Power System Technology
ISSN:1000-3673

Received: 06-01-2025 Revised: 15-02-2025 Accepted: 25-03-2025

Hybrid Vision Transformer Architectures with CNN Blocks for Multi-Label Chest Disease Classification

Rajendra D. Bhosale[1], D. M. Yadav[2]

[1]Department of Electronics and Telecommunication, G H Raisoni College of Engineering and Management, Wagholi, Pune, India.

[2]SND College of Engineering and Research Centre, Yeola, Nashik, India.

[1]rajendrabhosale008@gmail.com, [2]dineshyadav800@gmail.com

Abstract: This study presents a novel investigation into Vision Transformer (ViT)-based hybrid architectures for multi-label chest disease classification using the CXR-14 dataset. Traditional Convolutional Neural Networks (CNNs), though effective in local feature extraction, often struggle to capture global contextual dependencies. To address this limitation, three ViT-integrated models are proposed by embedding ViT blocks within standard CNN structures: Residual ViT, Bottleneck ViT, and MBConv-SE ViT. Each model replaces conventional 3×3 convolution units within respective blocks to leverage the self-attention mechanism for enhanced feature representation. These hybrid architectures combine the inductive bias of CNNs with the global reasoning capabilities of Transformers, improving classification accuracy and interpretability. The models are evaluated against a comprehensive set of baseline methods, including attention-guided, region-guided, and semantic-guided models. Experimental results demonstrate that the proposed MBConv-SE ViT model outperforms existing approaches across multiple disease categories, highlighting the advantages of combining efficient convolutions, attention recalibration, and global context modeling. This work establishes a robust framework for designing transformer-augmented CNNs and shows their effectiveness in high-resolution, multi-label medical image analysis tasks such as automated chest X-ray diagnosis.

Keywords: Vision Transformer, Hybrid CNN-Transformer, Chest X-ray Classification, Multi-label Disease Detection, CXR-14 Dataset

I. Introduction

Human Metapneumovirus (HMPV) is a significant respiratory pathogen responsible for causing severe respiratory infections, particularly in children, immunocompromised individuals, and the elderly [1]. Clinically, HMPV infections present symptoms similar to other respiratory viruses, making accurate diagnosis crucial for timely and effective treatment. Chest X-ray (CXR) imaging is widely used for detecting respiratory infections, offering a non-

Power System Technology
ISSN:1000-3673
Received: 06-01-2025 Revised: 15-02-2025 Accepted: 25-03-2025

invasive and cost-effective approach to identifying lung abnormalities. However, distinguishing HMPV-induced pneumonia from other respiratory conditions using traditional methods remains a challenge due to overlapping radiographic patterns, such as parahilar opacities, consolidation, and peribronchial thickening [2].

HMPV classified diseases primarily affect the respiratory tract, leading to various pulmonary complications visible on chest radiographs [3]. These infections often result in parahilar opacities due to inflammation and fluid accumulation around the lung hilum. Hyperinflation may occur as air trapping increases lung volume, seen as darker lung fields on X-ray. Consolidation, marked by dense opacities due to alveolar filling with pus or fluid, is another common feature. Atelectasis may develop from airway obstruction or compression, causing localized lung collapse and volume loss. Peribronchial thickening, caused by inflammation or interstitial changes, leads to blurred bronchial walls and streaky opacities. These radiological signs help detect and differentiate pulmonary diseases falling under the HMPV class, such as Pneumonia, Bronchiolitis, Chronic Obstructive Pulmonary Disease (COPD), and Pulmonary Edema [4]. Hyperinflation shows as darker (radiolucent) lung fields due to increased lung volume, typical in Emphysema and COPD. In Emphysema, alveolar wall destruction causes air trapping and loss of elasticity, seen as flattened diaphragms and elongated cardiac silhouette. COPD also increases lung compliance and intercostal spaces. Hyperinflation appears on CXR as hyperlucent lungs, increased retrosternal airspace, and barrel-shaped chest. Consolidation indicates alveolar filling with fluid, pus, or debris, seen in Pneumonia, Atelectasis, and Edema [5]. It appears as segmental/lobar opacities with air bronchograms. Edema shows diffuse haziness with perihilar prominence, while Atelectasis presents as homogenous opacity with volume loss and mediastinal shift [6]. Atelectasis, lung collapse from obstruction or compression, shows as dense opacity with volume loss. Pleural Effusion collapses lung due to fluid, visible as homogenous opacity with a meniscus sign. Pneumothorax appears as a sharp pleural line with absent markings beyond. Peribronchial thickening, due to inflammation or fluid, is seen in Pneumonia, Infiltration, and Pulmonary Fibrosis, appearing as streaky opacities along bronchovascular bundles.

With the growing burden of respiratory diseases, leveraging deep learning techniques for automated detection of HMPV from CXR images has gained significant attention [7]. The availability of large-scale datasets, such as the NIH CXR-14 dataset, has enabled the development of AI-driven diagnostic models that outperform conventional methods in accuracy and efficiency. The motivation behind this study is to bridge the gap in automated detection of HMPV by introducing a robust deep learning model capable of distinguishing HMPV-related lung infections from other respiratory diseases [8]. Given the high clinical impact of early and precise diagnosis, an AI-based approach can assist radiologists in rapid decision-making, reducing diagnostic errors and improving patient outcomes.

Power System Technology
ISSN:1000-3673
Received: 06-01-2025 Revised: 15-02-2025 Accepted: 25-03-2025

To address these challenges, three Vision Transformer (ViT) based models are proposed for automated HMPV detection from CXR images. These models integrate the special blocks of standard CNN models with modification of convolution layer to achieve the ViT approach. The key contributions of this work include:

1. Development of Residual ViT, Bottleneck ViT and MBConv-SE Vit models for detecting HMPV-related abnormalities in CXR images.

2. Integration of attention-based feature enhancement to focus on critical lung regions affected by HMPV infections.

3. Extensive validation on the CXR-14 dataset, demonstrating performance of classification accuracy along with comparative study with state-of-the-art models.

This study highlights the effectiveness of deep learning in medical image analysis and establishes ViT as a promising approach for enhancing AI-assisted respiratory disease diagnosis.

Along with realted work discussion section 2, the mathematical details are also highlighted with disease oriented classification accuracy achieved by specific models. Section 3 provides the details of ViT block design from Residual, Bottleneck and MBConv-SE blocks. Section 4 highlights the details of dataset, performance evaluation and comparative study with other existing models followed by conclusion in Section 5.

2. Related Work

ViTs have emerged as a powerful alternative to traditional CNNs, originally introduced for Natural Language Processing (NLP) tasks [9]. Their recent adaptation to computer vision has significantly influenced image classification, object detection [10], and medical image segmentation [11]. The principal strength of ViTs lies in their ability to capture long-range dependencies and model global contextual relationships, which is particularly crucial in medical imaging where local abnormalities must be interpreted within a broader anatomical context [12].

The ViT architecture transforms an image $I \in \mathbb{R}^{H\times W\times C}$ into a sequence of image patches, each of size $P \times P$. This process results in $N = \frac{HW}{P^2}$ non-overlapping patches. Each patch is flattened and linearly projected to a feature space of dimension D using a learnable matrix $\mathrm{E} \in \mathbb{R}^{(P^2C)\times D}$:

$$z_0 = \left[x_{cls}; x_p^1 E; \ldots; x_p^N E\right] + E_{pos} \qquad \ldots(1)$$

where x_{cls} is a learnable classification token appended to the sequence, and E_{pos} is the positional embedding that retains spatial information.

