U-Net Performance in Lung Segmentation on Chest X-Rays featuring Multiple Pathologies

Viorel DEDIU*, Andreea UDREA

Department of Automatic Control and Systems Engineering, Faculty of Automatic Control and Computer Science, National University of Science and Technology Polytechnic of Bucharest, 313 Splaiul Independenței, Bucharest, 060042, Romania
viorel.dediu@stud.acs.upb.ro (**Corresponding author*), andreea.udrea@upb.ro

Abstract: Automated lung segmentation in chest X-rays is essential for diagnosing lung diseases, yet ensuring model generalization across different datasets remains a challenge. This study evaluates the performance of U-Net on three public datasets (Darwin, Montgomery, Shenzhen) covering multiple pathologies and introduces a fine-tuning approach using self-generated pseudo-labels to correct label inconsistencies across datasets. The findings of this study show that model training on a single dataset results in a poor generalization (a Dice score below 90%). However, by first training a baseline model on all the three datasets and then applying the proposed fine-tuning strategy, the segmentation accuracy improves significantly, increasing the Dice score from 96.13% to 99.16%. This performance matches that of larger models while maintaining a compact architecture. Additionally, the impact of preprocessing techniques (CLAHE, Gaussian, Laplacian filtering) was assessed and it was found that their contribution to segmentation accuracy is minimal (a Dice score below 1%), which highlights that dataset diversity and label consistency are more influential factors than preprocessing. These findings provide an effective alternative to increasing model size while ensuring a high segmentation accuracy.

Keywords: Lung segmentation, U-Net, X-ray, Model generalization across multiple pathologies.

1. Introduction

Image segmentation plays a crucial role in the diagnosis of lung diseases by providing an accurate and efficient method for correctly identifying the organ of interest (Selvan et al., 2020) and different disorders (Lei et al., 2020).

Accurate organ delineation despite the existing pathologies is crucial, for example to determine the overall lung volume such that a viable organ percentage can be computed (e.g. in the case of COVID lesions, just a part of the lung still functions correctly, and this information is crucial for further treatment planning (Lee et al., 2020)). Also, lung segmentation helps as a preliminary step when the detection of a different disease is of interest, as it improves the results (Gordaliza et al., 2018).

The chest radiography (X-ray) is one of the most used non-invasive methods for assessing lung diseases and the automatic processing of this type of images is of high interest, due to the large number of procedures of this sort (Shen, Wu & Suk, 2017). However, due to their higher noise levels (e.g. the overlapping of organs such as the ribs and clavicle), X-rays are challenging (compared to CTs for example) when automatic segmentation is of interest. Substantial research efforts lead to high performances of X-ray lung segmentation algorithms varying from rule-based algorithms to neural network models (Agrawal & Choudhar, 2023).

The study of Azimi et al. (2022) examines how precise lung segmentation enhances the classification performance. By isolating lung regions, a CNN-based model (e.g. ResNet-50) improved the accuracy from 88% to 92% for pneumonia and from 84% to 90% for COVID-19 detection. This highlights that accurate segmentation reduces noise and helps models focus on relevant areas.

X-ray lung segmentation methods perform well on healthy lungs, but their performance may vary when applied to patients with diseases such as COVID-19 or tuberculosis. This is because such conditions can cause significant changes in the lung structure, such as opacities, consolidations, cavities or other abnormalities, which are not detected as part of the segmented lung.

Therefore, ensuring model generalization across datasets with different diseases, acquisition conditions and labelling styles remains an open challenge.

This study evaluates the generalization capability of the U-Net architecture for lung segmentation across multiple datasets, which include diverse imaging conditions and pathologies.

https://doi.org/10.24846/v34i2y202504

Unlike prior studies that primarily focused on modifying network architectures, this paper investigates how different preprocessing strategies (CLAHE, Gaussian, Laplace filtering), training strategies and label consistency influence the segmentation performance.

The key novelty of this work lies in proposing a fine-tuning strategy using self-generated pseudo-labels to mitigate labelling inconsistencies across datasets. This approach aims to improve segmentation performance while maintaining a lightweight architecture, making it more suitable for real-world clinical applications.

