Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Two-Stage Framework Combining U-Net++ Lung Segmentation and DenseNet121 for Multi-Class Classification from Chest X-Ray

Mohammed Borhan Uddin

mohammedborhan15@gmail.com

Chittagong University of Engineering & Technology

Sanjida Tasneem

International Islamic University Chittagong

Abdur Rahaman

International Islamic University Chittagong

Shahariar Nafiz

International Islamic University Chittagong

Md. Nurul Alam

International Islamic University Chittagong

Mohammad Shamsul Arefin

Chittagong University of Engineering & Technology

Research Article

Keywords:

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11759481/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Two-Stage Framework Combining U-Net++ Lung Segmentation and DenseNet121 for Multi-Class Classification from Chest X-Ray

Mohammed Borhan Uddin[a,*], Sanjida Tasneem[b,1], Abdur Rahaman[b,1], Shahariar Nafiz[b], Md. Nurul Alam[b], Mohammad Shamsul Arefin[a]

[a]Department of Computer Science and Engineering, Chittagong University of Engineering and Technology, Chattogram 4349, Bangladesh

[b]Department of Computer Science and Engineering, International Islamic University Chittagong, Chattogram 4318, Bangladesh

Emails: mohammedborhan15@gmail.com, sanjidatasneem24@gmail.com, mabdurrahaman7@gmail.com, shahariarnafiziiuc@gmail.com, nurulalamreaz@gmail.com, sarefin@cuet.ac.bd

***Abstract*—Chest X-rays (CXRs) are widely used to detect lung diseases, but deep learning classifiers may rely on visual cues outside the lungs rather than relevant clinical features. This study evaluates three lung segmentation models (U-Net, U-Net++, and fine-tuned SAM ViT-B) and four ImageNet-pretrained classifiers (DenseNet121, ResNet50, VGG19, and Xception) using the COVID-19 Radiography Database of 21,165 images. U-Net++ achieved the best segmentation performance, while DenseNet121 performed best in classification. Based on these findings, we developed a two-stage framework that uses U-Net++ to segment the lungs and DenseNet121 to classify the extracted regions as COVID-19, Lung Opacity, Normal, or Viral Pneumonia. The framework was tested on unlabeled DICOM radiographs from a diagnostic center without additional fine-tuning. A medical doctor reviewed 15 cases flagged as low-confidence during an earlier run, revealing disagreements in which abnormal cases were classified as Normal, sometimes with very high confidence. Grad-CAM and insertion/deletion curves were used to examine the image regions influencing the model's predictions. These findings demonstrate the feasibility of integrating lung segmentation and disease classification for native DICOM images while highlighting the need for careful clinical validation beyond benchmark performance.**

Keywords: Chest X-ray, Lung Segmentation, U-Net++, DenseNet121, Explainable AI, Grad-CAM, COVID-19, DICOM, Deep Learning

1. Introduction

Lung diseases such as coronavirus disease 2019 (COVID-19), viral pneumonia, and bacterial pneumonia remain important global health concerns. Chest X-ray (CXR) imaging is widely used to assess suspected pulmonary disease because it is relatively inexpensive, rapid, and broadly available. However, different pulmonary conditions may produce similar radiographic appearances, and interpretation can vary among readers, particularly for less experienced readers [1].

Convolutional neural networks (CNNs) and transfer learning are widely used for disease classification from chest X-ray (CXR) images [2], [3]. However, models trained on complete radiographs may learn from non-pulmonary cues, such as text labels, laterality markers, and imaging artifacts, rather than relying exclusively on disease-related lung features [4]. Lung segmentation is one approach to focusing classification on relevant pulmonary regions, and recent studies have combined lung segmentation networks with deep learning classifiers for CXR disease prediction [5], [6]. Explainable artificial intelligence (XAI) methods, including Grad-CAM, can help visualize image regions that contribute to a model's predictions [7]–[9].

Several research questions remain open. In particular, the effect of segmentation-model selection on the performance of a complete segmentation–classification framework requires further investigation. Comparing segmentation models under a consistent classification setup can help clarify how lung-mask quality and model choice affect downstream predictions. Visual explanations also need to be interpreted in the context of the image regions provided to the classifier, since non-pulmonary structures may influence predictions when the entire radiograph is used. Furthermore, many studies primarily report results on public datasets, leaving uncertainty about model behavior on images obtained in real clinical settings, including images stored in the Digital Imaging and Communications in Medicine (DICOM) format.

In this study, we systematically compare three segmentation models – U-Net, U-Net++, and the fine-tuned Segment Anything Model (SAM, ViT-B) – alongside four ImageNet-pretrained classifiers: DenseNet121, ResNet50, VGG19, and Xception, using the publicly available COVID-19 Radiography Database. Based on the comparative experimental results, U-Net++ and DenseNet121 are selected to develop a two-stage framework that integrates lung segmentation with chest X-ray classification. To improve the interpretability of the classification results, we analyze DenseNet121 predictions using Grad-CAM and insertion/deletion curves. Furthermore, we assess

*Corresponding authors. *E-mail addresses:* mohammedborhan15@gmail.com (M.B. Uddin)

[1]These authors contributed equally to this work.

the framework's applicability to real-world clinical data by evaluating it, without additional fine-tuning, on 341 unlabeled DICOM chest X-ray images obtained from a diagnostic center. This evaluation includes prediction-confidence analysis, lung-mask coverage assessment, and physician review of selected cases.

The main contributions of this work are:

1) A comparative evaluation of U-Net, U-Net++, and fine-tuned SAM (ViT-B) for lung segmentation using segmentation-performance metrics, parameter count, and computational cost.
2) A comparison of four transfer-learning classifiers under a consistent training protocol, followed by the development of a two-stage framework combining U-Net++ for lung segmentation with DenseNet121 for disease classification.
3) An explainability analysis of the DenseNet121 classifier using Grad-CAM and insertion/deletion curves, alongside an analysis of its training and validation curves.
4) An evaluation of the framework on 341 real-world, unlabeled DICOM chest radiographs, including prediction-confidence analysis, lung-mask coverage assessment, and a physician's review of selected cases, including cases in which the framework's predictions differed from the physician's assessment.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the data, preprocessing, models, and the two-stage framework. Section 4 reports the experiments, including the DICOM evaluation and the explainability analysis. Section 5 discusses the findings and their limits, and Section 6 concludes.

2. Related Work

2.1 Lung Segmentation Architectures

Encoder–decoder networks are widely used for lung field segmentation in chest X-ray images. A 2026 comparative study of ten segmentation architectures, including U-Net, U-Net++, Dense-UNet, and Swin U-Net, reported that Dense-UNet achieved the highest Dice score (98.48%) and IoU (97.00%) [10]. Recent research has also investigated lightweight residual U-Net architectures incorporating a convolutional block attention module (CBAM) and atrous spatial pyramid pooling (ASPP). The proposed architecture achieved Dice scores of 97.49%, 98.72%, and 99.08% on the SZ, JSRT, and MC datasets, respectively [11]. Beyond conventional segmentation networks, foundation models such as the Segment Anything Model (SAM) have attracted interest because of their general-purpose segmentation capabilities. However, applying SAM to medical images remains challenging because of differences between natural and medical images, and effective adaptation may require task-specific fine-tuning and additional computational resources [12], [13]. These findings highlight the importance of balancing segmentation accuracy, computational efficiency, and suitability for medical images when selecting a segmentation architecture for an integrated lung disease classification framework.

2.2 CNN-Based and Transfer-Learning Classification

Transfer learning using ImageNet-pretrained convolutional neural networks (CNNs) is widely adopted for chest X-ray (CXR) disease classification. Residual networks (ResNet) employ skip connections to facilitate the training of deeper neural networks and have been applied to explainable pneumonia detection using CXR images [14]. DenseNet architectures promote feature reuse through dense connections between layers; DenseNet121 was one of the CNNs (with ResNet50 and EfficientNetV2-S) used to distinguish mycoplasma from viral pneumonia on pediatric radiographs [15]. Comparative studies have also evaluated multiple pretrained CNN architectures, including ResNet50, DenseNet201, VGG16, and VGG19, for pneumonia detection. One such study of ten pretrained CNNs reported its highest accuracy (98.50%) for a fine-tuned AlexNet on a single-hospital pneumonia-versus-healthy dataset [3]. However, classification accuracy alone is insufficient when selecting a backbone for an integrated segmentation and classification framework. Computational requirements, feature representation, and compatibility with the segmentation component must also be considered. Therefore, model selection should account for both classification performance and the practical requirements of the overall framework.

2.3 Attention Mechanisms and Explainability

Attention mechanisms and explainable artificial intelligence (XAI) techniques have gained increasing attention in medical imaging because they can help improve the interpretability of deep learning predictions. Grad-CAM generates class-discriminative localization maps that highlight image regions contributing to a model's predictions, allowing users to examine the visual evidence associated with a classification result [16]. Veeramani et al. proposed the NextGen XAI-TRANS framework, which integrates Inception-based transfer learning with an improved U-Net for lung segmentation and uses LIME and Grad-CAM to explain chest X-ray classification predictions. The authors reported a maximum multiclass accuracy of 97% and a 4.75% relative improvement that they attribute to their XAI-TRANS pipeline [9]. These studies demonstrate the potential of integrating segmentation, classification, and explainability to support more interpretable medical image analysis.

