Hierarchical Transformer Residual Model for Pneumonia Detection and Lesion Mapping

Anupama Prasanth
College of Computer Studies, University of Technology Bahrain, Salmabad, Kingdom of Bahrain

Abstract*—Pneumonia, a potentially fatal infection and a common disease-causing culprit among children and the elderly, still remains as a prevalent threat even after years of research on tackling it. Rapid and proper identification is crucial for timely treatment and improved results. While thoracic radiographs are widely employed in pneumonia diagnosis, real-world clinical assessment is frequently questioned by factors such as subtle radiographic patterns, overlapping symptoms, subjective manual judgement and dependency on expert radiologists. The study proposes a hybrid deep learning model integrating ResNet50 and the Swin Transformer, coupled with an auxiliary segmentation decoder to facilitate both classification and lesion localization in chest X-ray images. ResNet50 acts as the backbone for hierarchical spatial feature extraction, capturing fine-grained local textures indicative of pulmonary abnormalities, and the Swin Transformer serves as the global attention-driven feature aggregator. The shifted window mechanism of the Swin Transformer maintains spatial hierarchy while facilitating effective contextual modelling. Global Average Pooling (GAP) and Multilayer Perceptron (MLP) form the classification head, yielding accurate predictions in classifying the images, while the segmentation decoder utilizes multiscale features to generate pixel-wise masks for pneumonia lesion regions. The model outperformed conventional methods with 98.4% classification accuracy, 98.2% precision, 99.2% recall and an F1-score of 98.7% with a 0.88 Dice Coefficient in segmentation. These results reflect the hybrid architecture's superior performance and its dual capacity for diagnostic prediction and lesion interpretability. The proposed model demonstrates promising results for deployment in real-world clinical workflows, especially in resource-constrained or high-patient-load environments.

Keywords*—Pneumonia detection; lesion segmentation; chest X-ray; ResNet50; swin transformer; global average pooling

I. Introduction

Pneumonia is a potentially life-threatening pulmonary infection causing inflammation of the alveoli, the tiny air sacs within the lungs, filling up with fluid or pus. Affecting over 150 million individuals and causing around 2.5 million deaths every year, it is the primary cause of infectious disease deaths among children under five years old globally, responsible for 70,000 deaths per year, a terrifying 14% of deaths [1]. A number of pathogens, including bacteria, viruses and fungi, could act as the agents of the fatal disease. Streptococcus pneumoniae remains the most common bacterial agent, and respiratory viruses such as influenza and SARS-CoV-2 also contribute significantly to disease incidence [2]. Prolonged coughing, chest pain, fever, chills and respiratory difficulties are common signs of pneumonia, and in extreme cases, additional symptoms like confusion, cyanosis and fast breathing can also occur. Delayed or inaccurate diagnosis results in disease progression, complications or even death, highlighting the critical need for timely and reliable diagnostic methodologies [3].

Clinical examination, auscultation, laboratory tests and manual chest radiography analysis are the basic traditional pneumonia detection methods. Clinical examination and auscultation have limited sensitivity and delayed turnaround times, are frequently used only as a manual screening process, and that too with high error rates in early pneumonia manifestations [4]. Due to the unavailability of specific biomarkers, laboratory tests frequently fail to pinpoint pneumonia from other respiratory illnesses. Even though chest X-rays offer a non-invasive and accessible tool for visual confirmation of pneumonia, their interpretation is subjective, depending heavily on the expertise of a radiologist [5]. Misinterpretation or variability in radiographic evaluations usually leads to delayed diagnosis or misdiagnosis. Traditional image analysis techniques have limited capacity to discern subtle pathological patterns or overlapping pulmonary anomalies, especially in early or atypical Pneumonia infections [6].

Deep learning (DL) based models have displayed promising results in automated pneumonia detection methods from chest radiographs in their initial phases. Convolutional neural networks (CNNs), DenseNet and Vision Transformers (ViT) frameworks have demonstrated considerable success in enhancing classification accuracy and reducing diagnostic turnaround time [7]. But the lack of spatial interpretability, being computationally complex and costly, hinder real-time deployment of most of the models. Limited interpretability and high resource requirements, and skilled hardware operation requirements make them irrelevant in rural and resource constrained regions. Moreover, few integrate simultaneous segmentation of pathological regions, which is vital for localized analysis and clinical insight.

The study proposes a novel ResNet50-Swin Transformer hybrid model, addressing the diagnostic challenges by combining robust local feature extraction with global contextual representation, enabling both accurate classification and precise segmentation of pneumonia from chest X-rays. This dual-capability framework enhances diagnostic reliability while maintaining computational efficiency rendering it ideal for practical clinical deployment. To bridge the gap between theoretical accuracy and practical feasibility, the proposed model aims to design a model that not only performs robustly in controlled settings but is also scalable and reliable in high-patient-load environments such as rural clinics or emergency care centers lacking expert radiologists. By reducing reliance on

manual interpretation and enabling automatic lesion localization, the hybrid framework holds the potential to significantly reduce diagnostic delays in telemedicine setups and frontline screening programs. The lightweight architectural design further enables deployment on edge devices such as mobile radiology units and AI-enabled X-ray machines, making it especially useful in developing countries or disaster-struck zones. In real-world hospital settings, such a model could assist junior clinicians or non-specialist health workers in triaging suspected pneumonia cases more efficiently, improving both speed and consistency of care.

The core research question addressed in this study is whether a hybrid model integrating convolutional and transformer-based architectures with an integrated segmentation decoder enhances both classification accuracy and lesion-level interpretability in pneumonia detection from chest radiographic images. The key objectives of the study are as mentioned below:

- Design and implement a hybrid DL framework that leverages the spatial feature extraction capabilities of ResNet50 with contextual attention mechanisms of the Swin Transformer for binary classification of thoracic radiograph images into affected and healthy cases.
- Incorporate an integrated segmentation decoder within the hybrid model to generate pixel-wise lesion maps, enabling accurate localization of pneumonia-affected regions and thereby improving diagnostic interpretability and clinical decision-making.

The further sections of the research are structured as follows: Section II offers a comprehensive analysis of the latest advances in Pneumonia detection research and highlights the current research constraints. Section III outlines the proposed methodology. Experimental findings are presented in Section IV, accompanied by a comprehensive analysis of the model performance in Section V. Section VI concludes the study by encapsulating the principal results and emphasizing prospective areas for further research.

II. Related Works

Ali et al. [8] proposed a DL-based diagnostic framework to identify pneumonia from thoracic radiograph images. Six CNN architectures were implemented and trained on the dataset. EfficientNetV2L outperformed others with an accuracy of 94.02%, followed by VGG16 (91.66%), Xception (90.7%), InceptionResNetV2 (88.94%), ResNet50 (87.98%) and a baseline CNN (87.78%). The high performance of EfficientNetV2L was attributed to the compound scaling strategy and efficient use of network parameters, enabling the framework for fine-grained image classification tasks. However, the absence of region-wise interpretability restricted the clinical transparency of model predictions.

Shaikh et al. [9] proposed MDEV, an ensemble model for thoracic radiograph image classification into Pneumonia affected and healthy categories constructed by concatenating four deep transfer learning models: MobileNet, DenseNet-201, EfficientNet-B0 and VGG-16. Each component was fine-tuned and trained on the selected dataset and achieved an accuracy of 92.15% on evaluation. The MDEV model operated in two hierarchical levels: in the first level, learners independently extracted features from pre-processed thoracic radiograph images using distinct deep networks and in the second level, learning was performed by a meta-learner that integrated the outputs. The architectural complexity led to reduced interpretability and higher computational requirements, hampering the real-time deployment in low-resource or point-of-care environments.

Barhoom et al. [10] proposed a CNN framework for the Pneumonia identification and classification using thoracic radiograph images. The study utilized the hierarchical feature extraction capabilities of CNN to distinguish between normal and infected lungs and utilized a dataset of chest radiographs, focusing on three output classes: bacterial pneumonia, viral pneumonia and normal lungs. Among the tested architectures, VGG16 demonstrated superior performance, attributed to its deep structure and effective extraction of low- to high-level spatial features. The CNN architecture allowed for scalable feature learning and enabled accurate triaging of pneumonia subtypes, clinically valuable for treatment planning. The absence of cross-validation or external dataset benchmarking limited the model's generalizability to broader clinical populations and varying imaging conditions.

Wang et al. [11] suggested a pneumonia classification framework based on DenseNet, addressing the structural complexity and uneven gray-level distribution in chest X-ray images. The core architecture utilized DenseNet due to its ability to propagate learned features across all layers, supporting parameter efficiency and improved local feature learning. Squeeze and Excitation (SE) blocks, a channel attention mechanism, were integrated into the network to emphasize pneumonia-related information in the feature maps. The integration of attention modules and pooling adjustments contributed to a more discriminative lesion focus. Evaluated on the Chest X-ray 2017 dataset, the modified DenseNet achieved an accuracy of 92.8%. The study's precision and recall balance was flawed, indicating the need for further optimization for proper classification.

