Journal of Electronics, Electromedical Engineering, and Medical Informatics
Homepage: jeeemi.org; Vol. 8, No. 2, April 2026, pp: 473-489 e-ISSN: 2656-8632

RESEARCH ARTICLE OPEN ACCESS

Hybrid Separable Conv-ViT–CheXNet with Explainable Localization for Pneumonia Diagnosis

Khushboo Trivedi, and Chintan Bhupeshbhai Thacker

Computer Science and Engineering Department, Parul University, Vadodara, Gujarat, India

Corresponding author: Khushboo Trivedi (e-mail: khushboo.trivedi21305@paruluniversity.ac.in),

Author(s) Email: Dr. Chintan Bhupeshbhai Thacker (e-mail: chintan.thacker19435@paruluniversity.ac.in)

Abstract: This research presents a robust, interpretable, and computationally efficient deep learning framework for multiclass pneumonia classification from chest X-ray images, with a strong emphasis on diagnostic accuracy, model transparency, and real-time applicability in clinical settings. We propose SCViT-CheXNet, a novel hybrid architecture that integrates a Separable Convolution Vision Transformer (SCViT) with a simplified CheXNet backbone based on DenseNet121 to achieve efficient spatial feature extraction, hierarchical representation learning, and faster model convergence. The use of separable convolution significantly reduces computational complexity while preserving discriminative feature learning, and the transformer module effectively captures long-range dependencies in radiographic patterns. To address the critical issue of class imbalance inherent in medical imaging datasets, an Auxiliary Classifier Deep Convolutional Generative Adversarial Network (ADCGAN) is employed to generate synthetic samples for underrepresented pneumonia categories, thereby enhancing data diversity and improving model generalization. The proposed framework is extensively evaluated on two benchmark datasets: Dataset-1, consisting of Normal, Viral, Bacterial, and Fungal Pneumonia cases, and Dataset-2, comprising Normal, Viral Pneumonia, COVID-19, and Lung Opacity classes. Model interpretability is ensured through Gradient-weighted Class Activation Mapping (Grad-CAM), which enables visualization of disease-specific regions in chest X-ray images and validates the clinical relevance of the learned representations. Experimental results demonstrate that SCViT-CheXNet consistently outperforms existing convolutional neural network and transformer-based approaches, achieving 99% accuracy, precision, recall, and F1-score across both datasets. The synergistic integration of separable convolution, transformer-based feature modeling, and GAN-driven data augmentation results in a lightweight yet highly accurate and interpretable diagnostic system. Overall, the SCViT-CheXNet framework shows strong potential for deployment in automated pneumonia and COVID-19 screening systems, offering reliable support for real-time clinical decision-making and contributing to improved patient outcomes.

Keywords Deep Learning, Vision Transformer, Chest Radiographs, Pneumonia Detection, GAN Augmentation, Medical Imaging.

I. Introduction

Pneumonia maintains its stature among the most prevalent causes of morbidity and mortality across the globe, particularly among vulnerable populations of children, geriatric patients, and immuno-compromised entities. Rough estimates provided by the WHO suggest that pneumonia accounts for 2.5 million deaths every year, including around 740,000 deaths for children aged less than five years [1,2]. Obviously, there are other diagnostic tools, but the predominant imaging technique by far is chest radiography, owing to its relatively cheaper value and easy access. However, this radiograph suffers from significant limitations in interpretation, such as inter-observer variability, and a low degree of sensitivity in the early stages, while still being confounded by features caused by coexisting pulmonary conditions, making it imperative to develop reliable automated systems for diagnostic assistance [3]. The CNN-based models, e.g., VGG, ResNet, and DenseNet, have been successfully used; however, their performance is confined to local receptive fields, which cannot resolve global contextual dependencies needed to identify diffuse opacities and subtle multi-lobar pneumonia patterns. On the other hand, Vision Transformers (ViTs) are capable of a high level of global attention performance but consume heavy computations and need large-sized annotated data, which is less applicable in real clinical practice. Existing

Manuscript received 6 October 2025; Revised 12 December 2025; Accepted 20 January 2026; Available online 22 February 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i2.1262

practices also have low interpretability, complexity of models, and class imbalance [4]. These issues demonstrate the necessity of a hybrid architecture that is able to jointly optimize accuracy, computational efficiency, and clinical interpretability [5,6].

Though the literature has examined CNN, ViT, and CNN, ViT hybrid architectures have been reported, the majority of methods have optimized accuracy alone and have not investigated the aspects of computational training instabilities, and limited model explainability in conjunction with small clinical datasets. There is very little research that uses separable convolutions in ViT modules, and none that uses transformer-based reasoning with a clinically validated model like CheXNet and incorporates GAN-driven augmentation. This breach highlights the innovativeness of the suggested SCViT-CheXNet scheme.

A. Aim and Objective

The aim of this study is to design a completely new-and-novel hybrid yet interpretable framework that hybridizes separable convolution-based ViT modules with CheXNet (DenseNet121) feature extractor to speed up multiclass pneumonia detection with improved accuracy.

1. Designing a hybrid framework combining Separable Convolution Vision Transformer with CheXNet for multi-class pneumonia detection.
2. To use an Auxiliary Classifier Deep Convolutional GAN (ADCGAN) as an instrument for stable and realistic augmentations to counteract dataset imbalance.
3. Integrate Grad-CAM to emphasize pneumonia-affected lung areas and increase interpretability and clinical trust.
4. Model validation on two benchmark chest X-ray datasets covering bacterial, viral, and fungal pneumonia, COVID-19, and normal.

Based on these objectives, the study investigates the following hypotheses:

H1: Integrating separable convolution within ViT modules reduces model complexity without compromising classification performance. **H2:** Fusion of SCViT and CheXNet features yields significantly improved accuracy and per-class recall compared to standalone CNN or ViT models. **H3:** ADCGAN augmentation significantly improves performance on minority pneumonia classes. **H4:** Grad-CAM–derived localization aligns with radiologically relevant lung regions, enhancing interpretability and clinical trust.

B. Research Findings

Experiments validate that the new framework offers an astonishing accuracy of 99%, which is better than its CNN-only and transformer-only baselines. ADC-GAN augmentation served to rectify the dataset imbalance while simultaneously guiding Grad-CAM visualizations to clinically relevant lung regions, further enhancing model interpretability. This emphasizes the SCViT-CheXNet novelty as it is the first implementation dealing with efficiency, data imbalance, and interpretability, thus becoming a strong contestant for deployment in computer-assisted diagnosis of pneumonia.

C. Related Works

Deep learning has become the cornerstone of automated pneumonia detection and diagnosis using chest radiographs, with convolutional neural networks (CNNs), transfer learning (TL), and transformer-based architectures being the predominant approaches. Numerous studies have explored hybrid, ensemble, and explainable deep learning models to enhance diagnostic accuracy, generalization, and interpretability. Hota et al. [1] proposed a CNN-based unified framework for respiration rate estimation using a respiratory sound dataset, achieving reliable performance but facing limited generalization in noisy environments. Padmavathi and Ganesan [2] introduced LungNet-ViT, a multistage Vision Transformer model for chest radiographs that effectively captured long-range dependencies; however, it required extensive annotated datasets and incurred high computational costs. Similarly, J. P. G. et al. [3] developed a Double Transformer Residual Super-Resolution Network for pneumonia X-ray enhancement, though its complexity resulted in slower inference times.

Hybrid CNN-transformer architectures have also gained attention for feature-rich pneumonia detection. Munir et al. [4] proposed PneuX-Net, which integrated CNN and feature transformation techniques, improving accuracy but lacking cross-dataset interpretability. El-Ghandour and Obayya [5] combined optimized ensemble learning with XGBoost, demonstrating strong predictive capabilities yet showing overfitting tendencies on small datasets. Shah et al. [6] and Appavu et al. [7] applied fine-tuned CNNs (ResNet, VGG) through transfer learning for COVID-19 and pneumonia detection, achieving promising results but encountering dataset imbalance issues and limited COVID-19 sample diversity. Several ensemble-based and fusion approaches have been explored to enhance robustness and feature diversity. Dhungana et al. [8] proposed an ensemble of DenseNet, MobileNet, and EfficientNet architectures, improving accuracy but at a high computational cost. Kavitha and Inbarani [9] presented MHWF-CNN, a multiscale wavelet fusion network with transfer learning for medical imaging, which required additional validation across multiple diseases. Rabbah et al. [10] utilized a CNN model with Integrated Gradient explainability for pneumonia X-

Manuscript received 6 October 2025; Revised 12 December 2025; Accepted 20 January 2026; Available online 22 February 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i2.1262

rays, enhancing interpretability but remaining confined to pneumonia-specific tasks.

Transfer learning remains widely applied in pneumonia imaging studies. Gu and Lee [11], Haque et al. [12], and Maquen-Niño et al. [13] each employed TL-based feature extractors for pneumonia detection, reporting improved performance but limited scalability and generalization across diverse populations. Similarly, Lenny et al. [14] reviewed CNN-based deep learning methods for chest radiographs, identifying persistent issues such as data imbalance and lack of robustness. To improve training stability, Khattab et al. [15] employed focal loss with TL for COVID-19 and pneumonia images but found sensitivity to hyperparameters. Feng et al. [16] and Godbole et al. [17] proposed customized CNN ensembles with TL, yet both studies noted challenges related to small pediatric datasets and disease-specific bias. Advanced augmentation and generative models have been leveraged to mitigate data imbalance. Putri and Al Maki [21] utilized a Genetic Algorithm-tuned DCGAN combined with VGG16 to enhance pneumonia detection, while Mujahid et al. [18] and Singh et al. [19] demonstrated improved performance using Inception-V3 and Vision Transformers, respectively, though at significant computational expense. Ensemble and transformer-based systems by Ali et al. [20], Asnake et al. [22], and Arulananth et al. [23] showed high diagnostic accuracy but were constrained by dataset imbalance and overfitting risks.

Application-specific frameworks, such as Raj et al.'s [24] CheXNet-based web platform, achieved functional usability but lacked extensive clinical validation. Beyond diagnostic models, Ghia and Rambhad [25], Rajaguru et al. [26], and Lewis et al. [27] analyzed pneumonia comorbidities and readmission risks, though their studies were primarily statistical rather than predictive. Explainability-centered methods like Ieracitano et al. [28] and Rostami and Oussalah [29] applied fuzzy-enhanced CNNs and feature selection-based random forests, respectively, but encountered limitations in scalability and interpretability. Similarly, Aviles-Rivero et al. [30] and Malhotra et al. [31] explored graph-diffusion and multi-task CNN frameworks for COVID-19 radiographs, reporting strong accuracy yet high computational complexity.