The remainder of this paper is organized as follows. Section 2 presents the related work, summarizing the key advancements in lung segmentation and contextualizing the proposed approach in the recent literature. Section 3 introduces the dataset and preprocessing pipeline, detailing the acquisition parameters, annotation strategies, and normalization techniques employed for preparing the chest X-rays dataset. Further on, Section 4 sets forth the proposed method, describing the architecture of the U-Net model, the training procedures, and the different testing scenarios implemented for evaluation purposes. Section 5 presents the experimental results, providing the quantitative metrics and qualitative assessments across various testing conditions. Section 6 discusses the implications of the findings of this paper, the limitations of the proposed approach, and potential directions for future work. Finally, Section 7 concludes this paper, highlighting the key contributions and practical significance of the proposed unsupervised segmentation technique and possible future enhancements for a better model performance.

2. Related Work

Neural networks have been extensively used for the lung segmentation task.

In (Rahman et al., 2021), an adapted version of the U-Net architecture was proposed for automatic lung segmentation in chest X-rays, achieving a Dice coefficient (DC) of 94.21% on the Montgomery County dataset.

An improved model for lung segmentation in chest X-rays was developed using the U-Net architecture with pre-trained EfficientNet-b4 as the encoder and residual blocks, and the LeakyReLU activation function in the decoder. The model obtained a Dice coefficient (DC) of 97.9% on the JSRT dataset and a DC of 97.7% on the Montgomery County dataset (Liu et al., 2022).

In (Kim & Lee, 2021), adding attention modules to a U-Net architecture increased the segmentation performances to a DC of 98.2% on the Montgomery dataset and to a DC of 95.4% on the Shenzhen dataset.

The work of Naqvi et al. (2022) presents a new method that enhances the standard U-Net architecture by incorporating morphological operations (e.g. dilation, erosion) as post-processing steps to refine segmentation and achieve a Dice Similarity Coefficient above 95% on the Montgomery and Shenzhen datasets.

The study of Bombiński et al. (2024) highlights a key challenge in automated lung segmentation: underestimation of lung regions due to anatomical variability and image noise. The findings show that for datasets like JSRT and Montgomery, many segmentation models achieve an IoU (Intersection over Union) below 85%, underscoring the need for more robust preprocessing techniques and advanced architectures to improve the segmentation accuracy.

A comparative analysis in (Hryniewska-Guzik et al., 2024) evaluated the performance of several deep learning architectures, including U-Net, ResNet, and DenseNet. The results showed that while the standard U-Net achieved an average Dice Similarity Coefficient (DSC) of 92%, more advanced models, such as U-Net++ and hybrid architectures incorporating Attention Mechanisms, reached a DSC of over 94% demonstrating the benefits of architectural improvements. Additionally, this study found that data augmentation and transfer learning significantly enhanced the segmentation robustness across diverse datasets.

3. Dataset and Preprocessing

3.1 Dataset Description

In the study of Danilov et al. (2022), three publicly available datasets containing lung X-rays images and the corresponding segmentation masks are used. The datasets are the following: the Darwin dataset (6,106 images), the Montgomery dataset (138 images), and the Shenzhen dataset (566 images).

Darwin dataset: This dataset primarily includes heart and lung opacities, which are useful for assessing the severity of viral pneumonia (e.g. COVID-19). It contains 6,106 images, out of which 1,397 depict viral pathologies, 2,591 bacterial pathologies, and 2,118 are normal images. The images vary in resolution and orientation, ranging from 156x156 pixels to 5600x4700 pixels. Some chest radiographs are of lower quality compared to standard X-rays. Lung segmentations were performed by human annotators using Darwin's Auto-Annotate AI and reviewed by expert radiologists (Anon, n.d.).

Shenzhen dataset: Published by the United States National Library of Medicine, this dataset contains normal chest X-rays and X-rays featuring tuberculosis manifestations, aimed at supporting research on the automated diagnosis of lung diseases, particularly tuberculosis. The data was collected from the Department of Health and Human Services (Maryland, USA) and Shenzhen No. 3 People's Hospital (Shenzhen, China) (Danilov et al., 2022). The dataset includes 566 images, 240 from tuberculosis patients and 326 from healthy individuals.