2.4 Hybrid Segmentation–Classification Frameworks

Hybrid segmentation–classification frameworks combine anatomical region localization with disease prediction. In chest X-ray (CXR) analysis, lung segmentation can help focus classification models on pulmonary structures while reducing the influence of irrelevant image regions and non-pulmonary artifacts. Previous studies have investigated the integration of lung segmentation and deep learning-based classification for lung disease analysis [5], [6]. Other studies have focused on improving lung segmentation using enhanced U-Net++ architectures with attention mechanisms and wavelet-based features [17]. However, high segmentation accuracy alone does not establish the reliability of an integrated framework under

varying imaging conditions. Similarly, strong classification performance on a public benchmark does not necessarily demonstrate generalizability to images acquired in clinical environments. These limitations highlight the need to evaluate segmentation and classification components together, assess the quality of model explanations, and investigate performance on real-world clinical images. In particular, evaluation using native Digital Imaging and Communications in Medicine (DICOM) images can provide additional insight into the practical applicability of a framework beyond its original benchmark dataset.

2.5 Research Gaps Addressed by This Study

The reviewed literature highlights several areas that warrant further investigation. First, although various segmentation and classification architectures have been evaluated for chest X-ray (CXR) analysis, selecting compatible models for an integrated framework requires consideration of predictive performance, segmentation quality, and computational constraints. Second, many studies report results primarily on public benchmark datasets, leaving uncertainty about how well their approaches perform on real-world clinical images acquired under different imaging conditions. Third, although visual explanation methods such as Grad-CAM are widely used to interpret classification predictions, further analysis is needed to examine whether the highlighted regions correspond to clinically relevant pulmonary structures. Addressing these gaps, the present study compares multiple lung segmentation and classification models and develops a two-stage framework combining U-Net++ for lung segmentation with DenseNet121 for disease classification. The framework's predictions are examined using Grad-CAM visualizations and insertion/deletion curves. In addition, the study evaluates the framework on 341 real-world, unlabeled DICOM chest X-ray images through confidence analysis and medical doctor review of selected cases. Table I summarizes representative studies on lung segmentation, CXR disease classification, and explainable artificial intelligence (XAI).

3. Methodology

In this section, we present the methodology for lung segmentation and multi-class chest X-ray (CXR) disease classification. We describe the datasets used, the model training process, the preprocessing steps, and the data augmentation techniques applied in this study. We also present the proposed two-stage framework, which combines the best-performing segmentation and classification models, and evaluate its performance on unseen real-world data. The overall workflow of the study is illustrated in Figure 1.

3.1 Dataset

The publicly available COVID-19 Radiography Database was used for model training and evaluation. In addition, 341 real-world radiographs from a diagnostic center were used to check how the models behave outside the benchmark. The two datasets are described below.

(1) Public benchmark – COVID-19 Radiography Database. This publicly available dataset from Kaggle [19] contains 21,165 chest X-ray images across four classes: COVID (class 0; 3,616), Lung Opacity (class 1; 6,012), Normal (class 2; 10,192), and Viral Pneumonia (class 3; 1,345) (Table II; Figures 2 and 3). The corresponding lung masks were used for training and evaluating the segmentation networks (Section 3.6). The classification networks (Section 3.7) were trained on the original, unsegmented chest X-ray images. Representative samples from the dataset are shown in Figure 4.

(2) Primary clinical dataset. This set consists of 341 unlabeled chest radiographs collected in native DICOM format from a diagnostic center. It was used exclusively for downstream deployment and validation of the two-stage framework (Section 3.8). No reference labels were available for this set; a medical doctor reviewed a subset of the predictions, as described in Section 4. The DICOM files were processed as described in Algorithm 1 before being passed to the proposed architecture.

3.2 Preprocessing and Dataset Split

For the public dataset, images were loaded from their class-labeled directories. For the primary DICOM dataset, the raw pixel array was extracted from each file using `pydicom`, converted to a floating-point representation, min–max normalized to the $[0, 255]$ range, and cast to `uint8`. Single-channel (grayscale) acquisitions were replicated into a 3-channel representation, and when a DICOM file contained multiple frames, the middle frame was selected. Two separate preprocessing branches were then applied according to downstream use. The segmentation branch converted the image to grayscale, resized it to 256×256, scaled the pixel values to $[0, 1]$, and stacked the image into three channels. The resulting image was then normalized to $[-1, 1]$ using a mean and standard deviation of 0.5 for each channel, consistent with the segmentation training transform. In the two-stage framework, the classification branch cropped the lung region defined by the predicted mask, resized the crop to 256×256 (the input size of the trained DenseNet121), normalized it to $[0, 1]$, and then re-normalized it using the ImageNet channel mean and standard deviation. The complete procedure is summarized in Algorithm 1.

The dataset was separated into training (70%, 14,815 images), validation (15%, 3,175 images) and test (15%, 3,175 images) sets (see Table III). Model parameters were trained on the training set, hyperparameters were tuned and overfitting was monitored on the validation set, and performance was evaluated on new data using the test set. The same splitting protocol was applied separately for segmentation (image–mask pairs) and classification (image–label pairs).

3.3 Data Augmentation

Data augmentation was applied to both segmentation and classification models using different augmentation strategies. Segmentation training used the Albumentations library, whereas classification training used Keras

TABLE I: Comparison of Related Studies on Lung Segmentation, Classification, and Explainability

Ref.	Method	Year	Seg.	Class.	XAI	Dataset / Task	Reported Metric / Score
[10]	Dense-UNet and U-Net++ comparison	2026	✓	×	×	Lung CXR	Dice: 98.48%
[7]	Hybrid attention with Grad-CAM and GAN augmentation	2026	×	✓	✓	Pneumonia CXR	Accuracy: 96.3%
[18]	Dual-attention network	2025	✓	×	×	Lung segmentation	Dice: 97.29%; IoU: 94.74%
[6]	DAA-UNet with ensemble classification	2025	✓	✓	×	Multi-class CXR	Accuracy: 96.87%
[8]	ASCE-XAI with anatomical segmentation and Grad-CAM	2025	✓	✓	✓	Pneumonia CXR	Accuracy: ∼95.0%

Note: Values are taken from the abstracts of the cited works and were not independently reproduced.

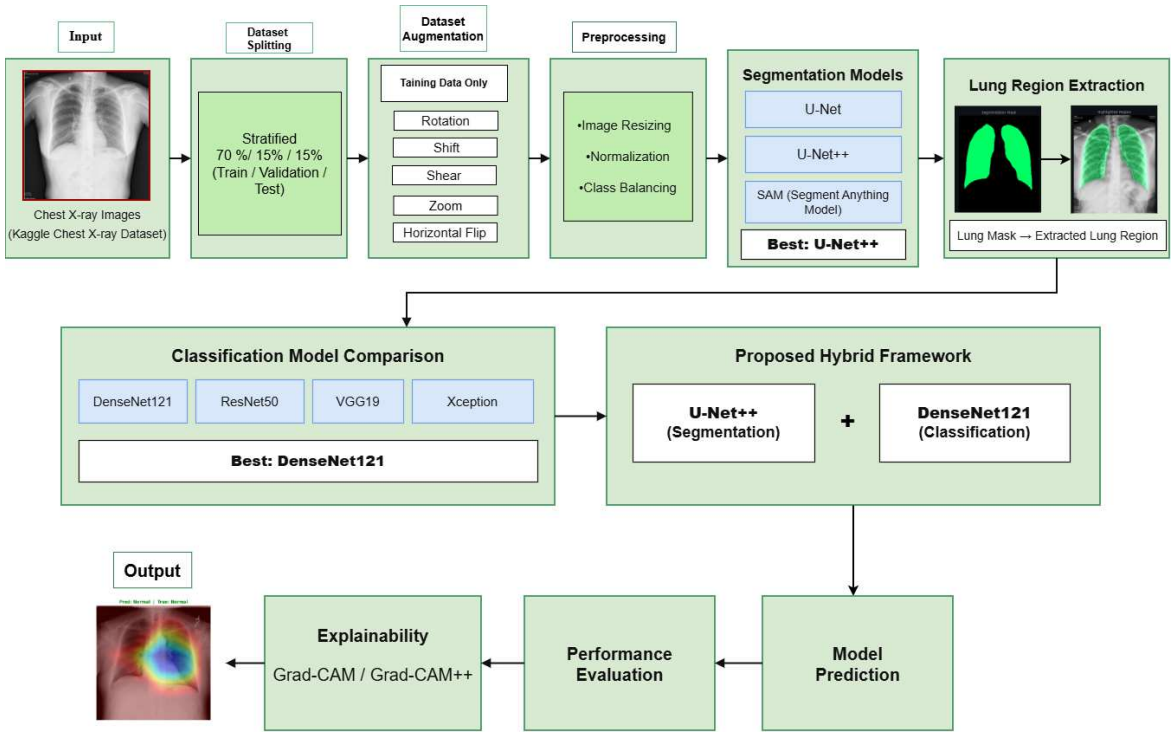


Fig. 1: Workflow of the Proposed CXR Classification Framework: Model Selection and Evaluation

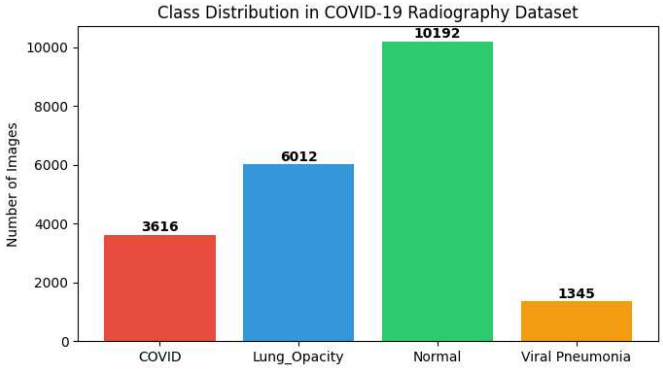


Fig. 2: Class Distribution of COVID-19 Radiography Database

TABLE II: Class distribution of the COVID-19 Radiography Database (public benchmark) and the number of test-set images per class.