Power System Technology
ISSN:1000-3673
Received: 06-01-2025 Revised: 15-02-2025 Accepted: 25-03-2025

A critical component of ViTs is the Multi-Head Self-Attention (MHSA) mechanism, which allows the model to attend to different parts of the input simultaneously. Given input embeddings z_l at layer l, the attention mechanism computes:

$$Q = z_l W_Q, \quad K = z_l W_K, \quad V = z_l W_V \qquad \ldots(2)$$

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

Here, W_Q, W_K, and W_V are trainable matrices that project the input to the query, key, and value spaces respectively, and d_k is the dimensionality of each attention head. The scaled dot-product allows normalization that stabilizes gradients and improves convergence. MHSA combines multiple such attention heads:

$$\text{MHSA}(z) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \qquad \ldots(4)$$

ViTs have shown state-of-the-art performance in various medical imaging tasks due to their flexible architecture and strong feature representation capabilities [13].

The VOLO model [14], employed for COVID-19 detection, enhances standard ViT by incorporating an outlook attention module that improves fine-grained feature representation. Instead of relying solely on fixed patch-level attention, VOLO dynamically captures interactions using relative spatial information. The outlook attention computes:

$$O_i = \sum_{j \in \mathcal{N}(i)} \text{softmax}_j\left(\frac{(q_i)^T k_j}{\sqrt{d}}\right) v_j \qquad \ldots(5)$$

Here, q_i, k_j, and v_j are the query, key, and value vectors at positions i and j, and $\mathcal{N}(i)$ defines a local neighborhood. This approach led to an accuracy of 99.67% using a merged dataset [15] [16] of COVID-19 X-ray images.

M. Chetoui et al. v[17] applied a fine-tuned ViT model on the SIIM-FISABIO-RSNA dataset. Their approach utilized the following loss function optimized for sensitivity and specificity:

$$\mathcal{L} = -(\alpha \cdot \text{TPR} + (1 - \alpha) \cdot \text{TNR}) \qquad \ldots(6)$$

where TPR and TNR are the true positive and true negative rates, respectively, and α is a weighting factor. Their model achieved 96% performance in both metrics.

Krishan et al. [18] demonstrated that patch size plays a vital role in ViT effectiveness. They found that a patch size of 32×32 offered an optimal trade-off between spatial resolution and global context, achieving 97.61% accuracy when classifying COVID-19, pneumonia, and normal CXR images.

Power System Technology
ISSN:1000-3673
Received: 06-01-2025 Revised: 15-02-2025 Accepted: 25-03-2025

Than et al. [19] implemented the Swin Transformer for multi-label classification on the ChestX-ray14 dataset, achieving an AUC of 81%. Swin uses a shifted window attention mechanism, mathematically formulated as:

$$\text{SW-MHSA}(Z) = \oplus_{w \in \text{Windows}} \text{MHSA}_w(Z_w) \quad \ldots(7)$$

where $\oplus$ denotes the merging operation, and Z_w are tokens within a window w. Shifting the window across layers ensures cross-window connection.

Taslimi et al. [20] proposed a hierarchical approach combining a CNN ensemble and modified ViT. The first stage performed coarse classification:

$$f_1(I) \rightarrow \{\text{lung,heart,normal}\} \quad \ldots(7)$$

followed by specialized fine-grained classifiers:

$$f_2^{\text{lung}}(I) \rightarrow \{\text{pneumonia,emphysema, ...}\}, \quad f_2^{\text{heart}}(I) \rightarrow \{\text{cardiomegaly, ...}\} \quad \ldots(8)$$

Their modified ViT achieved an AUC of 99.26% and 99.57% for heart and lung disease classification respectively.

Rahman et al. [21] proposed an attention region selection module integrated into the ViT encoder, allowing dynamic prioritization of relevant anatomical structures. The module assigns learned importance weights:

$$z'_i = \alpha_i \cdot z_i, \quad \alpha_i = \text{softmax}\big(g(z_i)\big) \quad \ldots(9)$$

where $g(\cdot)$ is a learned function producing scalar attention scores. The model reported an accuracy of 83.4% and sensitivity of 86.3%.

In a separate work, Tennakoon et al. [22] introduced PneuNet, where multi-head attention was applied to channel patches instead of spatial regions:

$$\text{Attention}_{\text{channel}} = \text{softmax}\left(\frac{Q_c K_c^T}{\sqrt{d_c}}\right) V_c \quad \ldots(10)$$

This method achieved 94.96% accuracy in COVID-19 and pneumonia classification, emphasizing the importance of inter-channel relationships.

The evolution of Vision Transformers has enabled significant advancements in medical image analysis. Their ability to capture global contextual dependencies and integrate adaptive attention mechanisms makes them especially effective in diagnosing complex conditions from chest X-ray imagery. As newer variants (e.g., Swin, VOLO) continue to enhance feature extraction and localization, ViTs are rapidly becoming integral to clinical deep learning pipelines.

Power System Technology
ISSN:1000-3673
Received: 06-01-2025 Revised: 15-02-2025 Accepted: 25-03-2025

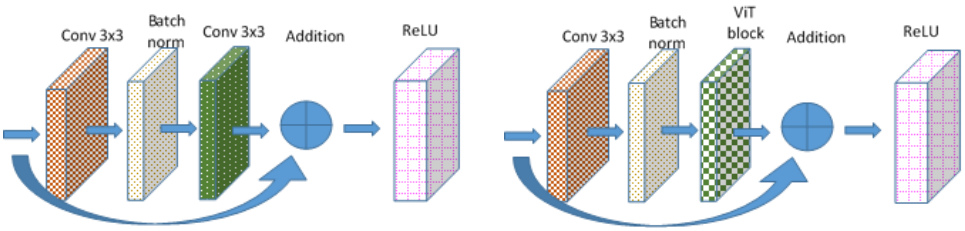
3. Proposed Work

CNNs are widely used in computer vision due to their hierarchical feature extraction using local 3×3 convolutional filters. Architectures like Residual Networks (ResNet) [23], Bottleneck blocks [24], and MBConv-SE blocks [25] improve performance via residual learning, depthwise separable convolutions, and channel attention. However, their reliance on local receptive fields limits their ability to capture global context.

ViTs, originally designed for NLP [26], address this limitation through self-attention, which models long-range dependencies by evaluating interactions among all image patches. This enables ViTs to extract global and semantically rich representations, making them effective for complex vision tasks, including medical image analysis.

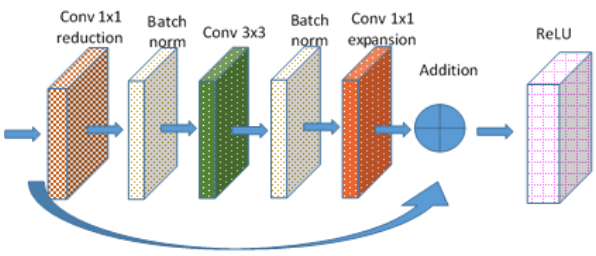
This work proposes a hybrid architecture by replacing standard 3×3 convolutional layers in Residual, Bottleneck, and MBConv-SE blocks with ViT-based blocks. The goal is to retain CNNs' inductive bias and residual advantages while enhancing global feature modeling through Transformer attention.