Mabrouk et al. [12] presented an ensemble learning (EL) approach for the computer-aided pneumonia classification utilizing lung radiograph images. The model integrated three pretrained architectures: DenseNet169, MobileNetV2 and ViT, fine-tuned on pediatric chest radiographs from the Chest X-ray Images (Pneumonia) dataset. Each model acted as a functional layer, independently extracting feature representations from the input images, and these features were passed through GAP for dimensionality reduction and subsequently combined to form the final predictive output. The proposed EL approach achieved 93.91% accuracy but was limited by the computational complexity, as the simultaneous use of three deep architectures increased inference time and resource consumption, hampering deployment in real-time or resource-limited clinical environments.

Ortiz-Toro et al. [13] explored the diagnostic potential of three textural image characterization techniques as input biomarkers for artificial intelligence (AI) models in pneumonia detection from thoracic radiograph images. Two datasets, Guangzhou Women and Children Medical Center pediatric

dataset (GWCMCx) and a composite dataset (Josep-NIH) including COVID-19 chest X-rays were utilized and three machine learning (ML) classifiers: Support Vector Machine (SVM), K-Nearest Neighbors (KNN) and Random Forest (RF), were employed in model evaluation. The study integrated handcrafted feature descriptors with conventional classifiers and demonstrated the utility of texture-based biomarkers in distinguishing pneumonia-affected lungs from normal cases with minimal preprocessing. On the GWCMCx dataset, the super pixel-based histon method yielded the highest accuracy at 91.3%, outperforming radiomics (83.3% accuracy) and fractal dimension (89.9%). On the Josep-NIH dataset, the fractal dimension achieved better accuracy, but the latter dataset used for evaluation was more biased, hampering the reliability.

Bhatt et al. [14] proposed an ensemble-based CNN framework for pneumonia diagnosis using lung radiographic images. The architecture comprised three parallel CNN models with distinct kernel sizes, each capturing features at varying spatial resolutions and the outputs were fused through a weighted ensemble mechanism that enabled dynamic thresholding, allowing clinicians to fine-tune diagnostic sensitivity based on clinical requirements. The framework maintained a lightweight footprint, and the ensemble increased feature diversity, robustness and allowed personalized classification thresholds for clinical flexibility with an accuracy of 84.12% and precision of 80.04%. The model incurred a higher computational cost than a single CNN, and the dataset size introduced a risk of overfitting.

Singh et al. [15] proposed an attention-aware CNN architecture for pneumonia detection in thoracic radiograph images. The study integrated channel and spatial attention modules within a deep neural network (DNN) to guide the learning process toward clinically relevant regions. A bottom-up and top-down feedforward attention mechanism was adopted, allowing both feedforward and feedback attention processes to coexist within each module. The dual-level attention provided enhanced feature discrimination and allowed the model to dynamically localize pneumonia-affected regions. The network achieved a 95.47% accuracy, outperforming both baseline CNN and a pretrained ResNet50 model. Comparative evaluation revealed that ResNet50 with integrated attention modules improved accuracy from 84.53% to 95.73%. The attention modules however, increased architectural complexity, resulting in extended training time and requiring more computational resources in deployment scenarios.

Ibrahim et al. [16] presented a DL framework employing a pretrained AlexNet model to segregate chest X-ray images across multiple diagnostic scenarios. By utilizing AlexNet's hierarchical feature extraction capabilities, the study handled both binary and multiclass classification tasks and demonstrated functional adaptability for a range of diagnostic applications in chest radiology. On evaluation through two-way, three-way and four-way classification tasks using publicly available datasets, the model achieved accuracy of 94.43% for non-COVID-19 viral pneumonia vs. normal and 91.43% for bacterial pneumonia vs. normal and 94.00% accuracy for three-class classification and four-class classification with 93.42% accuracy. The heterogeneity of image sources across public datasets introduced domain shifts affecting consistency and performance in real-world clinical environments.

Avola et al. [17] evaluated twelve DNN architectures using transfer learning for the pneumonia classification from lung radiographic images. Employing two datasets with chest radiographs of patients diagnosed with pneumonia caused by bacterial, generic viral or SARS-CoV-2 infections and unaffected cases. The architecture was adapted for transfer learning by replacing the final classification layers of each model with task-specific dense layers and leveraged hierarchical feature extraction to identify localized pathological patterns. On evaluation in the combined dataset, MobileNetV2 and MobileNetV3 achieved the highest accuracies of above 0.80, followed by AlexNet (0.78), DenseNet and ResNet variants (0.75–0.79) and lower accuracies for SqueezeNet (0.33) and NASNet (0.55). Significant performance degradation was observed for most models under data-scarce conditions.

Zhu et al. [18] suggested a multi-task DL approach for classification of pneumonia types and segmenting associated lesions from chest CT scans involving 181 patients, categorized as lobar, lobular or interstitial pneumonia, while ground truth labels for segmentation were created through manual annotation. The network handled segmentation and classification tasks, streamlining the diagnostic workflow by using shared feature extraction pathways. The multitask architecture enabled the model to contextualize lesion morphology in parallel with disease categorization and achieved a classification accuracy of 92.7%. Bias was introduced by using data from a single medical center that hampered generalizability across diverse populations, imaging protocols and scanner types.

An et al. [19] proposed a deep CNN for pneumonia classification integrating EfficientNetB0 and DenseNet121 as feature extractors. The architecture incorporated multi-head self-attention mechanisms to enhance feature representation and discrimination across varying spatial regions of chest X-ray images. A cross-channel attention-based feature fusion strategy is implemented to combine features from both networks and employed residual blocks and dynamic pooling layers for refining the learning. The attention-enhanced architecture allowed for selective amplification of disease-relevant features, improving the interpretability and clinical reliability of the predictions. The study utilized a chest X-ray dataset that included both normal and pneumonia-affected cases and achieved an accuracy of 95.19%. The exclusive use of static two-dimensional images restricted diagnostic depth compared to three-dimensional volumetric data.

Bhandari et al. [20] presented a DL framework for classifying chest radiograph images into four distinct categories, with pneumonia as one category. The study utilized a single CNN architecture and utilized a publicly available dataset of 7,132 images. The model was integrated with an explainable AI (XAI) framework using SHAP, LIME and Grad-CAM, which offered visual and statistical information on the decision-making method of the classifier. These XAI techniques provided class-discriminative feature maps and model interpretability, which were further corroborated by medical experts, and the interpretive capability added trustworthiness to the diagnostic tool. The system achieved an accuracy of 94.31% through ten-

fold cross-validation, but the dataset size limitations hampered the generalization of the study.

Xue et al. [21] evaluated DL techniques for the segregation of COVID-19 and pneumonia utilizing both thoracic radiograph and computed tomography scan images. The research employed an ensemble of deep transfer learning models: ResNet152, ResNet50, DenseNet121 and an enhanced version of VGG16 across five different datasets. Two datasets contained chest radiographs, while the remaining three included CT imaging data and achieved precision and accuracy ranging from 95% to 96%. The integration of deep residual and densely connected layers allowed for hierarchical feature extraction across both modalities, and transfer learning addressed data scarcity and computational efficiency. Despite its better performance, the study acknowledged a critical limitation related to inconsistencies in ground truth annotations within the datasets, affecting the reliability and generalizability of learned features across different institutions and clinical settings.

Barakat et al. [22] suggested an ML-based framework designed for the early detection of pediatric pneumonia using thoracic radiograph images. The model segmented the input chest X-ray into 64 fixed regions of interest (ROIs) for focused and granular feature extraction. The extracted features were mapped back to their anatomical locations, supporting interpretable diagnostics. On evaluation employing various ML models, the quadratic SVM classifier, using a 64-region-of-interest (ROI) feature extraction scheme, achieved the highest performance with an accuracy of 97.58%, with a lesser classification time compared to transfer learning benchmarks. Extensive reliance on hand-engineered feature extraction and manually defined ROIs that did not generalize across varying image resolutions, anatomical variations and datasets introduced the risk of omitting subtle pathological indicators outside the preselected ROIs and affected model adaptability in practical clinical environments.