Class	Total images	Test-set images
COVID	3,616	542
Lung Opacity	6,012	902
Normal	10,192	1,529
Viral Pneumonia	1,345	202
Total	**21,165**	**3,175**

TABLE III: Train/validation/test split of the public dataset.

Partition	Percentage	Images
Training	70%	14,815
Validation	15%	3,175
Test	15%	3,175

`ImageDataGenerator`. The augmentation parameters for both models are summarized in Table IV. Example augmented chest X-ray images are shown in Figure 5. Validation and test images were not augmented.

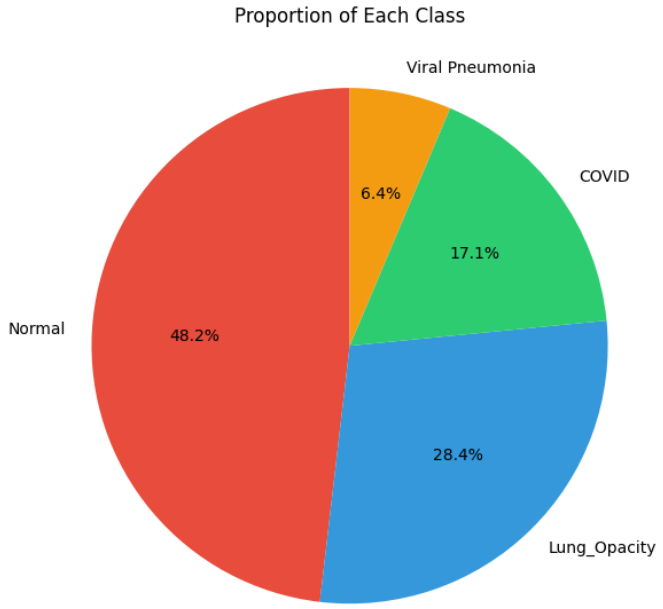


Fig. 3: Proportion of each class of the COVID-19 Radiography Database.

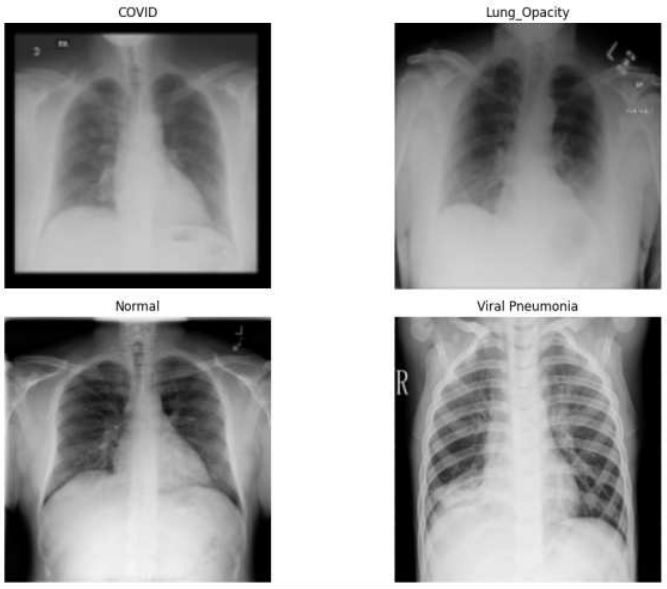


Fig. 4: Sample images of the dataset.

3.4 Model Training

Deep learning models were used for both training and evaluation. U-Net, U-Net++, and SAM were used for segmentation, and VGG19, ResNet50, DenseNet121, and Xception were implemented for classification. This section describes the training procedures and implementation details of these models. The dataset was first split into training, validation, and test sets. The models were trained on the training set, validated on the validation set, and finally tested on the test set. The primary dataset was additionally used for real-world validation.

Algorithm 1 DICOM Preprocessing for Segmentation and Classification

Require: DICOM file path
Ensure: Segmentation input tensor, classification input array
Load DICOM image: read DICOM file, extract pixel array, convert to float
if image has multiple frames **then**
    Select middle frame
end if
Normalize image: find min and max pixel values, scale values to range $[0, 255]$, convert to `uint8`
if image is grayscale **then**
    Convert to 3-channel RGB
else
    Keep first 3 channels
end if
Segmentation preprocessing (U-Net++):
Convert to grayscale
Resize to 256×256
Normalize to $[0, 1]$
Stack into 3 channels
Convert to tensor of shape $(1, 3, H, W)$
output segmentation input tensor
Classification preprocessing (DenseNet121):
Crop the lung bounding box of the predicted mask (5% margin)
Resize to 256×256
Normalize to $[0, 1]$
Apply ImageNet mean and std normalization
Expand dimensions to $(1, H, W, 3)$
output classification input array

TABLE IV: Data augmentation parameters for segmentation and classification.

Transform	Parameter	Description
Segmentation (Albumentations)		
HorizontalFlip	$p = 0.5$	Random horizontal flip
VerticalFlip	$p = 0.2$	Random vertical flip
Rotate	$\pm 15^\circ$, $p = 0.5$	Random rotation
ShiftScaleRotate	0.05, 0.05, $\pm 10^\circ$	Random shift, scale, and rotation ($p = 0.5$)
Random BrightnessContrast	$p = 0.3$	Random brightness and contrast adjustment
Classification (Keras ImageDataGenerator)		
rotation_range	20	Rotation up to $\pm 20^\circ$
width_shift_range	0.2	Horizontal shift up to 20%
height_shift_range	0.2	Vertical shift up to 20%
shear_range	0.2	Shear transformation
zoom_range	0.2	Random zoom in or out
horizontal_flip	True	Random horizontal flip
fill_mode	nearest	Fills empty pixels using nearest pixel values
rescale	$1/255$	Rescales pixel values to $[0, 1]$

3.5 Model Training Summary

Table V and Table VI summarize the training configurations of the segmentation and classification models, respectively.

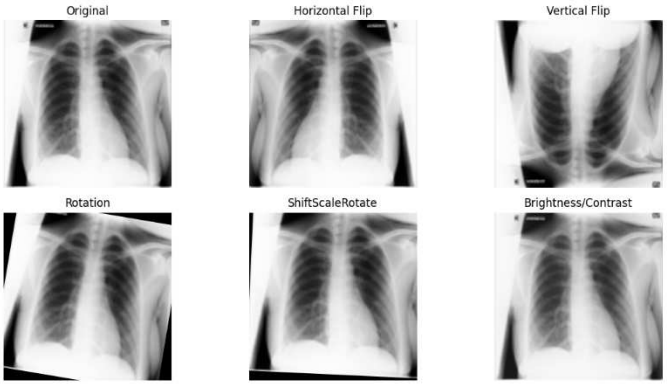


Fig. 5: Example augmented views of chest X-ray images.

TABLE V: Model Training summary for segmentation models

Model	Params.	Optimizer	Epochs	LR	Complexity	Cost
U-Net	31.04 M	Adam	20	1×10^{-4}	48.35 GMac	42
U-Net++	**36.17 M**	**Adam**	**20**	1×10^{-4}	**119.35 GMac**	**64**
SAM	4.06 M	Adam	20	1×10^{-4}	371.22 GMac	93

Abbreviations: Params. = number of parameters (M = million); LR = learning rate; Complexity = computational complexity in GMac (giga multiply–accumulate operations); Cost = Unit Cost (Approximate).

TABLE VI: Model Training summary for classification models

Model	Params.	Optimizer	Epochs	LR	Cost
VGG19	36.80 M	SGD	30	1×10^{-3}	6
DenseNet121	**40.59 M**	**SGD**	**30**	1×10^{-3}	**6.2**
ResNet50	90.70 M	SGD	30	1×10^{-3}	12
Xception	87.97 M	SGD	30	1×10^{-3}	10

Abbreviations: Params. = number of parameters (M = million); LR = learning rate; Cost = Unit Cost (Approximate).

3.6 Segmentation Architectures

Three segmentation backbones were trained and evaluated under matched conditions (20 epochs, Adam optimizer, learning rate $= 1 \times 10^{-4}$; Section 4):

- **U-Net**: An encoder–decoder network with skip connections between corresponding encoder and decoder resolutions [20].
- **U-Net++**: A U-Net variant with nested, dense skip pathways between the encoder and decoder sub-networks, designed to reduce the semantic gap between encoder and decoder feature maps at different resolutions [17], [21].
- **SAM (ViT-B)**: A Vision Transformer-based, promptable segmentation foundation model fine-tuned for lung-mask extraction rather than used in a zero-shot setting [12].