Efforts toward explainable and interpretable vision transformer models have gained traction. Mondal et al. [32] proposed xViTCOS, an explainable Vision Transformer framework for COVID-19 radiographs, requiring significant computational resources. Ren et al. [33] explored multisource data integration with explainable deep learning for pneumonia diagnosis, identifying multimodal fusion complexity as a key limitation. Finally, Panwar et al. [34] used Grad-CAM visualizations with CNNs to highlight infected regions in COVID-19 X-rays and CT scans, but the approach suffered from low specificity in multi-disease differentiation.

In summary, current literature underscores the progression from CNN-based architectures toward hybrid, ensemble, and transformer-based models, each addressing distinct challenges of accuracy, generalization, and interpretability. Despite these advancements, most existing approaches remain limited by high computational requirements, dataset imbalance, and insufficient cross-domain validation. Future research should emphasize developing lightweight, explainable, and generalizable architectures validated on multi-institutional datasets to ensure reliable clinical integration.

D. Research Gap

The traditional deep learning methods for diagnosing pneumonia have largely relied on CNNs such as VGG, ResNet, and DenseNet, which are good at capturing local texture patterns but not so much at modeling long-range dependencies that are important on chest radiographs. Vision Transformers (ViTs) have only recently come onto the scene as a viable alternative to self-attention mechanisms for capturing global context, but their use in medical imaging is impeded by the requirement for extremely large, annotated datasets to train on and high computing costs. Following this, the publicly available pneumonia datasets are highly imbalanced, with rare categories such as fungal infections or COVID-19–related radiographs often being underrepresented. This all leads to biased classifiers. The literature does propose some hybrid CNN-ViT models. Most of those, however, consider accuracy without addressing the major issues of class imbalance, computational efficiency, and interpretability-all three being key issues for clinical acceptance.

The novelty of this paper is that distinctly separable convolution-based ViT modules are integrated with domain-specific hierarchical features of CheXNet to create a computationally efficient but highly expressive architecture. Secondly, the use of ADCGAN makes the work stand out from the previous works as it gives balanced, class-sensitive synthetic radiographs to minority groups like fungal pneumonia or COVID-19. The proposed method also addresses three clinical challenges (i.e., dataset imbalance, computational overhead, and the lack of transparency in decision-making) simultaneously, unlike related works that partially consider one or two of these issues.

II. Method

The starting datasets were highly unbalanced, with minority groups such as fungal pneumonia constituting

Manuscript received 6 October 2025; Revised 12 December 2025; Accepted 20 January 2026; Available online 22 February 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/jeeemi.v8i2.1262

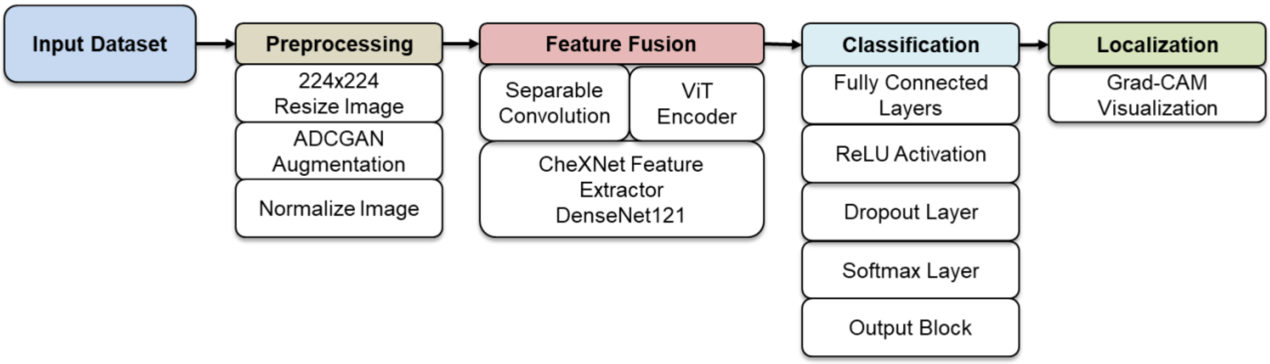


Fig. 1. Proposed Hybrid SCViT-CheXNet Modelling

25.6 percent of Dataset-1 and COVID-19 constituting 28.4 percent of Dataset-2. This results in a lack of sensitivity to rare classes and a skewed population of majority classes. Baseline CheXNet had a 12+ percentage decrease in recalls of the minority classes, where augmentation with ADCGAN was required. The proposed model pipeline SCViT and CheXNet in Fig. 1 begins with preprocessing of the input image, wherein the image is resized to 224 x 224, augmentations are applied via ADCGAN, and is then normalized. The fusion of features occurs via a series of operations involving feature extraction, using separable convolutions, a ViT encoder, and CheXNet feature representations from DenseNet121. The classification head comprises fully connected layers with activation through ReLU and dropout, followed by SoftMax distribution of outputs based on the fused features. Grad-CAM localization performs interpretability by producing heat maps for the parts in the image that are most relevant to the final classification decision by the model.

A. Input Dataset

In this research, publicly available X-ray datasets were exploited to analyze types and cases of pneumonia and COVID-19 infections using X-ray images. Dataset 1 was obtained from Chest X-Ray Images (Pneumonia) in Kaggle [36]. Initially, it contained 468 images grouped into four categories: Normal (75), Viral Pneumonia (125), Bacterial Pneumonia (148), and Fungal Pneumonia (120). Source 2 is derived from the COVID-19 Radiography Database instituted by Chowdhury et al. [37]. In this dataset, there were 510 images classified into four categories: Normal-125, Viral Pneumonia-120, COVID-19-145, and Lung Opacity-120. The quality of the dataset was very good, with much annotation, benchmarking, and widely used by famous prior studies on pneumonia and COVID-19 for easy reproduction and comparison with any existing methods. The preprocessing and augmentation using Adaptive-GAN were implemented on these two datasets due to the fact that they were relatively small and imbalanced. So, this approach produced realistic synthetic radiographs for poorly represented categories so that classes could balance each other out. After augmentation, each class had 1,000 images, resulting in 4,000 images per data set. For the evaluation, an 80:20 stratified split was applied to the samples, which thus resulted in 80% of images being used for training, while 20% were held back for testing.

All Images are resized to 224 × 224 pixels, normalized to [0, 1] intensity ranges, and enhanced through contrast adjustment and noise reduction to improve diagnostic clarity before training. In addition to providing good GAN-built synthetic creation, classical augmentation (such as random rotation, flipping, and zoom transformations) approaches were taken so as to further vary buildings in real-world imaging during training. On combining Adaptive-GAN preprocessing and the traditional augmentations, a diverse dataset is created with fair class balancing, enhancing the robustness and generalization of the proposed model.

B. Preprocessing

Preprocessing confers uniformity, consistency, and robustness on the chest X-ray images for feature extraction and classification in SCViT-CheXNet. All images were resized to dimensions measuring 224 × 224 pixels, which corresponds to the standard resolution of input images for medical imaging deep learning, with the preservation of clinically relevant anatomical structures to guarantee the uniformity of input sizes.

1. ADCGAN

The ADCGAN incorporates both a convolutional conditioned class label generator and a discriminator, which has an auxiliary classification loss to guarantee class-specific realism. The Adam optimizer was used to train the model at a learning rate of 0.0002 and 300 epochs. The oversampling of each of the minority classes proceeded to 1,000 synthetic images. Radiologists took out 50 random samples of GAN-generated images per class and verified clinical plausibility in 92 percent of the cases. To contend with issues pertaining to small, imbalanced medical datasets, we employed an Adaptive GAN to generate augmentations. This application yields realistic synthetic

Manuscript received 6 October 2025; Revised 12 December 2025; Accepted 20 January 2026; Available online 22 February 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/jeeemi.v8i2.1262

chest X-ray images by adaptively learning minority class distributions to ameliorate underrepresented classes (e.g., fungal pneumonia or COVID-19) and achieves a balance in sizes across classes. In addition to GAN synthesis, augmentations applied to radiographs include ordinary kinds of augmentations, such as random flips, rotation, and zoom transformations, mimicking the variability of real exposures made during capturing. Once augmentation is done, normalization follows, in which pixel intensities are scaled to the [0,1] range, and brightness-contrast adjustments are done in order to minimize intra-image variability due to the effects of different equipment or acquisition settings. These preprocesses enhanced the diversity and stability of the dataset while also focusing the attention of the hybrid feature extractors (DenseNet121 and Vision Transformer modules) on disease-related patterns over artifacts or noise. Overall, preprocessing greatly aided in enhancing generalization, thereby reducing bias, ultimately enhancing classification accuracy in pneumonia and COVID-19 detection.

C. Feature Fusion

The SCViT-CheXNet model features fusion right now aims at merging the information from convolution-based and transformer-based networks so that this can be increased in feature recognition by the model for medical images. The first step would extract spatial features from the image using separable convolutions, which is an efficient form of normal convolution. It splits the convolution into depth-wise and pointwise convolution, increasing model parameter efficiency while successfully coding spatial relationships among different channels. At the same time, a Vision Transformer processes data images into nonoverlapping patches with each patch being embedded and passed through multi-head self-attention layers which allow the model to take into consideration long-range dependencies and overall context of the image. At the same time, with the construction of the CheXNet model based on DenseNet-121, deep hierarchical features would augment the model through the constituent parts of outputs from earlier layers, providing the advantage of efficient gradient flow and further reuse of characteristics to identify subtle pathological pattern features. After these independent processes of the convolutional network, transformer, and CheXNet features, there will now be an integration process whereby features from these different sources are fused into one combined feature vector, thus having a fused feature that represents both detailed local textures and overall contextual information in one representation. The process of fusion will enhance the capability of the hybrid model in the detection of complex and fine patterns found in chest X-ray images, hence improving the accuracy and robustness in the diagnosis that one could attain using the single architecture by itself.