A. Research Gap

Despite substantial research in pneumonia detection using AI such as EfficientNetV2L, DenseNet, ResNet50, ViT etc., certain challenges still exist [8] [12] [21]. Most frameworks focus on overall accuracy improvement but fall short to address segmentation of pathological regions, model explainability and clinical adaptability properly [16] [21]. Even though MDEV and ensemble CNNs attempt to enhance feature representation, computational overhead hinders their real-time deployment [9] [14]. Attention-based architectures and multi-head fusion exhibits promising performance but is burdened by the model complexity, making training and interpretability more challenging. Furthermore, segmentation is often excluded, limiting end-to-end diagnostic support and limited external validation also reduces generalizability across diverse clinical environments. A cohesive model that balances both local and global feature extraction capabilities is absent in the majority of the research. Only a few number of existing methods have investigated designs that can carry out simultaneous classification and segmentation with great computing efficiency and diagnostic confidence. Hence, there is a pressing need for a unified, lightweight and interpretable hybrid framework that can simultaneously perform accurate classification and segmentation while being computationally feasible for practical medical applications.

III. Materials and Methods

The proposed model integrates ResNet50 and Swin transformer effectively for the classification and segmentation of thoracic radiographs for Pneumonia. The Chest X-Ray images (Pneumonia) dataset is utilized in the study, and the methodological pipeline consists of a ResNet50 backbone and Swin Transformer modules for hybrid feature extraction after meticulous preprocessing and data augmentation. A segmentation decoder for lesion localization is employed for the segmentation task, and the classification head utilizes global average pooling (GAP) and a multilayer perceptron head (MLP). The proposed model's block diagram is illustrated in Fig. 1.

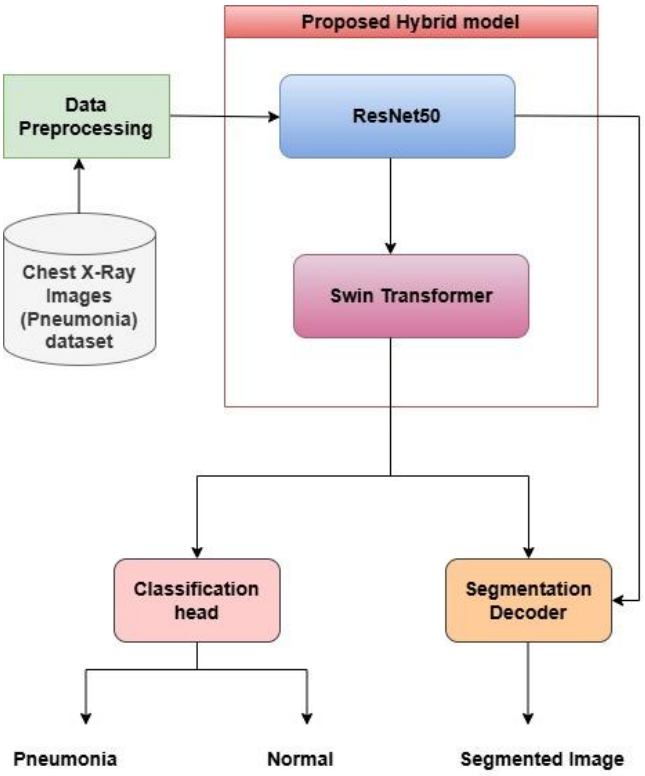

Fig. 1. Block diagram.

B. Dataset Description

The study utilizes Chest X-Ray images (Pneumonia), the openly accessible Kaggle dataset with images gathered from Guangzhou Women and Children's Medical Centre in China [23]. It has 5,863 anterior-posterior (AP) view chest radiographs of paediatric patients aged between one and five years categorized either as Normal or Pneumonia. Fig. 2 represents the number of images present in the three primary directories: train, test and validation, each with two subfolders corresponding to the diagnostic categories. Normal X-rays show clear lungs free of any abnormal opacities. Viral pneumonia primarily presents as more diffuse, bilateral interstitial opacities, while bacterial pneumonia usually exhibits focal lobar consolidation, especially in the upper lobes.

All chest X-rays underwent quality screening to remove unclear or low-resolution scans to guarantee diagnostic accuracy and image credibility. Two expert radiologists independently confirmed the image diagnose and the evaluation was examined by third expert to identify potential discrepancies in classification. This pediatric dataset presents real-world complexity due to the subtle and variable nature of radiographic pneumonia manifestations in young children, making it an ideal benchmark for evaluating AI-based diagnostic systems. Its expert-verified labels, well-structured format and diversity in infection types make it well-suited for both training and evaluation of models focused on classification and lesion localization. Furthermore, the public accessibility facilitates reproducibility, benchmarking and practical deployment of AI solutions in resource-limited or high-volume clinical settings. Fig. 3 illustrates the random sample images in the dataset.

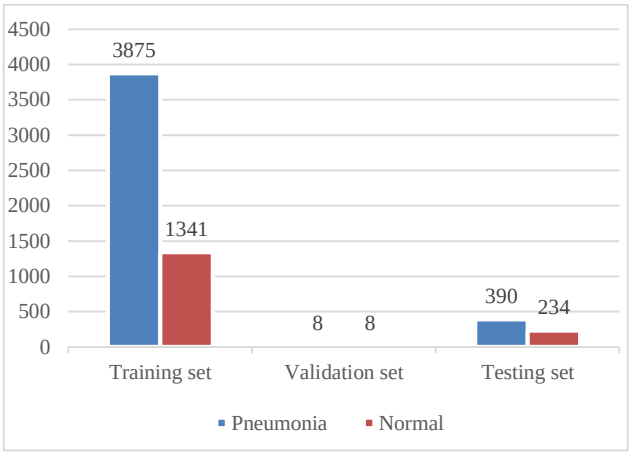

Fig. 2. Number of images in each directory.

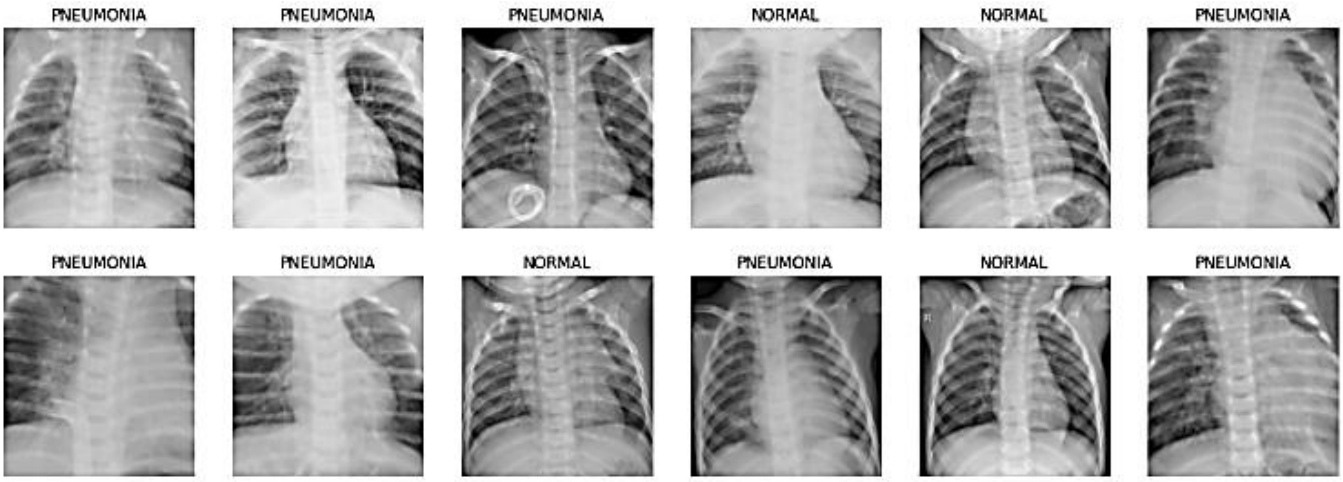

Fig. 3. Random sample images in the dataset.

C. Data Preprocessing

A systematic preprocessing and augmentation strategy is designed for the chest X-ray images prior to inputting them into the model. Initially, each image is resized to a standardized spatial dimension of 224×224 pixels to maintain compatibility with both ResNet50 and Swin Transformer input requirements. The original greyscale chest X-ray image is then converted to a 3-channel RGB image through channel-wise expansion. This transformation is necessary to match the pretrained convolutional backbones that anticipate three-channel inputs. Subsequently, the pixel intensity values $I \in [0,255]$ are normalized to a continuous range of $[0,1]$ using the rescaling operation, as shown in Eq. (1):

$$I_{normalized} = \frac{I}{255} \quad (1)$$

Data augmentation is introduced in the training set to increase the generalization capability and to reduce overfitting. The augmentation operations include horizontal and vertical flipping, random zooming with a zoom range up to 0.3, shearing with a shear angle of 10 degrees and rotation within ±20 degrees. By scaling the intensity in a range of [0.5,2.0], intensity variations are also applied, along with spatial translations. As certain transformations resulted in vacant pixel areas, the nearest neighbour approach is utilized to fill in missing values, as shown in Eq. (2):

$$I'(x,y) = \arg \min_{(x',y')} \|(x,y)(x',y')\| \; ; \; \text{such that I}'(x',y') \neq 0 \quad (2)$$

The images are then organized into mini-batches and passed into the model pipeline. Labels are encoded into binary numerical form, as in Eq. (3):

$$y = \begin{cases} 0 \; ; Normal \\ 1 \; ; Pneumonia \end{cases} \quad (3)$$

In addition to providing enough variation to the training set for reliable feature extraction, the preprocessing pipeline assures the spatial integrity and semantic information preservation of the thoracic radiograph images.