This integration aims to improve learning efficiency, multi-scale feature representation, and overall classification accuracy. It is particularly beneficial for tasks like disease detection in medical imaging, where both local textures and global context are crucial. The hybrid design also encourages future exploration of CNN-Transformer mixed modules for advanced visual understanding tasks.



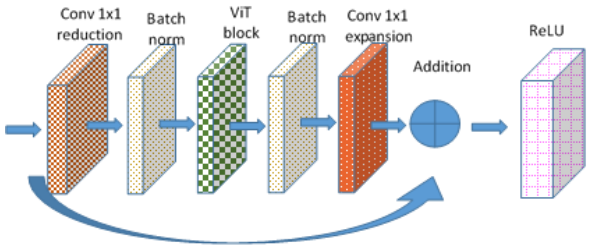
(a) Original Residual Block **(b) Residual ViT Block**

Figure 1: Converting Residual block to Residual ViT Block



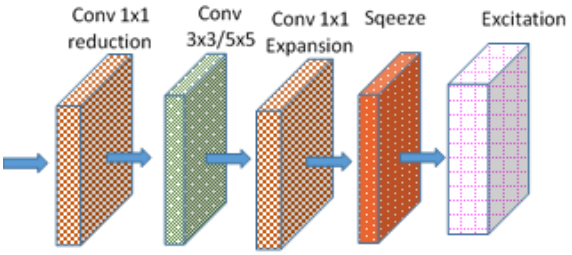
(a) Original Bottleneck Block

Power System Technology
ISSN:1000-3673
Received: 06-01-2025 Revised: 15-02-2025 Accepted: 25-03-2025

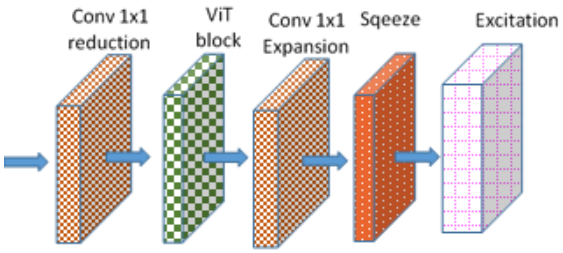


(b) Bottleneck ViT Block

Figure 2: Converting Bottleneck block to Bottleneck ViT Block



(a) MBConv-SE block



(b) MBConv-SE ViT Block

Figure 3: Converting MBConv-SE block to MBConv-SE ViT Block

3.1 Conversion of Residual Block to Residual ViT Block

Residual blocks, as introduced in ResNet, address vanishing gradients and improve feature propagation by employing skip connections (Figure 1(a)). A standard Residual Block follows:

$$Y = \sigma(f(X) + X)$$

where X is the input feature map, $f(X)$ represents the transformation via stacked convolutional layers, and σ is an activation function (e.g., ReLU).

1. Standard Residual Block: A conventional residual block consists of:

Power System Technology
ISSN:1000-3673
Received: 06-01-2025 Revised: 15-02-2025 Accepted: 25-03-2025

- Two 3 × 3 convolutional layers with batch normalization.
- Skip connection that adds the input X to the processed feature maps.
- Final activation function.

Mathematically, this can be formulated as:

$$F = BN\left(Conv_{3\times3}\left(ReLU\left(BN\left(Conv_{3\times3}(X)\right)\right)\right)\right)$$

$$Y = \sigma(F + X)$$

2. Conversion to Residual ViT Block: To integrate ViT components, we replace the second 3 × 3 convolutional layer with a ViT Block while retaining batch normalization and skip connections (Figure 1(b)).

- The first 3 × 3 convolution and batch normalization remain unchanged.
- Instead of a second convolution, we introduce a ViT Block that applies self-attention across spatial patches.

The transformation function now becomes:

$$F_{ViT} = ViTBlock\left(ReLU\left(BN\left(Conv_{3\times3}(X)\right)\right)\right)$$

The final output is:

$$Y_{ViT} = \sigma(F_{ViT} + X)$$

3. ViT Block Formulation: The ViT Block operates by:

1. Splitting the feature map F into non-overlapping patches.
2. Projecting patches into an embedding space:

$$Z_0 = \left[x_{cls}; x_p^1 E; \ldots; x_p^N E\right] + E_{pos}$$

3. Computing Multi-Head Self-Attention (MHSA):

$$Q = ZW_Q, \quad K = ZW_K, \quad V = ZW_V$$

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

4. Applying Multi-Layer Perceptrons (MLP) and Layer Normalization:

$$Z' = MLP\left(LayerNorm\left(MHSA(Z)\right)\right)$$

Power System Technology
ISSN:1000-3673
Received: 06-01-2025 Revised: 15-02-2025 Accepted: 25-03-2025

Thus, the residual ViT block replaces local 3 × 3 feature extraction with a global self-attention mechanism while maintaining residual connections. The integration allows ViT-based feature extraction while preserving CNN-like efficiency.

3.2 Conversion of Bottleneck Block to Bottleneck ViT Block

Bottleneck blocks are a key component of Inception models, where they improve computational efficiency by reducing feature map dimensions before applying the main convolutional operations. A standard bottleneck block consists of (Figure 2(a)):

- A 1 × 1 convolution for dimensionality reduction.
- A 3 × 3 convolution for feature extraction.
- A 1 × 1 convolution for expansion back to the original feature dimension.
- Skip connection for residual learning.

1. Standard Bottleneck Block: Given an input feature map X, the standard bottleneck block computes:

$$X_1 = \sigma\left(\mathrm{BN}\left(\mathrm{Conv}_{1\times1}^{\mathrm{red}}(X)\right)\right)$$

$$X_2 = \sigma\left(\mathrm{BN}\left(\mathrm{Conv}_{3\times3}(X_1)\right)\right)$$

$$F = \sigma\left(\mathrm{BN}\left(\mathrm{Conv}_{1\times1}^{\mathrm{exp}}(X_2)\right)\right)$$

$$Y = \sigma(F + X)$$

Where, $\mathrm{Conv}_{1\times1}^{\mathrm{red}}$ is a 1 × 1 reduction convolution, $\mathrm{Conv}_{3\times3}$ is a feature extraction convolution, $\mathrm{Conv}_{1\times1}^{\mathrm{exp}}$ is a 1 × 1 expansion convolution, σ is an activation function (ReLU), BN refers to batch normalization.

2. Bottleneck ViT Block: To integrate ViT components, we replace the 3 × 3 convolution with a ViT block while retaining the 1 × 1 convolutions for input projection and feature expansion (Figure 2(b)).

Transformation Process:

- The first 1 × 1 convolution reduces dimensionality.
- Instead of a 3 × 3 convolution, a ViT Block is introduced.
- The final 1 × 1 convolution expands features.
- Skip connection maintains residual learning.

