





Article

# Efficient and Accurate Pneumonia Detection Using a Novel Multi-Scale Transformer Approach

Alireza Saber<sup>1</sup> , Amirreza Fateh<sup>2</sup> , Pouria Parhami<sup>1</sup>, Alimohammad Siahkarzadeh<sup>1</sup>, Mansoor Fateh<sup>1,\*</sup>   
and Saideh Ferdowsi<sup>3,\*</sup> 

- <sup>1</sup> Faculty of Computer Engineering, Shahrood University of Technology, Shahrood 36199-95161, Iran; alireza\_saber@shahroodut.ac.ir (A.S.); p.parhami@gmail.com (P.P.); alimohamad2005@gmail.com (A.S.)  
<sup>2</sup> School of Computer Engineering, Iran University of Science and Technology (IUST), Tehran 13114-16846, Iran; amirreza\_fateh@comp.iust.ac.ir  
<sup>3</sup> School of Mathematics, Statistics and Actuarial Science, University of Essex, Colchester CO4 3SQ, UK  
\* Correspondence: mansoor\_fateh@shahroodut.ac.ir (M.F.); s.ferdowsi@essex.ac.uk (S.F.)

## Abstract

Pneumonia, a prevalent respiratory infection, remains a leading cause of morbidity and mortality worldwide, particularly among vulnerable populations. Chest X-rays serve as a primary tool for pneumonia detection; however, variations in imaging conditions and subtle visual indicators complicate consistent interpretation. Automated tools can enhance traditional methods by improving diagnostic reliability and supporting clinical decision-making. In this study, we propose a novel multi-scale transformer approach for pneumonia detection that integrates lung segmentation and classification into a unified framework. Our method introduces a lightweight transformer-enhanced TransUNet for precise lung segmentation, achieving a Dice score of 95.68% on the “Chest X-ray Masks and Labels” dataset with fewer parameters than traditional transformers. For classification, we employ pre-trained ResNet models (ResNet-50 and ResNet-101) to extract multi-scale feature maps, which are then processed through a convolutional Residual Attention Module and a modified transformer module to enhance pneumonia detection. This integration of multi-scale feature extraction and lightweight attention mechanisms ensures robust performance, making our method suitable for resource-constrained clinical environments. Our approach achieves 93.75% accuracy on the “Kermany” dataset and 96.04% accuracy on the “Cohen” dataset, outperforming existing methods while maintaining computational efficiency.

**Keywords:** transformer; multi scale; pneumonia; classification; segmentation



Academic Editor: Cristiana Corsi

Received: 13 October 2025

Revised: 19 November 2025

Accepted: 24 November 2025

Published: 27 November 2025

**Citation:** Saber, A.; Fateh, A.; Parhami, P.; Siahkarzadeh, A.; Fateh, M.; Ferdowsi, S. Efficient and Accurate Pneumonia Detection Using a Novel Multi-Scale Transformer Approach. *Sensors* **2025**, *25*, 7233. <https://doi.org/10.3390/s25237233>

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Pneumonia is a serious respiratory condition that causes inflammation in one or both lungs, leading to symptoms such as fever, cough, and difficulty breathing. This illness is particularly dangerous for young children, accounting for approximately 15% of mortality in children under the age of five [1]. The disease is more prevalent in developing countries, where limited access to healthcare, pollution, overcrowding, and poor living conditions exacerbate its effects [2].

Early and accurate diagnosis is essential for effective treatment; however, pneumonia can be challenging to identify due to its similarity to other lung diseases [3]. Chest X-rays are commonly used for diagnosis due to their cost-effectiveness and non-invasive nature [4]. Nevertheless, the interpretation of these images can vary significantly, underscoring the necessity for consistent and automated diagnostic tools.

Recent advancements in deep learning, particularly in convolutional neural networks (CNNs), have shown significant promise in improving pneumonia diagnosis from chest X-rays [5]. These models can analyze medical images with remarkable precision, often outperforming human radiologists in both consistency and speed. Recent innovations, such as attention mechanisms, have further enhanced diagnostic accuracy [6]. Additionally, transformers have demonstrated significant potential in medical imaging tasks due to their ability to model long-range dependencies and identify complex patterns [7]. These advancements highlight the potential of AI to complement radiologists' expertise and enhance patient outcomes.

Lung segmentation is a crucial preprocessing step for improving the accuracy of pneumonia detection in chest X-rays. However, this task faces several challenges. First, the presence of artifacts, overlapping anatomical structures, and low contrast in chest X-rays can make it difficult to accurately delineate lung boundaries [8]. Traditional segmentation models, such as U-Net, while effective, often struggle to capture fine-grained details and contextual information, resulting in suboptimal performance in complex cases [9]. Additionally, the high variability in lung shapes and sizes across patients further complicates the segmentation process. Although transformer-based models have shown promise in addressing these issues, they often come with high computational costs and large parameter counts, making them unsuitable for resource-constrained environments.

The classification of pneumonia from chest X-rays also presents significant challenges. First, subtle visual indicators of pneumonia, such as small opacities or localized consolidations, can be easily missed by traditional CNN-based models like ResNet and DenseNet, which primarily focus on local features. While transformer-based models excel at capturing global contextual information, they require large amounts of labeled data and extensive computational resources, limiting their practicality in real-world clinical settings [10,11]. Moreover, the domain shift between pre-trained models (e.g., those trained on ImageNet) and medical imaging datasets often results in suboptimal performance, necessitating advanced techniques like transfer learning and domain adaptation [12]. These challenges highlight the need for a computationally efficient and accurate model that can effectively leverage both local and global features for pneumonia diagnosis.

To address these challenges, we propose an innovative approach that leverages deep learning through an integrated lightweight transformer. The proposed transformer significantly reduces the number of parameters compared to traditional transformers while maintaining lower model complexity. Our method begins with lung segmentation using a TransUNet model, which integrates transformer-based attention mechanisms into the U-Net architecture. The TransUNet model is trained on the "Chest X-ray Masks and Labels" dataset [13,14] to accurately segment lung regions in the images. Once trained, this pre-trained model is used with frozen weights to predict lung masks for our target datasets, "Kermany" [15] and "Cohen" [16]. This segmentation step isolates the lung regions, thereby enhancing the subsequent classification task.

For classification, we utilize pre-trained ResNet models, specifically ResNet-50 and ResNet-101, as the foundation for feature extraction. By extracting multi-scale feature maps from various stages of the ResNet models, we can leverage multiple feature spaces, which enhances the accuracy of our detection. This is achieved through a customized transformer module that employs a cross-attention mechanism, allowing us to make decisions based on more than one feature space. This transformer has been optimized to minimize the number of parameters while preserving performance. By concentrating on the relevant lung regions and integrating multi-scale information, our approach aims to achieve high diagnostic accuracy for pneumonia detection. This architecture reduces the computational

load and ensures robust and reliable performance, making it suitable for deployment in resource-limited settings.

Our proposed method offers the following key contributions:

- Development of a novel transformer structure that significantly reduces complexity compared to traditional transformer-based models while maintaining high performance;
- Introduction of a novel TransUNet architecture for the segmentation task, achieving a Dice score of 95.68% on the “Chest X-ray Masks and Labels” dataset;
- Introducing a convolutional Residual Attention Module (CRAM) that enriches feature representation by integrating multi-layer residual learning with lightweight attention mechanisms;
- Incorporation of multi-scale feature extraction, enabling enhanced performance through the utilization of multiple feature spaces;
- Achieving high accuracy rates of 93.75% on the “Kermany” dataset and 96.04% on the “Cohen” dataset.

This paper is structured as follows: Section 2 reviews related work. Section 3 presents the details of our proposed architecture, including design choices, the integration of multi-scale feature extraction, attention mechanism and transformer. Section 4 outlines the experimental setup, describes the datasets, and reports both quantitative and qualitative results, including performance metrics and interpretability analysis. Finally, Section 5 concludes the paper and discusses potential directions for future research.

## 2. Related Work

In recent years, the focus of research on diagnosing and categorizing lung diseases, including pneumonia, through medical imaging has intensified, driven by advances in machine learning and deep learning technologies [17]. Precisely segmenting lung areas in chest X-ray (CXR) images is essential for reliable disease identification and thorough analysis. This section examines deep learning techniques for segmenting and diagnosing lung diseases in chest X-ray (CXR) images. For the segmentation task, we focus on the U-Net architecture and its variations, including attention mechanisms and transformer blocks, which have significantly advanced lung disease segmentation. In the classification task, we categorize approaches into basic deep learning models, transfer learning, fine-tuning, and custom models, emphasizing how these advanced techniques have progressively improved diagnostic outcomes.

### 2.1. Segmentation

#### 2.1.1. U-Net for CXR Segmentation

The U-Net architecture, with its encoder–decoder structure and skip connections, has emerged as a leading method for CXR segmentation. This setup, which captures high-level semantic information and low-level details, is crucial for accurately outlining lung boundaries. Studies have consistently shown U-Net’s effectiveness in segmenting lung regions with high accuracy, a factor that significantly supports the potential of this technology in improving diagnostic outcomes [18,19]. U-Net, introduced by Ronneberger et al., has become a fundamental tool in medical image segmentation [20]. Additionally, Liu et al. [21] employed a pre-trained EfficientNet-B4 and developed an enhanced version of U-Net for identifying and segmenting lung regions.

However, traditional U-Net architectures face several limitations that impede their effectiveness in complex segmentation tasks. These limitations include the inability to leverage multi-scale information, which is essential for capturing fine-grained details, and difficulties in extracting rich contextual information, particularly for small or complex anatomical structures [22]. Furthermore, the simple skip connections in U-Net may transfer

irrelevant or noisy features, leading to ambiguity in feature representation and reduced segmentation accuracy [23]. These challenges are especially problematic in chest X-rays, where overlapping structures and low contrast exacerbate noise.

### 2.1.2. U-Net Enhancements with Transformers

To address traditional U-Net limitations, advanced architectures that enhance U-Net's ability to capture multi-scale and contextual information are needed. Recent research has significantly advanced lung segmentation by enhancing the U-Net architecture with attention mechanisms. Oktay et al. [24] introduced mechanisms that enable the model to concentrate on the most crucial areas within chest X-rays using Attention Gates (AGs). This innovation enhances segmentation accuracy and sensitivity to disease characteristics.

Azad et al. and Chen et al. extended the U-Net framework with transformers, demonstrating significant improvements in capturing intricate details and achieving top-tier results in lung segmentation tasks [25].

The incorporation of transformer modules has marked a landmark in lung segmentation research. Transformer architectures, known for capturing long-range dependencies and contextual information from text, have been successfully integrated into U-Net variants, leading to notable improvements in segmentation accuracy. For instance, Chen et al. [26] created a hybrid CNN-Transformer model for medical image segmentation, merging the strengths of CNNs and transformers to enhance accuracy and robustness in lung tissue segmentation.

## 2.2. Classification

### 2.2.1. Classical Approaches for CXR Classification

Early methods for classifying chest X-ray (CXR) images primarily depended on traditional machine learning techniques, employing classifiers such as Support Vector Machines (SVMs), K-nearest Neighbors (k-NNs), and Random Forests. For example, Stokes et al. used logistic regression, decision trees, and SVM to categorize patients' clinical data into bronchitis or pneumonia, with decision trees yielding the highest recall value of 80% and an AUC of 93% [27]. Chandra et al. used a multi-layer perceptron (MLP) to segment lung regions from CXR images, reaching an accuracy of 95.39% [28]. However, these methods, which heavily relied on symptomatic data, had limited accuracy and were evaluated on small datasets [29,30].