D. Model Development

1) ResNet50: ResNet50 is a fifty-layer deep CNN architecture built upon the principle of residual learning, which resolves the vanishing gradient problem by employing shortcut connections [24]. A stem layer is followed by four sequential convolutional stages, each containing a stack of convolutional blocks and identity blocks. Each block includes a sequence of

convolutional layers with kernel sizes set to 1×1, 3×3 and 1×1, followed by batch normalization and ReLU activation. The defining feature of ResNet50 is the residual connection, where the input x to a block is added to its output F(x), forming the final output, as in Eq. (4):

$$y = F(x, \{W_i\}) + x \quad (4)$$

where, $\{W_i\}$ represents a set of learnable parameters of the convolutional layers. This residual mapping promotes deeper network construction while maintaining productivity and enables the network to learn identity functions. Rich local characteristics can be extracted from images using the ResNet50 backbone, which progressively minimizes spatial dimensions while deepening feature representations. In the proposed hybrid model, only the convolutional and residual blocks of ResNet50 are utilized. Fig. 4 illustrates the general architecture of ResNet50.

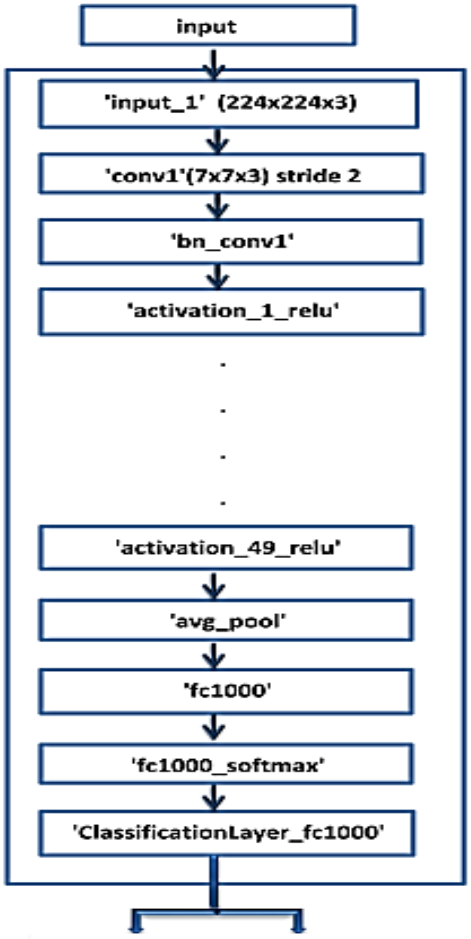

Fig. 4. General architecture of ResNet50.

An input image is initially passed through a convolutional layer with a 7×7 kernel and a stride of 2 to reduce spatial dimensions while capturing broad features. This is followed by batch normalization and a ReLU activation to standardize and introduce non-linearity. The network then progresses through a sequence of 49 residual layers, each built using bottleneck blocks that incorporate skip connections. These connections help preserve input information and facilitate efficient gradient flow, enabling deep network training without degradation. Each residual block performs convolutional transformations and combines the original input with the transformed output. Following the final activation layer, the architecture applies GAP to reduce the spatial dimensions, resulting in a 1×1×2048 feature vector. The vector is then carried to a fully connected dense layer with 1000 units and finally through a softmax activation, producing class probabilities. In the proposed model, the softmax and classification layers are discarded as ResNet50 is used solely for deep feature extraction.

2) Swin transformer: The Swin Transformer (Shifted Window Transformer) is a hierarchical ViT that applies self-attention within non-overlapping local windows and shifts these windows across layers to enable cross-window connections [25]. It is composed of a series of Swin Transformer Blocks, each including Layer Normalization (LN), Multi-head Self-Attention (MSA) within shifted local windows, MLP layers with GELU activation and Residual connections. Fig. 5 represents the basic architecture of the Swin transformer.

Initially, the input image $I \in \mathbb{R}^{H \times W \times 3}$ is partitioned into patches and embedded into a lower-dimensional space, as shown in Eq. (5):

$$x_0 = PatchEmbed(I), x_0 \in \mathbb{R}^{\frac{H}{P} \times \frac{W}{P} \times C} \quad (5)$$

where, P is the patch size and C is the embedding dimension. The embeddings pass through a series of stages composed of Swin Transformer Blocks. The spatial resolution is minimized and the feature dimension is boosted in each stage, resulting in a hierarchical representation. Window-based MSA (W-MSA) followed by Shifted Window MSA (SW-MSA) is applied in each block further. The attention is computed within each window w , as in Eq. (6):

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d}} + B\right)V \quad (6)$$

where Q , K and V represent the query, key and value matrices respectively, computed from the patch tokens confined within a local window and d, the dimension of the key vectors and B, the relative position bias. Each attention output is passed through a LN layer to stabilize the training dynamics, as shown in Eq. (7):

$$LN(x) = \frac{x - \mu}{\sqrt{\sigma^2 + \epsilon}} \cdot \gamma + \beta \quad (7)$$

where, μ, σ represent the mean and standard deviation of the input feature vector respectively and γ, β represents learnable affine parameters. This is followed by a Feed-Forward Network (FFN) comprising two linear transformations with a GELU non-linearity in between, as in Eq. (8):

$$FFN(x) = W_2 \cdot GELU(W_1 \cdot x + b_1) + b_2 \quad (8)$$

where, W_1, W_2 represents weight matrices, b_1, b_2 are biases.

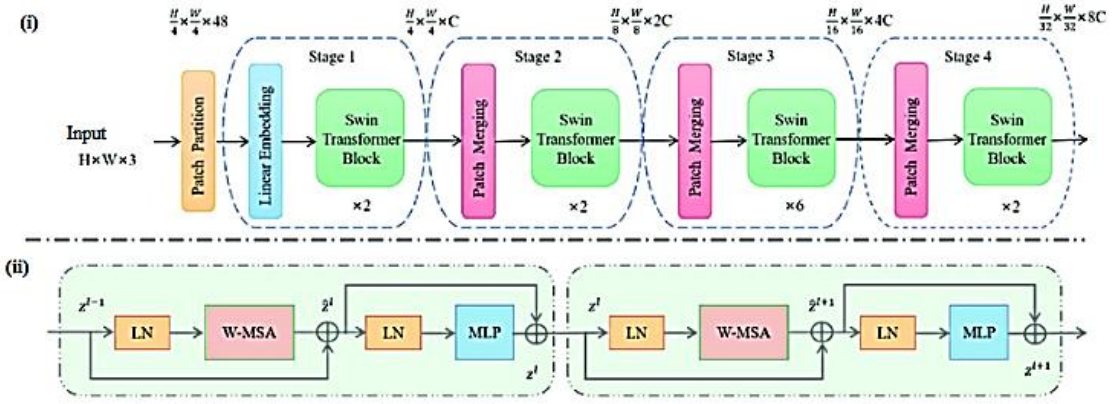

Fig. 5. Swin transformer: (i) Basic architecture (ii) Swin transformer block.