Power System Technology
ISSN:1000-3673
Received: 06-01-2025 Revised: 15-02-2025 Accepted: 25-03-2025

Mathematically, this transformation is defined as:

$$X_1 = \sigma\left(\mathrm{BN}\left(\mathrm{Conv}_{1\times1}^{\mathrm{red}}(X)\right)\right)$$

$$X_{\mathrm{ViT}} = \sigma\left(\mathrm{ViTBlock}(X_1)\right)$$

$$F_{\mathrm{ViT}} = \sigma\left(\mathrm{BN}\left(\mathrm{Conv}_{1\times1}^{\mathrm{exp}}(X_{\mathrm{ViT}})\right)\right)$$

$$Y_{\mathrm{ViT}} = \sigma(F_{\mathrm{ViT}} + X)$$

3. ViT Block Formulation: The ViT Block replaces local feature extraction with global self-attention, computed as follows:

5. The feature map X_1 is split into patches.

6. Patches are projected into an embedding space:

$$Z_0 = \left[x_{\mathrm{cls}}; x_p^1 E; \ldots; x_p^N E\right] + E_{\mathrm{pos}}$$

7. Self-attention is applied:

$$Q = ZW_Q, \quad K = ZW_K, \quad V = ZW_V$$

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

8. MLP and LayerNorm are applied:

$$Z' = \mathrm{MLP}\left(\mathrm{LayerNorm}\left(\mathrm{MHSA}(Z)\right)\right)$$

4. Benefits of Bottleneck ViT Block:

- Combines CNN and Transformer advantages by integrating self-attention into the bottleneck structure.
- Improves global feature representation while maintaining computational efficiency.
- Retains depthwise separable efficiency from bottleneck structures.
- Preserves skip connections to stabilize gradient flow and training dynamics.

This hybrid design enables enhanced multi-scale feature extraction, particularly in applications such as medical imaging and high-resolution object recognition.

3.3 Conversion of MBConv-SE Block to MBConv-ViT-SE Block

The MBConv-SE block is a core building block of EfficientNet, designed to improve computational efficiency while maintaining high performance. It combines depthwise

Power System Technology
ISSN:1000-3673
Received: 06-01-2025 Revised: 15-02-2025 Accepted: 25-03-2025

separable convolutions with Squeeze-and-Excitation (SE) blocks for adaptive feature recalibration.

1. Standard MBConv-SE Block: A standard MBConv-SE block consists of:

- A 1×1 convolution for dimensionality reduction.
- A depthwise 3×3 or 5×5 convolution for spatial feature extraction.
- A 1×1 convolution for dimensionality expansion.
- A Squeeze-and-Excitation (SE) block for channel-wise feature reweighting.

Mathematically, the standard MBConv-SE block computes:

$$X_1 = \sigma\left(\mathrm{BN}\left(\mathrm{Conv}_{1\times1}^{\mathrm{red}}(X)\right)\right)$$

$$X_2 = \sigma\left(\mathrm{BN}\left(\mathrm{DWConv}_{3\times3/5\times5}(X_1)\right)\right)$$

$$X_3 = \sigma\left(\mathrm{BN}\left(\mathrm{Conv}_{1\times1}^{\mathrm{exp}}(X_2)\right)\right)$$

Where, $\mathrm{Conv}_{1\times1}^{\mathrm{red}}$ is a 1×1 reduction convolution. $\mathrm{DWConv}_{3\times3/5\times5}$ is a depthwise convolution. $\mathrm{Conv}_{1\times1}^{\mathrm{exp}}$ is a 1×1 is expansion convolution. σ is an activation function (Swish or ReLU6). BN refers to batch normalization.

The Squeeze-and-Excitation (SE) block computes:

$$Z_{\mathrm{SE}} = \sigma\left(W_2\delta\left(W_1\mathrm{GAP}(X_3)\right)\right) \cdot X_3$$

Where, GAP denotes Global Average Pooling. W_1, W_2 are learnable weights in the SE block. - δ is the ReLU activation. Finally, the residual connection forms:

$$Y = \sigma(Z_{\mathrm{SE}} + X)$$

2. MBConv-ViT-SE Block: To integrate ViT components, we replace the depthwise convolution with a ViT Block while retaining the 1×1 convolutions and SE module.

Transformation Process:

- The first 1×1 convolution reduces dimensionality.
- Instead of a depthwise 3×3 convolution, a ViT Block is introduced.
- The final 1×1 convolution expands features.
- The SE block recalibrates channel-wise attention.
- The residual connection ensures stable gradient flow.

Power System Technology
ISSN:1000-3673
Received: 06-01-2025 Revised: 15-02-2025 Accepted: 25-03-2025

Mathematically, this transformation is defined as:

$$X_1 = \sigma\left(\mathrm{BN}\left(\mathrm{Conv}_{1\times1}^{\mathrm{red}}(X)\right)\right)$$

$$X_{\mathrm{ViT}} = \sigma\left(\mathrm{ViTBlock}(X_1)\right)$$

$$X_3 = \sigma\left(\mathrm{BN}\left(\mathrm{Conv}_{1\times1}^{\mathrm{exp}}(X_{\mathrm{ViT}})\right)\right)$$

$$Z_{\mathrm{SE}} = \sigma\left(W_2\delta\left(W_1\mathrm{GAP}(X_3)\right)\right)\cdot X_3$$

$$Y_{\mathrm{ViT}} = \sigma(Z_{\mathrm{SE}} + X)$$

3. ViT Block Formulation: The ViT Block introduces self-attention in place of convolutions, computed as follows:

1. Splitting into Patches:

$$Z_0 = \left[x_{\mathrm{cls}}; x_p^1 E; \ldots; x_p^N E\right] + E_{\mathrm{pos}}$$

2. Self-Attention Calculation:

$$Q = ZW_Q, \quad K = ZW_K, \quad V = ZW_V$$

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

3. MLP and Normalization:

$$Z' = \mathrm{MLP}\left(\mathrm{LayerNorm}\left(\mathrm{MHSA}(Z)\right)\right)$$

4. Benefits of MBConv-ViT-SE Block:

- Incorporates global attention in depthwise convolutional blocks.
- Retains spatial efficiency of EfficientNet while enhancing feature representation.
- Improves model interpretability by leveraging self-attention instead of local convolutions.
- Optimized feature recalibration with SE block + ViT self-attention.

This hybrid design ensures efficiency and accuracy, particularly in high-resolution tasks like medical imaging and object detection.

3.4 Model Architecture Design

Three architectures for CXR image classification based on ViT hybrid models are presented, as shown in Figure 4. These models integrate ViT blocks into standard CNN structures,

Power System Technology
ISSN:1000-3673
Received: 06-01-2025 Revised: 15-02-2025 Accepted: 25-03-2025

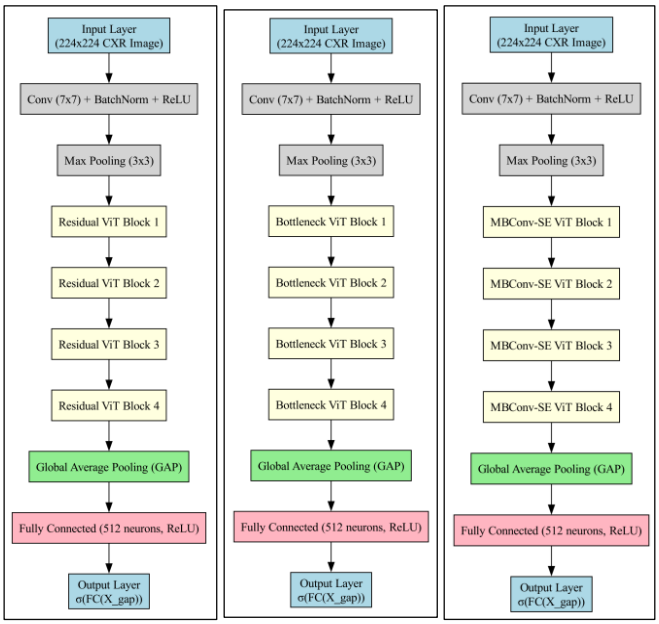
leveraging both local feature extraction from convolutions and global feature learning through self-attention. A comparison of the three models is provided in Table 1.