### 2.2.2. Deep Learning Models

The beginning of deep learning, especially convolutional Neural Networks (CNNs), has significantly transformed medical image analysis by providing superior accuracy and robustness [31]. For instance, Stephen et al. designed a custom CNN model from scratch, achieving a training accuracy of 95.31% and a validation accuracy of 93.73% [32]. Similarly, Sharma et al. created a straightforward CNN architecture that reached a 90.68% accuracy rate on the "Kermay" dataset using data augmentation [15]. However, relying solely on data augmentation does not introduce substantially new information, restricting the model's ability to learn a wide range of complex patterns from the training data.

### 2.2.3. Transfer Learning

Pre-trained CNNs have become the standard for image classification tasks, including CXR analysis. These models leverage large datasets and transfer learning to enhance performance on specific medical imaging tasks. Transfer learning, where pre-trained models are adapted and refined for new, specific tasks, has achieved significant results. For instance, Rajpurkar et al. utilized DenseNet-121 on the ChestX-ray8 dataset, comprising

112,150 frontal CXR images, achieving an F1-score of 76.8%. This study highlighted the potential of transfer learning in medical image classification [33].

#### 2.2.4. Ensemble Approaches

Ensemble learning, which combines the outputs of multiple CNN models, has shown considerable promise. For instance, Ukwuoma et al. [34] proposed two ensemble methods: ensemble group A (DenseNet201, VGG16, and GoogleNet) and ensemble group B (DenseNet201, InceptionResNetV2, and Xception). These models, followed by a self-attention layer and a multi-layer perceptron (MLP) for disease identification, achieved 97.22% accuracy for binary classification, and 97.2% and 96.4% for multi-class classification, respectively. Jaiswal et al. [35] used a mask region-based CNN for pneumonia detection through segmentation, employing an ensemble of ResNet-50 and ResNet-101 for image thresholding.

Despite their success, pre-trained models such as ResNet have inherent limitations. While ResNet models are powerful, they often struggle to independently capture all the discriminative features required for specific tasks, particularly in complex medical imaging scenarios like pneumonia detection [35]. This limitation is evident in studies where ResNet architectures require complementary support from other models or advanced techniques, such as snapshot ensembling and weighted averaging, to achieve optimal performance [36]. Furthermore, ResNet's reliance on local feature extraction through convolutional layers can hinder its ability to model long-range dependencies and global contextual information, which are crucial for accurate classification in medical images [25]. These shortcomings highlight the need for more robust frameworks, such as transformers, which excel at capturing global context and intricate patterns, thereby addressing the limitations of traditional CNN-based models like ResNet.

#### 2.2.5. Transformers

Recent advancements in medical image classification have harnessed transformer architectures alongside deep learning, yielding impressive outcomes [37]. Wang et al. [38] unveiled TransPath, a hybrid model merging CNN and transformer architectures, highlighting the potential of such integrations. They proved the efficacy of self-supervised pretraining on extensive datasets like TCGA and PAIP, followed by fine-tuning on specific medical image datasets, resulting in solid performance: 89.68% accuracy on MHIIST, 95.85% on NCT-CRC-HE, and 89.91% on PatchCamelyon. Transformer-based models have garnered attention for their capacity to capture long-range dependencies in images. Wu et al. [39] introduced a Swin Transformer-based model for pulmonary nodule classification, successfully adapting the architecture to the smaller scale of medical image datasets and achieving significant results.

In recent years, transformer-based models have continued to evolve, with a particular emphasis on improving efficiency and accuracy in medical imaging tasks. The Swin Transformer V2 [40] has emerged as a powerful architecture for various medical imaging tasks, including pneumonia detection. It achieves superior performance by leveraging hierarchical feature extraction and shifted window mechanisms, which allow it to capture both local and global patterns in chest X-rays. In a recent study, the Swin Transformer V2 achieved an accuracy of 98.6% on a diverse chest X-ray dataset (a modified dataset built by their own), outperforming traditional CNNs like ResNet and DenseNet [41]. This highlights its potential for clinical applications where high diagnostic accuracy is essential.

Hybrid architectures that combine CNNs and transformers have demonstrated remarkable success in pneumonia detection. For instance, a hybrid model integrating ResNet34 with a Multi-Axis Vision Transformer achieved a state-of-the-art accuracy of 94.87% on the Kaggle pediatric pneumonia dataset. This model leverages the local feature extraction

capabilities of CNNs and the global context modeling of transformers, resulting in fewer misclassifications and improved robustness [42].

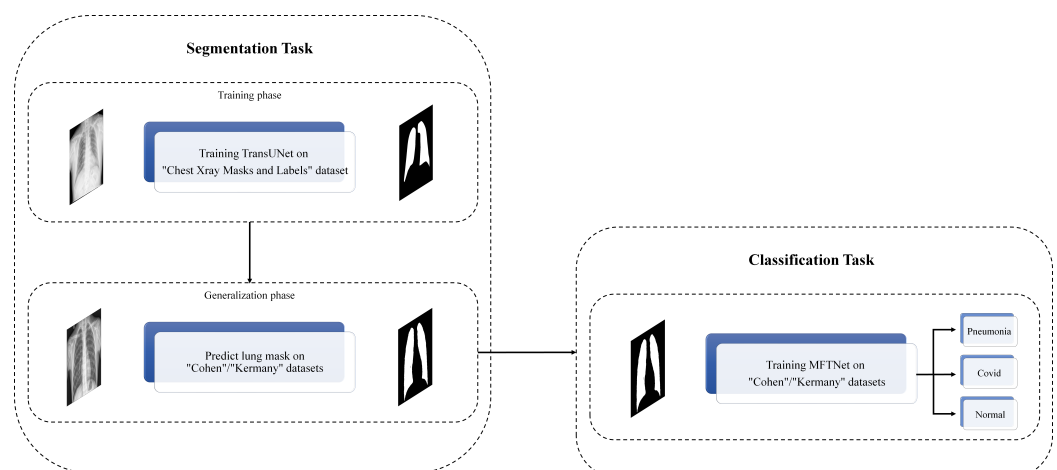
Many existing transformer architectures, such as Vision Transformers (ViTs) and Swin Transformers, require large computational resources and extensive training data, making them unsuitable for resource-constrained clinical environments [40]. Additionally, these models often struggle to effectively combine local feature extraction (a strength of CNNs) with global context modeling (a strength of transformers), leading to suboptimal performance in tasks like pneumonia detection, where both fine-grained details and global patterns are critical [42]. Furthermore, the high parameter counts and complexity of traditional transformers can result in longer training times and higher hardware requirements, limiting their practicality in real-world applications [43].

### 3. Proposed Method

#### 3.1. Overview

In this study, we propose a novel approach for segmentation and classification of Pneumonia Chest X-ray images by leveraging the power of deep learning and transformer-based attention mechanisms. Our method utilizes pre-trained ResNet models, specifically ResNet-50 and ResNet-101, as the backbone for feature extraction. These models are well-known for their ability to capture intricate patterns and features in images due to their deep architecture and residual connections. The block diagram of the proposed method is illustrated in Figure 1.

Our approach begins with a segmentation step where we employ a TransUNet model, which integrates transformer-based attention mechanisms into the popular U-Net architecture. This model is trained on the “Chest Xray Masks and Labels” dataset [13,14] to accurately segment lung regions in the images. By predicting masks for the “Cohen” dataset [16] using this pre-trained TransUNet, we can isolate the regions of interest, enhancing the subsequent classification task. The segmentation step provides us with precise lung masks, ensuring that our classification model focuses on the relevant areas of the X-ray images. This preprocessing step is crucial for improving the overall accuracy of the system by reducing background noise and irrelevant features.



**Figure 1.** The block diagram of the proposed method.

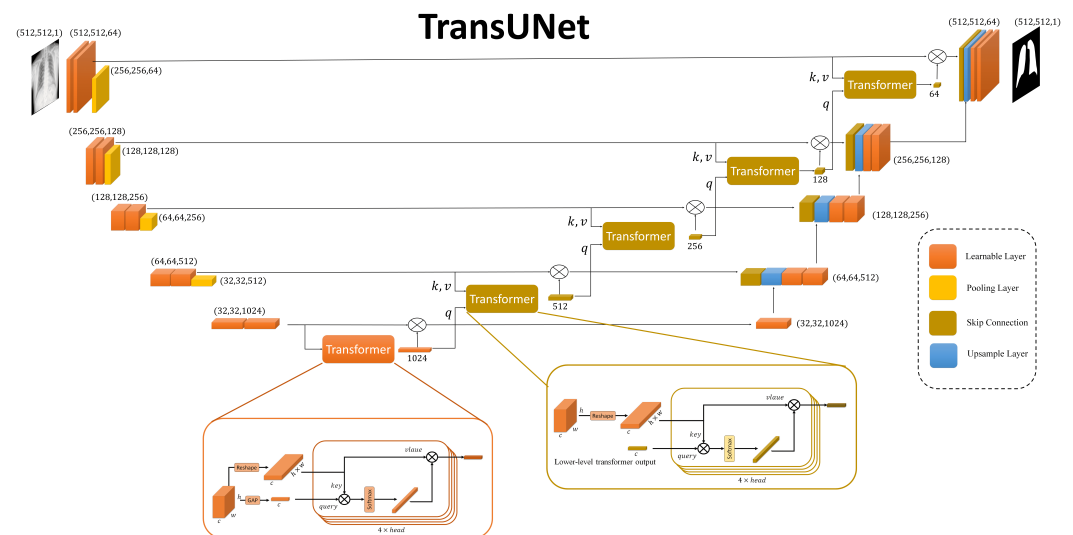
Our classification approach extracts multi-scale feature maps from three key stages of the ResNet models: the outputs of Block 2, Block 3, and Block 4. These stages provide a rich set of features at different scales, which are crucial for accurately identifying Pneumonia in chest X-rays. The extracted features are first refined through CRAM, which enhances discriminative features through dual attention mechanisms and residual learning. These

enhanced feature maps are then processed through a specialized transformer module that employs an attention mechanism, further refining the representation by allowing the network to focus on the most relevant parts of the image.

After the attention processing, the feature maps are concatenated to form a comprehensive representation of the input image. This combined feature map is subsequently fed into fully connected layers to perform the final classification. The overall architecture is designed to effectively integrate multi-scale information and attention mechanisms, thereby improving the classification accuracy.

### 3.2. Segmentation Task

In our proposed method, the segmentation task is pivotal for isolating lung regions in chest X-ray images, thereby enhancing the accuracy of pneumonia classification. For this purpose, we have designed a TransUNet model, which uniquely combines the strengths of the U-Net architecture with advanced techniques. The overall architecture of our TransUNet is illustrated in Figure 2.



**Figure 2.** The overview illustration of TransUNet Architecture. This includes, combination of transformers in each skip connection of the classic U-net.

#### 3.2.1. TransUNet Architecture

The TransUNet architecture can be divided into three main components: the encoder, the bottleneck, and the decoder. The encoder consists of a series of convolutional layers designed to capture hierarchical features from the input image. Each stage of the encoder includes a double convolution block, which performs two consecutive convolutions followed by batch normalization and ReLU activation. This setup helps in learning complex features at multiple levels. The encoder progressively reduces the spatial dimensions while increasing the depth of the feature maps through max-pooling operations.

At the bottleneck stage, the most abstract features of the input image are captured. This layer consists of a double convolution block. The bottleneck also incorporates an embedding layer and a positional encoding mechanism, which prepare the feature maps for the subsequent transformer module. The detailed structure and function of the transformer module will be discussed later in the classification subsection.

Also, the transformer modules integrate into each skip connection between the encoder and decoder. These transformers enhance the model's ability to capture global contextual information at each resolution level. In this design, the query for each transformer's attention mechanism is derived from the output of the transformer at the preceding, lower level. The transformer's output is then element-wise multiplied with the original

skip connection feature maps before being passed to the decoder. This ensures that the information passed to the decoder not only retains local details but also incorporates refined global representations.