3) Proposed ResNet50- Swin Transformer hybrid model: The proposed hybrid architecture integrates the advantages of both convolutional and transformer-based approaches by employing ResNet50 for localized feature extraction and the Swin Transformer for global context modelling. ResNet50 captures low and mid-level spatial features from the input thoracic radiograph images, identifying pathological patterns in the process. The output feature map of ResNet50 undergoes processing to be compatible with the input requirement of the Swin transformer. A convolutional adaptation layer consisting of a 1×1 followed by batch normalization and ReLU activation is utilized to transform the feature dimensions to the patch embedding size of the Swin Transformer, as shown in Eq. (9):

$$F_{adapted} = RELU\big(BN(W_c * F_{ResNet} + b)\big) \quad (9)$$

where, $W_c \in \mathbb{R}^{1\times1\times c_1\times c}$ represents the kernel for dimension adjustment. c_1, c denotes the number of input and output channels, respectively, F_{ResNet} represents the feature map output from ResNet50, of shape $[H, W, c_1]$, b denotes the learnable bias added after convolution and $*$ represents the convolution operation. Once adapted, the feature maps are tokenized into non-overlapping patches and embedded, forming the input to the Swin Transformer. The window-based and shifted window-based self-attention mechanisms of the transformer architecture allow it to gather long-range dependencies and semantic patterns across the image. Both local features from ResNet50 and global contextual cues from the transformer are therefore jointly modelled. The output of the final Swin Transformer block is bifurcated: one branch passes through a GAP layer followed by an MLP for binary classification (pneumonia or normal), while the other branch serves as the input to a segmentation decoder. The decoder reconstructs high-resolution spatial masks of infected lung regions through a cascade of upsampling and convolutional layers.

a) Classification head: Global Average Pooling (GAP) is a downsampling operation that reduces each feature map to a single scalar by computing the average of all spatial elements converting feature maps into compact feature vectors while preserving spatially aggregated information [26]. In the proposed model, the GAP layer condenses the spatial dimensions of the feature map output of the Swin Transformer into a single feature vector. This operation preserves the most salient global contextual features and drastically minimises the number of parameters and computation before classification. The resultant feature map of dimension $H \times W \times C$ of the Swin transformer block is spatially reduced to produce a compact descriptor vector, as in Eq. (10):

$$x_{GAP} = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} F_{Swin}[i,j] \quad (10)$$

The condensed descriptor is then passed into a final MLP layer that serves as the final decision-making unit by transforming the aggregated features into a scalar output that facilitates binary classification. Its learnable weights and bias terms are optimized during training to minimize loss and improve diagnostic accuracy. A sigmoid function is applied to generate a probability score $\hat{y} \in [0,1]$ activated output is as shown in Eq. (11):

$$\hat{y} = \sigma(W \cdot x_{GAP} + b) \quad (11)$$

where, W and b denote the weight and bias of the dense layer, and the sigmoid function is denoted by σ.

b) Segmented decoder: The segmentation decoder is responsible for reconstructing pixel-wise masks from high-dimensional feature representations learned by the encoder, a combined sequence of ResNet50 and Swin Transformer. During the forward pass, selected intermediate feature maps from ResNet50 are cached and routed via skip connections to the corresponding stages in the decoder. This facilitates the recovery of fine spatial details lost during downsampling and enhances localization precision. The decoder adopts an up-sampling architecture that consists of bilinear upsampling followed by convolutional blocks that refine spatial granularity. The up-sampled feature at level l is obtained, as shown in Eq. (12):

$$F_{up}^{(l)} = Upsample\left(F_{dec}^{(l+1)}\right) \quad (12)$$

where, $F_{dec}^{(l+1)}$ is the decoder feature map from a deeper layer $l+1$. The output is concatenated with the corresponding encoder feature map $F_{enc}^{(l)}$ from ResNet50, as shown in Eq. (13):

$$F_{concat}^{(l)} = Concat\left(F_{up}^{(l)}, F_{enc}^{(l)}\right) \quad (13)$$

The concatenated feature is processed through a convolutional refinement block, as in Eq. (14):

$$F_{dec}^{(l)} = ReLU\left(BN\left(W^{(l)} * F_{concat}^{(l)} + b^{(l)}\right)\right) \quad (14)$$

where, $W^{(l)}$ and $b^{(l)}$ are the convolution weights and biases, BN represents batch normalization, and $ReLU$ is the non-linear activation function. The decoding process is iteratively performed through successive layers until the output spatial resolution is comparable to that of the input image. Finally, a 1×1 convolutional layer is applied to map the refined feature maps to a single-channel output, as shown in Eq. (15):

$$S = \sigma(W_{seg} * F_{dec}^{(l)} + b_{seg}) \quad (15)$$

where, W_{seg} and b_{seg} denote weights and bias of the final layer. The segmentation output $S \in [0,1]^{H\times W}$ indicates the prospect of pneumonia-affected regions for each pixel in the input image. The architecture enables effective integration of global context via Swin Transformer and spatial precision via ResNet skip features, resulting in reliable and interpretable pneumonia lesion segmentation. Fig. 6 illustrates the basic architecture of the proposed ResNet50-Swin Transformer hybrid model. The detailed algorithm for the proposed model is given below (see Algorithm 1).

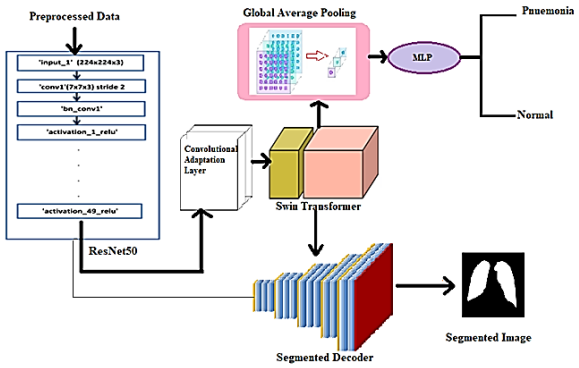

Fig. 6. Basic architecture of the proposed ResNet50-swin transformer hybrid model.

Algorithm 1: ResNet50-Swin Transformer hybrid model

Input:

Chest X-ray image I ∈ $\mathbb{R}^{H\times W\times 3}$

Label vector $y_i \in [0,1]$, indicating Pneumonia (1) or Normal (0)

Output:

Predicted class label $\hat{y} \in [0,1]$

Segmentation mask $S \in \mathbb{R}^{H\times W}$

Begin:

Data collection

Load Chest X-ray image (Pneumonia) dataset

Pre-processing

Resize the images into 224x224

Normalize pixel values to range [0,1]: $I_{normalized} = \frac{I}{255}$

Label encoding in the images:

$$y = \begin{cases} 0 \ ; Normal \\ 1 \ ; Pneumonia \end{cases}$$

Data Augmentation on the training dataset

ResNet50

Local feature extraction and final output generation, $y = F(x,\{W_i\}) + x$

Store feature maps for decoder skip connections.

Feature Adaptation

Transform F_{ResNet} to match Swin Transformer input size:

$F_{adapted} = RELU\big(BN(W_c * F_{ResNet} + b)\big)$

Swin Transformer

Embed $F_{adapted}$ into non-overlapping patches.

Apply Swin Transformer blocks with shifted window attention and FFN: $Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d}} + B\right)V$

Apply Layer Norm → FFN → Layer Norm sequence

$LN(x) = \frac{x-\mu}{\sqrt{\sigma^2+\epsilon}} \cdot \gamma + \beta \ ; \ FFN(x) = W_2 \cdot GELU(W_1 \cdot x + b_1) + b_2$

Store final feature map F_{Swin}

Classification Head

Apply Global Average Pooling:

$x_{GAP} = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} F_{Swin}[i,j]$

Multilayer Perceptron apply sigmoid function:

$\hat{y} = \sigma(W \cdot x_{GAP} + b)$

Segmentation Decoder

Input: F_{Swin} and skip connections from ResNet50.

For each decoder stage:

Up sample previous output

Concatenate with corresponding skip connection

Apply convolution → BatchNorm → ReLU

Final segmentation masks=σ (Conv1×1 (Decoder Output))

Model Compilation and Training

Compile model with loss = Binary crossentropy, learning rate = 0.0001, optimizer = Adam, Epochs =75

Train model: model.fit (X_train, y_train)

Evaluation and Model Saving

Evaluate model: model. evaluate (X_test, y_test)

Tune hyperparameters

Save the model

End

E. Simulation Map

The proposed hybrid model was implemented in a high-level deep learning framework and trained on a workstation equipped with a high-performance GPU (NVIDIA RTX 3090, 24GB VRAM), 128GB RAM and an Intel Xeon processor. Python was used as the programming language, and model development was carried out using the Keras API built on top of TensorFlow, chosen for its modular design, scalability and ease of customization for hybrid DL architectures. Training and experimentation were conducted on the Google Colaboratory (Colab) platform, leveraging its cloud-based infrastructure and access to high-performance GPUs. The training process was optimized by carefully tuning key hyperparameters, which significantly influenced model convergence and generalization. Table I shows the full list of hyperparameters and training settings employed in the study.