All three were evaluated using the Dice coefficient and Intersection-over-Union (IoU):

$$\text{Dice} = \frac{2|P \cap G|}{|P| + |G|}, \qquad \text{IoU} = \frac{|P \cap G|}{|P \cup G|} \tag{1}$$

where P and G denote the predicted and ground-truth binary lung masks, respectively. The network with the best Dice/IoU (Table VIII) was carried forward into the two-stage framework (Section 3.8).

3.7 Classification Architectures

Four ImageNet-pretrained convolutional backbones were fine-tuned independently for the four-class task, using an identical data framework, split and training schedule (30 epochs; Section 4), so that the backbones can be compared under the same experimental protocol:

- **DenseNet121** [22]: dense inter-layer connectivity, in which feature maps from all preceding layers within a dense block are concatenated.
- **ResNet50** [23]: a 50-layer residual network that uses identity skip connections across bottleneck blocks.
- **VGG19** [24]: a 19-layer sequential convolutional network with uniform 3×3 kernels.
- **Xception** [25]: depthwise-separable convolutions throughout, which factorize spatial and channel-wise filtering.

In each case, the pretrained convolutional base was retained, and a new classification head (a Flatten layer, dropout of 0.5, a 512-unit fully connected layer and a final 4-way softmax) was trained for the target task. Classification performance was reported using accuracy, precision, recall, macro-F1, and Cohen's κ (Table IX).

3.8 Proposed Two-Stage Framework

The framework pairs the best segmentation model, U-Net++, with the best classifier, DenseNet121. The U-Net++ checkpoint holds 36,179,871 stored parameters and buffers, all of them non-zero, and the DenseNet121 classifier has 40.59 M parameters (Table VI). As shown in Figure 6, each radiograph goes through two steps:

1) **Segmentation.** The preprocessed $256 \times 256 \times 3$ image is passed through the trained U-Net++, which outputs a per-pixel lung-probability map. The map is thresholded at $t = 0.3$ to give a binary lung mask M_{seg}, which is resized back to the original image resolution.
2) **Region extraction and classification.** The bounding box of M_{seg}, enlarged by 5% on each side, is used to crop the lung region from the radiograph (the full image is used if the mask is empty). The crop is resized to 256×256, scaled to $[0, 1]$, standardized with the ImageNet channel mean and standard deviation, and passed through the trained DenseNet121 to obtain probabilities for {COVID, Lung Opacity, Normal, Viral Pneumonia}.

3.9 Explainability: Grad-CAM

We applied Grad-CAM [16] to the trained DenseNet121 classifier. The gradient of the predicted-class score was taken with respect to the output of the DenseNet121 convolutional base (an $8 \times 8 \times 1024$ feature map), averaged over the spatial dimensions to give one weight per channel, and used to form a weighted sum of the feature maps. The result was passed

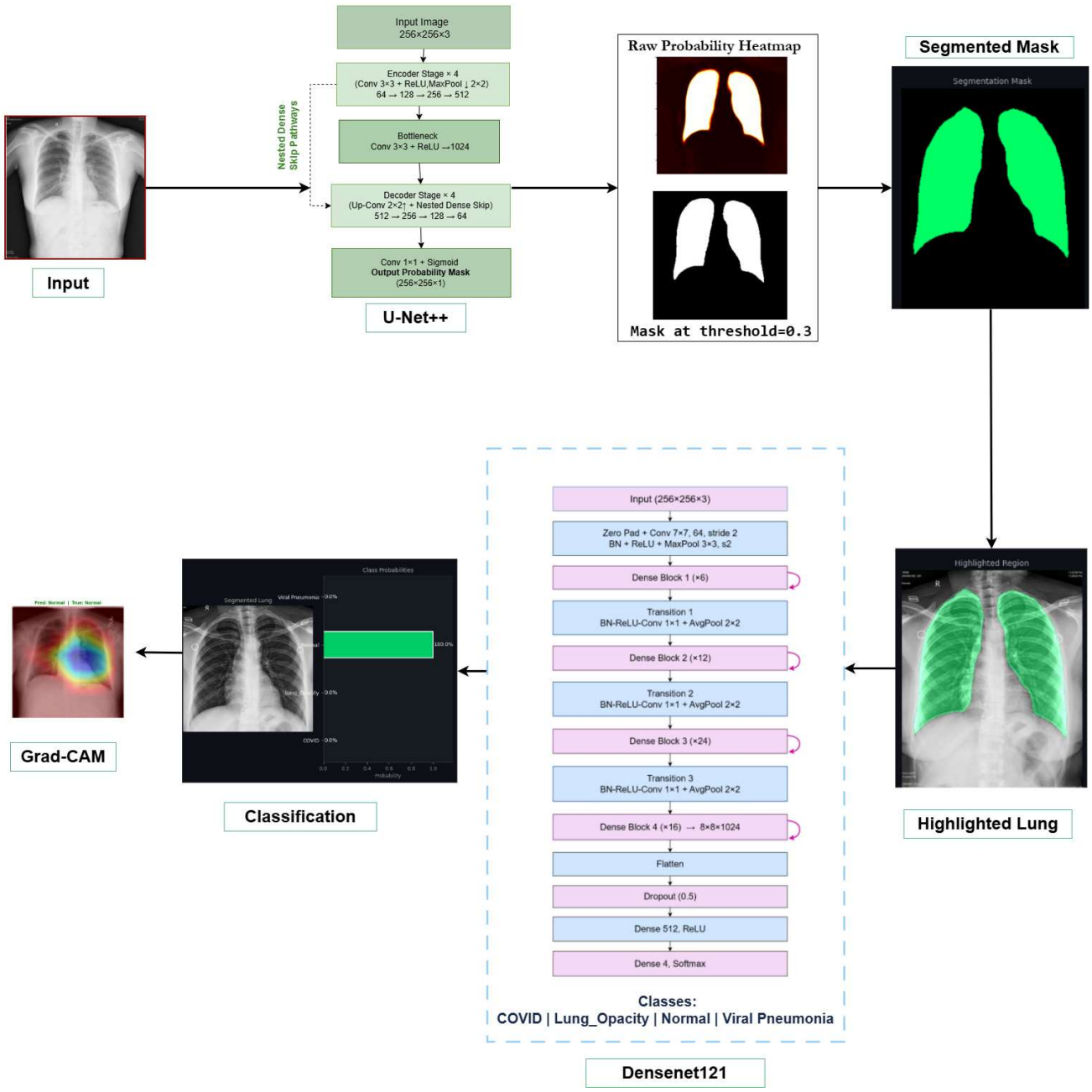


Fig. 6: Proposed Two-Stage Lung Segmentation and Disease Classification Framework.

through a ReLU, normalized to $[0, 1]$, resized to the input resolution and overlaid on the radiograph with a jet colormap at 0.4 opacity, so warm colors mean stronger class-specific activation. Maps were generated for three randomly drawn images from the public test set (Figure 13). Because the classifier was trained on full radiographs, these maps were computed on full radiographs and not on the extracted lung regions. To go beyond visual inspection, we also computed insertion and deletion curves for each image: pixels were ranked by heatmap value and then progressively revealed on a Gaussian-blurred copy of the image (insertion) or progressively blurred out (deletion) in 100 steps, the confidence of the predicted class was recorded at every step, and the area under each curve (AUC) was computed.

4. Experimental Setup and Results

The COVID-19 Radiography Database (70%/15%/15% stratified split, Table III) was used for training and held-out evaluation, and the 341 unlabeled DICOM radiographs were used only for deployment-style testing. Segmentation was scored with Dice and IoU, and classification with accuracy, macro-averaged precision, recall and F1, and Cohen's κ. Table VII lists the configuration used in the experiments, and Figure 2 shows the class distribution of the public dataset.

TABLE VII: Experimental Setup Summary

Parameter	Configuration
Software	TensorFlow/Keras for classification; PyTorch for U-Net and U-Net++; `segment-anything` for SAM; `pydicom` for DICOM I/O
Compute environment	Google Colab GPUs; Tesla T4 for framework inference
Input size	256×256 for U-Net, U-Net++, and all classifiers; 1024×1024 for SAM
Segmentation training	Adam, learning rate 1×10^{-4}, 20 epochs, BCE-with-logits loss; batch size 32 for U-Net and U-Net++, and 12 for SAM; cosine annealing to 1×10^{-6} for SAM
SAM (ViT-B)	Image and prompt encoders frozen; mask decoder fine-tuned without prompts
Segmentation augmentation	Albumentations: HorizontalFlip (0.5), VerticalFlip (0.2), Rotate ($\pm 15^{\circ}$, 0.5), ShiftScaleRotate (0.5), RandomBrightnessContrast (0.3)
Classification training	ImageNet-initialized backbone with Flatten–Dropout(0.5)–Dense(512)–softmax head; SGD, learning rate 1×10^{-3}, momentum 0.9; `ReduceLROnPlateau` on validation accuracy (factor 0.5, patience 5, minimum learning rate 1×10^{-5}); 30 epochs, batch size 32, categorical cross-entropy
Classification input	Pixel rescaling to $[0, 1]$ using $1/255$; training augmentation with rotation (20°), width/height shifts (0.2), shear (0.2), zoom (0.2), and horizontal flipping
Model selection	Final-epoch weights for classifiers; best validation Dice for segmentation
Mask threshold (t)	0.3 for the U-Net++ probability map in the two-stage framework; 0.5 for Dice/IoU evaluation
Data split	70:15:15 (training, validation, testing), stratified by class, seed 42, fixed across all experiments

4.1 Segmentation Results

Table VIII compares the three segmentation models. U-Net++ obtained a test Dice of 0.9872 and an IoU of 0.9747. For U-Net, the reported Dice (0.9865) and IoU (0.9733) were computed on the validation set rather than the test set, and SAM obtained 0.9741 and 0.9496. Because the U-Net values come from a different split and all results are from single runs, the comparison between U-Net++ and U-Net is only indicative, and the two models are close in practice. SAM, trained with a frozen image encoder, requires the most computation (371.22 GMac) and obtains the lowest scores. Hence, U-Net++ was incorporated into the two-stage framework. The predicted masks of the three models for the same image are shown in Fig. 7.