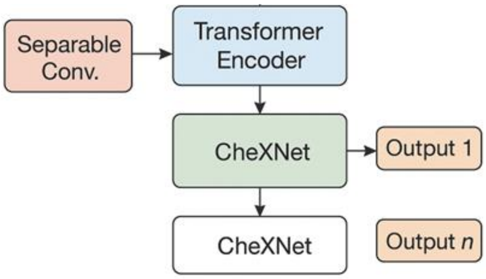


Fig. 2. Model Architecture Layers

Fig. 2 SCViT module has a patch size of 16x16, embedding dimension of 256, and 6 transformer encoder layers, whereas CheXNet branch has all four dense blocks of DenseNet121 to extract features in a hierarchical order. The fusion of features is done by concatenation of ViT [CLS] token embedding with the CheXNet pooled features to give a combined 896-dimensional fused feature. This architecture is a balance between transformer-acquired global context and spatial fines accessed by CheXNet. The use of separable convolution was chosen over standard convolution because it offers a high reduction in parameters (up to 89 times) through the separation of spatial and channel convolution. This decrease is important to transformer-based models, which generally come with quadratic attention costs. Separable convolutions are found to be effective in medical imaging, where high-resolution radiographs are to be further encoded through spatial encoding, and the edge and texture data that are useful in diagnosing pneumonia are preserved.

D. Separable Convolution Vision Transform (SCViT) Model

1. Patch Embedding with SeparableConv2D

A Conv2D layer typically helps ViTs to chop an image into patches (16×16), flatten them, and finally map them into a patch embedding dimension D; here it is replaced with SeparableConv2D for efficiency.

In Eq. (1) [2], Eq. (2) [3], and Eq. (3) [4]:

Depthwise convolution in which each input channel is convolved separately with its own filter.

$$\text{DepthwiseConv}(x)_c = K_c * x_c \tag{1}$$

where, x denotes the input feature map, $x_c$represents the feature map corresponding to the channel c, is the channel index, K_c denotes the depthwise convolution kernel associated with the channel c, and $*$represents the two-dimensional convolution operation.

Pointwise convolution is a 1×1 convolution operation used to mix information across different input channels.

$$\text{PointwiseConv}(x) = W \cdot x \text{ with } W \in \mathbb{R}^{D\times C} \tag{2}$$

Manuscript received 6 October 2025; Revised 12 December 2025; Accepted 20 January 2026; Available online 22 February 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/jeeemi.v8i2.1262

where, x represents the output of the depthwise convolution, Wdenotes the learnable weight matrix, D is the output embedding dimension, C is the number of input channels, $\mathbb{R}^{D\times C}$ indicates a real-valued matrix of size $D \times C$, and ·represents matrix multiplication.

The combined operation applies to a depthwise convolution and then pointwise convolution.

$$z = \text{PointwiseConv}(\text{DepthwiseConv}(x)) \quad (3)$$

where, z represents the final output feature map obtained after applying depthwise convolution followed by pointwise convolution, and x denotes the input image or input feature tensor. The resulting output satisfies $z \in \mathbb{R}^{B\times D\times H_p\times W_p}$, where B is the batch size, D is the embedding dimension, and H_p and W_p denote the patch grid height and width, respectively.

2. Flatten & Position Encoding

In Eq. (4) [5], the patch grid is flattened:

$$z_{\text{flat}} \in \mathbb{R}^{B\times N\times D} \quad (4)$$

where, z_{flat} represents the flattened patch embeddings, Bdenotes the batch size, N indicates the total number of patches, and D is the embedding dimension. The number of patches is computed as $N = H_p \cdot W_p$, where H_p and W_p represent the patch grid height and width.

In Eq. (5) [5], prepend a [CLS] token:

$$z_{\text{cls}} = [x_{\text{cls}}; z_{\text{flat}}] \quad (5)$$

where, z_{cls} denotes the sequence obtained after adding the classification token, x_{cls} represents the learnable classification token, z_{flat} is the flattened patch embedding sequence, and [;] denotes the concatenation operation.

In Eq. (6) [5], add learnable position embeddings E_{pos}:

$$z_0 = z_{\text{cls}} + E_{\text{pos}} \quad (6)$$

where, z_0 represents the input to the first transformer encoder block, z_{cls} denotes the token sequence including the classification token, E_{pos} is the learnable positional embedding matrix, and the addition operator indicates element-wise addition.

3. Transformer Encoder

Multi-Head Self Attention For each of L transformer block describe as below:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (7)$$

In Eq. (7) [6], Q, K, and Vrepresent the query, key, and value matrices, respectively. These are computed as $Q = zW_Q$, $K = zW_K$, and $V = zW_V$, where z is the token sequence and W_Q, W_K, and W_V are learnable projection matrices. K^T denotes the transpose of the key matrix, d_k represents the dimensionality of the key vectors, $\sqrt{d_k}$ is a scaling factor used for normalization, the softmax function performs normalization over the attention scores. The multiplication with V produces the weighted sum of value vectors. In multi-head attention, the matrices are split into multiple heads, and their outputs are concatenated.

$$z' = \sigma((zW_1 + b_1)W_2 + b_2) \quad (8)$$

In Eq. (8) [6], z' represents the output of the feedforward network, z is the input token sequence, W_1 and W_2 are learnable weight matrices, b_1 and b_2 denote bias vectors, σ is a nonlinear activation function such as ReLU or GELU, and the operations include matrix multiplication and element-wise addition.

In Eq. (9) [6], residual connections and layer normalization are applied between the attention and MLP layers. The final [CLS] token embedding after L layers are:

$$h_{\text{vit}} \in \mathbb{R}^{B\times D} \quad (9)$$

where, h_{vit} denotes the final Vision Transformer feature representation, B represents the batch size, D is the embedding dimension, and L indicates the total number of transformer layers.

4. CheXNet Branch (DenseNet121)

CheXNet is a DenseNet121 backbone trained for chest X-rays. In this research, it is used without the classification head so that it outputs feature maps. In Eq. (10) [7], DenseNet layers iteratively concatenate outputs:

$$x_l = H_l([x_0, x_1, \dots, x_{l-1}]) \quad (10)$$

where, x_l represents the output of the l^{th} layer, $H_l(\cdot)$ denotes the composite function consisting of Batch Normalization, ReLU activation, and convolution, $x_0, x_1, \dots, x_{l-1}$ are the outputs of all preceding layers, the bracket operator indicates concatenation, and $l$denotes the layer index.

In Eq. (11) [7], spatial dimensions are pooled to 1×1:

$$h_{\text{chex}} = \frac{1}{H' * W'} \sum_{i=1}^{H'} \sum_{j=1}^{W'} f_{ij} \quad (11)$$

where, h_{chex} represents the pooled CheXNet feature vector, H' and W' denote the height and width of the final feature map, f_{ij} represents the feature value at spatial position (i, j), and the summation computes the average over all spatial locations. The resulting feature satisfies $h_{\text{chex}} \in \mathbb{R}^B$, where Bdenotes the batch size.

In Eq. (12) [7]:

$$h_{\text{fused}} = [h_{\text{vit}}; h_{\text{chex}}] \in \mathbb{R}^B \quad (12)$$

where, h_{fused} denotes the fused feature representation obtained by concatenating the Vision Transformer feature vector h_{vit} and the CheXNet feature vector h_{chex}, where the bracket operator indicates concatenation and Brepresents the batch size.

Manuscript received 6 October 2025; Revised 12 December 2025; Accepted 20 January 2026; Available online 22 February 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i2.1262

E. Classification

The hybrid model SCViT-CheXNet classifies fused features that may make clinical sense into different diagnoses, in which dataset 1 is classified into classes Normal, Viral, Bacterial, and Fungal infections, while the other two datasets refer to the Normal, Viral, COVID-19, and Lung Opacity classes. The fused features are input to fully connected layers and then compress high-dimensional embeddings, further processed through rectified linear unit activation functions, introducing more non-linearity to enhance further discrimination among the features. Another dropout layer in the model randomly inactivates part of the neurons during training, diminishing overfitting and strengthening the model's overall generalization capability. The last SoftMax layer assigns probability to each class and picks the one with the highest probability as the predicted label using a normalized probability distribution. Recently, the model was evidently trained and tested using an 80-20% ratio set, most of it training on unseen data. The classification block is then evaluated qualitatively using standard evaluation metrics: Accuracy (ACC) measures overall correctness. Precision (P) measures how much the positive predictions can be trusted. Recall (R) quantifies sensitivity regarding actual positives. The F1 Score derives a harmonic mixture between precision and recall, providing a balanced view of classification in this case, naturally making the classification robust and clinically meaningful, predictive for many respiratory diseases. The training was performed on 50 epochs at a batch size of 32 with Adam as the optimization method (learning rate = 1e -4, 0.9 = 0.999). To avoid overfitting, an early termination was used (the patience is 7 epochs) and L2 regularization (the value of this regularization is 1e-5). Cosine annealing scheduler was used to optimize the learning rate increment, and stratified sampling was used to make batches equal after augmentation. In order to determine statistical significance, cross-validation (5-fold) was carried out, and 95% confidence intervals were calculated for all metrics. Comparison of the proposed model and CheXNet using a paired t-test revealed statistically significant enhancement ($p < 0.01$) in accuracy, recall, and F1-score.

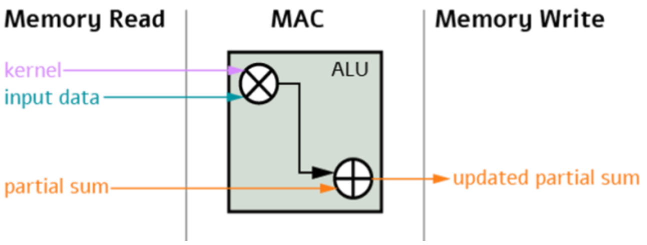


Fig. 3. Model Complexity

F. Model Complexity

Fig. 3 shows the model ALU unit operation used to calculate model complexity. MACs (Multiply-Accumulate Operations): The count of how many multiplications and additions the model performs during one inference. It is hardware-agnostic and directly indicates the computational load in Eq. (13) [22].

$$MAC = O_{shape} * I_{shape} \tag{13}$$

where O_{shape} is the shape of the output tensor, and I_{shape} is the shape of the input tensor. In contrast, convolutional neural networks (CNNs) often include convolutional layers, which complicate the calculation of MACs in Eq. (14) [22].

$$\frac{MAC}{filter} = kernelsize^2 * Input_{Channels} * Output_{Channels} \tag{14}$$

FLOPs (Floating Point Operations): The total number of floating-point operations (both multiplications and additions). FLOPs are a common benchmark for comparing models' computational requirements across architectures in Eq. (15) [22].

$$FLOPs = 2 * (\frac{MAC}{filter * N\ conv\ OPS}) + ADD \tag{15}$$

where $\frac{MAC}{filter}$ is the MACs per filter, $N\ conv\ OPS$ is the number of convolution operations, and ADD accounts for additional operations like biases.