TABLE I. HYPERPARAMETER SPECIFICATIONS

Hyper parameters	Values
Optimizer	ADAM
Swin Transformer Layers	4
Loss function	Binary cross-entropy (Classification) Dice loss (Segmentation)
Activation function	Sigmoid
Batch size	16
Epochs	75
Learning Rate	0.0001
Dropout	0.3

IV. RESULTS

A set of standard evaluation metrics, as illustrated in Eq. (16) to Eq. (20), was employed to comprehensively assess the proposed model's performance. True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN) are mathematically computed using the confusion matrix core elements. Different metrics offer unique insights into the model performance. Overall correctness by accuracy, precision and recall highlights the ability of the model to correctly detect pneumonia without excessive misses or false alarms.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (16)$$

$$Precision = \frac{TP}{TP+FP} \quad (17)$$

$$Recall = \frac{TP}{TP+FN} \quad (18)$$

$$F1-score = 2 \times \frac{precision \times Recall}{Precision+Recall} \quad (19)$$

The Dice Coefficient evaluates the similarity between predicted segmentation and the ground truth and measures the degree of overlap between two binary masks and ranges from 0 to 1.

$$Dice(A,B) = \frac{2|A \cap B|}{|A|+|B|} \quad (20)$$

where, A and B denote the sets of predicted and ground truth pixels, respectively, labelled as positive. A dice coefficient close to 1 indicates better spatial agreement between prediction and ground truth. The accuracy and loss plots serve as fundamental diagnostic tools to monitor the model's learning dynamics over successive training epochs. A steadily increasing accuracy curve alongside a consistently decreasing loss trajectory indicates effective gradient optimization and convergence toward a generalized solution. Any divergence between training and validation curves may suggest overfitting or underfitting, necessitating hyperparameter adjustments.

The accuracy plot of the proposed hybrid model, represented in Fig. 7, demonstrates a strong and consistent learning curve over 75 epochs, highlighting effective training convergence. A rapid rise in training and validation accuracies is observed within the initial 20 epochs, indicative of efficient low-level feature acquisition. As the epochs progress, the model's accuracy stabilizes with the training accuracy approaching 99.5%, while the validation accuracy maintains close proximity, around 98.4%. The minimal gap between the two curves signifies robust generalization and minimal overfitting and affirms the model's potential in retaining discriminative features and achieving high classification reliability on unseen chest X-ray data.

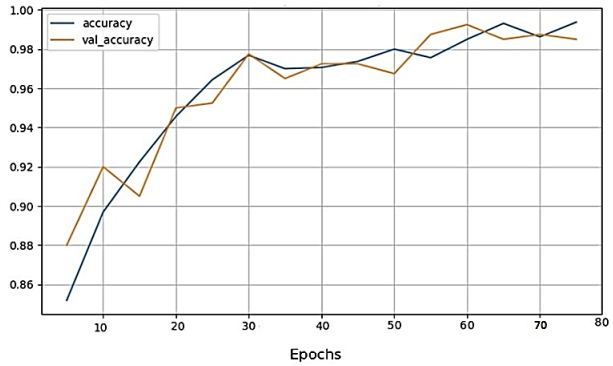

Fig. 7. Accuracy plot of ResNet50-Swin Transformer hybrid model.

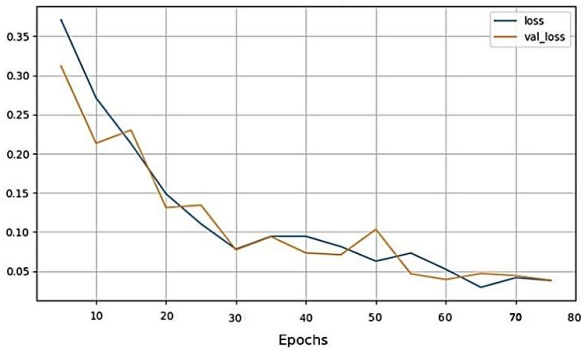

Fig. 8. Loss plot of ResNet50-Swin Transformer hybrid model.

Fig. 8 illustrates the loss plot for the proposed model that demonstrates a steady and significant decrease in both training and validation loss over 75 epochs, reflecting effective model convergence. In the initial training phase, a steep decline in loss is observed during the first 15 epochs, indicating rapid optimization of weights and improved learning of discriminative features. Beyond epoch 20, the loss values continue to decrease gradually and stabilize around 0.04, signifying that the model is

minimizing the error function efficiently. The close alignment between training and validation loss trajectories throughout the epochs confirms the absence of overfitting and affirms the generalization capability across unseen data.

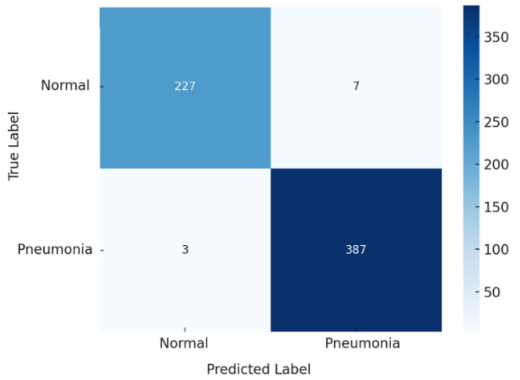

Fig. 9. Confusion matrix of the proposed model.

Fig. 9 represents the confusion matrix highlighting the classification performance of the hybrid model on the test dataset. Out of 234 normal images, the framework achieved a true negative rate of around 97%, and in the pneumonia category, 387 out of 390 pneumonia images were correctly classified, exhibiting better accuracy. The results indicate strong predictive performance and class discrimination, particularly in identifying pneumonia cases, and the minimal number of misclassifications further emphasizes the model's reliability for binary medical image segregation.

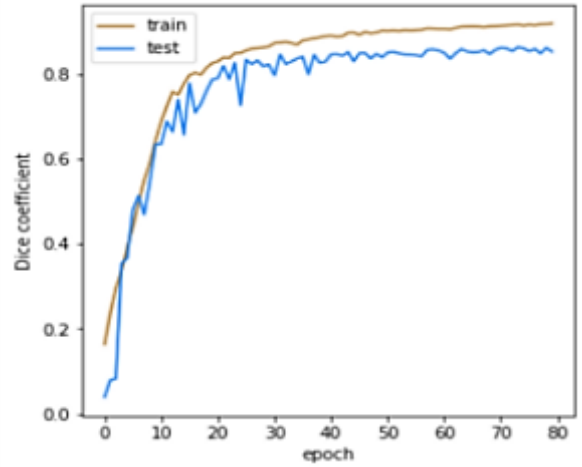

Fig. 10. Dice coefficient plot.

The Dice Coefficient plot illustrated in Fig. 10 demonstrates the segmentation performance of the proposed model over 75 training epochs. Both training and test curves exhibit a steep ascent within the initial 20 epochs, indicating rapid learning and effective optimization of segmentation boundaries. Beyond epoch 25, the Dice Coefficient stabilizes around 0.88 for the test set, while the training score plateaus slightly higher, suggesting excellent overlap between predicted and ground truth segmentation masks with minimal overfitting. The narrow gap between training and test curves throughout the learning process reflects the model's excellent consistency across datasets. The stable performance at later epochs confirms that the model effectively captures spatial and structural patterns necessary for precise medical image segmentation, reinforcing its reliability in real-world diagnostic contexts.

	precision	recall	f1-score
Pneumonia	0.98	0.98	0.97
Normal	0.98	0.99	0.98
accuracy			0.98
macro avg	0.98	0.99	0.98
weighted avg	0.98	0.99	0.98

Fig. 11. Classification report of ResNet50-Swin Transformer hybrid model.

Fig. 11 represents the classification report for the proposed framework, revealing outstanding predictive performance across both classes. With a precision of 0.98 for both classes, the model has a high proportion of accurately detected positive instances relative to all predicted positives. The recall for the Normal class reaches 0.99, while Pneumonia records a recall of 0.98, demonstrating the excellent sensitivity in detecting true positives, particularly for pneumonia diagnosis. The F1-score stands at 0.97 for Pneumonia and 0.98 for Normal, highlighting the model's balanced and reliable classification performance, and the overall accuracy of 0.984 reaffirms the model's robustness. These results collectively validate the clinical reliability of the proposed model in distinguishing pneumonia from normal thoracic radiograph images with great precision and generalizability. Fig. 12 represents the evaluation metrics of the proposed model.

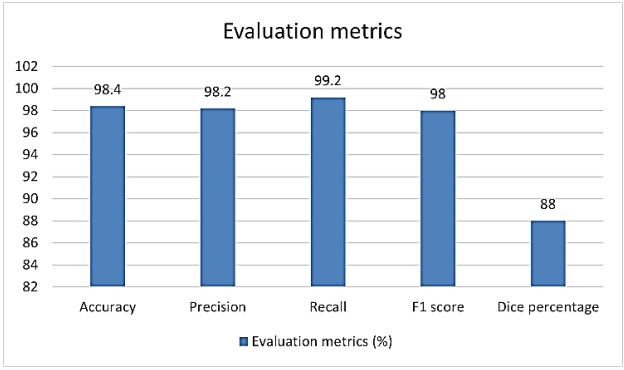

Fig. 12. Evaluation metrics for the proposed model.