TABLE VIII: Quantitative Comparison of Segmentation Models on the Public Dataset Using Dice and IoU Metrics.

Model	Params	MACs	Train Dice	Test Dice	Test IoU
U-Net	31.04M	48.35G	0.9838	0.9865†	0.9733†
U-Net++	**36.17M**	**119.35G**	**0.9838**	**0.9872**	**0.9747**
SAM (ViT-B)	4.06M*	371.22G	0.9716	0.9741	0.9496

**Trainable parameters only.* †*Computed on the validation set, not the test set.*

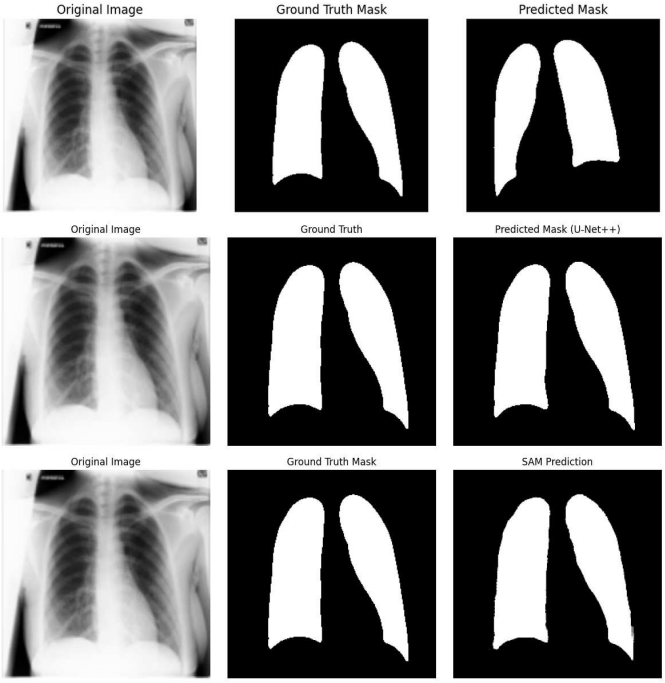


Fig. 7: Predicted lung masks of U-Net (top), U-Net++ (middle) and SAM (bottom) for the same example image, shown with the original radiograph and the ground-truth mask.

TABLE IX: Quantitative Comparison of Classification Models on the Public Test Set (3,175 Images).

Model	Acc.	Prec.	Rec.	F1	κ
DenseNet121	**0.9694**	**0.9772**	**0.9672**	**0.9721**	**0.9532**
VGG19	0.9682	0.9734	0.9669	0.9700	0.9512
ResNet50	0.9650	0.9739	0.9619	0.9676	0.9463
Xception	0.9518	0.9612	0.9464	0.9536	0.9260

4.2 Classification Results

Table IX reports the four classifiers, each trained on full, unsegmented radiographs. DenseNet121 reached the highest accuracy (96.94%), macro-F1 (0.9721) and κ (0.9532), followed by VGG19 (96.82%), ResNet50 (96.50%) and Xception (95.18%). The gaps among the top three are small: DenseNet121 classified 4 more test images correctly than VGG19 and 14 more than ResNet50, out of 3,175, and each model was trained once, so we do not treat the ranking as statistically established. Most of the remaining errors involve Lung Opacity and Normal (Figure 8). For DenseNet121, the per-class recall was 0.996 for COVID, 0.946 for Lung Opacity, 0.977 for Normal and 0.951 for Viral Pneumonia (540/542, 853/902, 1493/1529 and 192/202 images). Its training curves (Figure 9) show training accuracy close to 1.0 after about ten epochs, while validation accuracy stays near 96.5% and validation loss slowly rises from about 0.17 to 0.28. This is a sign of overfitting; since the final-epoch weights were used, the reported test accuracy reflects that.

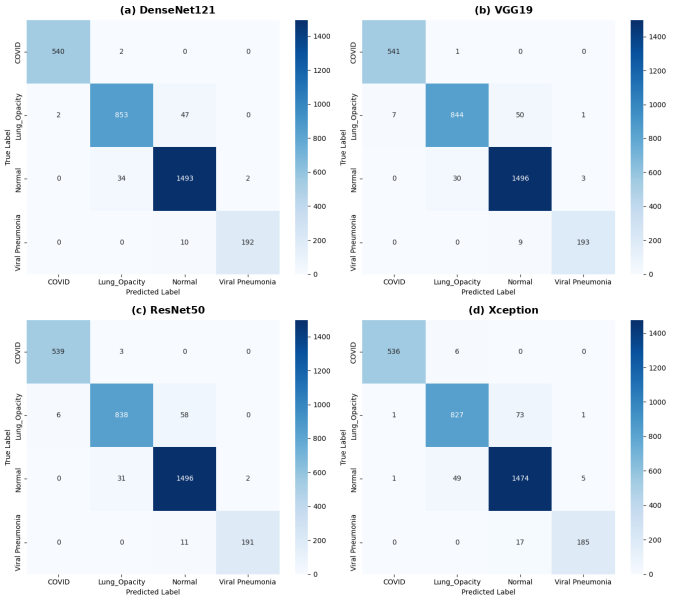


Fig. 8: Test-set confusion matrices of the four classification backbones.

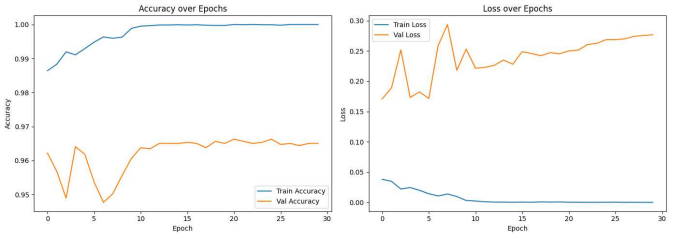


Fig. 9: DenseNet121 training and validation accuracy (left) and loss (right) over 30 epochs.

4.3 Proposed Framework on Real-World DICOM Images

The best-performing models from the benchmark experiments were selected to construct the two-stage framework: U-Net++ for lung segmentation, based on its superior Dice and IoU scores (Table VIII), and DenseNet121 for classification, based on its highest accuracy, macro-F1 score, and Cohen's κ (Table IX). Neither model was fine-tuned on the DICOM images. The resulting framework processed all 341 DICOM files without execution errors (Table X). The mean confidence of the top class was 97.3%, and the U-Net++ masks covered 24.6% of the image area on average. Of the 341 files, 291 (85.3%) were labeled Normal, 34 (10.0%) Viral Pneumonia and 16 (4.7%) Lung Opacity; none was labeled COVID-19. Confidence was highest for Normal predictions (mean 98.9%) and lower for Lung Opacity (89.5%) and Viral Pneumonia (86.9%). Since this set has no reference labels, these counts describe what the model output and say nothing about its accuracy. Figure 11 shows the output for three images predicted as Normal (Fig. 11a), Viral Pneumonia (Fig. 11b) and Lung Opacity (Fig. 11c), and Figure 12 summarizes the predicted classes, confidence and mask coverage for the whole set. Text

TABLE X: Framework output on the 341-image DICOM set.

Item	Value
Images processed	341
Execution errors	0
Average prediction confidence	97.3%
Average lung-mask coverage	24.6%
Predicted Normal	291 (85.3%)
Predicted Viral Pneumonia	34 (10.0%)
Predicted Lung Opacity	16 (4.7%)
Predicted COVID	0 (0.0%)

TABLE XI: Insertion and deletion AUCs for the three Grad-CAM examples.

Image	Pred	True	Ins AUC	Del AUC
1	Normal	Normal	0.8761	0.1665
2	Lung_Opacity	Lung_Opacity	0.9188	0.8367
3	Normal	Normal	0.8349	0.5568
		Mean	**0.8766**	**0.5200**

burned into the DICOM images was blurred in all figures. Figure 10 shows the raw U-Net++ probability map and the thresholded mask ($t = 0.3$) for one DICOM image. The lung fields are segmented cleanly; the small isolated region in the lower-left corner of the mask corresponds to a feature of the radiograph itself rather than to lung tissue.