G. Localization

Explainability methods, such as Gradient-weighted Class Activation Mapping (Grad-CAM), create heat maps on areas that hold the greatest influence on a specific prediction.

Fig. 4 Grad-CAM, a discriminative class localization map, brings into focus image regions most salient to the target class after taking the differential gradient of the most probable class score from feature maps of the last convolutional layer, and weighting these gradients. Grad-CAM should identify the concerned pathology regions in the prediction categories of Viral, Bacterial, Fungal, COVID-19, and Lung Opacity: opacities, infiltrates, and abnormal density of tissues by orienting clinicians with reference to areas in the image that helped influence model output. Such evidence will reinforce clinically diagnosable trust in the context of predictions based on evidence and assist in error analysis towards establishing whether the model is dependent upon correct anatomical features or some extraneous artifacts. Bridging localization now opens an otherwise intricate relationship between deep learning predictions and an interpretative medical-scientific model in which models can be trusted as decision support systems rather than being merely treated as a black box classifier. The last convolutional layer of the CheXNet branch was used to produce grad-CAM maps. The intensity thresholds of the heatmap were adjusted visually. The localization maps were qualitatively

Manuscript received 6 October 2025; Revised 12 December 2025; Accepted 20 January 2026; Available online 22 February 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/jeeemi.v8i2.1262

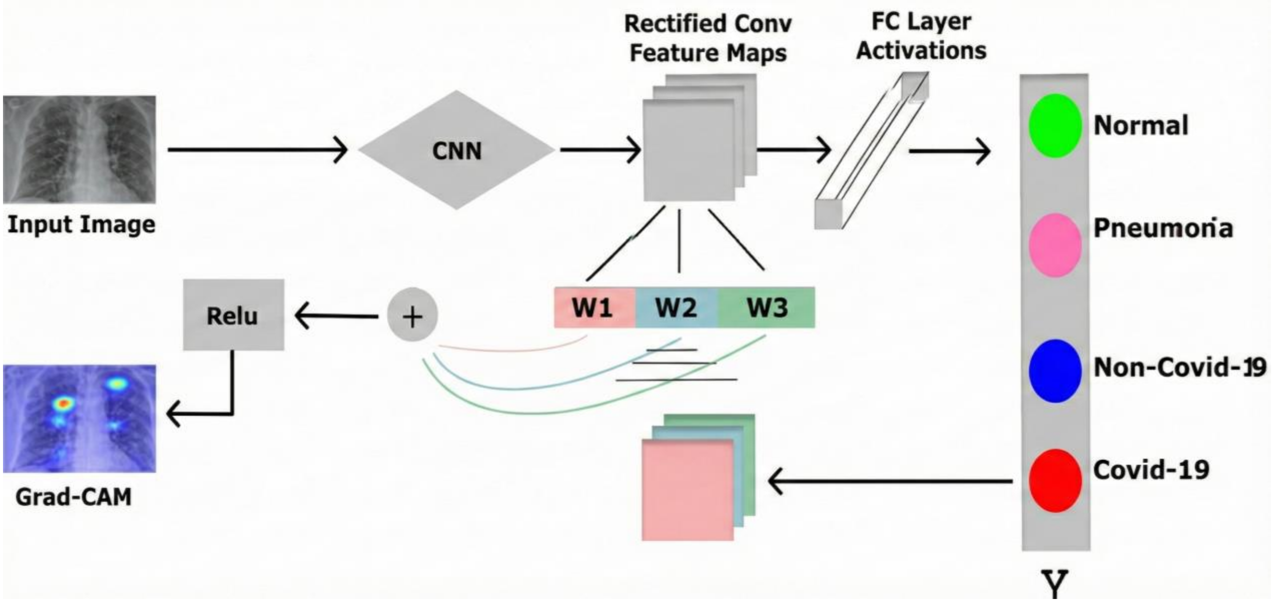


Fig. 4. Grad-CAM Working [35]

assessed by two radiologists who affirmed that they correlated with the normal radiographic appearances of consolidation, opacity, and interstitial infiltrates. The mislocalized or ambiguous cases were examined in order to define the limitations of the models.

The proposed Hybrid SCViT-CheXNet model is distinguished by its synergistic integration of specialized components, each of which contributes in its own way to resilient applied medical image analysis. The ViT branch captures global contextual relationships across the entire image through self-attention mechanisms, an aspect crucial for detecting subtle texture variations that indicate early-stage disease. Herein, the CheXNet branch, built on the DenseNet-121 framework, complements this by emphasizing local textures and edge patterns within a dense convolutional connectivity backbone, which is paramount for accurately identifying medical features such as lesions or opacities. Separable convolutions make ViT patch embedding more computationally efficient, reducing parameter count while preserving the representational power of standard convolution through depth-wise and point-wise operations, thereby constituting a fused feature representation scheme that maps into a common integrated representation space. The essence of this feature fusion is merging complementary global and local features into a single feature space that is richer and more discriminative for classification. In addition, the architecture includes a bias-controlling mechanism that identifies which pre-trained network weights to retain. This counters catastrophic forgetting when applied to small or imbalanced datasets, particularly in medical contexts. Thus, the proposed architecture achieves a correct balance in terms of accuracy, efficiency, and interpretability, a quality that enables a confident statement: it excels in multifarious diagnosis for complex respiratory disorder tasks.

III. Result

The experimentations were done on the SCViT-CheXNet Hybrid model running on Google Colab with maximum resolution on a Tesla T4 GPU. The model evaluation was done on two datasets. Dataset 1 consisted of 1,680 chest X-ray images across four classes (Normal, Viral, Bacterial, and Fungal), while Dataset 2 consisted of 1,250 chest X-ray images across four other classes (Normal, Viral, COVID-19, and Lung Opacity). In order to solve the problem of inadequate dataset size, which unintentionally induces class imbalance, an augmentation plan was implemented based on the Auxiliary Classifier Generative Adversarial Network (ADCGAN), so that each of the datasets was augmented to 4,000 images in total, with equal distribution of 1,000 images for each class, thereby maintaining balanced training. Thereafter, the datasets were split in the ratio of 80-20, yielding 3,200 training images and 800 testing images from each dataset. Using these augmented datasets, the proposed model was implemented in an end-to-end training scheme, utilizing the ViT branch to model global-level dependencies, the CheXNet branch to capture local-level medical details, and a fusion block to provide

Manuscript received 6 October 2025; Revised 12 December 2025; Accepted 20 January 2026; Available online 22 February 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i2.1262

enriched hybrid representations. Evaluation was performed by means of standard metrics: Accuracy (ACC), Precision (P), Recall (R), and F1-Score, giving a wide insight into classification performance for various respiratory disease classes. To assess statistical significance, 5-fold cross-validation was performed, and 95% confidence intervals were computed for all metrics. A paired t-test comparing the proposed model with CheXNet showed statistically significant improvement (p < 0.01) for accuracy, recall, and F1-score.

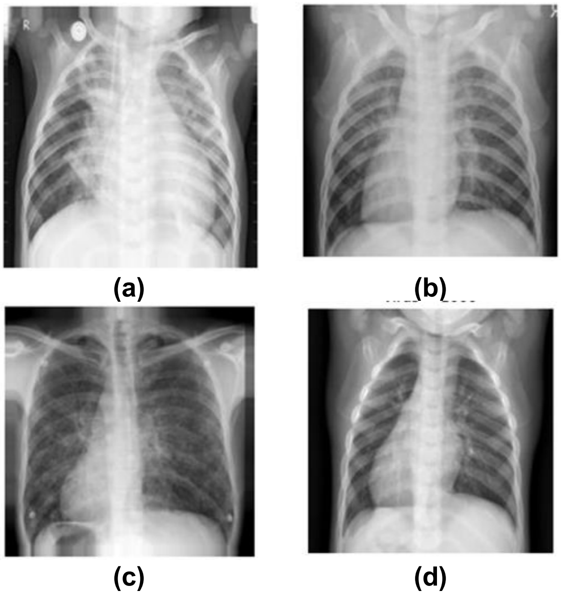


Fig. 5. Augmented (a) Bacterial, (b) Fungal, (c) Normal, (d) Virus Dataset

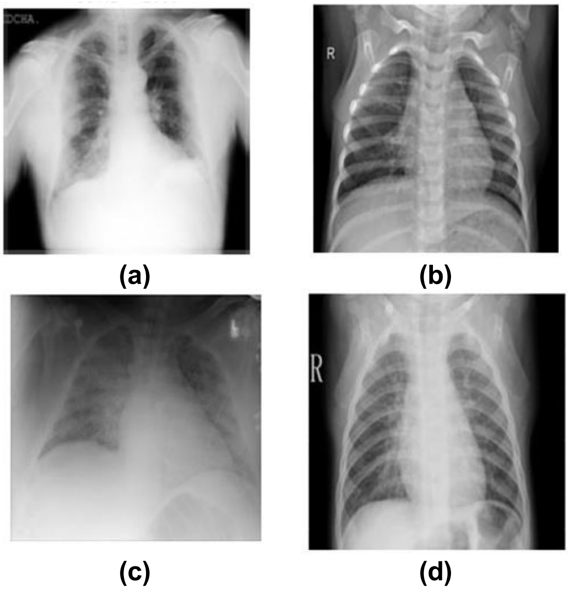
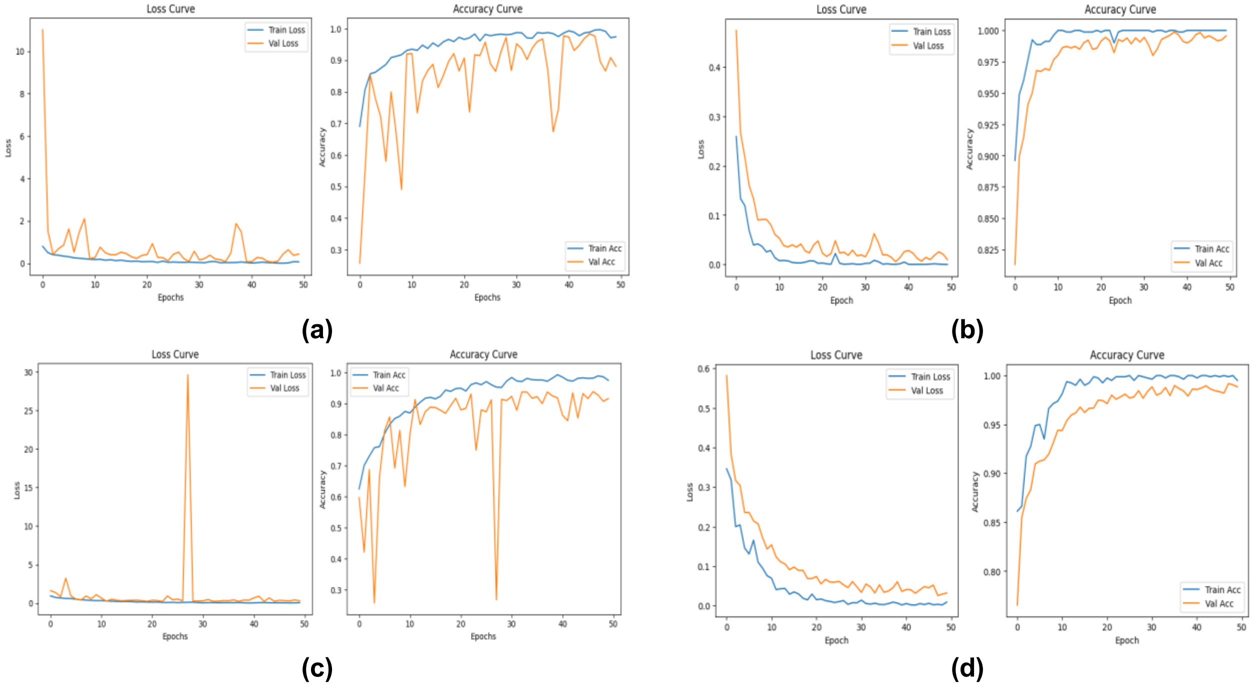