The visual outputs generated by the proposed hybrid model are illustrated in Fig. 13 and Fig. 14, effectively demonstrating its dual capabilities in both classification and segmentation of chest X-ray images. As evident from the classification report, the framework accurately distinguishes between pneumonia and normal cases, with predictions aligning closely with the actual ground truth labels, while the segmentation results highlight the effectiveness in accurately localizing lung regions. The original mask and the segmented output exhibit strong overlap, with the segmented regions closely approximating the annotated ground truth. While some minor deviations and false-positive activations are observed at the lung peripheries, the overall shape, structure and spatial consistency of the lung regions are

well-preserved. These outcomes confirm the robustness of the segmentation decoder in delineating pulmonary areas, thereby supporting further clinical interpretability and diagnostic localization.

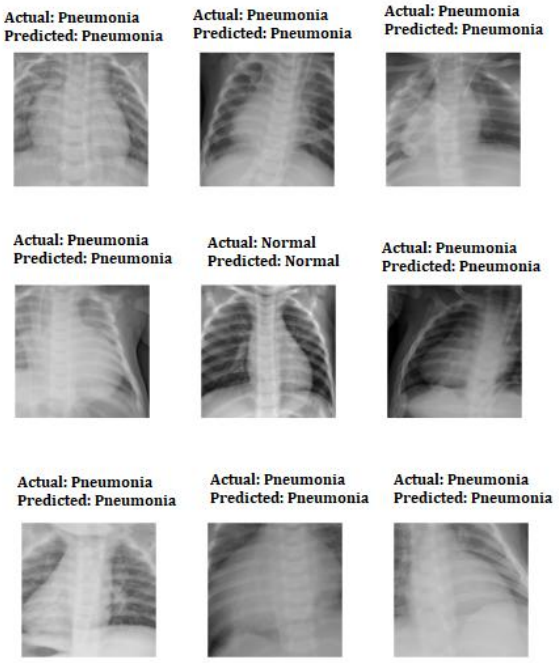

Fig. 13. Classification results.

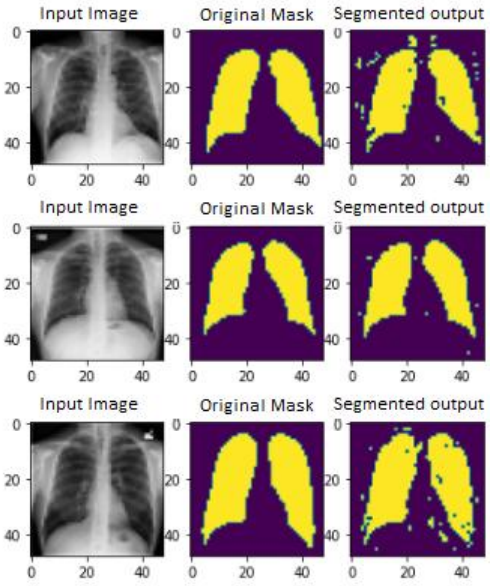

Fig. 14. Segmented output.

TABLE II. ACCURACY COMPARISON OF RESNET50-SWIN TRANSFORMER HYBRID MODEL WITH OTHER METHODS

Author [Ref]	Model	Accuracy
Ali et al. [8]	EfficientNetV2L	94.02%
Shaikh et al. [9]	MDEV	92.15%
Wang et al. [11]	DenseNet with SE blocks	92.8%
Mabrouk et al. [12]	Ensemble Learning	93.91%
Ortiz-Toro et al. [13]	KNN, SVM, RF	91.3%
Bhatt et al. [14]	Ensemble CNNs	84.12%
Singh et al. [15]	Attention-aware CNN	95.47%
Ibrahim et al. [16]	AlexNet	94.43%
Avola et al. [17]	MobileNetV3	80%
Zhu et al. [18]	Multi-task DL	92.7%
An et al. [19]	EfficientNetB0 + DenseNet121	95.19%
Bhandari et al. [20]	CNN + XAI	94.31%
Xue et al. [21]	Ensemble	96%
Barakat et al. [22]	Quadratic SVM	97.58%
Proposed model	**ResNet50-Swin Transformer hybrid**	**98.4%**

Table II and Fig. 15 illustrate the comparative analysis of the proposed hybrid framework with conventional pneumonia detection models, revealing that even though some approaches demonstrate comparable classification accuracy, they often suffer from drawbacks in terms of interpretability, architectural complexity and generalizability. MDEV and ensemble CNN architecture models were efficient in capturing diverse features, but are hampered by computational overhead due to their multi-network designs restricting real-time installation, especially in limited resource settings. Radiomics and SVM-based frameworks, relying on hand-crafted features or region-of-interest segmentation struggles with adaptability across varying anatomical structures and image qualities and were seen as inefficient in most clinical scenarios. The better performing models such as EfficientNetB0 + DenseNet121 or attention-aware CNNs, with accuracies in the range of 94 to 96%, require complex attention modules or fine-tuned fusion strategies increasing training time and computational complexity limiting their adaptability. The proposed ResNet50-Swin Transformer hybrid model achieves superior accuracy of 98.4%, combining the convolutional strength of ResNet50 for low-level feature extraction with the hierarchical attention-driven capabilities of Swin Transformer for contextual learning. The framework not only enhances classification precision but also enables segmentation functionality within a unified architecture, delivering both diagnostic accuracy and spatial interpretability in a computationally efficient manner.

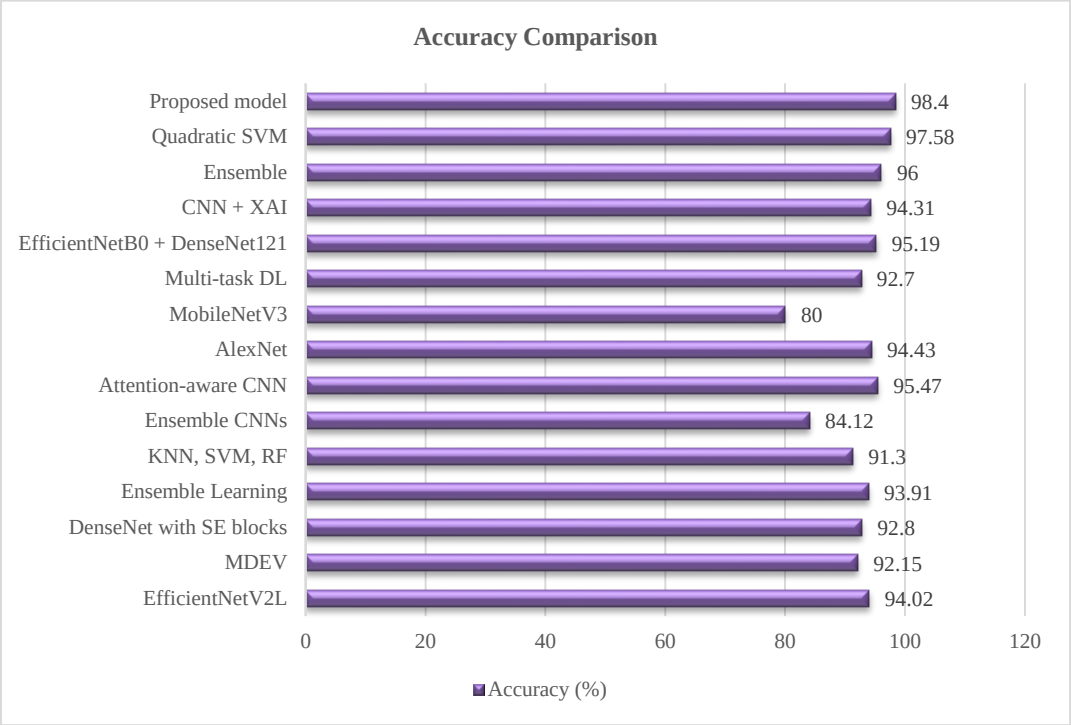

Fig. 15. Accuracy comparison of ResNet50-Swin Transformer hybrid model with other methods.

V. Discussions

The hybrid ResNet50-Swin Transformer framework demonstrates a novel architecture for dual-task learning, enabling both binary classification and pixel-level lesion localization of pneumonia from thoracic radiographs. Unlike conventional models that excel in feature extraction but often lack contextual awareness, the incorporation of the Swin Transformer effectively models global dependencies through its hierarchical window-based attention mechanism. This context-aware learning facilitates improved representation of complex pulmonary structures, especially in cases with overlapping anatomical patterns or ambiguous radiographic findings. The framework's superior classification accuracy (98.4%) and high Dice Coefficient (0.88) for lesion segmentation reinforce the robustness in both predictive performance and clinical interpretability. Compared to existing models that often perform one task effectively but compromise the other, the dual-capability approach ensures diagnostic relevance without incurring excessive computational burden. This efficiency renders it particularly suitable for scalable deployment in clinical settings where radiologist availability or diagnostic infrastructure is constrained. Moreover, the segmentation decoder not only assists in validating classification outputs but also promotes explainability; bridging the trust gap between black-box AI models and medical practitioners. These visual insights are vital for treatment planning and follow-up monitoring, as the hybrid framework offers a pragmatic solution for intelligent diagnostic assistance.