4.4 Explainability and Clinical Review

Figure 13 shows Grad-CAM overlays for three randomly drawn, correctly classified test images (two Normal and one Lung Opacity); warm colors mark the regions with the strongest class-specific activation. The maps are coarse because they come from the 8×8 feature grid of the last convolutional block. In the two Normal examples the activation sits mostly in the central chest around the heart and hilar region and not in the peripheral lung fields, whereas in the Lung Opacity example it spreads over both lower lung zones.

The insertion and deletion AUCs for the three examples are listed in Table XI. The insertion AUCs were 0.876, 0.919 and 0.835 (mean 0.877), and the deletion AUCs were 0.166, 0.837 and 0.557 (mean 0.520). A faithful explanation should give a high insertion AUC and a low deletion AUC. Only the first example does so clearly. In the Lung Opacity example the deletion curve stays near 1.0 until roughly 80% of the pixels have been removed, which means that removing the highlighted pixels barely changed the prediction. With only three images, these values illustrate how the metrics behave and should not be read as evidence about the explanations in general.

Clinical review. Fifteen DICOM images predicted with low confidence in a previous run were reviewed by a medical doctor (MBBS). In the final U-Net++ and DenseNet121 framework, the model agreed with the doctor's diagnosis in 11 out of 15 cases (73.3%) and disagreed in 4 cases (Table XII). All four disagreements were abnormal cases labeled as Normal. Three of these had confidence scores of at least 99.8%.

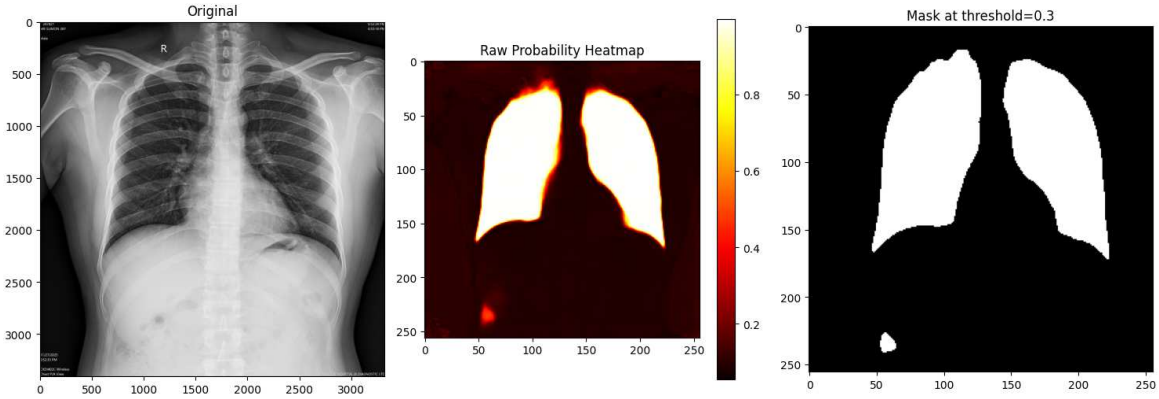


Fig. 10: DICOM radiograph (left), raw U-Net++ probability map (middle) and binary lung mask at threshold $t = 0.3$ (right).

These findings highlight that high prediction confidence does not necessarily indicate a correct classification. Figure 14 shows the framework output for the first six reviewed cases (Table XII, Nos. 1–6). Since the cases were selected based on earlier model confidence rather than random sampling, the results are descriptive and should not be interpreted as an estimate of clinical accuracy.

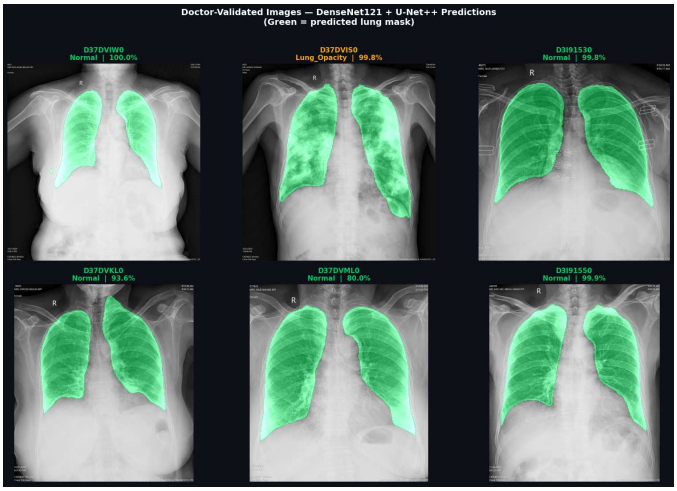


Fig. 14: U-Net++ and DenseNet121 predictions for the first six doctor-reviewed DICOM cases (Table XII, Nos. 1–6). Green indicates the predicted lung mask; the title above each image shows the image ID, predicted class and confidence.

5. Discussion

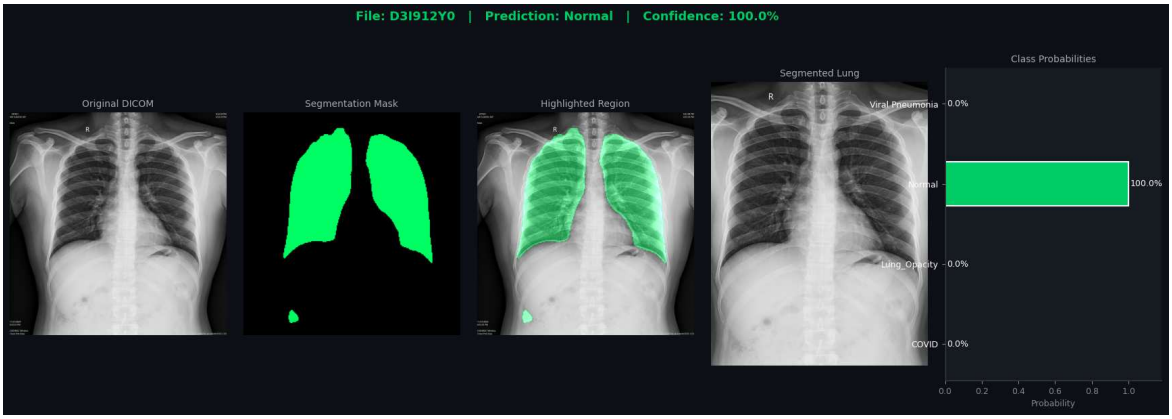
The benchmark results provided a basis for comparing the segmentation and classification models and selecting the best-performing models for the proposed framework. U-Net++ achieved a test Dice score of 0.9872 and an IoU of 0.9747, slightly higher than the U-Net scores of 0.9865 and 0.9733, respectively, which were reported on the validation set. U-Net++ also required more computation, with 119.35 GMac compared with 48.35 GMac for U-Net. All four classification models achieved accuracies above 95%, with DenseNet121, VGG19, and ResNet50 differing by only 0.44 percentage points. DenseNet121 was selected for the proposed framework because it achieved the highest accuracy, macro-F1 score, and Cohen's κ. However, since the models were evaluated using single runs, these small differences should be interpreted cautiously, as the ranking may change with a different random seed.

TABLE XII: Comparison of Doctor's Diagnoses and DenseNet121 Predictions Across Fifteen Cases.

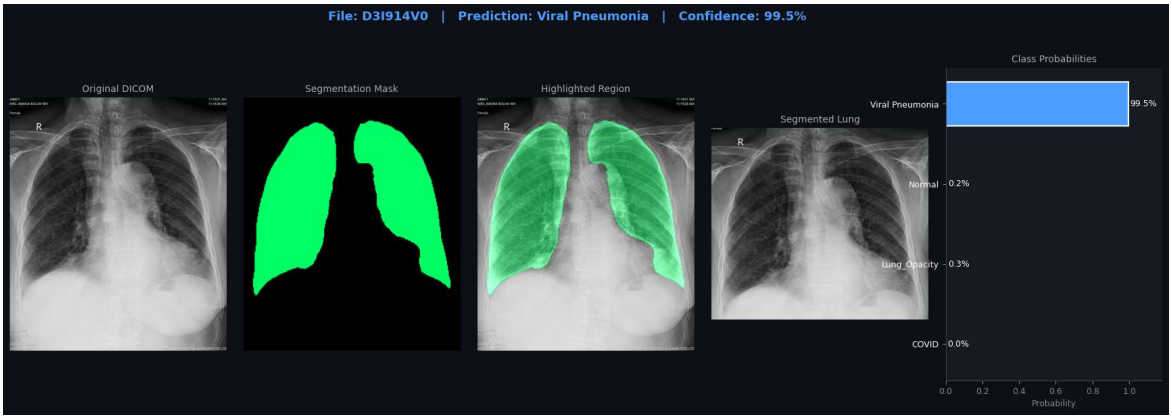
No.	Image ID	Doctor	Model Prediction	Conf. (%)	Agree
1	D37DVIW0	Opacity	Normal	100.0	No
2	D37DVIS0	Opacity	Lung Opacity	99.8	Yes
3	D3I91530	Opacity	Normal	99.8	No
4	D37DVKL0	Normal	Normal	93.6	Yes
5	D37DVML0	Normal	Normal	80.0	Yes
6	D3I91550	Normal	Normal	99.9	Yes
7	D35HD300	Pneumonia	Normal	100.0	No
8	D35HD2M0	Normal	Normal	63.2	Yes
9	D37DVN00	Opacity	Lung Opacity	56.9	Yes
10	D35HCZ10	Pneumonia	Viral Pneumonia	67.9	Yes
11	D35HD2U0	Pneumonia	Viral Pneumonia	57.9	Yes
12	D37DV160	Normal	Normal	77.1	Yes
13	R35HD000	Normal	Normal	86.6	Yes
14	D37DVJA0	Pneumonia	Normal	66.6	No
15	D3I91320	Pneumonia	Viral Pneumonia	64.6	Yes

All real-world DICOM images were processed successfully, and the predicted lung masks covered an average of 24.6% of each image. However, without ground-truth labels, it was not possible to determine whether the predicted class distribution reflected the actual cases. During the doctor's review, the model's predictions matched the doctor's assessment in 11 of 15 cases (73.3%). In the remaining four cases, the doctor identified opacity or pneumonia, but the model classified the images as Normal. Notably, three of these predictions had confidence scores of at least 99.8%. This shows that even highly confident predictions can be wrong, so relying on confidence scores alone may not be enough to detect potential classification errors.