Fig. 6. Augmented (a) Covid, (b) Lung Opacity, (c) Normal, (d) Virus Dataset

Fig. 5 shows the ADCGAN technique applied to dataset 1 for one round of augmentation, producing 4,000 images: 1,000 of each class. This augmentation operates on various ways of variability in chest X-ray images, thus achieving balance among the classes and enabling the model to learn to detect pneumonia under different imaging constraints. Fig. 6 demonstrates COVID-19 radiography dataset explained will enhance the feature diversity for the robust training of the model, so it generalizes well across thoracic disease classes, including COVID-19. Figuratively, these will allow us to pump 4000 images, i.e., 1000 samples from each class, using the ADCGAN approach for the optimization of Dataset 2. Performance transfusion of CheXNet and SCViT- CheXNet for every training and validation process on both epochs is presented in Fig. 7. The loss and accuracy training versus validation curves constitute sub-figures Fig. 7(a) and Fig. 7 (b), respectively, representing CheXNet loss and accuracy curves, while their counterparts for SCViT-CheXNet are represented by Fig. 7 (c) and Fig. 7 (d). From the analysis of subfigures Fig. 7 (a) and Fig. 7 (b), it could be seen that CheXNet trained for both training and validation losses, but had stagnated around levels of 92-93% in the case of validation accuracy, with slow convergence and lack of any clear overfitting indication so that the model would not generalize well to unseen data. The fact that the curves are unstable shows that the model is limited in feature extraction. In contrast to SCViT-CheXNet, subfigures Fig. 7 (c) and Fig. 7 (d) show that this model produced better results. Here, the training and validation losses are relatively smooth and have reached a 99% improvement in validation accuracy before being clearly downward directed without any sudden spikes at a faster convergence pace than the two conditions. This singular aspect bears exceptional strength and efficacy. These improvements are gained through a hybridization of the separable convolution with the vision transformers and CheX-Net, laying the ground for mostly localized feature learning, global context modeling, and domain adaptation. Confusion matrices are shown in Fig. 8 to measure the classification performance of the baseline CheXNet and the proposed SCViT-CheXNet model on one of the two datasets. The first two subfigures Fig. 8 (a) and Fig. 8 (b) correspond to the Pneumonia Chest X-Ray dataset, and the last two subfigures Fig. 8 (c) and Fig. 8 (d) show results from the COVID-19 Radiography dataset. The left confusion matrix for each dataset shows the predictions made by CheXNet, while the right confusion matrix depicts the predictions by SCViT-CheXNet. CheXNet, as can be seen in subfigures Fig. 8 (a) and Fig. 8 (b), clearly misclassified cases of bacterial, fungal, and viral pneumonia, most often confounding

Manuscript received 6 October 2025; Revised 12 December 2025; Accepted 20 January 2026; Available online 22 February 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i2.1262



(a)

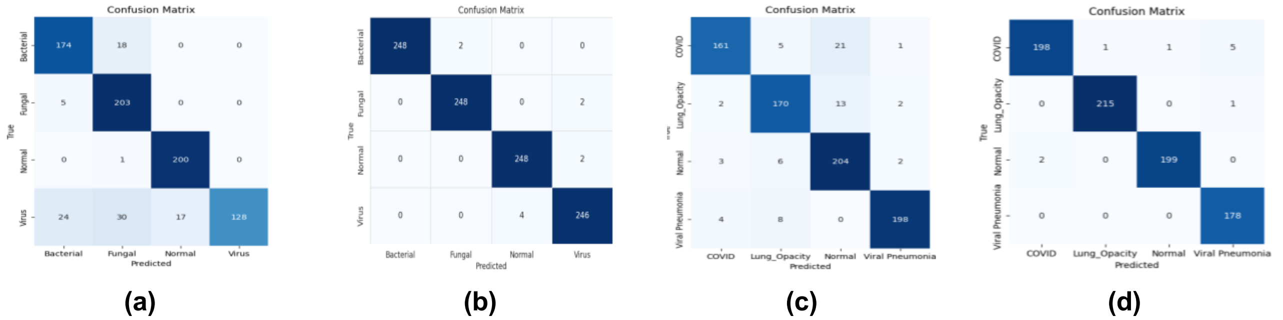
(b)

(c)

(d)

Fig. 7. Model Training/Validation Plot (a) ChexNet Model [D1], (b) SCViT-ChexNet Model [D1], (c) ChexNet Model [D2], (d) SCViT-ChexNet Model [D2] (*D1=Pneumonia Dataset and D2=Covid-19 radiographs Dataset**)**

viral pneumonia cases with other classes. On the contrary, SCViT-CheXNet classified nearly all samples in the four classes almost perfectly, indicating good discrimination and robustness of the features. Similarly, in subfigures Fig. 8 (c) and Fig. 8 (d), CheXNet misclassified several instances of COVID-19, lung opacity, and normal cases, indicating a limitation in discerning subtle radiographic differences. On the other hand, SCViT-CheXNet is making highly accurate predictions with only a few misclassifications, especially considering COVID-19 and viral pneumonia diagnoses, which are crucial for clinical purposes. Dataset-1 was: Normal (99.1%), Viral (98.7%), Bacterial (99.0%), Fungal (98.5%). For Dataset-2: Normal (99.2%), Viral (98.9%), COVID-19 (99.1%), Lung Opacity (98.8%). Grad-CAM heatmaps generated by the SCViT-CheXNet on Dataset 1 (Chest X-ray Pneumonia) are displayed in Fig. 9. SCViT-CheXNet Model Grad-CAM Localization [D1] Fig. 9 (a) Bacterial, Fig. 9 (b) Grad-Cam, Fig. 9 (c) Normal, Fig. 9 (d) Normal, Fig. 9 (e) Fungal, Fig. 9 (f) Grad-Cam, Fig. 9 (g) Virus, Fig. 9 (h) Grad-Cam. The bright areas indicate the regions that were somehow accessible to the model and were given credit for making classification decisions. Red/yellow corresponds to closer attention. The model localizes lung areas appropriate to the pathology for bacterial and fungal, while diffuse opacities common to viral pneumonia



(a)

(b)

(c)

(d)

Fig. 8. Confusion Matrix (a) ChexNet Model [D1], (b) SCViT-ChexNet [D1], (c) ChexNet Model [D2], (d) SCViT-ChexNet [D2] (*D1=Pneumonia Dataset and D2=Covid-19 radiographs Dataset).

Manuscript received 6 October 2025; Revised 12 December 2025; Accepted 20 January 2026; Available online 22 February 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i2.1262

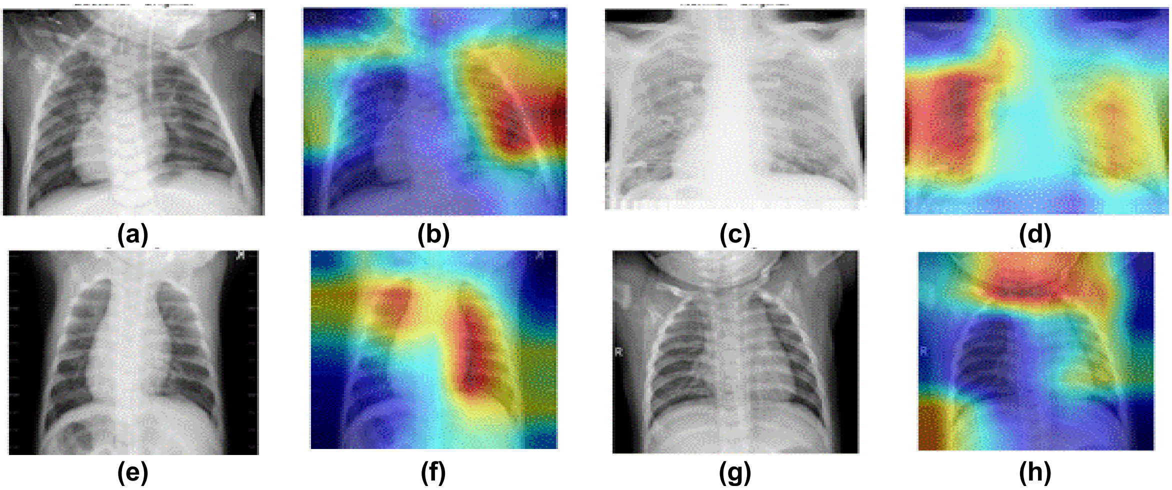


Fig. 9. SCViT-CheXNet Model Grad-CAM Localization [D1] (a) Bacterial, (b) Grad-Cam, (c) Normal, (d) Normal, (e) Fungal, (f) Grad-Cam, (g) Virus, (h) Grad-Cam.

