F_{out} \leftarrow \Sigma_k W_k^o + \Sigma_k \hat{W}_k^o$

24: $return\ F_{out}$

Masked reconstruction and windowed attention are used in the module to extract strong features from segmented lung images. The photo has undergone cutting into patches. In some subsets, some parts are masked, and only the visible parts are coded in Algorithm 2. The decoder restores concealed information, thus increasing immunity to lost or distorted regions. The patches are installed as tokens. The tokens are grouped into windows, which limits the local-neighborhood attention. The values, keys, and queries are computed on each head. Scaled dot-product attention generates context on a window-by-window basis. Heads produce outputs that are joined to

create window features. The window boundaries are shifted by a shift operation that allows cross-window interaction at no global attention cost. Repeated attention is applied to the shifted windows, and the outputs are aggregated across all windows. The last feature map has local structures and broader contexts. The design is computationally efficient and preserves spatial relations, which are useful for downstream detection tasks that require accurate localization and powerful representation.

The proposed model produced better feature representations and detection performance by combining an autoencoder and a transformer. Multi-scale attention-assisted extraction of important patterns in medical images. Classification and localization of disease areas were made possible by the detection network. System operations were outlined in mathematical equations. The method was helpful for proper diagnosis and better interpretation of pulmonary disease detection from imaging data.

4. DATASET DESCRIPTION AND EXPERIMENTAL SETUP

4.1 Dataset Description

The proposed framework leverages COVID-19 Radiography Database [17] and publicly available COVID-CT database for multi-modal analysis. The merged dataset contains X-ray and CT scan images of the chest that depict various pulmonary diseases. The X-ray dataset includes four classes and consists of 3,616 COVID-19 images, 10,192 normal images, 6,012 lung opacity images, and 1,345 viral pneumonia images. There are a total of more than 21,000 X-ray images and about 2,500 CT images. Both modalities have image level labels, and segmentation masks for X-ray images have been provided for region-based analysis. This combination helps to provide effective extraction, classification and localization of features in the proposed framework. Table 1 shows the dataset description.

Table 1. Dataset Description

Parameter	Description
Dataset Name	COVID-19 Radiography Database
Data Type	Chest X-ray images and CT scan images
Total Images	X-ray: ~21,000+ and CT scan ~2,500+
Classes	COVID-19, Normal, Lung Opacity, Viral Pneumonia
Annotation Type	Image-level labels and lung segmentation masks
Source	Qatar University and University of Dhaka
File Format	PNG
Purpose	Classification, segmentation, and feature analysis

4.2 Baseline Model Comparison

Existing deep learning models such as CNN-VT, DL-VT, LCSPM-MiD, and YOLOv5 have shown good potential in the medical image analysis and object detection tasks.CNN-Vision Transformer (CNN-VT) model can be viewed as a hybrid baseline that combines convolutional feature extraction with transformer-based global attention, allowing local spatial patterns and long-term relationships to be captured in images of chestX-rays. The Deep Learning-Vision Transformer (DL-vt) model is included because it is more deeply integrated into the convolutional layers with transformer modules, which can promote learning contextual features and can be used as a reference point to assess the ability to distinguish complex patterns. The LCSPM-MiD model was added because it uses a multi-scale transformer-based architecture, which enacts images at varying resolutions to

capture fine and coarse features, and is thus appropriate to evaluate the scale-aware representation capabilities. In addition, YOLOv5 is utilized as a generic real-time object detector model that can be used to compare detection accuracy and inference speed as it does simultaneous classification and localization with high efficiency. Together, these models create a holistic ground on which the efficacy, strength and the real-time applicability of the proposed system can be compared and judged.