The decoder reconstructs the segmented output by progressively upsampling the feature maps and concatenating them with the outputs of the transformer-augmented skip connections. Each upsampling step is followed by a double convolution block to refine the features and reduce the number of channels. This structure allows the decoder to restore the spatial resolution of the feature maps while retaining the detailed information captured by the encoder. Finally, a convolutional layer with a single output channel generates the segmentation mask for the lung regions.

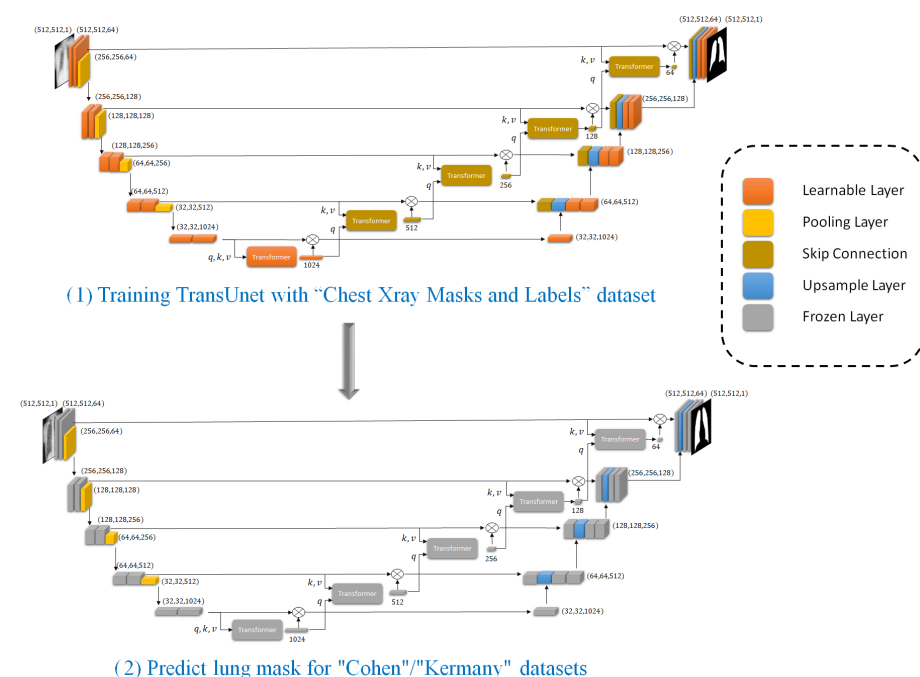
### 3.2.2. Training the TransUNet Model

Initially, we resize all images to  $512 \times 512$  pixels. The TransUNet model is trained on the “Chest X-ray Masks and Labels” dataset, which provides paired X-ray images and corresponding lung masks. By training on this dataset, the TransUNet model learns to accurately segment lung regions in chest X-rays, ensuring that the subsequent classification step focuses on the relevant areas, thereby improving the overall accuracy of the system.

By effectively combining the robust feature extraction capabilities of the U-Net architecture with advanced processing techniques, the TransUNet model provides a powerful solution for the segmentation task in our proposed method.

### 3.2.3. Applying the Trained TransUNet to Cohen/Kermany Datasets

As shown in Figure 3, after successfully training the TransUNet model on the “Chest X-ray Masks and Labels” dataset, we utilize this pre-trained segmentation model to predict lung masks for the “Cohen”/“Kermany” datasets. This step is crucial for enhancing the accuracy of the classification task by focusing on the lung regions within the X-ray images.



**Figure 3.** Part (1) includes training TransUNet and part (2) includes applying trained TransUNet to predict lung masks on “Cohen”/“Kermany” datasets that will help the classifier to focus on the most relevant anatomical regions.

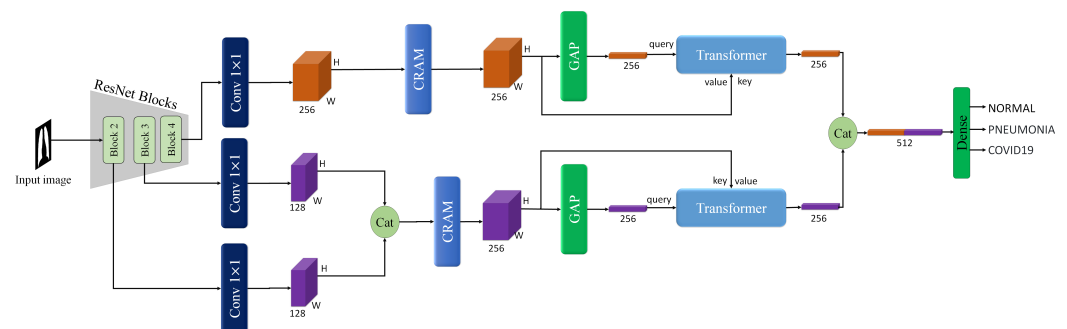
The “Cohen”/“Kermany” datasets, which contain chest X-ray images, require preprocessing to ensure that our classification model focuses on the most relevant regions. To

achieve this, we apply the trained TransUNet model to segment the lung areas from these images. The “Cohen” dataset is first preprocessed to match the input requirements of the TransUNet model. This involves standardizing the image dimensions and normalizing the pixel values to ensure consistency with the training data used for the TransUNet model.

Using the pre-trained TransUNet model, we generate lung masks for each X-ray image in the “Cohen”/“Kermany” datasets. The segmentation model outputs binary masks that highlight the lung regions while suppressing the background. By utilizing the pre-trained TransUNet model to segment the lung regions in the “Cohen” dataset, we effectively preprocess the data to improve the performance of our classification model. This segmentation step filters out noise and irrelevant features, allowing the classifier to concentrate on the lung areas, thereby enhancing the overall accuracy and robustness of our proposed method.

### 3.3. Classification Task

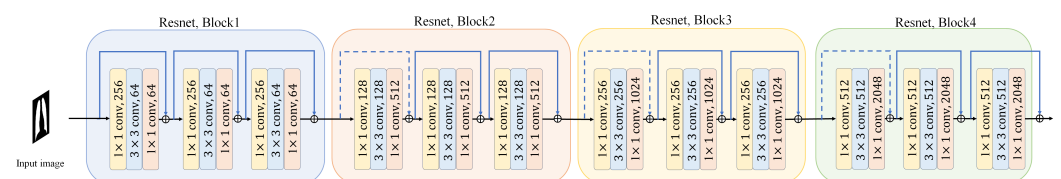
Following the segmentation of lung regions using the TransUNet model, the next step in our proposed method is the classification task. This task involves accurately identifying the presence of pneumonia or COVID-19 in the preprocessed chest X-ray images. By focusing on the lung regions isolated during the segmentation phase, we enhance the classification model’s ability to detect relevant features indicative of pneumonia or COVID-19, thereby improving diagnostic accuracy. The overview of the proposed method on the classification task is shown in Figure 4.



**Figure 4.** The overview of the proposed method on the classification task, in which the classifier operates on lung-extracted images produced by the segmentation stage.

#### 3.3.1. Backbone

The backbone of our proposed method utilizes a pre-trained ResNet model, specifically ResNet-50 or ResNet-101, to extract multi-scale feature maps from the input chest X-ray images. An illustration of the ResNet architecture is provided in Figure 5. Initially, we resize all images to  $512 \times 512$  pixels. We focus on the outputs from Block 2, Block 3, and Block 4 of the ResNet, denoted as  $B^2$ ,  $B^3$ , and  $B^4$ , respectively. Each of these blocks provides feature maps of size  $(c, h, w)$ , where  $h = w = 64$ . The numbers of channels ( $c$ ) for these blocks are 512, 1024, and 2048, respectively.



**Figure 5.** Illustration of the ResNet architecture blocks which have been used as the backbone for classification task.

The selection of Block 2, Block 3, and Block 4 is motivated by their unique contributions to the classification task. Block 2 captures low-level features such as edges and textures, which are essential for identifying subtle patterns in the lung regions. Block 3 extracts mid-level features, including more complex structures like lung lobes and localized consolidations, which are critical for detecting early signs of pneumonia. Block 4 provides high-level semantic features, such as global contextual information and disease-specific patterns, which are necessary for accurate classification. By combining these multi-scale features, our model can effectively capture both fine-grained details and global context, leading to improved diagnostic performance.

To handle the complexity and standardize the feature maps for subsequent processing, we apply  $1 \times 1$  convolution operations to reduce the number of channels. Specifically, for the output of Block 4 ( $B^4$ ), as shown in Equation (1), we reduce the channels to 64 using a  $1 \times 1$  convolution.

$$B'^4 = C_{1 \times 1}(B^4) \quad (1)$$

where  $C_{1 \times 1}$  denotes the  $1 \times 1$  convolution operation.

For the outputs of Blocks 2 and 3 ( $B^2$  and  $B^3$ ), as shown in Equations (2) and (3), we use separate  $1 \times 1$  convolutions to reduce the number of channels for each to 32.

$$B'^2 = C_{1 \times 1}(B^2) \quad (2)$$

$$B'^3 = C_{1 \times 1}(B^3) \quad (3)$$

where  $C_{1 \times 1}$  in each equation indicates a reduction in the number of channels to 32.

After reducing the channels, we concatenate the feature maps from Block 2 and Block 3 to form a merged feature map (Equation (4)).

$$B^{merged} = cat(B'^2, B'^3) \quad (4)$$

where  $cat$  denotes the concatenation operation.

Thus, we have two main feature maps with a size of (64,64,64) for further processing:

- $B'^4$
- $B^{merged}$

### 3.3.2. Convolutional Residual Attention Module (CRAM)

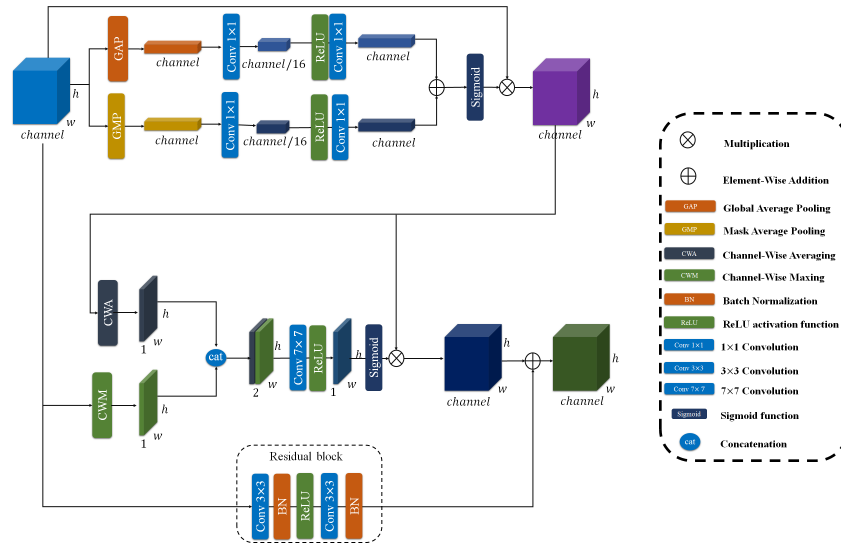
To further refine the multi-scale feature representations extracted from the ResNet backbone, we propose a CRAM module. As shown in Figure 6, this module operates on two parallel pathways to enhance feature discriminability while preserving spatial relationships. The CRAM is applied to both feature streams: the high-level semantic features from Block 4 ( $B'^4$ ) and the merged multi-scale features from Blocks 2 and 3 ( $B^{merged}$ ).

The CRAM architecture consists of two complementary components working together. For an input feature map  $X \in \mathbb{R}^{C \times H \times W}$ , the module processes it through both pathways simultaneously.

**1. Dual-Attention Mechanism:** This module incorporates a sophisticated attention mechanism that operates through two distinct dimensions.