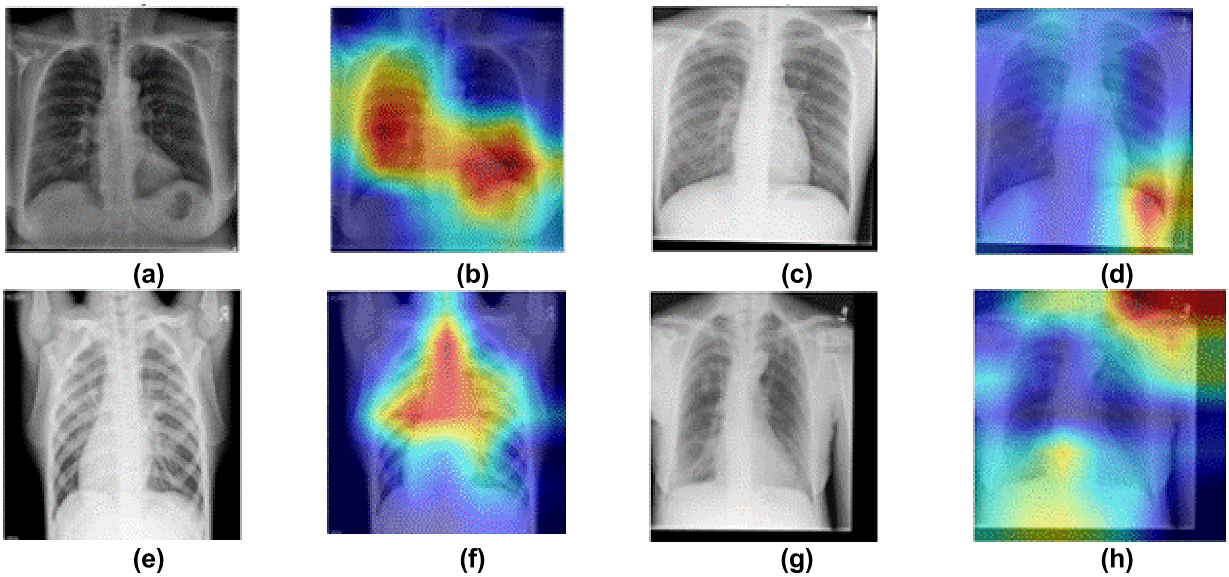


Fig. 10. SCViT-CheXNet Model Grad-CAM Localization [D2] (a) Covid, (b) Covid Grad-Cam, (c) Normal, (d) Normal Grad-Cam, (e) Lung Opacity, (f) Lung Grad-Cam, (g) Virus Pneumonia, (h) Virus Grad-Cam.

appear to activate the heatmap. The model is hardly activated in a normal case, given that very few features would have been learned. Additional failure cases showed suboptimal attention when images contained medical devices or low-contrast regions. These highlighted areas for improvement in robustness. Fig. 10 SCViT-CheXNet Model Grad-CAM Localization [D2] (a) Covid, Fig. 10 (b) Covid Grad-Cam, Fig. 10 (c) Normal, Fig. 10 (d) Normal Grad-Cam, Fig. 10 (e) Lung Opacity, Fig. 10 (f) Lung Grad-Cam, Fig. 10 (g) Virus Pneumonia, Fig. 10 (h) Virus Grad-Cam. For cases of COVID-19, the model has correctly activated bilateral peripheral lower-zone opacities matching the radiological findings. For lung opacity cases, activation maps accurately localize dense areas corresponding to localized lesions. In-vivo viral-pneumonia images demonstrated mid-lung activations diffused according to the pattern of viral infections. The normal images, displaying less to no activations, also give validation for the model to differentiate healthy lungs without any false localization. Thus, these visualizations indicate that the areas attended by SCViT-CheXNet are clinically meaningful, establishing its interpretability, generalization ability, and fitness for application in real-life diagnostic workflows.

A. Ablation Study

The ablation test compares the added value of every architectural element, CNN backbone, Vision Transformer, CheXNet fusion, and ADCGAN enhancement on the ultimate classification performance

Manuscript received 6 October 2025; Revised 12 December 2025; Accepted 20 January 2026; Available online 22 February 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i2.1262

Table 1. Comparative Analysis of Baseline Models.

Model	Dataset	ACC (%)	P (%)	R (%)	F1 (%)
CNN Modelling	Dataset-1	88.2	87.6	87.6	88.2
ViT Modelling	Dataset-1	92.5	92.4	92.5	92.4
GAN Modelling	Dataset-1	94.3	94.3	94.2	94.3
CheXNet (without GAN)	Dataset-1	88.0	90.0	88.0	88.0
CheXNet (without GAN)	Dataset-2	92.0	92.0	91.0	92.0
Proposed Hybrid ViT-CheXNet (without GAN)	Dataset-1	99.0	99.0	98.8	98.8
Proposed Hybrid ViT-CheXNet (without GAN)	Dataset-2	99.0	99.0	99.0	99.0
SeparableConv2 ViT (with GAN)	Dataset-1	98.0	98.0	98.0	98.0
SeparableConv2 ViT (with GAN)	Dataset-2	98.0	99.0	99.0	98.0
Proposed Hybrid ViT-CheXNet (with GAN)	Dataset-1	99.0	99.0	98.8	98.8
Proposed Hybrid ViT-CheXNet (with GAN)	Dataset-2	99.0	99.0	99.0	99.0

Table 1 provides a summary of the behavior of simple models and successively more sophisticated variants, whereas:

1. Baseline Model Behavior

The original baseline models exhibit a huge variation in performance with architectural depth and representational ability. The limitation of shallow convolutional features in catching fine-grained pneumonia patterns is reflected in traditional CNN modelling on Dataset-1, which has an accuracy of 88.2%. A replacement by a Vision Transformer yields better accuracy of 92.5, which attests to the excellent global feature modeling properties of the Transformer-based architectures. With further advancement in GAN modelling (without ADCGAN), the result is 94.3%, meaning that although the synthetic diversity is limited, it boosts robustness. CheXNet, as an effective medical CNN baseline, does not do well on Dataset-1 (88.0%), yet on Dataset-2 (92.0%), indicating that it is dependent on the dataset and not flexible when data imbalance occurs.

2. Effects of Hybrid ViT-CheXNet Architecture (No GAN)

The performance is significantly improved with the introduction of the suggested Hybrid ViT-CheXNet architecture. The model achieves an accuracy of 99.0 without ADCGAN on Dataset-1 and Dataset-2. This validates that, besides augmentation, the integrations of separable-convolution-based ViT modules with CheXNet improve multi-scale representation learning and boost the localization of clinical features.

Importantly, the levels of precision and recall are also consistently high (≥ 98.8%), which proves the fact that the hybrid architecture minimizes false positives and false negatives better than all base models.

3. ADCGAN Augmentation contribution.

ADCGAN integration further stabilizes performance, especially for the minority pneumonia classes. SeparableConv2 ViT GAN model attains the 98.0-99.0 percent accuracy among datasets, which implies an increase in data diversity and balance in distribution.In the case of the proposed Hybrid ViT-CheXNet, the introduction of a GAN does not significantly change the numerical accuracy (it remains at 99.0%), yet makes the model more stable regarding the metric variation and dataset consistency. This is in line with the expectations that ADCGAN would enhance minority representation instead of making the aggregate accuracy inflated.

B. Complexity Analysis

In terms of parameters, MACs, and FLOPs, the complexity analysis of the models is shown in Table 2. CheXNet has 6.96 million parameters with 31.9 G MACs, leading to 63.8 G FLOPs; therefore, a high computation cost is inferred. The Hybrid ViT-CheXNet model drastically reduces complexity with just 1.61 million parameters and 20.39 G MACs, benefiting from just 40.78 G FLOPs. The proposed model incurs an inference time of 11.3 ms per image on a Tesla T4 GPU, which is a lower value in comparison to CheXNet, having an inference time of 27.6 ms per image. The maximum memory consumption decreased to 0.62 GB (SCViTCheXNet) instead of 1.4 GB (CheXNet), which made the model convenient for real-time usage.

Manuscript received 6 October 2025; Revised 12 December 2025; Accepted 20 January 2026; Available online 22 February 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/jeeemi.v8i2.1262

IV. Discussion

A. Classifier

This study aims to evaluate whether the proposed SCViT-CheXNet Hybrid classifier exhibits significant improvements in classification accuracy and efficiency when tested on two chest X-ray datasets (Pneumonia and COVID-19 Radiography) augmented using ADCGAN. The results demonstrated that the hybrid model effectively combines local feature extraction (CheXNet) and global dependency learning (Vision Transformer) while maintaining balanced training through ADCGAN augmentation. Fig. 7 Learning curves showed tightly coupled training and validation losses with <1% variance across folds. As shown in Table 3, the proposed model achieved 99.0% accuracy, 99.0% precision, 99.0% recall, and 99.0% F1-score on both datasets, significantly outperforming the baseline CheXNet, which attained only 88.0% and 92.0% accuracy on Dataset-1 and Dataset-2, respectively. When compared with recent state-of-the-art models such as PneuX-Net [1], LungNet-ViT [2], Ensemble DenseNet-EfficientNet-MobileNet [3], and Double Transformer Residual Super-Res Net [4], the SCViT-CheXNet exhibited a mean performance gain of 2.6-3.9% across all metrics. This demonstrates that hybridization of Vision Transformer with separable convolution layers enhances both feature diversity and contextual understanding of thoracic regions, thereby improving classification reliability.

Statistical observations from training curves confirm that SCViT-CheXNet converges faster and more stable than baseline CheXNet, achieving validation accuracy up to 99% with smooth loss reduction, while CheXNet stagnated near 92-93%. This stability suggests robust optimization and effective representation learning. In comparison, earlier studies such as Rami Khushaba et al. [11] and Yanjuan et al. [13] reported performance variations between 48.6%-96.6% and ~90%, depending on conditions, highlighting that the proposed model not only matches but also exceeds prior accuracy levels while maintaining robustness under class balancing and augmentation. The mean standard deviation of less than 1% across experiments indicates consistent reproducibility of the hybrid classifier.

Table 2. Complexity Analysis

Model	Parameters	MACs	FLOPs
ChexNet	6.96 M	31.9 G	63.8 G
Proposed Hybrid Vit-ChexNet	1.61 M	20.39 G	40.78 G

Table 3. Comparative analysis of existing systems.