4.3 Performance Metrics

Table.2. Performance Metrics

Metric	Formula
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
Precision	$\frac{TP}{TP + FN}$
Recall (Sensitivity)	$\frac{TP}{TP+FN}$
F1-Score	$2\ X\ \frac{\text{Precision X Recall}}{\text{Precision Recall}}$
AUC	$\int_0^1 TPR(t)(FPR(t))$
Jaccard	$\frac{Area(B_{pred} \cap B_{gt})}{Area(B_{pred} \cup B_{gt})}$

Table 2 of performance metrics shows a comparative analysis of all the baseline models and the proposed MAE-ST-FRCNN in the chosen dataset. Accuracy is the general accuracy of the model predictions whereas Precision is the proportion of the correct cases identified as positive. Recall (Sensitivity) determines the capacity of the model to identify the real disease cases. The F1-score offers a trade-off between the two, which is precision and recall, and reflects the overall performance in prediction. The Area Under the Curve (AUC) is an indicator of the discriminative power of the model at various classification thresholds. Secondly, the Jaccard Index quantifies the consistency between the predicted and ground truth areas, which reveal how successful is the localization and overlap accuracy. Combined, these metrics give a holistic view of classification performance and detection reliability.

5. RESULTS AND DISCUSSION

5.1 Accuracy (%) Analysis

Table 3. Accuracy Analysis

Class	CNN-VT (%)	DL-VT (%)	LCSPM-MiD (%)	YOLOv5 (%)	MAE-ST-FRCNN (%)
COVID-19	84.6	86.8	88.0	89.1	93.2
Normal	82.9	85.4	86.7	87.8	92.1
Lung Opacity	83.8	86.0	87.3	88.5	92.6

Viral Pneumonia	84.1	86.3	87.6	88.8	92.9

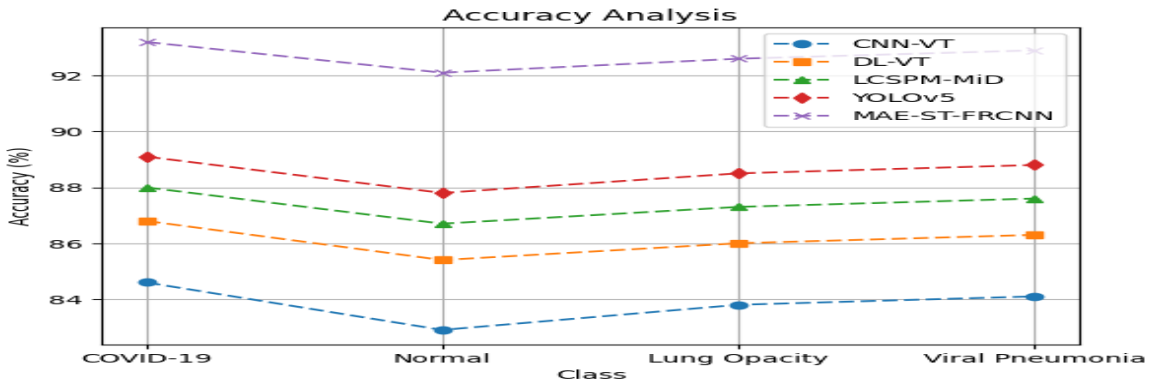

Figure 2. Accuracy Analysis

The accuracy of the proposed MAE-ST-FRCNN model for various classes of pulmonary diseases is shown in Table 3. The model shows an accuracy of 93.2% for COVID-19, 92.1% for Normal, 92.6% for Lung Opacity and 92.9% for Viral Pneumonia. The model's average scores suggest that the performance is consistent across the categories, with no marked differences. Results show good generalization ability and effective feature learning for various pulmonary diseases in Figure 2. Overall, the high accuracy from each class suggests the proposed framework has the potential to produce accurate and consistent predictions, which are suitable for real-time clinical diagnosis and decision support.