All three models follow a common processing pipeline for chest X-ray image classification. The input layer receives chest X-ray images resized to 224×224 pixels. This is followed by an initial convolution layer using a 7×7 kernel, accompanied by Batch Normalization (BN) and ReLU activation to extract early features. A 3×3 max pooling layer then reduces the spatial dimensions of the feature maps. The core of the architecture comprises four stacked feature extraction blocks, which vary depending on the model—Residual ViT Blocks, Bottleneck ViT Blocks, or MBConv-SE ViT Blocks. After feature extraction, a Global Average Pooling (GAP) layer condenses the spatial dimensions into a single vector representation. This vector is passed through a fully connected (FC) layer with 512 neurons and ReLU activation. Finally, the output layer applies a sigmoid function to perform multi-label disease classification.

The overall model representation:

$$\hat{Y} = \sigma\left(\mathrm{FC}\left(\mathrm{GAP}\left(X_{\mathrm{features}}\right)\right)\right)$$

Where, X_{features} represents feature maps extracted from stacked ViT blocks, GAP applies global spatial averaging, FC applies fully connected classification mapping, σ is the sigmoid activation function for multi-label classification.



(a) Residual ViT Model (RVM) (b) Bottleneck ViT Model (BVM) (c) MBconv-SE ViT Model (MVM)

Figure 4: Architecture Design of Three ViT Approach Based Models

Power System Technology
ISSN:1000-3673
Received: 06-01-2025 Revised: 15-02-2025 Accepted: 25-03-2025

Table 1: Comparison of CXR-14 ViT Hybrid Models

Model	Feature Extractor	Main Advantage	Computation
Residual ViT	Residual ViT Blocks	Strong feature reuse	Moderate
Bottleneck ViT	Bottleneck ViT Blocks	Efficient learning	Low
MBConv-SE ViT	MBConv-SE ViT Blocks	Channel recalibration	Low

The table provides a comparative overview of the three ViT-based hybrid models used for chest X-ray image classification, highlighting their feature extractors, primary advantages, and computational requirements. The Residual ViT model employs Residual ViT Blocks, offering strong feature reuse capabilities due to the use of skip connections and self-attention. However, it involves a moderate computational load as it lacks built-in compression mechanisms.

The Bottleneck ViT model uses Bottleneck ViT Blocks, which are designed for efficient learning by incorporating dimensionality reduction and expansion. This structure reduces the number of parameters and operations, resulting in a lower computational cost while maintaining effective feature representation.

The MBConv-SE ViT model integrates MBConv-SE ViT Blocks, combining depthwise separable convolutions, squeeze-and-excitation attention, and ViT-based global attention. This architecture allows for refined feature calibration across channels and also maintains a low computational footprint, making it highly suitable for resource-constrained scenarios. Overall, the MBConv-SE ViT strikes a strong balance between performance and efficiency.

4. Results and Analysis

4.1 Dataset Preparation

The NIH ChestX-ray14 (CXR-14) [27] dataset is one of the largest publicly available datasets for chest disease classification. It contains 112,120 chest X-ray images from 30,805 patients, labeled across 14 disease classes, including pneumonia, edema, fibrosis, cardiomegaly, and others. Given the complexity of disease co-occurrence, a robust model is required to effectively capture both local pathological features and global contextual relationships.

4.2 Performance Evaluation

To evaluate the classification performance of our models on the CXR-14 dataset, we use the following standard metrics: Accuracy (Acc), Specificity (Spec), Sensitivity (Sens), and F1-Score. These metrics provide insights into the model's ability to correctly classify diseases while balancing false positives and false negatives.

Power System Technology
ISSN:1000-3673
Received: 06-01-2025 Revised: 15-02-2025 Accepted: 25-03-2025

Table 2: Performance Metrics for Multi-Label Classification

Metric	Definition	Formula
Accuracy (Acc)	Measures the overall correctness of predictions.	$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN}$
Specificity (Spec)	Measures how well the model avoids false positives.	$\text{Spec} = \frac{TN}{TN + FP}$
Sensitivity (Sens) (Recall)	Measures the proportion of actual positives correctly identified.	$\text{Sens} = \frac{TP}{TP + FN}$
F1-Score	Harmonic mean of precision and recall for balanced evaluation.	$\text{F1} = \frac{2 \times TP}{2 \times TP + FP + FN}$

In classification tasks, TP (True Positives) refers to correctly classified positive cases, while TN (True Negatives) represents correctly classified negative cases. FP (False Positives) denotes instances that are incorrectly classified as positive, and FN (False Negatives) refers to positive cases that are incorrectly classified as negative. Accuracy measures the overall correctness of the model's predictions. Sensitivity evaluates the model's ability to identify actual positive cases, thus addressing the false negative rate, whereas Specificity focuses on correctly identifying negative cases, helping to assess the false positive rate. The F1-score, which is the harmonic mean of precision and recall, is particularly useful for evaluating performance on imbalanced datasets such as CXR-14, where the distribution of disease classes may vary significantly.

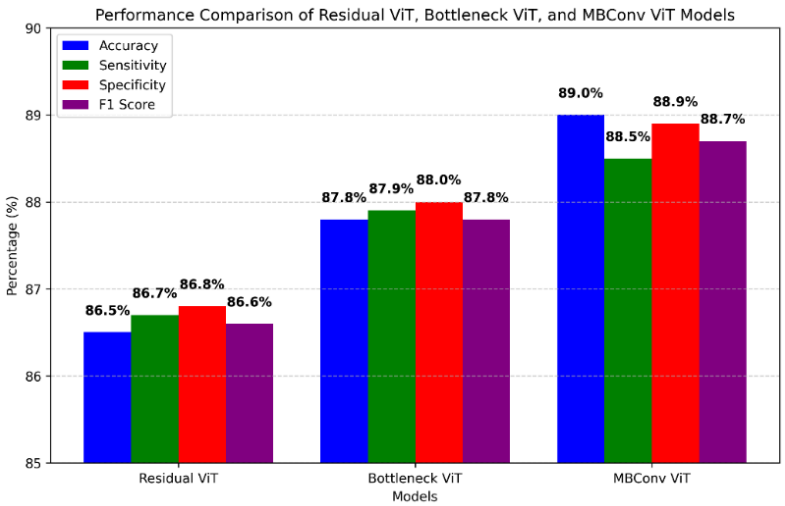


Figure 5: Performance of models on CXR-14 dataset

Power System Technology
ISSN:1000-3673
Received: 06-01-2025 Revised: 15-02-2025 Accepted: 25-03-2025