**Channel Attention:** This component generates a channel attention vector that emphasizes informative feature channels. To build our channel attention, we first define a Sequential convolution denoted as SC. This SC consists of two  $1 \times 1$  convolutional layers with a ReLU activation in between, which reduces the channel dimension to 1/16 of its original size and then expands it back to the initial number of channels, as shown in Equation (5).

$$SC = conv_{1 \times 1}(Relu(conv_{1 \times 1}(x))) \quad (5)$$



**Figure 6.** The overview of CRAM module. As illustrated, this includes channel attention and spatial attention.

After defining SC, we construct the channel attention mechanism to generate the attention vector, denoted as  $F_{CA}$ . In this process, both global average pooling (GAP) and global max pooling (GMP) are applied to the input feature map, and each pooled feature is passed through the SC module to capture different channel-wise statistics. The outputs of GAP and GMP branches are then summed and go through a Sigmoid function to produce the final channel attention map. The refined feature map is obtained by performing element-wise multiplication between the input feature map and the attention vector, as described in Equation (6).

$$F_{CA}(x) = \sigma(GAP(SC) + GMP(SC)) \times x \quad (6)$$

where  $\sigma$  represents Sigmoid activation function.

**Spatial Attention:** This component focuses on identifying spatially significant regions within the feature maps. As shown in Equations (7) and (8), it computes both average-pooled and max-pooled features across the channel dimension, concatenates them, and processes them through a convolutional layer with a sigmoid activation to produce spatial attention maps that highlight important spatial locations.

$$F_{SA}(x) = \sigma(ReLu(conv_{7 \times 7}(cat_{SA}))) \quad (7)$$

$$cat_{SA} = cat(CWA(F_{CA}(x)), CWM(F_{CA}(x))) \quad (8)$$

where  $F_{CA}$  denotes the output of the channel attention part, CWA and CWM represent Channel-Wise Averaging and Channel-Wise Maxing respectively,  $cat$  denotes concatenation.

The final output of the dual-attention mechanism is the element-wise product of the spatially and channel-wise refined features, as formalized in Equation (9).

$$DAM(X) = F_{SA}(x) \times F_{CA}(x) \quad (9)$$

**2. Parallel convolutional Pathway:** A complementary convolutional branch, formalized in Equation (10), transforms the input through two successive  $3 \times 3$  convolutional layers with batch normalization and ReLU non-linearities. This pathway facilitates the learning of complex feature mappings while ensuring stable gradient flow via skip connections.

$$\text{Re}(X) = \text{BN}(\text{conv}_{3 \times 3}(\text{Relu}(\text{BN}(\text{conv}_{3 \times 3}(X)))))) \quad (10)$$

where BN denotes Batch Normalization.

The outputs from both pathways are integrated through element-wise addition to produce the final enhanced representation, as formalized in Equation (11). This composite feature enhancement is subsequently applied to both multi-scale feature streams in our architecture, yielding the refined feature maps specified in Equations (12) and (13).

$$\text{CRAM}(X) = \text{DAM}(X) + \text{Re}(X) \quad (11)$$

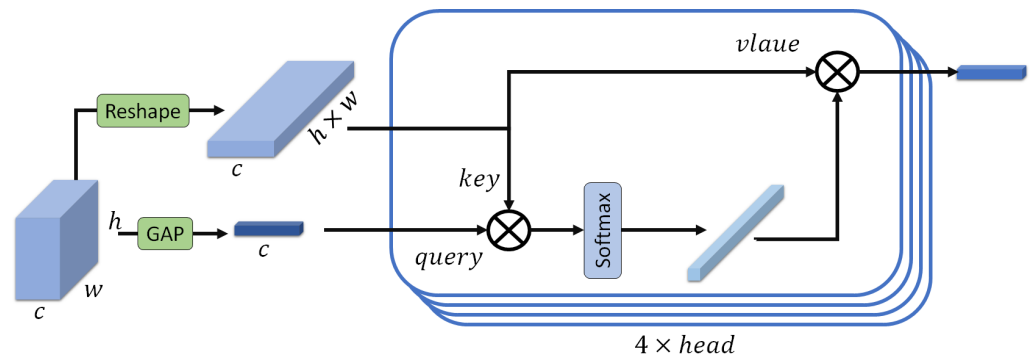
$$E^{\text{merged}} = \text{CRAM}(B^{\text{merged}}) \quad (12)$$

$$E^4 = \text{CRAM}(B'^4) \quad (13)$$

where  $E^{\text{merged}}$  denotes the enhanced multi-scale features from  $B'^2$  and  $B'^3$ , while  $E^4$  represents the refined high-level semantic features from  $B'^4$ .

### 3.3.3. Transformer

In our proposed method, a transformer is employed to enhance the feature representation obtained from the ResNet backbone, and a similar transformer architecture is used within the TransUNet model. For the classification task, we leverage this transformer to refine the multi-scale feature maps extracted from the ResNet backbone. We begin with two feature maps of size  $(64, 64, 64)$  derived from the ResNet backbone:  $E^4$  and  $E^{\text{merged}}$ . Each of these feature maps is fed into a separate transformer, although the structure of the transformers is identical. The entire process is visually represented in Figure 7, providing a detailed overview of the transformer's operation on the feature maps. For simplicity, we will describe the process for  $E^4$ .



**Figure 7.** The overview of the transformer.

#### Global Average Pooling (GAP)

We apply global average pooling to  $E^4$  to create a feature vector  $V^4$  (Equation (14)).

$$V^4 = \text{GAP}(E^4) \quad (14)$$

This vector serves as the query for the transformer with size of  $(64, 1, 1)$ .

#### Reshaping for Key and Value

The feature map  $E^4$  is then reshaped to form the key and value inputs for the transformer. Specifically, as shown in Equation (15),  $E^4$  is reshaped from  $(c, h, w)$  to  $(c, h \times w)$ .

$$E_{flat}^4 = \text{Reshape}(E^4) \quad (15)$$

### Attention Mechanism

Query( $Q$ ): The query is obtained from the feature vector  $V^4$  created by global average pooling.

Key( $K$ ) and Value( $V$ ): Both the key and value are derived from the reshaped feature map  $E_{flat}^4$  with a size of  $(64, 64 \times 64)$ .

### Scaled Dot-Product Attention

The attention scores are computed as the dot product of the query and key, followed by a softmax operation to obtain the attention weights (Equation (16)). Each Transformer block in our model includes a multi-head self-attention layer with four heads. The choice of four attention heads is motivated by several factors. Using too few heads (e.g., a single head) limits the model's ability to capture diverse and complementary feature relationships, as it may focus on only one dominant pattern. In contrast, using too many heads can make attention maps noisy, especially when training on limited data. Moreover, several state-of-the-art studies have also adopted four attention heads when working with datasets of a similar scale to ours.

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) \cdot V \quad (16)$$

where  $d_k$  is the dimensionality of the key, which is computed as the number of channels (64 here) divided by the number of attention heads (4 here), and  $Q \cdot K^T$  represents the dot product of the query and the transposed key. The result is a weighted sum of the value vectors, producing an output feature vector  $F^4$  of size equal to the channel dimension.

#### 3.3.4. Output Feature Vector

The output of the transformer for  $E^4$  is a feature vector  $F^4$  of size equal to the number of channels (64 in this case).

The same process is applied to  $E^{merged}$ , resulting in another feature vector  $F^{merged}$  of size 64.

#### 3.3.5. Find Correct Class

After processing the feature maps  $F^4$  and  $F^{merged}$  through transformers, we concatenate these outputs to form a unified feature representation (Equation (17)).

$$F^{concat} = \text{cat}(F^4, F^{merged}) \quad (17)$$

The concatenated feature vector  $F^{concat}$  is then flattened into a one-dimensional vector. The flattened feature vector is processed through a dense (fully connected) layer followed by a sigmoid activation function for binary classification.

#### 3.3.6. Loss Function

We employ binary cross-entropy loss to train the classifier. This loss function measures the discrepancy between predicted probabilities and true labels for binary classification tasks (Equation (18)).

$$\text{Loss} = -\frac{1}{N} \sum_{i=1}^N [y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)], \quad (18)$$

where  $N$  is the number of samples,  $y_i$  is the true label (0 or 1), and  $\hat{y}_i$  is the predicted probability.

## 4. Experimental Result

### 4.1. Dataset

In this research, we utilized several datasets to effectively train and evaluate our models for both segmentation and classification tasks. For training and validating the TransUNet model, we used the “Chest Xray Masks and Labels” dataset [13,14]. This dataset contains 714 chest X-ray images, accompanied by their masks. Due to data limitations, our segmentation model was trained on 690 images and their corresponding masks, with 24 images reserved for validation purposes. We acknowledge that the relatively small number of training samples may constrain the model’s generalization capability, highlighting a key limitation of the segmentation study.

For the classification task, we used two datasets. COVID-19 Image Data Collection provided by “Cohen” [16]. This dataset comprises a total of 6432 images, including three classes: Pneumonia, COVID-19, and Normal. The dataset is notably challenging due to its class imbalance and the complexity introduced by the three distinct classes. The distribution of images in the training set is as follows: 3418 images of Pneumonia, 1266 images of Normal, and 460 images of COVID-19. The remaining data were used for testing as follows: Pneumonia class includes 855 images, 317 images for the Normal class, and 116 images for the COVID-19 class.

Pediatric Pneumonia Chest X-ray Dataset provided by Kermany et al. [15]. This dataset includes 5856 images, with 5232 images used for training and the remaining images reserved for testing. The dataset presents a significant challenge due to its class imbalance, with 3883 images labeled as Pneumonia and 1349 images as Normal. Additionally, the images in this dataset are from children, who often experience discomfort during X-ray procedures. This discomfort can impact the quality and consistency of the images, and the physiological differences between children and adults add an extra layer of complexity to the classification task. Both datasets contribute valuable and complementary challenges to our classification task, ensuring that our model is robust and capable of handling various real-world scenarios.

### 4.2. Data Augmentation

To enhance model generalization and reduce overfitting, many data augmentation techniques were applied to both classification and segmentation tasks. Specifically, we used random rotation and color jitter. Furthermore, all images were resized to  $512 \times 512$  pixels for consistency across datasets. In the classification tasks, images were additionally normalized using ImageNet statistics to align with the ResNet backbone.

### 4.3. Experimental Setting

We implemented our proposed method using PyTorch version 1.8.1. In the classification task, we utilized pre-trained ResNet-50 and ResNet-101 models, which were kept frozen during training to preserve their learned representations. Notably, the learnable parameters of our method amount to only 2.29 million. The model was trained for 30 epochs, and this process was repeated five times to ensure the robustness of the results. The average of these results was then reported.

The Adam optimizer was employed for training, with a learning rate set to  $10^{-5}$ . All input images were resized to  $512 \times 512$  pixels, and the batch size was set to 64. The model used about 16.2 GB of GPU RAM during these processes. All experiments were conducted on an NVIDIA RTX 4090 GPU.

#### 4.4. Evaluation Metrics

To evaluate the effectiveness of our proposed approach, we use several performance metrics, including:

**Accuracy:** This metric reflects the ratio of correctly identified instances to the total number of instances. It is determined using Equation (19).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (19)$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. Accuracy offers a broad overview of the classifier's performance but can be deceptive when dealing with imbalanced datasets.

**Precision:** This metric quantifies the ratio of correctly predicted positive instances to the total predicted positives. It is represented by Equation (20).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (20)$$

Precision is especially valuable when the consequence of false positives is significant.

**Recall:** Also referred to as Sensitivity or True Positive Rate, Recall measures the ratio of correctly predicted positive instances to the total actual positives. It is expressed by Equation (21).

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (21)$$

Recall is vital in situations where the cost of missing a positive instance (false negatives) is high, ensuring that most positive instances are detected.

**F1 Score:** The F1 Score represents the harmonic mean of Precision and Recall, offering a balance between these two metrics. It is particularly advantageous when handling imbalanced datasets. The F1 Score is determined using Equation (22).