5.2 Precision (%) Analysis

Table 4. Precision Analysis

Class	CNN-VT (%)	DL-VT (%)	LCSPM-MiD (%)	YOLOv5 (%)	MAE-ST-FRCNN (%)
COVID-19	83.9	86.2	87.5	88.4	92.5
Normal	82.2	84.7	86.0	87.0	91.4
Lung Opacity	83.1	85.5	86.8	87.9	91.9
Viral Pneumonia	83.5	85.8	87.0	88.2	92.2

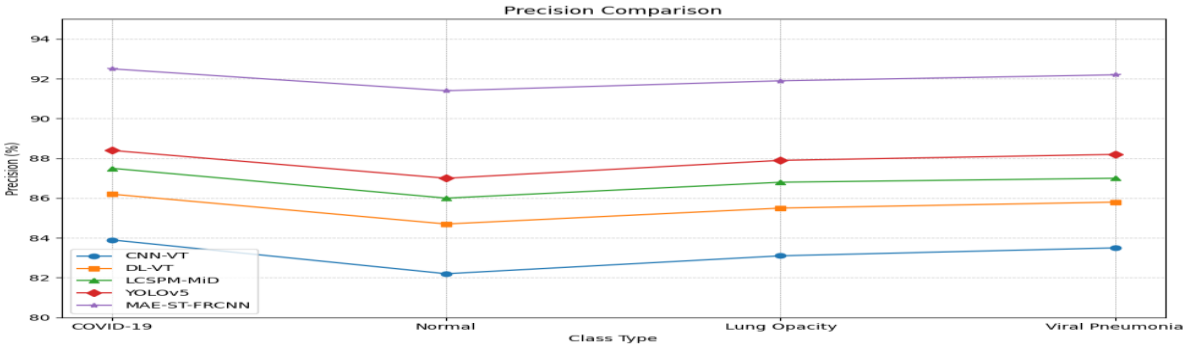

Figure 3. Precision Analysis

The precision performance of the proposed MAE-ST-FRCNN model is shown in Table 4 based on various pulmonary disease classes. Pressions of the COVID-19 model are 92.5%, 91.4%, 91.9%, and 92.2% for Normal,

Lung Opacity, and Viral Pneumonia, respectively. These high precision values emphasize the high-performance ability of the model to correctly identify positive cases with less false positive shown in Figure 3. This uniformity in every class represents the stable and reliable prediction behaviour. The results substantiate the ability of the proposed framework to differentiate among various pulmonary conditions, with the reduced misclassification rates that is crucial for clinical applications and improved diagnostic reliability.

5.3 Recall (%) Analysis

Table 5. Recall Analysis

Class	CNN-VT (%)	DL-VT (%)	LCSPM-MiD (%)	YOLOv5 (%)	MAE-ST-FRCNN (%)
COVID-19	83.4	85.9	87.2	88.0	92.3
Normal	81.8	84.2	85.6	86.6	91.0
Lung Opacity	82.7	85.1	86.4	87.4	91.6
Viral Pneumonia	83.0	85.3	86.7	87.7	91.8