Montgomery dataset: The dataset contains 138 images: 58 from patients diagnosed with tuberculosis and 80 from healthy individuals (Jaeger et al., 2014).

3.2 Image Preprocessing

All images are grayscale images (they have only a one-color channel representing light intensity). For this study, they were resized to 256x256 pixels to ensure uniformity across the dataset, simplifying multi-layer convolution and pooling computations while reducing computational complexity of larger images (e.g. 5600x4700 pixels). Standardizing the image size also helped eliminate inconsistencies and provided a uniform basis for training and evaluation.

The following preprocessing strategies were combined and tested for evaluating their impact on the final models' performances:

a. *CLAHE (Contrast Limited Adaptive Limited Histogram Equalization)* - a preprocessing technique used to enhance image contrast by applying local histogram equalization while limiting noise amplification was employed in order to enhance the contrast locally. This approach increases the level of local detail and prevents contrast overload by limiting contrast enhancement, thus avoiding noise or artifacts (Zuiderveld, 1994);

b. *Gaussian filter* was used to reduce noise and smooth fine variations in the images. This eliminates artifacts and improves the overall image quality. The filter applies the Gaussian function to the image by convolution, calculating a weighted average of the neighbourhood pixels, where pixels closer to the center pixel have a greater influence (Marr & Hildreth, 1980);

c. *Laplace filter* is an edge detection filter that highlights areas with rapid intensity variations in the image. It helps to emphasize fine details and anatomical structures such as organ outlines or lesions. The filter is based on the Laplacian operator (a second-order derivative operator) and measures the rate at which the intensity gradient changes (Haralick, 1984);

After filtering, all images were normalized to [0, 1]. This step was performed to prepare the image for input to the deep machine learning model.

No data augmentation techniques were applied.

All preprocessing was performed before model training, ensuring that every dataset underwent consistent transformations.

4. Method

4.1 The U-Net Convolutional Neural Network Model

The objective of this study is to analyse the performance of a classical U-Net neural network (Ronneberger, Fischer & Brox, 2015) used in the

segmentation process on three different datasets containing radiological images of a varying resolution and quality.

The architecture employed in this study is shown in Figure 1. It consists of four levels of convolutional and pooling layers, followed by a central convolutional layer, and then four levels of upsampling and concatenation. The final layer is an output layer with a sigmoid activation function. Each convolution layer contains 32, 64, 128, 256, and 512 filters, respectively, with a kernel size of 3x3.

4.2 Training and Testing Scenarios

The model is trained using the datasets presented in the previous section and the following scenarios (which leads to models with different performances):

Scenario 1. The Shenzhen and Montgomery datasets were used in the training and validation process, while the testing was done with the Darwin dataset:

Train set - 562 images

Validation set- 142 images

Test set- 6.106 images

This scenario tests the model's performance on chest X-rays featuring normal lungs, COVID-19 pathology, or viral and bacterial pneumonia, using a model trained on X-rays of healthy individuals and tuberculosis patients.

Scenario 2. The Darwin dataset was randomly split as follows:

Train set - 4884 images

Validation set - 1222 images

The Shenzhen and Montgomery datasets, as described in the previous section, were used for testing:

Test set 1 - Montgomery dataset - 138 images

Test set 2 – Shenzhen dataset – 566 images

Scenario 2 evaluates the proposed model's performance on chest X-rays featuring normal lungs or tuberculosis pathology, using a model