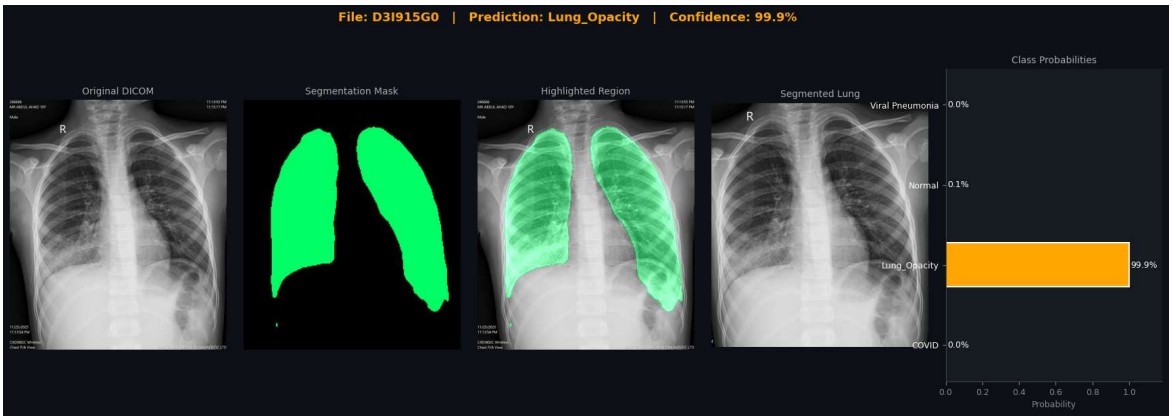
Several differences between the benchmark experiments and the DICOM evaluation may have contributed to these



(a) Normal prediction.



(b) Viral Pneumonia prediction.



(c) Lung Opacity prediction.

Fig. 11: Two-stage framework output for three DICOM images. Each row shows the original radiograph, predicted lung mask, overlay, extracted lung region (bounding-box crop) and class probabilities.

observations. During training, the classifiers received full radiographs, whereas the deployed framework classified lung regions extracted by U-Net++. The preprocessing procedures also differed: classification inputs at inference were normalized using ImageNet statistics, while training used the training-set mean and standard deviation. Similarly, the segmentation model was trained on images scaled to $[-1, 1]$ and evaluated using a threshold of 0.5, whereas the DICOM pipeline scaled images to $[0, 1]$ and applied a threshold of 0.3. Differences in scanners, image contrast, and encoding may have introduced additional variation. Since these factors were not examined through separate ablation experiments, their individual effects on classification performance remain unclear.

The explainability analysis offered limited insight into model predictions, as it covered only three examples. Grad-CAM maps were coarse due to the 8×8 feature grid. In two Normal cases, activations focused mainly on central chest regions, including the cardiac and hilar areas, while the Lung

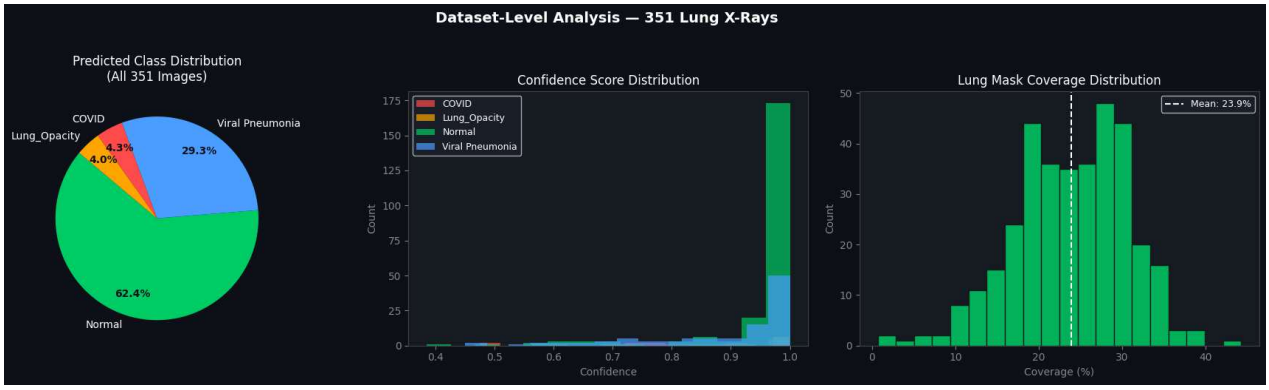


Fig. 12: Predicted classes, confidence, and lung-mask coverage for the 341 DICOM images.

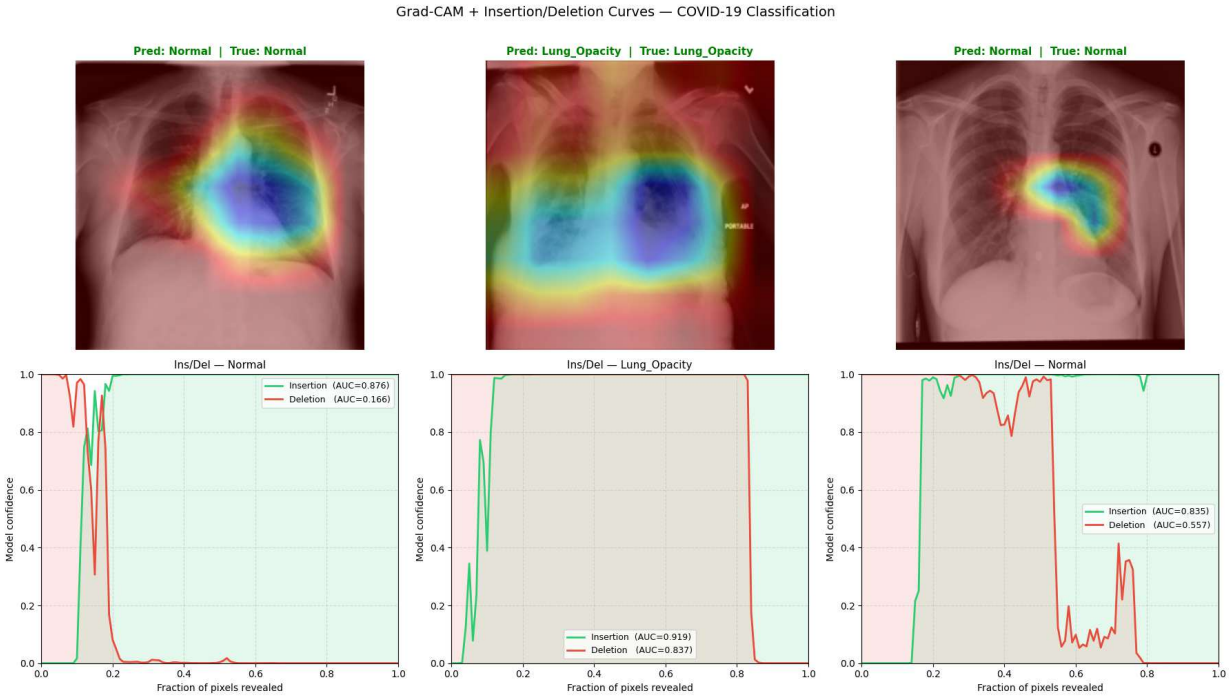


Fig. 13: Grad-CAM overlays (top) and insertion/deletion curves (bottom) for three correctly classified test images of the DenseNet121 classifier, computed on full radiographs. Predicted and true classes are shown above each image; warm colors indicate stronger class-specific activation. Insertion and deletion AUCs are given in the legends.

Opacity case showed activation across both lower lung zones. Insertion and deletion AUCs varied considerably, with a high deletion AUC of 0.837 suggesting limited prediction change after removing highlighted pixels in one case. These findings should not be generalized beyond the selected examples, and further analysis is needed to assess classifier attention specifically within segmented lung regions.

This study has several limitations. The doctor's review was limited to 15 cases selected based on low-confidence predictions from an earlier run, as the reviewing doctor had limited time available. Therefore, the review may not fully represent the clinical performance of the proposed framework. In addition, differences between training and inference preprocessing were not investigated separately, and the lung masks were not post-processed. Small spurious regions outside the lungs could enlarge the bounding box used for cropping, but this effect was not evaluated separately. Future work should align the training and inference pipelines, train and evaluate the classifier using segmented lung regions, and validate the framework on larger datasets from multiple centers. A larger expert review involving randomly sampled cases would provide a more representative assessment of clinical performance. Further explainability analysis should also focus directly on the segmented lung regions to better understand which features influence the model's predictions.