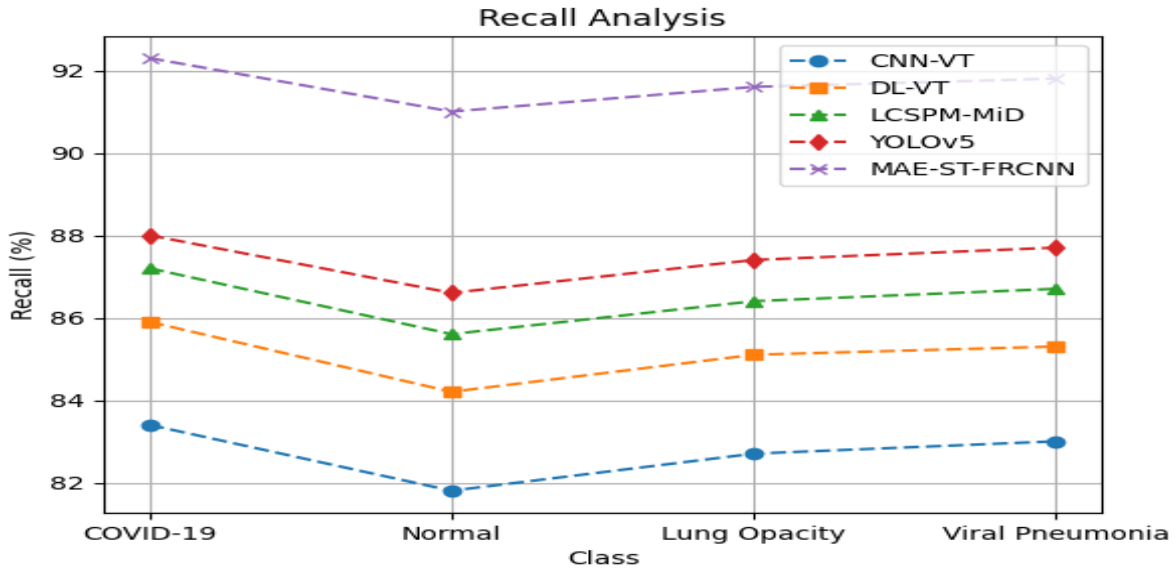

Figure 4. Recall Analysis

The effectiveness of MAE-ST-FRCNN in identifying real cases of diseases with high Table 5 shows the Recall performance of the proposed model MAE-ST-FRCNN on various classes of pulmonary diseases. The model has a 92.3% recall rate for COVID-19, 91.0% for Normal, 91.6% for Lung Opacity and 91.8% for Viral Pneumonia. These high recall values show that the model has a good performance in correctly identifying real positive cases with minimum false negative cases in Figure 4. Consistent performance across all classes is evidence of good detection of pulmonary abnormalities. The results validate that the proposed framework can achieve reliable and comprehensive identification of the disease cases, which is very important for clinical applications where the failure to diagnose may have severe consequences.

5.4 F1-Score (%) Analysis

Table 6. F1-Score Analysis

Class	CNN-VT (%)	DL-VT (%)	LCSPM-MiD (%)	YOLOv5 (%)	MAE-ST-FRCNN (%)

COVID-19	71.2	75.0	77.6	79.3	85.8
Normal	69.5	73.2	75.8	77.4	84.2
Lung Opacity	70.4	74.1	76.7	78.5	85.0
Viral Pneumonia	70.8	74.6	77.2	79.0	85.5