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (22)$$

This score provides a single metric that accounts for both false positives and false negatives, reflecting the classifier's overall performance.

**MCC:** The Matthews Correlation Coefficient (MCC) is a robust metric for evaluating segmentation performance, especially in imbalanced datasets. It considers true positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs) to provide a balanced measure of classification quality. As stated in [44], MCC is particularly useful for tasks where class imbalance is a concern, as it accounts for all four categories of the confusion matrix. MCC is calculated using Equation (23).

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP}) \times (\text{TP} + \text{FN}) \times (\text{TN} + \text{FP}) \times (\text{TN} + \text{FN})}} \quad (23)$$

**Dice Coefficient:** The Dice Coefficient, also known as the F1-score for segmentation tasks, measures the overlap between the predicted segmentation and the ground truth. It is particularly useful for evaluating the accuracy of region-based segmentation, such as lung segmentation in chest X-rays. The Dice Coefficient is calculated using Equation (24):

$$\text{Dice} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}} \quad (24)$$

#### 4.5. Comparison with State-of-the-Art

##### 4.5.1. Segmentation

Our lightweight transformer-enhanced TransUNet, initially designed to improve the accuracy of pneumonia classification by providing precise lung segmentation, also demonstrates superior performance in the segmentation task itself. While the primary goal of our segmentation model was to isolate lung regions for better classification, as shown in Table 1, its ability to outperform state-of-the-art segmentation methods highlights the effectiveness of our design choices. Although some works such as RU-UNET [45] achieved better precision than our model in terms of full metrics and especially Dice, which is more important for segmentation tasks, our model outperforms those methods. By integrating transformer-based attention mechanisms into the U-Net architecture, our model captures both local fine-grained details and global contextual information, which are essential for accurate lung segmentation. To ensure the robustness and reliability of our results, we trained and tested the model three times, with consistent performance across all runs. This dual capability, enhancing both segmentation and classification, sets our approach apart from existing methods, which often focus on one task at the expense of the other.

**Table 1.** Performance on “Chest X-ray Masks and Labels” dataset for segmentation task. Numbers in bold represent the best performance, while underlined values denote the second-best performance.

Model	Dice	Accuracy	Precision	Recall	F1-Score	MCC	Parameter
Unet [20]	93.46%	96.88%	97.41%	89.83%	93.46%	91.55%	31.2M
RU-Net [45]	92.07%	96.34%	<b>99.58%</b>	85.61%	92.07%	90.13%	44.2M
ResNet34-Unet [46]	93.83%	97.06%	98.13%	89.89%	93.83%	92.06%	28.5M
BCDU-Net [47]	94.14%	97.20%	98.25%	90.37%	94.14%	92.44%	65.2M
ResBCDUnet [48]	94.34%	97.31%	<u>98.89%</u>	90.20%	94.34%	92.75%	-
NasNet [49]	94.95%	97.52%	96.55%	<u>93.42%</u>	94.95%	93.33%	-
DABT-U-Net [50]	95.11%	97.64%	98.25%	92.16%	95.11%	93.64%	-
ABANet [51]	95.25%	97.71%	98.53%	92.18%	95.25%	93.84%	-
FusionLungNet [52]	<u>95.29%</u>	<u>97.73%</u>	98.66%	92.14%	<u>95.29%</u>	<u>93.89%</u>	48.1M
Our	<b>95.7% ± 0.5</b>	<b>97.9% ± 0.5</b>	97.5% ± 0.6	<b>93.9% ± 0.6</b>	<b>95.7% ± 0.5</b>	<b>94.3% ± 0.6</b>	43.3M

##### 4.5.2. Classification

We evaluated the performance of our proposed method on the “Cohen” and “Kermany” datasets, comparing it against several state-of-the-art methods. As shown in Tables 2 and 3, our approach outperformed existing methods across all evaluation metrics, including accuracy, precision, recall, and F1-score, using both ResNet-50 and ResNet-101 backbones. On the “Cohen” dataset, our method achieved an accuracy of 96.04% with ResNet-50 and 95.19% with ResNet-101, surpassing the best-performing existing method (Zhao et al. [53]) by a significant margin. Similarly, on the “Kermany” dataset, our method achieved an accuracy of 93.75% with ResNet-101 and 91.67% with ResNet-50, outperforming recent transformer-based models like MP-ViT [54] and ViT [55].

**Table 2.** Performance on “Cohen” dataset. Numbers in bold represent the best performance, while underlined values denote the second-best performance.

Models	Accuracy	Precision	Recall	F1-Score
Densenet121 *	87.8%	53.9%	71.0%	61.27%
Densenet169 *	87.1%	32.3%	65.6%	43.28%
Densenet201 *	88.4%	51.9%	79.0%	62.64%
Mobilenet_v2 *	86.9%	33.4%	75.0%	46.21%
ResNet-50 *	87.1%	38.4%	71.0%	49.84%
ResNet-101 *	87.9%	33.5%	73.0%	45.92%
Swin Transformer V2 [40]	92.55%	93.79%	88.46%	90.82%
Goodwinet al. (Ensemble learning) [56]	89.4%	53.3%	80.0%	63.97%
Gadza et al. [57]	84.9%	77.4%	90.6%	83.48%
Zhao et al. (Channel-Attention Capsule) [53]	90.43%	90.81%	90.43%	90.40%
CNN-based [58]	92.52%	-	-	-
CNN-based [58]	91.05%	-	-	-
Proposed method (ResNet-50 as backbone)	<b>96.04%</b> ± 0.5	<b>96.70%</b> ± 0.5	<b>94.90%</b> ± 0.6	<b>95.77%</b> ± 0.5
Proposed method (ResNet-101 as backbone)	<u>95.19%</u> ± 0.6	<u>95.59%</u> ± 0.5	<u>94.19%</u> ± 0.6	<u>94.86%</u> ± 0.6

Models marked with “\*” have results directly reported from [56].

**Table 3.** Performance on “Kermany” dataset. Numbers in bold represent the best performance, while underlined values denote the second-best performance.

Models	Accuracy	Precision	Recall	F1-Score
Yadav et al. (VGG16 as backbone) [59]	88.50%	-	-	-
Ayan et al. (VGG16 as backbone) [60]	87.98%	82.72%	85.90%	84.28%
Chattopadhyay et al. [61]	81.7%	-	-	80.6%
Bhatt et al. (CNN) [62]	85.58%	83.33%	96.15%	89.29%
Reshan et al. (MobileNet as backbone) [63]	90.85%	91.41%	95.28%	91.41%
Reshan et al. (ResNet152B2 as backbone) [63]	84.65%	82.38%	99.21%	90.02%
Reshan et al. (DenseNet121 as backbone) [63]	88.90%	88.33%	<b>96.87%</b>	92.41%
Reshan et al. (Xception as backbone) [63]	87.59%	91.75%	90.32%	91.03%
Reshan et al. (EfficientNet as backbone) [63]	51.02%	86.21%	45.85%	90.10%
Jiang et al. (MP-ViT) [54]	91.19%	91.82%	89.36%	90.34%
ViT in [55]	<u>92.45%</u>	<u>92.47%</u>	92.44%	92.47%
Proposed method (ResNet-50 as backbone)	91.67% ± 0.6	92.04% ± 0.5	94.87% ± 0.5	<u>93.43%</u> ± 0.6
Proposed method (ResNet-101 as backbone)	<b>93.75%</b> ± 0.5	<b>93.98%</b> ± 0.5	<u>96.16%</u> ± 0.5	<b>95.05%</b> ± 0.5

The superior performance of our method can be attributed to several key design choices. First, by leveraging pre-trained ResNet models, we extract robust multi-scale

feature representations that capture both low-level and high-level patterns in chest X-rays. Freezing the pre-trained layers allows the model to focus on learning task-specific features in the newly added layers, reducing overfitting and improving generalization. Second, CRAM refines the extracted features through dual attention mechanisms, enhancing discriminative local patterns while suppressing irrelevant information. Our lightweight transformer module then enhances the model’s ability to capture global contextual information, which is critical for distinguishing subtle pneumonia-related patterns from complex backgrounds. This integration of local and global feature extraction ensures robust performance across diverse datasets.

Furthermore, our method demonstrates consistent performance across both datasets, highlighting its generalizability and adaptability to different clinical settings. For instance, on the “Cohen” dataset, our method achieved an F1-score of 95.77%, significantly higher than the 83.48% reported by Gadza et al. [57]. Similarly, on the “Kermany” dataset, our method achieved an F1-score of 95.05%, outperforming the 92.41% achieved by Reshan et al. [63] using DenseNet121. These results underscore the effectiveness of our approach in addressing the challenges of pneumonia classification, such as class imbalance, subtle visual indicators, and domain shifts between datasets.

In summary, our proposed method not only achieves state-of-the-art performance but also demonstrates computational efficiency and generalizability, making it a promising solution for real-world clinical applications. The integration of pre-trained ResNet models with a lightweight transformer module provides a robust framework for accurate and efficient pneumonia detection, addressing the limitations of existing methods.

#### 4.6. Ablation Study

To evaluate the impact of the segmentation component on the performance of our classification model, we conducted an ablation study using the “Cohen” dataset. This study compares the classification results obtained with and without the segmentation step, providing insights into the effectiveness of incorporating lung masks generated by the TransUNet model.

The ablation study involves evaluating the classification performance of our model with two different input scenarios: (1) Original Images: Classification is performed directly on the raw chest X-ray images from the “Cohen” dataset. (2) Predicted Masks: Classification is performed on the chest X-ray images after segmenting the lung regions using the TransUNet model. The images used for classification are limited to the areas highlighted by the predicted lung masks.

The results of the ablation study are summarized in Table 4. The table displays classification metrics, including accuracy, precision, recall, and F1-score, for both ResNet-50 and ResNet-101 backbones under the two different input scenarios.