6. Conclusion

This study evaluated three lung segmentation models and four classification models using the COVID-19 Radiography Database. U-Net++ achieved the best segmentation results, while DenseNet121 performed best in classification. These models were combined into a two-stage framework that generates a lung mask and a four-class prediction directly from DICOM chest X-rays. The framework processed all 341 DICOM files from an external diagnostic center without execution errors.

The clinical review, however, highlighted important limitations. Of the 15 cases reviewed by a medical doctor, the framework agreed with the doctor in 11 and disagreed in 4. All four disagreements were abnormal cases classified as Normal by the model, three of them with confidence scores of at least 99.8%. These findings highlight that strong benchmark performance and high prediction confidence do not necessarily guarantee reliable results on real-world clinical images. Future work should focus on training the classifier with segmented lung regions, maintaining consistent preprocessing between training and inference, and evaluating the framework on larger, labeled datasets from multiple centers with more extensive expert review. Such steps are necessary to better understand its reliability and potential role in clinical decision support.

CRediT Authorship Contribution Statement

Mohammed Borhan Uddin: Conceptualization, Software, Methodology, Formal analysis, Investigation, Data curation, Visualization, Supervision, Project administration, Validation, Writing – original draft.
Sanjida Tasneem: Software, Methodology, Formal analysis, Investigation, Data curation, Visualization, Validation, Writing – original draft, review & editing.
Abdur Rahaman: Software, Methodology, Formal analysis, Investigation, Data curation, Writing – review & editing.
Shahariar Nafiz: Software, Methodology, Formal analysis, Investigation, Data curation, Visualization, Writing – review & editing.
Md. Nurul Alam: Software, Methodology, Formal analysis, Investigation, Data curation, Visualization, Writing – review & editing.
Mohammad Shamsul Arefin: Investigation, Supervision, Validation. Writing – review & editing.

Data Availability

The COVID-19 Radiography Database used for training and evaluation is publicly available on Kaggle at https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database. The 341 DICOM chest radiographs from the diagnostic center are available from the authors upon reasonable request.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] D. Li, L. M. Pehrson, L. Tøttrup, M. Fraccaro, R. Bonnevie, J. Thrane, P. J. Sørensen, A. Rykkje, T. T. Andersen, H. Steglich-Arnholm, D. M. R. Stærk, L. Borgwardt, K. L. Hansen, S. Darkner, J. F. Carlsen, and M. B. Nielsen, "Inter- and intra-observer agreement when using a diagnostic labeling scheme for annotating findings on chest x-rays—an early step in the development of a deep learning-based decision support system," *Diagnostics*, vol. 12, no. 12, 2022, art. no. 3112. [Online]. Available: https://doi.org/10.3390/diagnostics12123112

[2] T.-A. Pham and V.-D. Hoang, "Chest X-ray image classification using transfer learning and hyperparameter customization for lung disease diagnosis," *Journal of Information and Telecommunication*, vol. 8, no. 4, pp. 587–601, 2024. [Online]. Available: https://doi.org/10.1080/24751839.2024.2317509

[3] Z. Kadiroglu, E. Deniz, and A. Senyigit, "A comparison of deep learning models for pneumonia detection from chest x-ray images," *Journal of the Faculty of Engineering and Architecture of Gazi University*, vol. 39, no. 2, pp. 729–740, 2024. [Online]. Available: https://doi.org/10.17341/gazimmfd.1204092

[4] L. O. Teixeira, R. M. Pereira, D. Bertolini, L. S. Oliveira, L. Nanni, G. D. C. Cavalcanti, and Y. M. G. Costa, "Impact of lung segmentation on the diagnosis and explanation of COVID-19 in chest X-ray images," *Sensors*, vol. 21, no. 21, p. 7116, 2021. [Online]. Available: https://doi.org/10.3390/s21217116

[5] C. A. Babu, J. Pandu, and G. R. S. Reddy, "Lung disease classification using NF-RVFL optimised by IROA with HybridGNet segmentation and DIAG augmented data in chest X-ray analysis," *International Journal of Signal and Imaging Systems Engineering*, vol. 14, no. 3, pp. 165–185, 2026. [Online]. Available: https://doi.org/10.1504/IJSISE.2026.152503

[6] V. K. Yadav and J. Singhai, "Improved multiclass lung disease classification using segmentation and deep learning from chest X-ray images," *IETE Journal of Research*, pp. 318–331, 2025. [Online]. Available: https://doi.org/10.1080/02564602.2025.2501936

[7] D. Mohanapriya and M. Subhashree, "Hybrid attention-based Grad-CAM with GAN-augmented training for interpretable medical image classification for pneumonia detection," in *2026 International Conference on System, Computation, Automation and Networking (ICSCAN)*, 2026. [Online]. Available: https://doi.org/10.1109/ICSCAN66520.2026.11588418

[8] S. K. Samarla and P. Maragathavalli, "ASCE-XAI: Anatomical segmentation and color enhancement guided explainable AI for pneumonia detection in chest radiographs," in *2025 5th International Conference on Soft Computing for Security Applications (ICSCSA)*, 2025. [Online]. Available: https://doi.org/10.1109/ICSCSA66339.2025.11170826

[9] N. Veeramani, R. S. Sherine, S. S. Prabha, S. Srinidhi, and P. Jayaraman, "NextGen lung disease diagnosis with explainable artificial intelligence," *Scientific Reports*, vol. 15, 2025, art. no. 33052. [Online]. Available: https://doi.org/10.1038/s41598-025-07603-4

[10] A. K. Dubey, A. Jain, S. Khan, M. Kaur, A. Panwar, M. A. Alawad, J. Khan, and M. N. Alam, "Comparative evaluation of deep learning models for lung segmentation in chest X-rays: applications in infectious disease screening," *BMC Infectious Diseases*, vol. 26, no. 1, 2026, art. no. 922. [Online]. Available: https://doi.org/10.1186/s12879-026-13041-y

[11] M. Jannat, S. A. Birahim, M. A. Hasan, T. Roy, L. Sultana, H. Sarker, S. Fairuz, and H. A. Abdallah, "Lung segmentation with lightweight convolutional attention residual U-Net," *Diagnostics*, vol. 15, no. 7, p. 854, 2025. [Online]. Available: https://doi.org/10.3390/diagnostics15070854

[12] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," 2023. [Online]. Available: https://arxiv.org/abs/2304.02643

[13] M. Ali, T. Wu, H. Hu, Q. Luo, D. Xu, W. Zheng, N. Jin, and C. Yang, "A review of the segment anything model (sam) for medical image analysis: Accomplishments and perspectives," *Computerized Medical Imaging and Graphics*, 2025.

[14] S. Gautam and A. C. Katuwal, "Explainable pneumonia detection from chest X-ray images using a ResNet-50 based deep learning framework," *Journal of Advanced College of Engineering and Management*, vol. 12, no. 1, 2026. [Online]. Available: https://doi.org/10.3126/jacem.v12i01.93909

[15] X.-h. Lan, Y.-x. Zhang, W.-h. Yuan, F. Shi, and W.-l. Guo, "Image-based deep learning in diagnosing mycoplasma pneumonia on pediatric chest X-rays," *BMC Pediatrics*, vol. 24, 2024, art. no. 720. [Online]. Available: https://doi.org/10.1186/s12887-024-05204-0

[16] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017, pp. 618–626. [Online]. Available: https://doi.org/10.1109/ICCV.2017.74

[17] F. A. A. Slimani and M. Bentourkia, "Improving deep learning U-Net++ by discrete wavelet and attention gate mechanisms for effective pathological lung segmentation in chest X-ray imaging," *Physical and Engineering Sciences in Medicine*, vol. 48, no. 1, pp. 59–73, 2025. [Online]. Available: https://doi.org/10.1007/s13246-024-01489-8

[18] R. T. Wahyuningrum, R. A. Fadillah, I. Yunita, B. D. Satoto, A. Muntasa, A. K. Sari, P. Rahardjo, D. A. Dewi, and A. Bauravindah, "Optimizing hyperparameters for dual-attention network in lung segmentation," *Intelligence-Based Medicine*, vol. 11, p. 100221, 2025. [Online]. Available: https://doi.org/10.1016/j.ibmed.2025.100221

[19] T. Rahman, "COVID-19 Radiography Database," Kaggle, 2020, accessed: 2026. [Online]. Available: https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database

[20] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, ser. Lecture Notes in Computer Science, vol. 9351. Springer, 2015, pp. 234–241.

[21] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, ser. Lecture Notes in Computer Science, vol. 11045. Springer, 2018, pp. 3–11.

[22] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4700–4708. [Online]. Available: https://doi.org/10.1109/CVPR.2017.243

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. [Online]. Available: https://doi.org/10.1109/CVPR.2016.90

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014. [Online]. Available: https://arxiv.org/abs/1409.1556

[25] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1251–1258. [Online]. Available: https://doi.org/10.1109/CVPR.2017.195