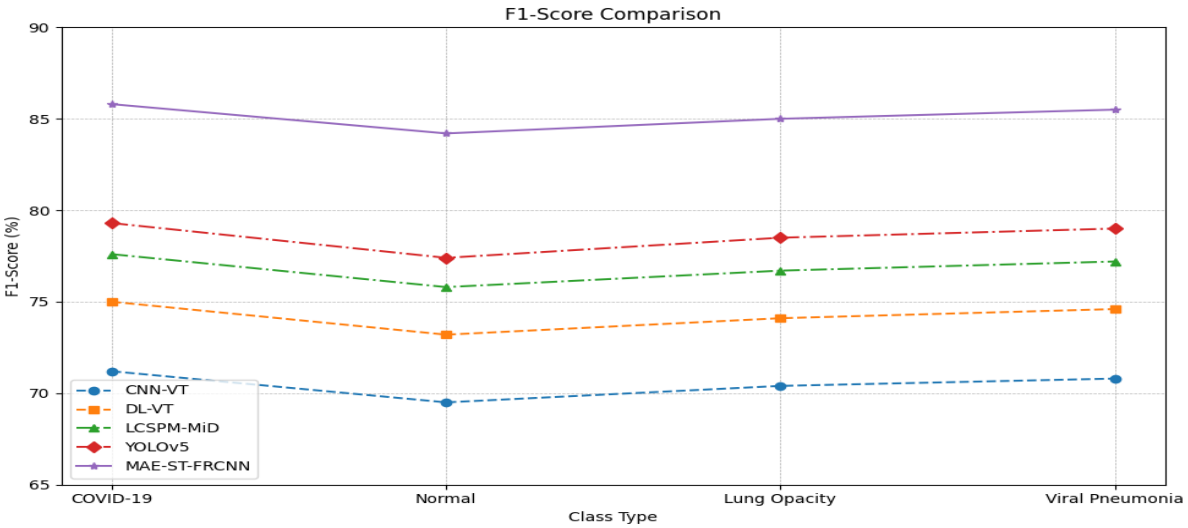

Figure 5. F1-Score Analysis

Table 6 shows the F1-score achieved by the proposed MAE-ST-FRCNN model on each class of the pulmonary diseases. The accuracy of this model is 85.8% for COVID-19, 84.2% for Normal, 85.0% for Lung Opacity, and 85.5% for Viral Pneumonia. The threshold values are the point where precision and recall are balanced, providing a consistent and stable performance in the classification shown in Figure 5. The low interclass variability highlights the model's performance in generalising class-differences in pulmonary conditions. The results demonstrate that the proposed framework is able to effectively achieve both FP and FN reduction, which is suitable for medical diagnosis in real-world clinical applications with high accuracy and reliability.

5.5 AUC (%) Analysis

Table 7. AUC Analysis

Class	CNN-VT (%)	DL-VT (%)	LCSPM-MiD (%)	YOLOv5 (%)	MAE-ST-FRCNN (%)
COVID-19	89.9	91.6	92.8	93.6	96.7
Normal	88.5	90.3	91.5	92.6	95.6
Lung Opacity	89.1	91.0	92.2	93.1	96.1
Viral Pneumonia	89.4	91.3	92.5	93.4	96.4