**Table 4.** Effect of segmentation on “Cohen” dataset.

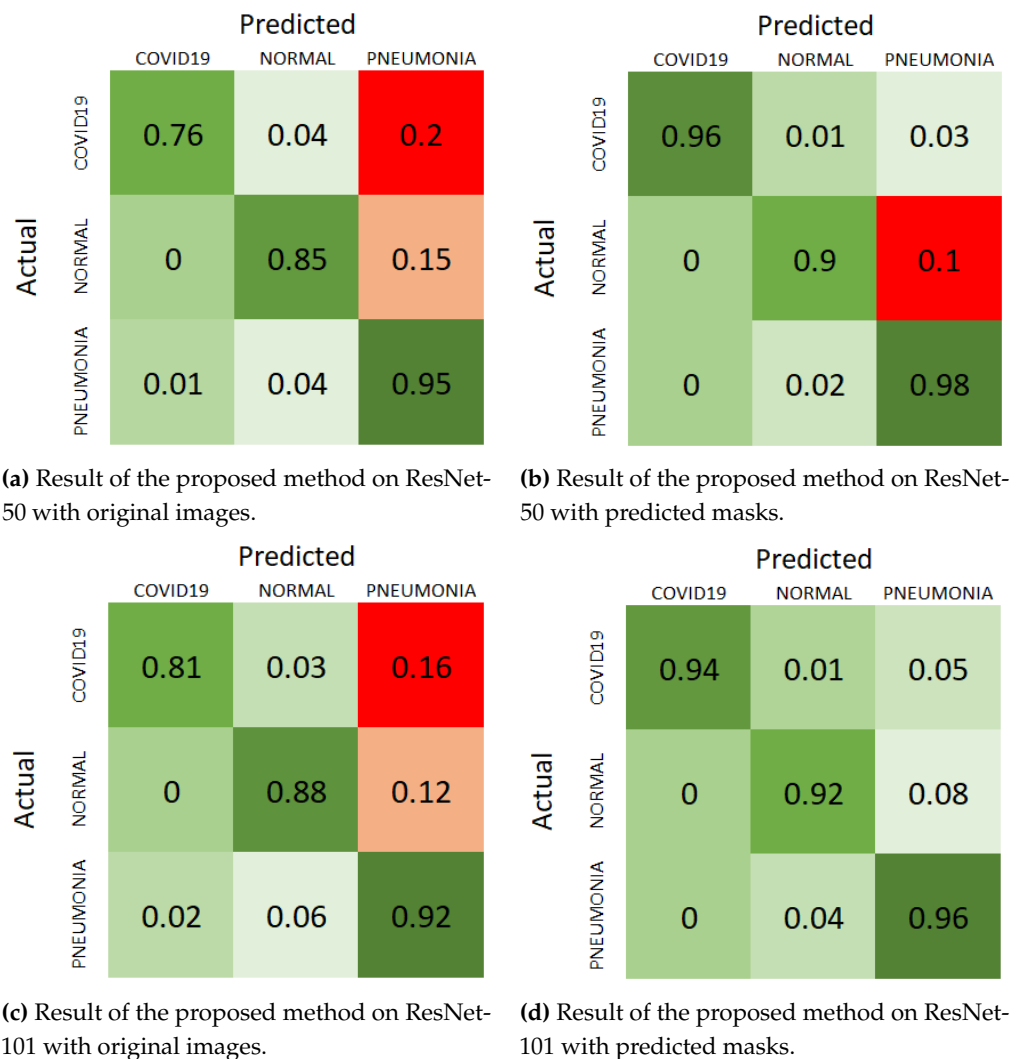
Backbones	Results of Proposed Method	Accuracy	Precision	Recall	F1-Score
ResNet-50	on original images	91.23%	90.94%	85.60%	88.19%
	on predicted masks	<b>96.04%</b>	<b>96.70%</b>	<b>94.90%</b>	<b>95.77%</b>
ResNet-101	on original images	90.22%	88.16%	87.04%	87.6%
	on predicted masks	<b>95.19%</b>	<b>95.59%</b>	<b>94.19%</b>	<b>94.86%</b>

The results clearly demonstrate the benefit of incorporating segmentation masks in the classification process. For both ResNet-50 and ResNet-101 backbones, the model achieves

higher accuracy, precision, recall, and F1-score when trained on images with predicted lung masks compared to the raw images. Specifically, the accuracy improves by almost five percentage points for both ResNet-50 and ResNet-101. Similarly, the precision, recall, and F1-score all show substantial improvements.

Furthermore, the segmentation step improves classification performance by ensuring that the model focuses exclusively on the clinically relevant lung regions and discards non-diagnostic background information. By isolating the lungs, the model can better concentrate on pathological areas that are essential for identifying pneumonia.

These findings underscore the effectiveness of the segmentation component in isolating relevant features within the lung regions, which enhances the classification model's ability to accurately diagnose pneumonia. By focusing on the segmented lung areas, the classification model benefits from reduced noise and more relevant information, leading to better overall performance. In addition to the classification metrics presented in Table 4, we further analyze the performance of our model using confusion matrices under different scenarios, as illustrated in Figure 8. For better and easier understanding, we used a normalized version of the confusion matrices, which provide a detailed breakdown of the classification results, showing the true positives, true negatives, false positives, and false negatives for each class.



**Figure 8.** Normalize confusion matrix of “Cohen” dataset with different scenarios.

For the ResNet-50 backbone, the confusion matrix in Figure 8a shows the results on original images. The model correctly classifies 0.76 COVID-19 cases, 0.81 normal cases,

and 0.95 pneumonia cases, with a notable number of misclassifications, particularly in the COVID-19 and pneumonia categories. When using predicted masks, as shown in Figure 8b, the model’s performance improves significantly, correctly classifying 0.96 COVID-19 cases, 0.9 normal cases, and 0.98 pneumonia cases. The number of misclassifications decreases across all categories, highlighting the benefit of segmentation in isolating relevant features.

For the ResNet-101 backbone, the confusion matrix in Figure 8c displays the results on original images, where the model correctly classifies 0.81 COVID-19 cases, 0.88 normal cases, and 0.92 pneumonia cases. However, the misclassifications are more pronounced compared to ResNet-50, particularly in the pneumonia category. With the predicted masks, as shown in Figure 8d, the performance improves, with correct classifications of 0.94 COVID-19 cases, 0.92 normal cases, and 0.96 pneumonia cases. This reduction in misclassifications further supports the effectiveness of incorporating segmentation masks.

These visual representations in the confusion matrices clearly demonstrate the improvement in classification performance when using the predicted masks. The consistent reduction in false positives and false negatives across both backbones underscores the robustness of the segmentation approach. By focusing on the lung regions and eliminating irrelevant background information, the segmentation component enhances the model’s ability to accurately diagnose pneumonia, resulting in better overall performance.

Furthermore, we investigate the contributions of key components within our model on the “Cohen” dataset, specifically focusing on the impact of multi-scale feature maps and the transformer module. The results are presented in Table 5, highlighting the performance metrics, including accuracy, precision, recall, and F1-score, across different configurations of the ResNet-50 and ResNet-101 backbones. Each row of the table demonstrates how the incorporation of each component affects the model’s performance, providing a comprehensive view of their individual and combined effects.

**Table 5.** The effect of using each key component on the “Cohen” dataset (on original images and without segmentation).

Backbones	Baseline	Multi-Scale Feature Maps	CRAM + Transformer	Accuracy	Precision	Recall	F1-Score	Training Time (per Image)	Inference Time	Learnable Parameters
ResNet-50	X			84.62% ± 0.04	75.68% ± 0.05	75.10% ± 0.06	75.39% ± 0.04	2.7 ms	0.8 ms	0.65 M
	X	X		87.73% ± 0.03	80.55% ± 0.04	79.62% ± 0.05	80.08% ± 0.04	2.8 ms	0.8 ms	0.85 M
	X		X	87.73% ± 0.03	80.63% ± 0.04	80.47% ± 0.03	80.55% ± 0.03	3.5 ms	1.2 ms	1.15 M
	X	X	X	<b>91.23% ± 0.02</b>	<b>90.94% ± 0.03</b>	<b>85.60% ± 0.03</b>	<b>88.19% ± 0.03</b>	12.2 ms	5.4 ms	2.29 M
ResNet-101	X			83.93% ± 0.04	74.21% ± 0.05	73.92% ± 0.04	74.06% ± 0.04	9.7 ms	2.9 ms	0.65 M
	X	X		85.56% ± 0.03	78.21% ± 0.04	80.40% ± 0.05	79.29% ± 0.04	12.3 ms	3.6 ms	0.85 M
	X		X	85.71% ± 0.03	78.59% ± 0.04	81.11% ± 0.04	79.83% ± 0.03	14.6 ms	5.3 ms	1.15 M
	X	X	X	<b>90.22% ± 0.03</b>	<b>88.16% ± 0.03</b>	<b>87.04% ± 0.03</b>	<b>87.60% ± 0.03</b>	43.7 ms	19.1 ms	2.29 M

The findings reveal a significant enhancement in model performance when both multi-scale feature maps and the transformer are employed alongside the baseline configuration.

For instance, with the ResNet-50 backbone, the accuracy improves from 84.62% (baseline) to 91.23% when all components are utilized. Similarly, the ResNet-101 backbone exhibits a notable increase in accuracy from 83.93% to 90.22%. These results underscore the effectiveness of our proposed innovations, illustrating that the integration of multi-scale feature maps and transformer elements not only enhances overall accuracy but also boosts precision, recall, and F1-score, which are crucial for the reliability of classification tasks. This highlights the importance of these key components in achieving improved performance in deep learning models for image analysis. Additionally, the inclusion of training time (per image) and learnable parameters underscores the computational efficiency and scalability of our approach, ensuring its suitability for resource-constrained environments.

We also evaluated three efficiency-related metrics to assess the Lightness of our proposed method: the number of parameters, training time per image, and inference time

per image. The results for the number of parameters count indicate that, although the number of parameters rises after adding each component but even the full model only uses 2.29 million parameters, maintaining a lightweight design. Moreover, the training time per image and inference times show only slight increases. When using the ResNet-50 backbone, training time per image rises from 0.27 ms to 1.22 ms, and inference time increases from 0.08 ms to 0.54 ms. Similarly, with the ResNet-101 backbone, training and inference times increase from 0.97 ms and 0.29 ms to 4.37 ms and 1.91 ms, respectively. Despite these increases, both configurations remain computationally feasible and efficient for clinical deployment.

Additionally, we conducted an ablation study to evaluate the contribution of each ResNet block (Blocks 2, 3, and 4) to the classification performance. Results, as shown in Table 6, demonstrate that intermediate layers (Blocks 2 and 3) play a significant role in improving performance, but their contribution is not as critical as that of Block 4. For instance, using only Block 4 achieves an accuracy of 84.62%, while adding Block 3 (without Block 2) improves the accuracy to 86.64%. Furthermore, we evaluate the condition where instead of using blocks 2 and 3, we concatenate blocks 3 and 4. In this situation, the model does not perform as well as when choosing blocks 2 and 3. This indicates that mid-level features from Block 3 provide additional discriminative information, enhancing the model's ability to detect subtle patterns in chest X-rays. However, the inclusion of Block 2, which captures low-level features, results in only a marginal improvement in accuracy (86.72%), suggesting that while low-level features are beneficial, they are less impactful compared to mid- and high-level features.

In our final design, we concatenate the feature maps from Blocks 2 and 3 because both capture mid-level semantic representations. Merging these two feature maps enables the model to integrate richer mid-level features and establish implicit relationships between them without introducing additional modules. In contrast, Block 4 generates high-level semantic features, which are better preserved in their original form at this stage. Processing Block 4 separately helps maintain its global contextual information and prevents spatial misalignment during subsequent fusion or decision phases.

**Table 6.** The effect of using different ResNet blocks on the “Cohen” dataset.

Backbones	Block 2	Block 3	Block 4	Accuracy	Precision	Recall	F1-Score
ResNet-50			X	84.62%	75.68%	75.10%	75.39%
	X		X	85.86%	77.03%	74.37%	75.68%
		X	X	86.64%	79.36%	76.15%	77.72%
	X	X	X	86.72%	79.02%	76.48%	77.73%
	X	Combined		86.96%	78.83%	79.23%	79.02%
	Combined		X	<b>87.73%</b>	<b>80.55%</b>	<b>79.62%</b>	<b>80.08%</b>

Interestingly, the best performance is achieved when using two feature maps: Block 4 ( $B^4$ ) and the merged feature map ( $B^{merged}$ ), which combines Blocks 2 and 3. This configuration achieves an accuracy of 87.73%, a precision of 80.55%, and an F1-score of 80.08%, outperforming the scenario where all three blocks are used independently. This result highlights the importance of separating high-level features (Block 4) from intermediate features (Blocks 2 and 3). While intermediate layers provide valuable contextual information, they are not as discriminative as the high-level features from Block 4. By merging Blocks 2 and 3 into a single feature map ( $B^{merged}$ ), we reduce redundancy and computational complexity while preserving the benefits of multi-scale feature extraction. This design choice ensures

that the model focuses on the most relevant features for pneumonia detection, leading to improved performance and efficiency.

We also investigate different strategies for our combined blocks. As shown in Table 7, we evaluate our model using concatenation or element-wise addition. Our chosen strategy achieved better performance in terms of classification metrics. Specifically, accuracy rose from 86.18% to 87.73% and F1-score increased from 76.28% to 80.08%.