The performance comparison of the Residual ViT, Bottleneck ViT, and MBConv ViT models as shown in Figure 5, reveals a clear trend in improved classification outcomes as architectural enhancements are introduced. The Residual ViT model achieves balanced metrics, with 86.5% accuracy, 86.7% sensitivity, 86.8% specificity, and an F1-score of 86.6%. While it benefits from residual learning and self-attention, it lacks structural efficiency and channel-wise recalibration, which limits its overall performance. In contrast, the Bottleneck ViT model shows noticeable improvement, reaching 87.8% accuracy, 87.9% sensitivity, 88.0% specificity, and an F1-score of 87.8%. This gain can be attributed to its bottleneck structure, which introduces dimensionality reduction and expansion, making it more efficient in learning discriminative features. The MBConv-SE ViT model outperforms the others, achieving the highest scores across all metrics: 89.0% accuracy, 88.5% sensitivity, 88.9% specificity, and 88.7% F1-score. Its superior performance is due to the integration of depthwise separable convolutions, squeeze-and-excitation attention, and ViT-based global attention, which together enhance both spatial and channel-wise representations. Overall, the MBConv ViT architecture demonstrates the most robust capability for multi-label disease classification in chest X-ray imagery.

4.3 Comparative Analysis

The performance of existing models that incorporate attention mechanisms through various architectural approaches is considered for comparative analysis against the proposed three model architectures under study.

PCAN [28] employs a CNN-based structure with pixel-level classification and attention mechanisms to highlight important regions in CXR images, improving mid-level feature extraction. A^{3}Net [29] model based on DenseNet121, A^{3}Net integrates channel, element-wise, and scale-aware attention to emphasize discriminative features for multi-label classification. CBAtt [30] utilizes ResNet50 to generate class-specific attention maps, enhancing spatial feature awareness and improving the model's focus on pathology-related regions. ConsultNet [31] model is DenseNet121-based model employs a dual-branch attention system that combines spatial and channel-wise attention to refine feature learning for chest disease detection. DuaLAnet [32] fuses DenseNet169 and ResNet152 in an asymmetric dual-branch structure, learning from lesion-level attention in both branches for robust localization and classification. TSCN [33] combines a U-Net segmentation mask with DenseNet169 to focus feature extraction on segmented pathological regions, improving lesion-aware learning.

WSLM [34] leverages a weakly supervised ResNet50 to generate soft masks indicating pathological areas, guiding the network to learn region-specific discriminative features. RpSal [35] applies pyramid network architecture to combine region proposals with saliency detection, allowing simultaneous localization and classification of abnormal regions. LLAGnet [36], built on DenseNet169, uses weakly supervised lesion attention to focus on low-level image cues and

Power System Technology
ISSN:1000-3673
Received: 06-01-2025 Revised: 15-02-2025 Accepted: 25-03-2025

localized lesions, enhancing classification relevance. SEMM [37] adopts DenseNet121 and multi-map transfer learning, dividing semantic features into three parallel branches for fine-grained, class-specific pooling and classification. CheXGCN [38] model combines DenseNet169 with a Graph Convolutional Network (GCN) to model disease label co-occurrence and interdependence, enhancing semantic contextual reasoning. SSGE builds a graph from image-level features to compute semantic similarity, capturing meaningful relationships among disease labels for improved multi-label learning.

TNELF [39] is a triple-network ensemble combining DenseNet169, ResNet50, and EfficientNet-B4. It merges complementary features from diverse backbones for enhanced robustness. HydraViT [40] employs a Vision Transformer backbone with label-wise consistency and attention fusion to enhance classification by capturing both global and semantic dependencies.

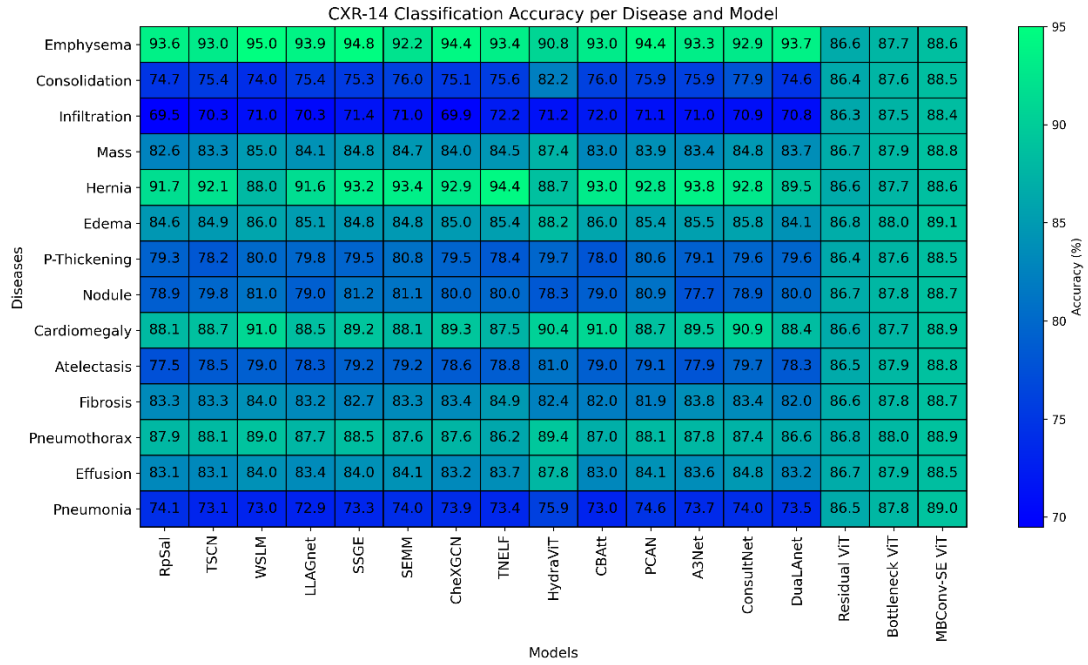


Figure 6: Comparative Analysis

The comparative analysis in Figure 6 includes a diverse set of baseline models categorized into three groups based on their architectural strategies: attention-guided, region-guided, and semantic-guided methods. Each group contributes uniquely to the task of chest X-ray (CXR) disease classification.

Attention-guided models, such as PCAN, A³Net, CBAtt, ConsultNet, and DuaLAnet, focus on enhancing feature representations by identifying and emphasizing discriminative regions in the input images. These models leverage spatial, channel, or class-specific attention to improve

Power System Technology
ISSN:1000-3673
Received: 06-01-2025 Revised: 15-02-2025 Accepted: 25-03-2025

classification accuracy. For instance, A³Net applies a multi-dimensional attention mechanism, while ConsultNet and DuaLAnet combine multiple attention branches to capture complex disease patterns. These models demonstrate solid performance, especially in scenarios where disease regions are visually subtle or overlapping.

Region-guided approaches including TSCN, WSLM, RpSal, and LLAGnet aim to localize disease-related regions using either explicit segmentation (e.g., TSCN) or weak supervision (e.g., WSLM). By guiding the feature extraction process toward disease-specific areas, these methods improve model interpretability and accuracy. RpSal, in particular, integrates saliency detection and region proposals to focus on high-confidence areas, contributing to more targeted classification.