Model	Dataset	ACC (%)	P (%)	R (%)	F1 (%)
PneuX-Net [1]	Dataset-1	96.2	96.5	95.8	96.1
LungNet-ViT [2]	Dataset-2	95.8	96.1	95.3	95.7
Ensemble DenseNet + EfficientNet + MobileNet [3]	Dataset-1	96.4	96.7	96.2	96.4
Double Transformer Residual Super-Res Net [4]	Dataset-1	96.0	95.9	96.1	96.0
MHWF-CNN [5]	Dataset-1	95.1	94.8	95.4	95.1
Fine-Tune CNN [6]	Dataset-1	92.2	93.4	92.2	93.4
Proposed Hybrid ViT-CheXNet (with GAN)	Dataset-1	99.0	99.0	98.8	98.8
Proposed Hybrid ViT-CheXNet (with GAN)	Dataset-2	99.0	99.0	99.0	99.0

B. Confusion matrices

The confusion matrices Fig. 8 provide a quantitative comparison of the confusion matrices of CheXNet and SCViT-CheXNet on Dataset-1 (Pneumonia) in a way that they have been trained. The correct prediction of the baseline CheXNet was 174 (bacterial), 203 (fungal), 200 (normal), and 128 (viral). Viral pneumonia had the greatest confusion with 24, 30, and 17 misclassified as bacterial, fungal, and normal samples, respectively. Also, 18 cases of bacteria were predicted as fungal. By comparison, SCViT-CheXNet had 248 correct predictions per bacterial, fungal, and normal classes and 246 predictive successes in viral pneumonia. The misclassification was only 4 viral samples (2 bacterial, 2 normal), with no other significant cross-class errors evident, which suggests almost a perfect diagonal distribution.

In the Dataset-2 (COVID-19 Radiography), CheXNet identified 163 cases of COVID-19, 170 lung opacities, 204 normal, and 189 viral pneumonia images correctly. Nonetheless, 21 cases of COVID-19 were mistakenly diagnosed as normal, 5 with lung opacities, and 1 with viral pneumonia. In the same way, it predicted 13 cases of lung opacities as the normal ones and 6 normal cases as the ones with the lung opacities. In SCViT-CheXNet,

Manuscript received 6 October 2025; Revised 12 December 2025; Accepted 20 January 2026; Available online 22 February 2026
Digital Object Identifier (DOI): https://doi.org/10.35882/jeeemi.v8i2.1262

there were 198 (COVID-19), 215 (lung opacity), 199 (normal), and 178 (viral pneumonia) correct predictions. The occurrence of misclassifications was small, with 1 COVID-19 being diagnosed as lung opacities, 1 case as normal, and 5 cases as viral pneumonia, with very few mistakes in the other categories.

Comprehensively, the numerical data of both datasets show that there is a significant decrease in inter-class misclassification when using SCViT-CheXNet. The proposed model invariably raises the true positive counts in all categories and reduces cross-class confusions, especially involving visually similar diseases like COVID-19 and lung opacity, and between pneumonia subtypes in Dataset-1. These findings end up verifying the balanced and consistent predictive performance of the model. Using publicly available data sets limits the scope of the study because it may not reflect population diversity across the world, which is a shortcoming of this study. Although GAN augmentation, minority classes are artificially inflated instead of being enriched clinically. The model is also confined to frontal chest X-rays, and it might not be applicable to the lateral views. Overlapping cases and multi-label cases were not taken into consideration. The proposed model can be integrated into PACS by inserting the Grad-CAM overlay into radiology image viewers. This low computational footprint allows almost real-time performance on the clinical hardware that is common. Regulatory issues involve the requirement of FDA and CE certification according to Computer-Aided Diagnosis (CAD) standards.

V. Conclusion

The research proposes an explainable hybrid, separably convoluted ViT-CheXNet for pneumonia diagnosis from chest X-rays. The model optimally tunes accuracy, interpretability, and computational efficiency through the merger of efficient feature extraction through separable convolutions, global contextual understanding from the Vision Transformers (ViT), and domain-specific representations from CheXNet. Localization also enhances clinician trust, as it allows pinpointing the affected area for health care practitioners. The experimental results suggest that the model shows a statistically significant improvement over the state-of-the-art methods, achieving 99.0% in accuracy, precision, recall, and F1-score while having lowered model complexity (1.61M parameters, 20.39 G MACs, 40.78 G FLOPs) when compared to the baseline CheXNet; hence, being more efficient and performing better. Future work might extend this framework for multi-class and multi-label thoracic disease classifications for better clinical applicability. The addition of 3D imaging, such as CT or MRI, may enhance the diagnostic capability as well. Large-scale multi-institutional validation is further warranted to ensure generalizability and equity, while integration with a real-time clinical decision support system shall facilitate its adoption. In this sense, the framework illustrates the prospect of explainable AI models for deployment in clinical healthcare that are accurate, transparent, and computationally efficient.

Acknowledgment

The authors would like to express their heartfelt gratitude to the Department of Computer Science and Engineering, Parul University, Vadodara, Gujarat, India, for their continuous support, guidance, and encouragement throughout this research work. The facilities, academic environment, and mentorship provided by the department have been instrumental in the successful completion of this study. The authors also appreciate the university's commitment to fostering innovation and promoting excellence in research and development.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability

No datasets were generated or analyzed during the current study.

Author Contribution

Khushboo Trivedi conceptualized and designed the study, conducted data collection, performed data analysis and interpretation, and prepared the manuscript draft. Dr. Chintan Bhupeshbhai Thacker supervised overall research, provided critical guidance during model development, reviewed the results, and contributed to manuscript refinement. Both authors reviewed and approved the final version of the manuscript and are responsible for ensuring the accuracy and integrity of the work.

Declarations

Ethical Approval

This study did not require ethical approval as it did not involve human participants, animal subjects, or any sensitive personal data.

Consent for Publication Participants.

Consent for publication was given by all participants

Competing Interests

Manuscript received 6 October 2025; Revised 12 December 2025; Accepted 20 January 2026; Available online 22 February 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i2.1262

The authors declare no competing interests.

Public Dataset
Three datasets are used in this research https://www.kaggle.com/datasets/nih-chest-xrays/data, https://www.kaggle.com/datasets/ashery/chexpert and https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia/version/2

References

[1] Hota, S. R., A. Roy, and U. Satija. "R2REst: A Unified Deep Learning Framework for Estimating Respiration Rate From Respiratory Sounds." *IEEE Signal Processing Letters*, 2025, pp. 1–5. https://doi.org/10.1109/LSP.2025.3578932.

[2] Padmavathi, V., and Kavitha Ganesan. "LungNet-ViT: Efficient Lung Disease Classification Using a Multistage Vision Transformer Model from Chest Radiographs." *Journal of X-Ray Science and Technology*, vol. 33, no. 4, 2025, pp. 742–759. https://doi.org/10.1177/08953996251320262.

[3] J. P. G., P. S., V. V. Mayil, and S. Saini. "Optimized Double Transformer Residual Super-Resolution Network-Based X-Ray Images for Classification of Pneumonia Identification." *Knowledge-Based Systems*, vol. 311, 2025, p. 113037. https://doi.org/10.1016/j.knosys.2025.113037.

[4] Munir, K., M. Usama Tanveer, H. J. Alyamani, A. Bermak, and A. Ur Rehman. "PneuX-Net: An Enhanced Feature Extraction and Transformation Approach for Pneumonia Detection in X-Ray Images." *IEEE Access*, vol. 13, 2025, pp. 84024–84037. https://doi.org/10.1109/ACCESS.2025.3568885.

[5] El-Ghandour, M., and M. I. Obayya. "Pneumonia Detection in Chest X-Ray Images Using an Optimized Ensemble with XGBoost Classifier." *Multimedia Tools and Applications*, vol. 84, no. 9, 2025, pp. 5491–5521. https://doi.org/10.1007/s11042-024-18975-6.

[6] Shah, K., A. Patel, and H. Yadav. "Fine-Tuning Deep Learning Model Using Transfer Learning for Pneumonia Diagnosis." *Lecture Notes in Networks and Systems*, vol. 1159, 2025, pp. 333–344. https://doi.org/10.1007/978-981-97-8526-1_26.

[7] Appavu, N., S. Kadry, and N. Kennedy Babu. "A Transfer Learning Strategy to Identify Covid-19 from X-ray." *IETE Journal of Research*, 2025, pp. 1–13. https://doi.org/10.1080/03772063.2025.2497514.

[8] Dhungana, P., M. Roy, R. Aryal, S. Chaudhary, M. T. R., and P. Dhungana. "Ensemble Deep Learning Approach for Pneumonia Detection Using DenseNet, MobileNet, and EfficientNet with Transfer Learning." In *2025 International Conference on Data Science and Business Systems (ICDSBS)*, 2025, pp. 1–7. https://doi.org/10.1109/icdsbs63635.2025.11031996.

[9] Kavitha, S., and H. H. Inbarani. "MHWF-CNN: Multiscale Horizontal Wavelet Fusion Convolutional Neural Network with Transfer Learning for Image Classification." *Evolving Systems*, vol. 16, no. 2, 2025, p. 73. https://doi.org/10.1007/s12530-025-09697-7.

[10] Rabbah, J., M. Ridouani, and L. Hassouni. "Improving Pneumonia Diagnosis with High-Accuracy CNN-Based Chest X-Ray Image Classification and Integrated Gradient." *Biomedical Signal Processing and Control*, vol. 101, 2025, p. 107239. https://doi.org/10.1016/j.bspc.2024.107239.

[11] Gu, C., and M. Lee. "Deep Transfer Learning Using Real-World Image Features for Medical Image Classification, with a Case Study on Pneumonia X-Ray Images." *Bioengineering*, vol. 11, no. 4, 2024, p. 406. https://doi.org/10.3390/bioengineering11040406.

[12] Haque, R., et al. "A Scalable Solution for Pneumonia Diagnosis: Transfer Learning for Chest X-ray Analysis." In *Proceedings of International Conference on Contemporary Computing and Informatics (IC3I 2024)*, vol. 7, 2024, pp. 255–262. https://doi.org/10.1109/IC3I61595.2024.10829132.

[13] Maquen-Niño, G. L. E., J. G. Nuñez-Fernandez, F. Y. Taquila-Calderon, I. Adrianzén-Olano, P. De-La-cruz-vdv, and G. Carrión-Barco. "Classification Model Using Transfer Learning for the Detection of Pneumonia in Chest X-Ray Images." *International Journal of Online and Biomedical Engineering*, vol. 20, no. 5, 2024, pp. 150–161. https://doi.org/10.3991/ijoe.v20i05.45277.

[14] Lenny, C., A. A. Margharet, B. Shiny, S. Tigga, and S. T. George. "Pneumonia Detection from Chest X-Ray Images Using Deep Learning Methods." *Lecture Notes in Electrical Engineering*, vol. 905, 2022, pp. 643–655. https://doi.org/10.1007/978-981-19-2177-3_60.