**Table 7.** The effect of different fusion strategies for combined blocks on the “Cohen” dataset.

Backbone	Concat	Element-Wised Addition	Accuracy	Precision	Recall	F1-Score
ResNet-50		X	86.18%	75.88%	76.68%	76.28%
	X		<b>87.73%</b>	<b>80.55%</b>	<b>79.62%</b>	<b>80.08%</b>

#### 4.7. Explainable AI Through Gradient-Weighted Class Activation Mapping

To enhance the interpretability of our model and provide clinical insights into its decision-making process, we employed Gradient-Weighted Class Activation Mapping (Grad-CAM) visualizations. Explainable AI is particularly crucial in medical diagnostics, where understanding the rationale behind a model’s predictions is essential for clinical adoption and trust. Grad-CAM generates heatmaps that highlight the regions of the input image that most significantly influenced the model’s classification decision, effectively creating a visual explanation for the prediction.

In the Grad-CAM visualizations, color intensity represents the level of model attention and contribution to the final prediction. Specifically, warm colors such as red indicate the highest and after that, orange and yellow regions had a strong positive influence on the model’s decision, while cooler colors such as blue indicate low contributions to the classification predictions.

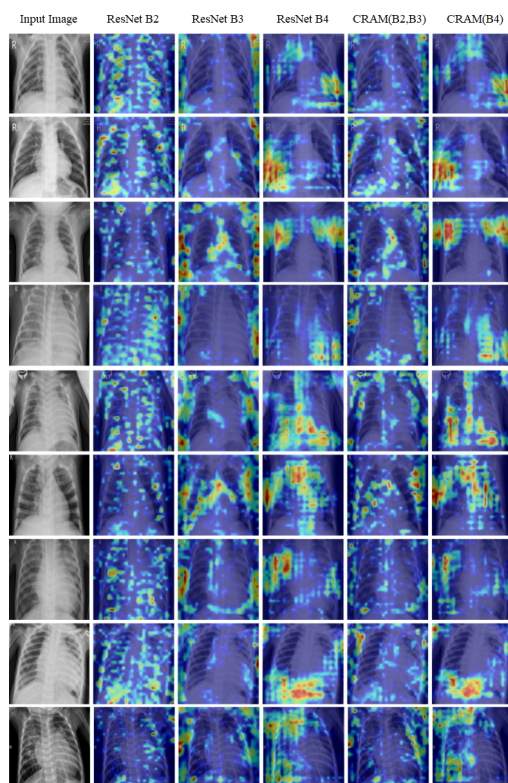
Based on information provided by clinician experts, certain anatomical regions of the chest are more informative for recognizing pneumonia from chest X-ray images [64,65]. Before discussing the Grad-CAM results, we briefly summarize these diagnostically relevant regions:

**Lower Lung Zones (especially posterior and basal regions):** Pneumonia frequently presents as areas of increased opacity or consolidation in the lower lobes, often more visible in the posterior or lateral aspects of the lungs. These appear as white and cloud-like regions, obscuring the normal dark lung fields.

**Perihilar Regions (central areas around the bronchi):** Radiologists assess the perihilar and mid-lung zones for patchy or reticular opacities, which may indicate bronchopneumonia or viral pneumonia.

**Costophrenic and Cardiophrenic Angles:** Blunting or loss of sharpness in these corners of the lungs can suggest pleural effusion or adjacent inflammatory changes secondary to pneumonia.

As shown in Figure 9, our Grad-CAM analysis reveals the hierarchical feature learning process across different network depths. The first column displays the original input chest X-ray image, providing the baseline for comparison. The subsequent three columns visualize the activation maps from the feature maps extracted at different ResNet blocks: Block 2 (capturing low-level features and basic patterns), Block 3 (capturing mid-level features and structural information), and Block 4 (capturing high-level semantic features and complex patterns). The final two columns present the activation maps from our CRAM: the concatenated feature map (merging Blocks 2 and 3) and the enhanced Block 4 feature map.



**Figure 9.** Grad-CAM visualizations on Kermany dataset samples.

Overall, these findings highlight that the most critical diagnostic regions are located in the middle and lower lung areas, particularly those encompassing the bone structures and surrounding soft tissue. As discussed in the following section, our model's Grad-CAM visualizations demonstrate similar behavior, focusing attention on these clinically relevant regions, which supports the interpretability and clinical alignment of the proposed framework.

## 5. Conclusions

This paper presents an innovative and efficient method for pneumonia detection utilizing a novel multi-scale transformer approach. By integrating lung segmentation using the TransUNet model with a specialized transformer module, our approach effectively isolates lung regions, thereby enhancing the performance of subsequent classification tasks. The proposed method demonstrates significant improvements in classification metrics, as evidenced by the ablation study on the “Cohen” dataset. Both ResNet-50 and ResNet-101 backbones benefited from the segmentation masks, showing increased accuracy, precision, recall, and F1-score. These improvements underscore the effectiveness of our approach in focusing on relevant lung features while reducing noise from irrelevant regions.

The high accuracy rates of 93.75% on the “Kermany” dataset and 96.04% on the “Cohen” dataset confirm the robustness and reliability of our model. The reduction in the number of parameters compared to other state-of-the-art transformer models highlights our contribution to creating a more efficient yet powerful diagnostic tool suitable for deployment in resource-constrained environments. Our work paves the way for future research in several areas. Future work could explore further optimization of the transformer module to enhance performance and reduce computational complexity. Additionally, in this work, we evaluate our model on two datasets that were both trained and tested. For more generalization, future works can test models on unseen datasets to further solve this limitation.

**Author Contributions:** Conceptualization, A.S. (Alireza Saber) and A.F.; data collection, A.S. (Alireza Saber); formal analysis, M.F. and S.F.; investigation, P.P. and A.S. (Alimohammad Siahkarzadeh); methodology, A.S. (Alireza Saber) and M.F.; project administration, M.F. and S.F.; resources, S.F.; software, A.S. (Alireza Saber) and P.P.; supervision, A.F., M.F.; validation, M.F. and S.F.; visualization, A.F. and A.S. (Alimohammad Siahkarzadeh); writing—original draft, A.S. (Alireza Saber) and A.F.; writing—review & editing, M.F., S.F. and A.F. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The datasets used in this study are publicly available and were sourced from the following repositories: Kermany dataset: <https://www.kaggle.com/datasets/andrewmvd/pediatric-pneumonia-chest-xray> (accessed on 23 November 2025). Cohen dataset: <https://github.com/ieee8023/covid-chestxray-dataset> (accessed on 23 November 2025). Chest Xray Masks and Labels dataset: <https://www.kaggle.com/datasets/nikhilpandey360/chest-xray-masks-and-labels> (accessed on 23 November 2025). Also The source code of this study is publicly available at: <https://github.com/amirrezafateh/Multi-Scale-Transformer-Pneumonia> (accessed on 23 November 2025).

**Acknowledgments:** During the preparation of this manuscript, the authors used (ChatGPT-4.5) only to improve grammar and readability. After that, all authors reviewed the contents and they take full responsibility for the content of this publication.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. UNICEF. Pneumonia. Available online: <https://data.unicef.org/topic/child-health/pneumonia/> (accessed on 15 November 2023).
2. World Health Organization. Pneumonia in Children. Available online: <https://www.who.int/news-room/fact-sheets/detail/pneumonia> (accessed on 8 November 2022).
3. Action, H.P. Key Facts: Poverty and Poor Health. Available online: <https://www.healthpovertyaction.org/news-events/key-facts-poverty-and-poor-health/> (accessed on 15 January 2018).
4. RadiologyInfo. Chest X-Ray. Available online: <https://www.radiologyinfo.org/en/info/chestrad> (accessed on 17 November 2022).
5. Lin, M.; Hou, B.; Mishra, S.; Yao, T.; Huo, Y.; Yang, Q.; Wang, F.; Shih, G.; Peng, Y. Enhancing thoracic disease detection using chest X-rays from PubMed Central Open Access. *Comput. Biol. Med.* **2023**, *159*, 106962. [[CrossRef](#)]
6. Askari, F.; Fateh, A.; Mohammadi, M.R. Enhancing few-shot image classification through learnable multi-scale embedding and attention mechanisms. *Neural Netw.* **2025**, *187*, 107339. [[CrossRef](#)] [[PubMed](#)]
7. Fateh, A.; Mohammadi, M.R.; Motlagh, M.R.J. MSDNet: Multi-Scale Decoder for Few-Shot Semantic Segmentation via Transformer-Guided Prototyping. *Image Vis. Comput.* **2025**, *162*, 105672. [[CrossRef](#)]
8. Song, Z.; Wu, W.; Wu, S. Multi-Scale Convolutional Attention and Structural Re-Parameterized Residual-Based 3D U-Net for Liver and Liver Tumor Segmentation from CT. *Sensors* **2025**, *25*, 1814. [[CrossRef](#)] [[PubMed](#)]
9. Junia, R.C.; Selvan, K. Deep learning-based automatic segmentation of COVID-19 in chest X-ray images using ensemble neural net sentinel algorithm. *Meas. Sens.* **2024**, *33*, 101117. [[CrossRef](#)]
10. Shavkatovich Buriboev, A.; Abduvaitov, A.; Jeon, H.S. Binary Classification of Pneumonia in Chest X-Ray Images Using Modified Contrast-Limited Adaptive Histogram Equalization Algorithm. *Sensors* **2025**, *25*, 3976. [[CrossRef](#)]
11. Buriboev, A.S.; Muhamediyeva, D.; Primova, H.; Sultanov, D.; Tashev, K.; Jeon, H.S. Concatenated CNN-based pneumonia detection using a fuzzy-enhanced dataset. *Sensors* **2024**, *24*, 6750. [[CrossRef](#)]
12. Kanwal, K.; Asif, M.; Khalid, S.G.; Liu, H.; Qurashi, A.G.; Abdullah, S. Current diagnostic techniques for pneumonia: A scoping review. *Sensors* **2024**, *24*, 4291. [[CrossRef](#)]
13. Candemir, S.; Jaeger, S.; Palaniappan, K.; Musco, J.P.; Singh, R.K.; Xue, Z.; Karargyris, A.; Antani, S.; Thoma, G.; McDonald, C.J. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE Trans. Med. Imaging* **2013**, *33*, 577–590. [[CrossRef](#)]