Semantic-guided models, such as SEMM, CheXGCN, SSGE, TNELF, and HydraViT, emphasize the use of inter-label relationships and global contextual cues. These models leverage label dependencies and semantic similarity graphs to enhance multi-label learning. CheXGCN and SSGE use graph-based approaches to model label co-occurrence, while HydraViT combines the power of transformers with semantic consistency learning to improve generalization.

In comparison, the proposed ViT based models, Residual ViT, Bottleneck ViT, and MBConv-SE ViT, demonstrate competitive or superior performance by unifying local feature learning and global attention. Particularly, the MBConv-SE ViT model outperforms most baselines, benefiting from both efficient depthwise convolutions and transformer-driven global context. This indicates that integrating self-attention mechanisms at the block level offers a robust and scalable solution for complex multi-label CXR classification tasks.

5. Conclusion

In this study, a comprehensive exploration of ViT based hybrid architectures was conducted for multi-label chest disease classification using the CXR-14 dataset. The primary contribution lies in the design of three novel ViT-integrated models: Residual ViT, Bottleneck ViT, and MBConv-SE ViT. Each model replaces the traditional 3×3 convolutional units within standard CNN building blocks Residual blocks, Bottleneck blocks, and MBConv-SE blocks with Transformer-based attention mechanisms. This integration enables the models to capture both local spatial patterns and long-range dependencies, addressing the limitations of conventional CNNs in modeling global context. The proposed architectures were rigorously evaluated and compared against a wide range of baseline models, grouped under attention-guided, region-guided, and semantic-guided strategies. Experimental results demonstrate that the MBConv-SE ViT model, in particular, achieves superior performance across multiple evaluation metrics, highlighting the effectiveness of combining channel recalibration, efficient convolution, and global self-attention. The comparative analysis also reveals that ViT-enhanced CNN blocks

Power System Technology
ISSN:1000-3673
Received: 06-01-2025 Revised: 15-02-2025 Accepted: 25-03-2025

improve both classification accuracy and generalization in complex multi-label medical image tasks. Overall, the work provides a strong foundation for designing scalable and interpretable ViT-based models in the medical imaging domain. It also emphasizes the potential of hybrid architectures that unify the strengths of CNN and Transformer paradigms for real-world diagnostic applications.

References:

[1] J. E. Schuster and J. V. Williams, "Human Metapneumovirus," *Antibodies Infect. Dis.*, pp. 237–247, Jul. 2023, doi: 10.1128/9781555817411.ch14.

[2] R. Geetha, M. Balasubramanian, and K. R. Devi, "COVIDetection: deep convolutional neural networks-based automatic detection of COVID-19 with chest x-ray images," *Res. Biomed. Eng.*, vol. 38, no. 3, pp. 955–964, Sep. 2022, doi: 10.1007/S42600-022-00230-2/FIGURES/9.

[3] J. P. Mazzoncini, C. B. Crowell, and C. S. Kang, "Human Metapneumovirus: An Emerging Respiratory Pathogen," *J. Emerg. Med.*, vol. 38, no. 4, p. 456, May 2008, doi: 10.1016/J.JEMERMED.2007.11.051.

[4] Q. Philippot *et al.*, "Human metapneumovirus infection is associated with a substantial morbidity and mortality burden in adult inpatients," *Heliyon*, vol. 10, no. 13, p. e33231, Jul. 2024, doi: 10.1016/J.HELIYON.2024.E33231.

[5] M. Epelman, "Chest," *Fundam. Pediatr. Imaging, Third Ed.*, pp. 27–70, Jan. 2022, doi: 10.1016/B978-0-12-822255-3.00003-4.

[6] M. Barile, "Pulmonary Edema: A Pictorial Review of Imaging Manifestations and Current Understanding of Mechanisms of Disease," *Eur. J. Radiol. Open*, vol. 7, p. 100274, Jan. 2020, doi: 10.1016/J.EJRO.2020.100274.

[7] R. Kumar, C. T. Pan, Y. M. Lin, S. Yow-Ling, T. S. Chung, and U. G. S. Janesha, "Enhanced Multi-Model Deep Learning for Rapid and Precise Diagnosis of Pulmonary Diseases Using Chest X-Ray Imaging," *Diagnostics (Basel, Switzerland)*, vol. 15, no. 3, Feb. 2025, doi: 10.3390/DIAGNOSTICS15030248.

[8] M. Shahriar, H. Apu, S. Islam, and T. Taharat Aurpa, "Explainable AI for Sentiment Analysis of Human Metapneumovirus (HMPV) Using XLNet," Feb. 2025, Accessed: Mar. 22, 2025. [Online]. Available: http://arxiv.org/abs/2502.01663

[9] S. Bbouzidi, G. Hcini, I. Jdey, and F. Drira, "Convolutional Neural Networks and Vision Transformers for Fashion MNIST Classification: A Literature Review," Jun. 2024, Accessed: Mar. 22, 2025. [Online]. Available: http://arxiv.org/abs/2406.03478

[10] Y. Shou, T. Meng, W. Ai, C. Xie, H. Liu, and Y. Wang, "Object Detection in Medical Images Based on Hierarchical Transformer and Mask Mechanism," *Comput. Intell. Neurosci.*, vol. 2022, p. 5863782, 2022, doi: 10.1155/2022/5863782.

[11] C. Wang, Y. Jin, J. Liang, Y. H. Yang, S. Nie, and Y. Hai, "Modif-SegUnet: Innovatively

Power System Technology
ISSN:1000-3673
Received: 06-01-2025 Revised: 15-02-2025 Accepted: 25-03-2025

Advancing Liver Cancer Diagnosis and Treatment through Efficient and Meaningful Segmentation of 3D Medical Images," *2023 IEEE Int. Conf. Bioinforma. Biomed.*, pp. 4966–4968, Dec. 2023, doi: 10.1109/BIBM58861.2023.10385862.

[12] L. Tanzi, A. Audisio, G. Cirrincione, A. Aprato, and E. Vezzetti, "Vision Transformer for femur fracture classification," *Injury*, vol. 53, no. 7, pp. 2625–2634, Jul. 2022, doi: 10.1016/J.INJURY.2022.04.013.

[13] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," *2021 IEEE/CVF Int. Conf. Comput. Vis.*, pp. 9992–10002, Oct. 2021, doi: 10.1109/ICCV48922.2021.00986.

[14] A. Liu, Z. Zhu, C. Liu, and Q. Yin, "Automatic Diagnosis of COVID-19 Using a tailored Transformer-Like Network," *J. Phys. Conf. Ser.*, vol. 2010, no. 1, p. 012175, Sep. 2021, doi: 10.1088/1742-6596/2010/1/012175.

[15] J. P. Cohen, P. Morrison, L. Dao, K. Roth, T. Q. Duong, and M. Ghassemi, "COVID-19 Image Data Collection: Prospective predictions are the future," Jun. 21, 2020, *arXiv*.

[16] M. E. H. Chowdhury *et al.*, "Can AI Help in Screening Viral and COVID-19 Pneumonia?," *IEEE Access*, vol. 8, pp. 132665–132676, 2020, doi: 10.1109/ACCESS.2020.3010287.