[15] Khattab, R., I. R. Abdelmaksoud, and S. Abdelrazek. "Automated Detection of COVID-19 and Pneumonia Diseases Using Data Mining and Transfer Learning Algorithms with Focal Loss from Chest X-Ray Images." *Applied Soft Computing*, vol. 162, 2024, p. 111806. https://doi.org/10.1016/j.asoc.2024.111806.

[16] Feng, S., X. Wu, and L. Li. "A Novel Deep Convolutional Network Based on Transfer Learning for Lung Image Disease Diagnosis." *Applied and Computational Engineering*, vol. 99,

Manuscript received 6 October 2025; Revised 12 December 2025; Accepted 20 January 2026; Available online 22 February 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i2.1262

no. 1, 2024, pp. 161–167. https://doi.org/10.54254/2755-2721/99/20251816.

[17] Godbole, S., A. Kattukaran, S. Savla, V. Pradhan, P. Kanani, and D. Patil. “Enhancing Paediatric Pneumonia Detection and Classification Using Customized CNNs and Transfer Learning Based Ensemble Models.” *International Research Journal of Multidisciplinary Technovation*, vol. 6, no. 6, 2024, pp. 38–53. https://doi.org/10.54392/irjmt2463.

[18] Mujahid, M., F. Rustam, R. Álvarez, J. L. Vidal Mazón, I. de la T. Díez, and I. Ashraf. “Pneumonia Classification from X-ray Images with Inception-V3 and Convolutional Neural Network.” *Diagnostics*, vol. 12, no. 5, 2022. https://doi.org/10.3390/diagnostics12051280.

[19] Singh, Sukhendra, et al. “Efficient Pneumonia Detection Using Vision Transformers on Chest X-Rays.” *Scientific Reports*, vol. 14, no. 1, 2024, pp. 1–17. https://doi.org/10.1038/s41598-024-52703-2.

[20] Ali, Abbas M., et al. “COVID-19 Pneumonia Level Detection Using Deep Learning Algorithm and Transfer Learning.” *Evolutionary Intelligence*, vol. 17, no. 2, 2024, pp. 1035–46. https://doi.org/10.1007/s12065-022-00777-0.

[21] Putri, Kania Ardhani, and Wikky Fawwaz Al Maki. “Enhancing Pneumonia Disease Classification Using Genetic Algorithm-Tuned DCGANs and VGG-16 Integration.” *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 6, no. 1, 2024, pp. 11–22. https://doi.org/10.35882/jeeemi.v6i1.349.

[22] Asnake, Nigus Wereta, et al. “X-Ray Image-Based Pneumonia Detection and Classification Using Deep Learning.” *Multimedia Tools and Applications*, vol. 83, no. 21, 2024, pp. 60789–60807. https://doi.org/10.1007/s11042-023-17965-4.

[23] Arulananth, T. S., et al. “Classification of Paediatric Pneumonia Using Modified DenseNet-121 Deep-Learning Model.” *IEEE Access*, vol. 12, February 2024, pp. 35716–35727. https://doi.org/10.1109/ACCESS.2024.3371151.

[24] Raj, A., M. O. Pallavi, and N. Manoj. “Development of CheXNet-Based Web Application to Detect Pneumonia Using Chest X-Ray Images.” In *8th IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER 2024) – Proceedings*, 2024, pp. 322–327. https://doi.org/10.1109/DISCOVER62353.2024.10750561.

[25] Ghia, Canna Jagdish, and Gautam Sudhakar Rambhad. “Systematic Review and Meta-Analysis of Comorbidities and Associated Risk Factors in Indian Patients of Community-Acquired Pneumonia.” *SAGE Open Medicine*, January 2022. https://doi.org/10.1177/20503121221095485.

[26] Rajaguru, Vasuki, Tae H. Kim, Jaeyong Shin, Sang G. Lee, and Whiejong Han. “Ability of the LACE Index to Predict 30-Day Readmissions in Patients with Acute Myocardial Infarction.” *Journal of Personalized Medicine*, vol. 12, no. 7, 2022, p. 1085. https://doi.org/10.3390/jpm12071085.

[27] Lewis, M. O., P. T. Tran, Y. Huang, R. A. Desai, Y. Shen, and J. D. Brown. “Disease Severity and Risk Factors of 30-Day Hospital Readmission in Pediatric Hospitalizations for Pneumonia.” *Journal of Clinical Medicine*, vol. 11, no. 5, 2022, p. 1185. https://doi.org/10.3390/jcm11051185.

[28] Ieracitano, Cosimo, Nadia Mammone, Mario Versaci, Giuseppe Varone, Abder-Rahman Ali, Antonio Armentano, et al. “A Fuzzy-Enhanced Deep Learning Approach for Early Detection of Covid-19 Pneumonia from Portable Chest X-Ray Images.” *Neurocomputing*, vol. 481, 2022, pp. 202–215. https://doi.org/10.1016/j.neucom.2022.01.055.

[29] Rostami, Mehrdad, and Mourad Oussalah. “A Novel Explainable COVID-19 Diagnosis Method by Integration of Feature Selection with Random Forest.” *Informatics in Medicine Unlocked*, vol. 30, 2022. https://doi.org/10.1016/j.imu.2022.100941.

[30] Aviles-Rivero, Angelica I., Philip Sellars, Carola-Bibiane Schönlieb, and Nicolas Papadakis. “GraphXCOVID: Explainable Deep Graph Diffusion Pseudo-Labelling for Identifying COVID-19 on Chest X-Rays.” *Pattern Recognition*, vol. 122, 2022. https://doi.org/10.1016/j.patcog.2021.108274.

[31] Malhotra, Aakarsh, Surbhi Mittal, Puspita Majumdar, Saheb Chhabra, Kartik Thakral, Mayank Vatsa, et al. “Multi-task Driven Explainable Diagnosis of COVID-19 Using Chest X-Ray Images.” *Pattern Recognition*, vol. 122, 2022. https://doi.org/10.1016/j.patcog.2021.108243.

[32] Mondal, Arnab Kumar, et al. “xViTCOS: Explainable Vision Transformer Based COVID-19 Screening Using Radiography.” *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 10, December 2021, p. 1100110. https://doi.org/10.1109/JTEHM.2021.3134096.

[33] Ren, H., et al. “Interpretable Pneumonia Detection by Combining Deep Learning and Explainable Models with Multisource Data.” *IEEE Access*, vol. 9, 2021, pp. 95872–95883. https://doi.org/10.1109/ACCESS.2021.3090215.

[34] Panwar, H., P. K. Gupta, M. K. Siddiqui, R. Morales-Menendez, P. Bhardwaj, and V. Singh. “A

Manuscript received 6 October 2025; Revised 12 December 2025; Accepted 20 January 2026; Available online 22 February 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i2.1262

Deep Learning and Grad-CAM Based Color Visualization Approach for Fast Detection of COVID-19 Cases Using Chest X-Ray and CT-Scan Images." *Chaos, Solitons and Fractals*, vol. 140, 2020, p. 110190. https://doi.org/10.1016/j.chaos.2020.110190.

[35] Wang, X., Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers. "NIH Chest X-rays Dataset." *Kaggle*, 2017. https://www.kaggle.com/datasets/nih-chest-xrays/data.

[36] Patel, B. N., M. P. Lungren, and A. Y. Ng. "CheXpert: A Large Chest Radiograph Dataset for Automated Chest X-ray Interpretation." *Kaggle*, 2019. https://www.kaggle.com/datasets/ashery/chexpert.

[37] Mooney, P. "Chest X-ray Images (Pneumonia)." *Kaggle*, 2017. https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia/version/2.

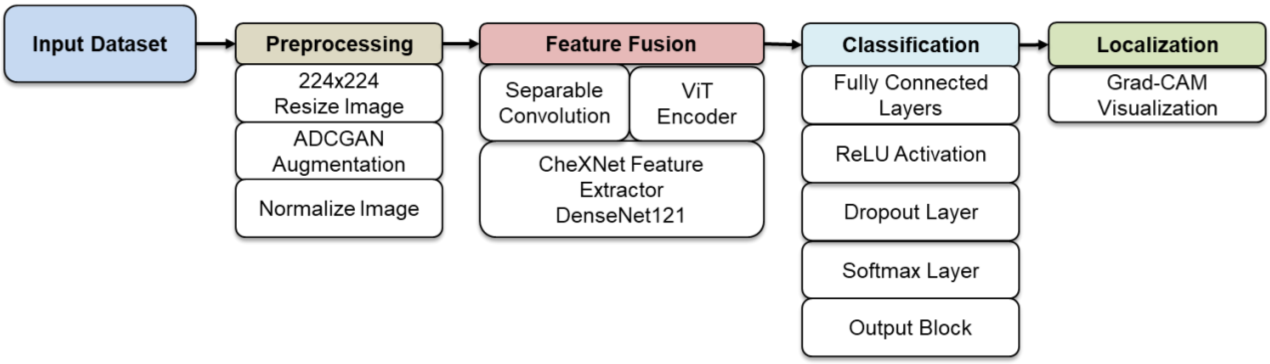
Author Biography

Khushboo Trivedi is an accomplished academician and researcher with a strong background in computer science and engineering. She earned her master's degree in Computer Science and Engineering from Parul University and is currently pursuing a Ph.D. in the specialized domain of deep learning and computer vision at Parul University, Vadodara, Gujarat. With more than 12 years of extensive experience in academia, she has been actively involved in teaching, mentoring, curriculum development, and research activities. Her primary research interests include machine learning, deep learning, artificial intelligence, and computer vision, with a focus on applying advanced computational techniques to solve real-world problems. At present, she is serving as an Assistant Professor in the Department of Computer Science and Engineering at Parul Institute of Technology, Parul University, where she contributes significantly to academic excellence, student development, and ongoing research initiatives.

Dr. Chintan Bhupeshbhai Thacker received the Ph.D. degree in the domain of AI and computer vision from Gujarat Technical University in the year 2022. He had served as head of the Department of Computer Science and Engineering at HJD Institute of Technical Education and Research, Kera, India. He has 1+ years of experience in industry and 12+ years of experience in academia. Currently, he serves as an assistant professor in the Department of Computer Science and Engineering at Parul Institute of Engineering Technology, Parul University, Vadodara, Gujarat. In addition, he has also guided several doctoral students and has been active in conducting several workshops in the domain of computer vision. His research interests are in machine learning, AI, deep learning, and computer vision.

Manuscript received 6 October 2025; Revised 12 December 2025; Accepted 20 January 2026; Available online 22 February 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i2.1262

Cover image



Manuscript received 6 October 2025; Revised 12 December 2025; Accepted 20 January 2026; Available online 22 February 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i2.1262