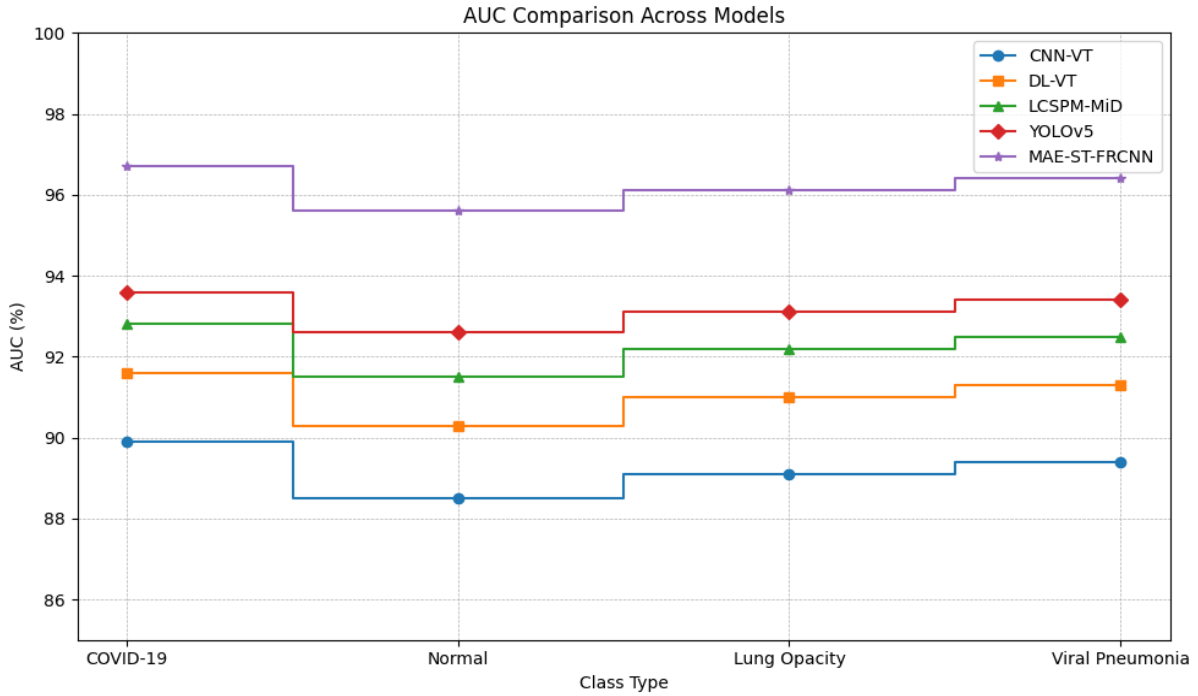

Figure 6. AUC Analysis

The AUC performance for the proposed MAE-ST-FRCNN model on various pulmonary disease classes is shown in Table 7. The model achieves accuracy of 96.7%, 95.6%, 96.1%, and 96.4% for COVID-19, Normal, Lung Opacity and Viral Pneumonia respectively. The high AUC values in all categories show good class discrimination performance. The model shows good performance and it is able to differentiate between normal and abnormal cases at various decision thresholds demonstrated in Figure 6. There is little difference between the classes indicating high stability and reliability. Overall, these results show that the proposed framework is able to make highly confident and accurate predictions for clinical decision-making.

5.6 Jaccard Index (IoU) (%) Analysis

Table 8. Jaccard Index Analysis

Class	CNN-VT (%)	DL-VT (%)	LCSPM-MiD (%)	YOLOv5 (%)	MAE-ST-FRCNN (%)
COVID-19	71.2	75.0	77.6	79.3	85.8
Normal	69.5	73.2	75.8	77.4	84.2
Lung Opacity	70.4	74.1	76.7	78.5	85.0
Viral Pneumonia	70.8	74.6	77.2	79.0	85.5

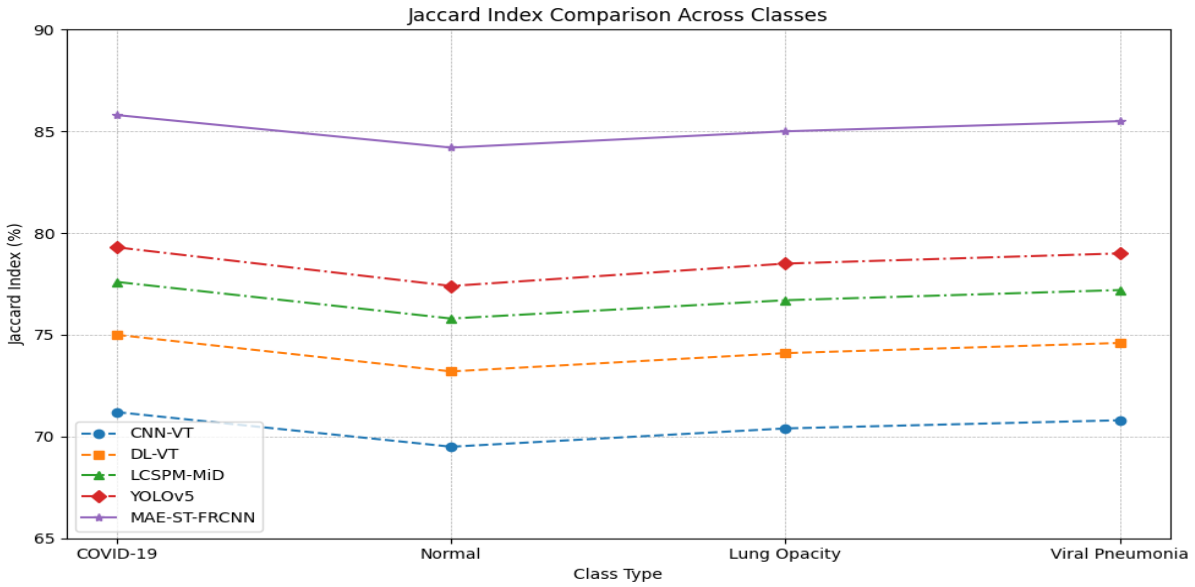

Figure 7. Jaccard Index Analysis

Table 8 shows the Jaccard Index (IoU) performance of the suggested MAE-ST-FRCNN model in the various pulmonary classification of disease. The model's accuracy is 85.8% for COVID-19, 84.2% for Normal, 85.0% for Lung Opacity, and 85.5% for Viral Pneumonia illustrated in Figure 7. The values suggest good overlap between predicted and actual regions which means there is good localization of abnormalities. The high scores in each class indicate that the model is good at accurately marking the affected areas within medical images. Such a high degree of localization accuracy is crucial to clinical diagnosis and treatment planning. In general, the outcomes demonstrate the reliability and accuracy of spatial identification of lung diseases as suggested by the proposed framework.