[17] J. M. Gorriz, Z. Dong, M. Chetoui, and M. A. Akhloufi, "Explainable Vision Transformers and Radiomics for COVID-19 Detection in Chest X-rays," *J. Clin. Med.*, vol. 11, no. 11, p. 3013, Jun. 2022, doi: 10.3390/JCM11113013.

[18] K. S. Krishnan and K. S. Krishnan, "Vision Transformer based COVID-19 Detection using Chest X-rays," *Proc. IEEE Int. Conf. Signal Process. Control*, vol. 2021-October, pp. 644–648, 2021, doi: 10.1109/ISPCC53510.2021.9609375.

[19] J. C. M. Than *et al.*, "Preliminary Study on Patch Sizes in Vision Transformers (ViT) for COVID-19 and Diseased Lungs Classification," *1st Natl. Biomed. Eng. Conf. NBEC 2021*, pp. 146–150, 2021, doi: 10.1109/NBEC53282.2021.9618751.

[20] S. Taslimi, S. Taslimi, N. Fathi, M. Salehi, and M. H. Rohban, "SwinCheX: Multi-label classification on chest X-ray images with transformers," Jun. 2022, Accessed: Mar. 22, 2025. [Online]. Available: https://arxiv.org/abs/2206.04246v1

[21] T. Rahman *et al.*, "Reliable tuberculosis detection using chest X-ray with deep learning, segmentation and visualization," *IEEE Access*, vol. 8, pp. 191586–191601, 2020, doi: 10.1109/ACCESS.2020.3031384.

[22] C. L. Tennakoon, A. L. Kulasekera, R. A. R. C. Gopura, and D. S. Chathuranga, "PneuNet Based Hybrid Soft Gripper for Multi-Shape Object Handling," *IEEE Access*, 2024, doi: 10.1109/ACCESS.2024.3515265.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 770–778, Dec. 2015, doi: 10.48550/arxiv.1512.03385.

Power System Technology
ISSN:1000-3673
Received: 06-01-2025 Revised: 15-02-2025 Accepted: 25-03-2025

[24] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 2818–2826, Dec. 2015, doi: 10.48550/arxiv.1512.00567.

[25] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *36th Int. Conf. Mach. Learn. ICML 2019*, vol. 2019-June, pp. 10691–10700, May 2019, Accessed: Aug. 17, 2023. [Online]. Available: https://arxiv.org/abs/1905.11946v5

[26] Y. K. Kim, J. Matías, D. Martino, and G. Sapiro, "Vision Transformers with Natural Language Semantics," Feb. 2024, Accessed: Mar. 22, 2025. [Online]. Available: https://arxiv.org/abs/2402.17863v1

[27] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 3462–3471, May 2017, doi: 10.1109/CVPR.2017.369.

[28] X. Zhu *et al.*, "PCAN: Pixel-wise classification and attention network for thoracic disease classification and weakly supervised localization," *Comput. Med. Imaging Graph.*, vol. 102, p. 102137, Dec. 2022, doi: 10.1016/J.COMPMEDIMAG.2022.102137.

[29] H. Wang, S. Wang, Z. Qin, Y. Zhang, R. Li, and Y. Xia, "Triple attention learning for classification of 14 thoracic diseases using chest radiography," *Med. Image Anal.*, vol. 67, p. 101846, Jan. 2021, doi: 10.1016/J.MEDIA.2020.101846.

[30] D. Sriker, H. Greenspan, and J. Goldberger, "Class-Based Attention Mechanism for Chest Radiograph Multi-Label Categorization," *Proc. - Int. Symp. Biomed. Imaging*, vol. 2022-March, 2022, doi: 10.1109/ISBI52829.2022.9761667.

[31] Q. Guan, Y. Huang, Y. Luo, P. Liu, M. Xu, and Y. Yang, "Discriminative Feature Learning for Thorax Disease Classification in Chest X-ray Images," *IEEE Trans. Image Process.*, vol. 30, pp. 2476–2487, 2021, doi: 10.1109/TIP.2021.3052711.

[32] V. Teixeira, L. Braz, H. Pedrini, and Z. Dias, "DuaLAnet: Dual Lesion Attention Network for Thoracic Disease Classification in Chest X-Rays," *Int. Conf. Syst. Signals, Image Process.*, vol. 2020-July, pp. 69–74, Jul. 2020, doi: 10.1109/IWSSIP48289.2020.9145037.

[33] B. Chen, Z. Zhang, J. Lin, Y. Chen, and G. Lu, "Two-stream collaborative network for multi-label chest X-ray Image classification with lung segmentation," *Pattern Recognit. Lett.*, vol. 135, pp. 221–227, Jul. 2020, doi: 10.1016/J.PATREC.2020.04.016.

[34] H. G. Jung, W. J. Nam, H. W. Kim, and S. W. Lee, "Weakly supervised thoracic disease localization via disease masks," *Neurocomputing*, vol. 517, pp. 34–43, Jan. 2023, doi: 10.1016/J.NEUCOM.2022.10.019.

Power System Technology
ISSN:1000-3673
Received: 06-01-2025 Revised: 15-02-2025 Accepted: 25-03-2025

[35] R. Hermoza, G. Maicas, J. C. Nascimento, and G. Carneiro, "Region Proposals for Saliency Map Refinement for Weakly-supervised Disease Localisation and Classification," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12266 LNCS, pp. 539–549, May 2020, doi: 10.1007/978-3-030-59725-2_52.

[36] B. Chen, J. Li, G. Lu, and D. Zhang, "Lesion Location Attention Guided Network for Multi-Label Thoracic Disease Classification in Chest X-Rays," *IEEE J. Biomed. Heal. Informatics*, vol. 24, no. 7, pp. 2016–2027, Jul. 2020, doi: 10.1109/JBHI.2019.2952597.

[37] C. Yan, J. Yao, R. Li, Z. Xu, and J. Huang, "Weakly Supervised Deep Learning for Thoracic Disease Classification and Localization on Chest X-rays," *ACM-BCB 2018 - Proc. 2018 ACM Int. Conf. Bioinformatics, Comput. Biol. Heal. Informatics*, vol. 18, pp. 103–110, Jul. 2018, doi: 10.1145/3233547.3233573.

[38] B. Chen, J. Li, G. Lu, H. Yu, and D. Zhang, "Label Co-Occurrence Learning With Graph Convolutional Networks for Multi-Label Chest X-Ray Image Classification," *IEEE J. Biomed. Heal. informatics*, vol. 24, no. 8, pp. 2292–2302, Aug. 2020, doi: 10.1109/JBHI.2020.2967084.

[39] M. Yang, H. Tanaka, and T. Ishida, "Performance improvement in multi-label thoracic abnormality classification of chest X-rays with noisy labels," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 18, no. 1, pp. 181–189, Jan. 2023, doi: 10.1007/S11548-022-02684-2/METRICS.

[40] Ş. Öztürk, M. Y. Turalı, and T. Çukur, "HydraViT: Adaptive multi-branch transformer for multi-label disease classification from Chest X-ray images," *Biomed. Signal Process. Control*, vol. 100, p. 106959, Feb. 2025, doi: 10.1016/J.BSPC.2024.106959.