VI. Conclusion

The study presents a novel hybrid DL model integrating ResNet50 with Swin Transformer to simultaneously perform classification and segmentation of pneumonia from thoracic radiograph images. The local feature extraction capability of ResNet50 and the hierarchical attention-driven global context modelling of the Swin Transformer are utilized to skillfully gather both fine-grained and contextual features. The model further branches into a classification head comprising global average pooling and a multilayer perceptron and a segmentation decoder that reconstructs pixel-level patches, providing interpretable diagnostic outputs. With 98.4% accuracy, a precision of 98.2%, a recall of 99.2% and an F1-score of 98% in classification and a Dice coefficient of 0.88 for segmentation, the model outperformed the conventional methods by broad margins.

The segmentation results illustrated accurate delineation of lung regions, aligning closely with the ground truth masks, indicating exceptional interpretability and potential applicability in aiding radiologists. Additionally, the model's dual functionality can streamline radiology workflows by reducing dependency on separate diagnostic and localization tools, thereby saving time and operational costs.

However, the study is not without limitations. The dataset used is limited to pediatric chest radiographs collected from a single institution, which may impact the model's generalizability to broader populations, including adults or those with coexisting pulmonary conditions. Additionally, the segmentation performance may degrade in cases of overlapping pathologies or poor image quality. These limitations call for broader validation to ensure clinical robustness.

Future extensions of this work may focus on multiclass classification involving multiple thoracic diseases, incorporating explainable AI modules to enhance transparency and validating the framework across multi-institutional and

multi-demographic datasets to ensure broader generalizability and adoption. Further, the integration into mobile diagnostic applications or point-of-care systems could significantly expand its reach to rural and underserved areas.

REFERENCES

[1] R. Karim, J. K. Afridi, S. R. Yar, M. B. Zaman, and B. K. Afridi, "Clinical findings and radiological evaluation of WHO-defined severe pneumonia among hospitalized children," Cureus, vol. 15, no. 1, Jan. 2023.C.

[2] Stotts, V. F. Corrales-Medina, and K. J. Rayner, "Pneumonia-induced inflammation, resolution and cardiovascular disease: causes, consequences and clinical opportunities," Circ. Res., vol. 132, no. 6, pp. 751–774, Mar. 2023.

[3] R. K. Seramo, S. M. Awol, Y. A. Wabe, and M. M. Ali, "Determinants of pneumonia among children attending public health facilities in Worabe town," Sci. Rep., vol. 12, no. 1, p. 6175, Apr. 2022.

[4] S. Ahmed, S. Sultana, A. M. Khan, M. S. Islam, G. M. Habib, I. M. McLane, and H. Nair, "Digital auscultation as a diagnostic aid to detect childhood pneumonia: A systematic review," J. Glob. Health, vol. 12, p. 04033, Dec. 2022.

[5] S. Ramgopal, L. Ambroggio, D. Lorenz, S. S. Shah, R. M. Ruddy, and T. A. Florin, "A prediction model for pediatric radiographic pneumonia," Pediatrics, vol. 149, no. 1, Jan. 2022.

[6] J. Zhang, Y. Zhu, Y. Zhou, F. Gao, X. Qiu, J. Li, and W. Lin, "Pediatric adenovirus pneumonia: clinical practice and current treatment," Front. Med., vol. 10, p. 1207568, Jan. 2023.

[7] K. Stokes, R. Castaldo, C. Federici, S. Pagliara, A. Maccaro, F. Cappuccio, and L. Pecchia, "The use of artificial intelligence systems in diagnosis of pneumonia via signs and symptoms: A systematic review," Biomed. Signal Process. Control, vol. 72, p. 103325, Oct. 2022.

[8] M. Ali, M. Shahroz, U. Akram, M. F. Mushtaq, S. C. Altamiranda, S. A. Obregon, and I. Ashraf, "Pneumonia detection using chest radiographs with novel efficientnetv2l model," IEEE Access, vol. 12, pp. xxxx–xxxx, Jan. 2024.

[9] M. Shaikh, I. F. Siddiqui, Q. Arain, J. Koo, M. A. Unar, and N. M. F. Qureshi, "MDEV model: A novel ensemble-based transfer learning approach for pneumonia classification using CXR images," Comput. Syst. Sci. Eng., vol. 46, no. 1, pp. 287–302, Jan. 2023.

[10] A. M. Barhoom and S. S. Abu-Naser, "Diagnosis of pneumonia using deep learning," Int. J. Acad. Eng. Res., vol. 6, no. 7, pp. 12–20, Jul. 2022.

[11] K. Wang, P. Jiang, J. Meng, and X. Jiang, "Attention-based DenseNet for pneumonia classification," IRBM, vol. 43, no. 5, pp. 479–485, Oct. 2022.

[12] A. Mabrouk, R. P. Diaz Redondo, A. Dahou, M. Abd Elaziz, and M. Kayed, "Pneumonia detection on chest X-ray images using ensemble of deep convolutional neural networks," Appl. Sci., vol. 12, no. 13, p. 6448, Jul. 2022.

[13] C. Ortiz-Toro, A. García-Pedrero, M. Lillo-Saavedra, and C. Gonzalo-Martin, "Automatic detection of pneumonia in chest X-ray images using textural features," Comput. Biol. Med., vol. 145, p. 105466, Jun. 2022.

[14] H. Bhatt and M. Shah, "A Convolutional Neural Network ensemble model for pneumonia detection using chest X-ray images," Healthc. Anal., vol. 3, p. 100176, Mar. 2023.

[15] S. Singh, S. S. Rawat, M. Gupta, B. K. Tripathi, F. Alanzi, A. Majumdar, and O. Thinnukool, "Deep attention network for pneumonia detection using chest X-ray images," Comput. Mater. Contin., vol. 74, pp. 1673–1691, May 2023.

[16] A. U. Ibrahim, M. Ozsoz, S. Serte, F. Al-Turjman, and P. S. Yakoi, "Pneumonia classification using deep learning from chest X-ray images during COVID-19," Cogn. Comput., vol. 16, no. 4, pp. 1589–1601, Apr. 2024.

[17] D. Avola, A. Bacciu, L. Cinque, A. Fagioli, M. R. Marini, and R. Taiello, "Study on transfer learning capabilities for pneumonia classification in chest-X-rays images," Comput. Methods Programs Biomed., vol. 221, p. 106833, Nov. 2022.

[18] Q. Zhu, P. Che, M. Li, W. Guo, K. Ye, W. Yin, and S. Li, "Artificial intelligence for segmentation and classification of lobar, lobular, and interstitial pneumonia using case-specific CT information," Quant. Imaging Med. Surg., vol. 14, no. 1, pp. 579–592, Jan. 2023.

[19] Q. An, W. Chen, and W. Shao, "A deep convolutional neural network for pneumonia detection in X-ray images with attention ensemble," Diagnostics, vol. 14, no. 4, p. 390, Feb. 2024.

[20] M. Bhandari, T. B. Shahi, B. Siku, and A. Neupane, "Explanatory classification of CXR images into COVID-19, pneumonia and tuberculosis using deep learning and XAI," Comput. Biol. Med., vol. 150, p. 106156, Nov. 2022.

[21] X. Xue, S. Chinnaperumal, G. M. Abdulsahib, R. R. Manyam, R. Marappan, S. K. Raju, and O. I. Khalaf, "Design and analysis of a deep learning ensemble framework model for the detection of COVID-19 and pneumonia using large-scale CT scan and X-ray image datasets," Bioengineering, vol. 10, no. 3, p. 363, Mar. 2023.

[22] N. Barakat, M. Awad, and B. A. Abu-Nabah, "A machine learning approach on chest X-rays for pediatric pneumonia detection," Digit. Health, vol. 9, p. 20552076231180008, Apr. 2023.

[23] https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia/data.

[24] M. B. Hossain, S. H. S. Iqbal, M. M. Islam, M. N. Akhtar, and I. H. Sarker, "Transfer learning with fine-tuned deep CNN ResNet50 model for classifying COVID-19 from chest X-ray images," Inform. Med. Unlocked, vol. 30, p. 100916, Oct. 2022.

[25] Y. Ma and W. Lv, "Identification of pneumonia in chest X-ray image based on transformer," Int. J. Antennas Propag., vol. 2022, no. 1, p. 5072666, Jan. 2022.

[26] S. Haseli Golzar, H. Bagherpour, and J. Amiri Parian, "A new method to optimize deep CNN model for classification of regular cucumber based on global average pooling," J. Food Process. Preserv., vol. 2024, no. 1, p. 5818803, Jan. 2024.