5.7 Overall Comparison of Parameters

Table 9. Comparison of all parameters

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)	Jaccard (%)
CNN-VT	83.8	83.2	82.7	70.5	89.2	70.5
DL-VT	86.1	85.6	85.1	74.2	91.1	74.2
LCSPM-MiD	87.4	86.8	86.5	76.8	92.3	76.8
YOLOv5	88.6	87.9	87.4	78.6	93.2	78.5
MAE-ST-FRCNN	92.7	92.0	91.6	85.1	96.2	85.1

Table 9 is a general comparison of all the models based on the average performance measurement. The accuracy, precision, and recall of the proposed MAE-ST-FRCNN are 92.7%, 92.0%, and 91.6%, respectively, which show that the proposed method is effective in terms of classification performance. It also achieves an F1-score of 85.1% which demonstrates a balanced precision/recall ratio. The model shows maximum AUC value as 96.2%, which demonstrates that the model has an excellent capacity to discriminate between classes. In addition, it has a Jaccard Index value of 85.1% which shows the higher localization accuracy. YOLOv5 achieves 88.6% accuracy and 78.5% Jaccard score, which outperforms other models and demonstrates the effectiveness and reliability of the proposed framework.

5.8 Ablation of Study

To analyze the contribution of each part in the proposed MAE-ST-FRCNN framework, an ablation study was carried out. The removal of the MAE module impacted the robustness of features and overall accuracy, demonstrating the importance of the self-supervised learning. The performance of multi-scale contextual understanding decreased when the Swin Transformer was removed, resulting in a reduction in Precision and F1 score. Likewise, the replacement of Faster R-CNN yielded similar results for localization, in which the localization ability of the model was weakened.

The Explainable Artificial Intelligence (XAI) is an important part of improving model explainability. Gradient-based visualization methods produce heatmaps that are useful to identify key areas that affect predictions in chest X-ray images and CT scans. When XAI is not used, the model becomes a black box and can be hard for clinicians to interpret and verify the correctness of the results. The lack of interpretability makes the clinical trust low and makes it less usable. Overall, the integration of MAE, Swin Transformer, Faster R-CNN and XAI achieves the best performance in detection, localization, and reliability.

6. CONCLUSION AND FUTURE WORKS

It was demonstrated that the proposed MAE-ST-FRCNN system was a robust and efficient tool for real-time recognition of pulmonary conditions, using chest X-ray and CT images. The model was very successful at local and global contextual dependencies by leveraging the Swin Transformer, which enables the application of Masked Autoencoders to obtain hierarchical features and Faster R-CNN to identify multi-modal objects with maximum precision in localization and classification. The results of the experiment showed that the experimental models were more accurate, had higher mean average precision and F1-Score, faster inference speed, and greater explainability than the baseline models. Model transparency and reliability among healthcare professionals improved, and clinically useful information was obtained through the application of XAI methods. Overall, the framework helps overcome the most significant limitations of classical CNN-based solutions and offers a stable, comprehensible solution that can be effectively integrated into clinical practice.

To further work with the model in the future, it can be extended to include other causes of pulmonary disease and to utilize larger, more diverse datasets to improve generalization across populations. Real-time hospital PACS systems can be integrated with automated clinical processes. Future research can explore designs for lightweight transformers to reduce the computational costs incurred by mobile or edge devices. The existing XAI module and multimodal fusion strategies should be complemented by quantitative interpretability metrics to ensure the clinical decision-making process is even stronger and to allow the system to be adaptable to the conditions of other healthcare settings.