14. Jaeger, S.; Karargyris, A.; Candemir, S.; Folio, L.; Siegelman, J.; Callaghan, F.; Xue, Z.; Palaniappan, K.; Singh, R.K.; Antani, S.; et al. Automatic tuberculosis screening using chest radiographs. *IEEE Trans. Med. Imaging* **2013**, *33*, 233–245. [[CrossRef](#)]
15. Kermany, D.; Zhang, K.; Goldbaum, M. Labeled optical coherence tomography (oct) and chest x-ray images for classification. *Mendeley Data* **2018**, *2*, 651.
16. Cohen, J.P.; Morrison, P.; Dao, L. COVID-19 image data collection. *arXiv* **2020**, arXiv:2003.11597.
17. Jennifer, J.S.; Sharmila, T.S. A neutrosophic set approach on chest X-rays for automatic lung infection detection. *Inf. Technol. Control.* **2023**, *52*, 37–52. [[CrossRef](#)]
18. Fateh, A.; Rezvani, M.; Tajary, A.; Fateh, M. Persian printed text line detection based on font size. *Multimed. Tools Appl.* **2023**, *82*, 2393–2418. [[CrossRef](#)]
19. Fateh, A.; Rezvani, M.; Tajary, A.; Fateh, M. Providing a voting-based method for combining deep neural network outputs to layout analysis of printed documents. *J. Mach. Vis. Image Process.* **2022**, *9*, 47–64.
20. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; pp. 234–241.
21. Liu, W.; Luo, J.; Yang, Y.; Wang, W.; Deng, J.; Yu, L. Automatic lung segmentation in chest X-ray images using improved U-Net. *Sci. Rep.* **2022**, *12*, 8649. [[CrossRef](#)]
22. Huang, H.; Lin, L.; Tong, R.; Hu, H.; Zhang, Q.; Iwamoto, Y.; Han, X.; Chen, Y.W.; Wu, J. Unet 3+: A full-scale connected unet for medical image segmentation. In Proceedings of the ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–9 May 2020; pp. 1055–1059.
23. Jha, D.; Riegler, M.A.; Johansen, D.; Halvorsen, P.; Johansen, H.D. Doubleu-net: A deep convolutional neural network for medical image segmentation. In Proceedings of the 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), Rochester, NY, USA, 28–30 July 2020; pp. 558–564.
24. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention u-net: Learning where to look for the pancreas. *arXiv* **2018**, arXiv:1804.03999. [[CrossRef](#)]
25. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306. [[CrossRef](#)]
26. Chen, Y.; Wang, T.; Tang, H.; Zhao, L.; Zhang, X.; Tan, T.; Gao, Q.; Du, M.; Tong, T. CoTrFuse: A novel framework by fusing CNN and transformer for medical image segmentation. *Phys. Med. Biol.* **2023**, *68*, 175027. [[CrossRef](#)] [[PubMed](#)]
27. Stokes, K.; Castaldo, R.; Franzese, M.; Salvatore, M.; Fico, G.; Pokvic, L.G.; Badnjevic, A.; Pecchia, L. A machine learning model for supporting symptom-based referral and diagnosis of bronchitis and pneumonia in limited resource settings. *Biocybern. Biomed. Eng.* **2021**, *41*, 1288–1302. [[CrossRef](#)]
28. Chandra, T.B.; Verma, K. Pneumonia detection on chest x-ray using machine learning paradigm. In Proceedings of the 3rd International Conference on Computer Vision and Image Processing, Jabalpur, India, 29 September–1 October 2018; pp. 21–33.
29. Wang, Y.; Liu, Z.L.; Yang, H.; Li, R.; Liao, S.J.; Huang, Y.; Peng, M.H.; Liu, X.; Si, G.Y.; He, Q.Z.; et al. Prediction of viral pneumonia based on machine learning models analyzing pulmonary inflammation index scores. *Comput. Biol. Med.* **2024**, *169*, 107905. [[CrossRef](#)] [[PubMed](#)]
30. Fateh, A.; Fateh, M.; Abolghasemi, V. Multilingual handwritten numeral recognition using a robust deep network joint with transfer learning. *Inf. Sci.* **2021**, *581*, 479–494. [[CrossRef](#)]
31. Alliou, H.; Mohammed, M.A.; Benameur, N.; Al-Khateeb, B.; Abdulkareem, K.H.; Garcia-Zapirain, B.; Damaševičius, R.; Maskeliūnas, R. A multi-agent deep reinforcement learning approach for enhancement of COVID-19 CT image segmentation. *J. Pers. Med.* **2022**, *12*, 309. [[CrossRef](#)]
32. Stephen, O.; Sain, M.; Maduh, U.J.; Jeong, D.U. An efficient deep learning approach to pneumonia classification in healthcare. *J. Healthc. Eng.* **2019**, *2019*, 4180949. [[CrossRef](#)] [[PubMed](#)]
33. Rajpurkar, P.; Irvin, J.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C.; Shpanskaya, K.; et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv* **2017**, arXiv:1711.05225.
34. Ukwuoma, C.C.; Qin, Z.; Heyat, M.B.B.; Akhtar, F.; Bamisile, O.; Muaad, A.Y.; Addo, D.; Al-Antari, M.A. A hybrid explainable ensemble transformer encoder for pneumonia identification from chest X-ray images. *J. Adv. Res.* **2023**, *48*, 191–211. [[CrossRef](#)]
35. Jaiswal, A.K.; Tiwari, P.; Kumar, S.; Gupta, D.; Khanna, A.; Rodrigues, J.J. Identifying pneumonia in chest X-rays: A deep learning approach. *Meas.* **2019**, *145*, 511–518. [[CrossRef](#)]
36. Gabruseva, T.; Poplavskiy, D.; Kalinin, A. Deep learning for automatic pneumonia detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 350–351.
37. Gholami, M.; Fateh, M.; Fateh, A. Text-Enhanced Semantic Segmentation via Contrastive Language-Image Pretraining Guided Multi-Modal Feature Fusion with Feature Refinement Approach. *Int. J. Eng.* **2026**, *39*, 1422–1437. [[CrossRef](#)]

38. Wang, X.; Yang, S.; Zhang, J.; Wang, M.; Zhang, J.; Huang, J.; Yang, W.; Han, X. Transpath: Transformer-based self-supervised learning for histopathological image classification. In Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2021: 24th International Conference, Strasbourg, France, 27 September–1 October 2021; pp. 186–195.
39. Wu, P.; Chen, J.; Wu, Y. Swin Transformer based benign and malignant pulmonary nodule classification. In Proceedings of the 5th International Conference on Computer Information Science and Application Technology (CISAT 2022), Chongqing, China, 29–31 July 2020; pp. 552–558.
40. Liu, Z.; Hu, H.; Lin, Y.; Yao, Z.; Xie, Z.; Wei, Y.; Ning, J.; Cao, Y.; Zhang, Z.; Dong, L.; et al. Swin transformer v2: Scaling up capacity and resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Los Angeles, CA, USA, 18–24 June 2022; pp. 12009–12019.
41. Mishra, V.R.; Malhotra, R. Empowering healthcare with Swin Transformer V2: Advancing pneumonia diagnosis through deep learning. In Proceedings of the 4th International Conference on Computational Methods in Science & Technology (ICCMST 2024), Mohali, India, 2–3 May 2024; pp. 99–105.
42. Angara, S.; Mannuru, N.R.; Mannuru, A.; Thirunagaru, S. A novel method to enhance pneumonia detection via a model-level ensembling of CNN and vision transformer. *arXiv* **2024**, arXiv:2401.02358. [[CrossRef](#)]
43. Mustapha, B.; Zhou, Y.; Shan, C.; Xiao, Z. Enhanced Pneumonia Detection in Chest X-Rays Using Hybrid Convolutional and Vision Transformer Networks. *Curr. Med. Imaging* **2025**, *21*, e15734056326685. [[CrossRef](#)]
44. Anbalagan, T.; Nath, M.K.; Vijayalakshmi, D.; Anbalagan, A. Analysis of various techniques for ECG signal in healthcare, past, present, and future. *Biomed. Eng. Adv.* **2023**, *6*, 100089. [[CrossRef](#)]
45. Alom, M.Z.; Yakopcic, C.; Taha, T.M.; Asari, V.K. Nuclei segmentation with recurrent residual convolutional neural networks based U-Net (R2U-Net). In Proceedings of the NAECON 2018-IEEE National Aerospace and Electronics Conference, Dayton, OH, USA, 23–26 July 2018; pp. 228–233.
46. Lau, S.L.; Chong, E.K.; Yang, X.; Wang, X. Automated pavement crack segmentation using u-net-based convolutional neural network. *IEEE Access* **2020**, *8*, 114892–114899. [[CrossRef](#)]
47. Azad, R.; Asadi-Aghbolaghi, M.; Fathy, M.; Escalera, S. Bi-directional ConvLSTM U-Net with densely connected convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27–28 October 2019.
48. Jalali, Y.; Fateh, M.; Rezvani, M.; Abolghasemi, V.; Anisi, M.H. ResBCDU-Net: A deep learning framework for lung CT image segmentation. *Sensors* **2021**, *21*, 268. [[CrossRef](#)]
49. Zhang, Y.; Davison, B.D.; Talghader, V.W.; Chen, Z.; Xiao, Z.; Kunkel, G.J. Automatic head overcoat thickness measure with NASNet-large-decoder net. In Proceedings of the Future Technologies Conference (FTC) 2021, Vancouver, BC, Canada, 28–29 October 2021; pp. 159–176.
50. Jalali, Y.; Fateh, M.; Rezvani, M. DABT-U-Net: Dual Attentive BConvLSTM U-Net with Transformers and Collaborative Patch-based Approach for Accurate Retinal Vessel Segmentation. *Int. J. Eng.* **2024**, *37*, 2051–2065. [[CrossRef](#)]
51. Rezvani, S.; Fateh, M.; Khosravi, H. ABANet: Attention boundary-aware network for image segmentation. *Expert Syst.* **2024**, *41*, e13625. [[CrossRef](#)]
52. Rezvani, S.; Fateh, M.; Jalali, Y.; Fateh, A. FusionLungNet: Multi-scale fusion convolution with refinement network for lung CT image segmentation. *Biomed. Signal Process. Control.* **2025**, *107*, 107858. [[CrossRef](#)]
53. Zhao, A.; Wu, H.; Chen, M.; Wang, N. DCACorrCapsNet: A deep channel-attention correlative capsule network for COVID-19 detection based on multi-source medical images. *IET Image Process.* **2023**, *17*, 988–1000. [[CrossRef](#)]
54. Jiang, Z.; Chen, L. Multisemantic level patch merger vision transformer for diagnosis of pneumonia. *Comput. Math. Methods Med.* **2022**, *2022*, 7852958. [[CrossRef](#)]
55. Mabrouk, A.; Diaz Redondo, R.P.; Dahou, A.; Abd Elaziz, M.; Kayed, M. Pneumonia detection on chest X-ray images using ensemble of deep convolutional neural networks. *Appl. Sci.* **2022**, *12*, 6448. [[CrossRef](#)]
56. Goodwin, B.D.; Jaskolski, C.; Zhong, C.; Asmani, H. Intra-model variability in COVID-19 classification using chest x-ray images. *arXiv* **2020**, arXiv:2005.02167.
57. Gazda, M.; Plavka, J.; Gazda, J.; Drotar, P. Self-supervised deep convolutional neural network for chest X-ray classification. *IEEE Access* **2021**, *9*, 151972–151982. [[CrossRef](#)]
58. Van, M.H.; Verma, P.; Wu, X. On large visual language models for medical imaging analysis: An empirical study. In Proceedings of the 2024 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE), Wilmington, DE, USA, 19–21 June 2024; pp. 172–176.
59. Yadav, S.S.; Jadhav, S.M. Deep convolutional neural network based medical image classification for disease diagnosis. *J. Big Data* **2019**, *6*, 1–18. [[CrossRef](#)]
60. Ayan, E.; Ünver, H.M. Diagnosis of pneumonia from chest X-ray images using deep learning. In Proceedings of the 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT), Istanbul, Turkey, 24–26 April 2019; pp. 1–5.

61. Chattopadhyay, S.; Ganguly, S.; Chaudhury, S.; Nag, S.; Chattopadhyay, S. Exploring Self-Supervised Representation Learning for Low-Resource Medical Image Analysis. In Proceedings of the 2023 IEEE International Conference on Image Processing (ICIP), Kuala Lumpur, Malaysia, 8–11 October 2023; pp. 1440–1444.
62. Bhatt, H.; Shah, M. A Convolutional Neural Network ensemble model for Pneumonia Detection using chest X-ray images. *Healthc. Anal.* **2023**, *3*, 100176. [[CrossRef](#)]
63. Reshan, M.S.A.; Gill, K.S.; Anand, V.; Gupta, S.; Alshahrani, H.; Sulaiman, A.; Shaikh, A. Detection of pneumonia from chest X-ray images utilizing mobilenet model. *Healthcare* **2023**, *11*, 1561. [[CrossRef](#)]
64. Zhang, S.; Fu, Y.; Fang, L.; Xu, Q.; Gu, S.; Zhou, H.; Zhou, J. Psittacosis pneumonia with the reversed halo sign: A case report and literature review. *BMC Infect. Dis.* **2025**, *25*, 717. [[CrossRef](#)]
65. Wahyudi, F.; Saputra, R.D. Hydropneumothorax Secondary to Community-Acquired Pneumonia with Broncholith: A Case Report. *Biosci. Med. J. Biomed. Transl. Res.* **2025**, *9*, 6005–6018.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.