References:

1. Fu, X., Lin, R., Du, W., Tavares, A., & Liang, Y. (2025). Explainable hybrid transformer for multi-classification of lung disease using chest X-rays. *Scientific Reports*, *15*(1), 6650.
2. Durgam, R., Panduri, B., Balaji, V., Khadidos, A. O., Khadidos, A. O., &Selvarajan, S. (2025). Enhancing lung cancer detection through integrated deep learning and transformer models. *Scientific Reports*, *15*(1), 15614.
3. Veeramani, N., SA, R. S., S, S. P., S, S., & Jayaraman, P. (2025). NextGen lung disease diagnosis with explainable artificial intelligence. *Scientific Reports*, *15*(1), 33052.
4. Ko, J., Park, S., & Woo, H. G. (2024). Optimization of vision transformer-based detection of lung diseases from chest X-ray images. *BMC Medical Informatics and Decision Making*, *24*(1), 191.
5. Anand, V., Shuaib, M., Khan, I., Ullah, M., & Alam, S. (2025). Secure pulmonary diagnosis using transformer-based approach to X-ray classification with KL divergence optimization. *Frontiers in Medicine*, *12*, 1716066.
6. Singh, S., Kumar, M., Kumar, A., Verma, B. K., Abhishek, K., &Selvarajan, S. (2024). Efficient pneumonia detection using Vision Transformers on chest X-rays. *Scientific reports*, *14*(1), 2487.

7. Naz, Z., Khan, M. U. G., Saba, T., Rehman, A., Nobanee, H., & Bahaj, S. A. (2023). An explainable AI-enabled framework for interpreting pulmonary diseases from chest radiographs. *Cancers*, *15*(1), 314.
8. Saber, A., Fateh, A., Parhami, P., Siahkarzadeh, A., Fateh, M., & Ferdowsi, S. (2025). Efficient and accurate pneumonia detection using a novel multi-scale transformer approach. *Sensors*, *25*(23), 7233.
9. Almughamisi, N., Abosamra, G., Albar, A., & Saleh, M. (2025). Detection of single and dual pulmonary diseases using an optimized vision transformer. *Journal of Intelligent Systems*, *34*(1), 20240419.
10. Adnan, K. M., Ghazal, T. M., Saleem, M., Farooq, M. S., Yeun, C. Y., Ahmad, M., & Lee, S. W. (2025). Secure and interpretable lung cancer prediction model using mapreduce private blockchain federated learning and XAI. *Scientific Reports*, *15*(1), 35693.
11. Mustapha, B., Zhou, Y., Shan, C., & Xiao, Z. (2025). Enhanced pneumonia detection in chest X-rays using hybrid convolutional and vision transformer networks. *Current Medical Imaging*, *21*(1), e15734056326685.
12. Fu, T. J., Lin, S., Wang, T., Chou, K. T., & Huang, S. F. (2024, July). Vision Transformer Based Detection Of Chronic Pulmonary Aspergillosis Lung Infections In Chest X-Ray Images. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 1-4). IEEE.
13. Alghadhban, A., Ramadan, R. A., &Alazmi, M. (2025). Advancing respiratory disease diagnosis: a deep learning and vision transformer-based approach with a novel X-ray dataset. *Computers in Biology and Medicine*, *194*, 110501.
14. Zhao, X., Li, J., Qi, M., Chen, X., Chen, W., Li, Y., ... & Zhang, C. (2025). MSTD: A multi-scale transformer-based method to diagnose benign and malignant lung nodules. *IEEE Access*, *13*, 16182-16195.
15. Zhu, P., Wang, T., Yang, F., Wang, M., & Zhang, Y. (2025). A transformer-based multi-scale deep learning model for lung cancer surgery optimization. *IEEE Access*.
16. Zhu, X., Lyu, S., Wang, X., & Zhao, Q. (2021). TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 2778-2788).
17. https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database