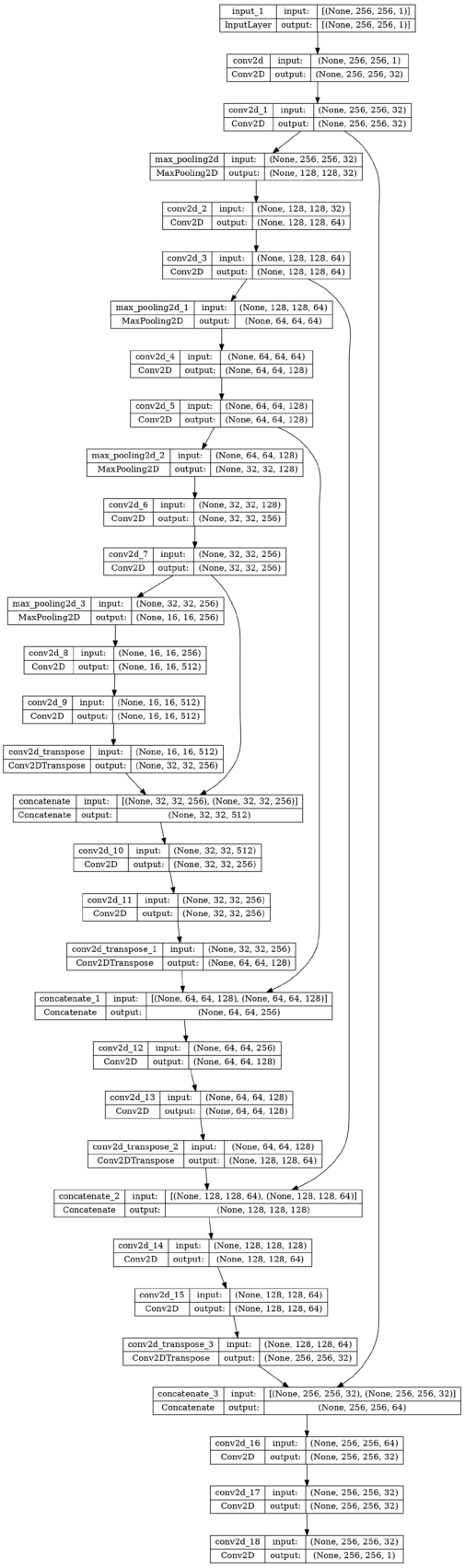


Figure 1. U-Net - graphical representation and structural details

trained on chest X-rays featuring COVID-19 or viral pneumonia.

The purpose of these two scenarios is to test if the segmentation performance using a model trained on specific pathologies can lead to comparable results when used on data related to another pathology.

Scenario 3a. The model was trained and tested on all tree datasets according to the following distribution:

Train set - 4,736 images

Validation set - 1,037 images

Test set - 1,037 images

The chest X-ray images from the Darwin, Montgomery and Shenzhen datasets were randomly split intro train (approx. 70%), validation (approx. 15%) and test (approx. 15%) sets, keeping as much as possible the same distribution of chest X-rays of healthy individuals vs. lung diseases as in the original datasets.

This scenario evaluates the model's performance on a homogeneous dataset containing all types of pathologies, including COVID-19, viral and bacterial pneumonia, and tuberculosis, as well as normal chest X-rays.

Scenario 3b. The model obtained in scenario 3a is fine-tunned by reusing the train images from the Montgomery and Shenzhen datasets but using the segmentation masks generated by the model obtained in scenario 3a instead of the original masks. This fine-tunning step was pursued due to an observation regarding quite different manual segmentation styles across datasets, namely the left lung segmentation in the Darwin dataset is performed such that it takes into consideration a larger lung area near the heart, while in the Shenzhen and Montgomery datasets the left lung segmentation is performed in a narrower fashion. From the authors` observation this difference can lead to a model that performs an under-segmentation of the left lung when it is affected by a disease in the proximity of the heart. This discrepancy confuses the model, leading to a systematic under-segmentation in cross-dataset tests. This fine-tunning step aims to diminish the effects of different manual segmentation styles across databases, while keeping the network size unchanged.

All the employed models were trained using an early stopping mechanism, starting with 50 epochs. During training, the Early Stopping callback continuously monitored the validation loss (val_loss). If the validation loss did not improve for 15 consecutive epochs, the callback would stop the training early to prevent overfitting.

The batch size for the training was set to 4, and a Nvidia GeForce RTX 3090 GPU was used for computational resources.

These scenarios, regarding lung segmentation based on U-Net on datasets including lung X-rays acquired from various devices and with a variety of pathologies are relevant from the following perspectives:

Understanding limitations: When a model is created using data containing just one type of pathology its performance can be very good, but most of such models described in the literature do not discuss the performance obtained for a different data distribution/other pathology.

Improved Generalization: The variability of the input data enhances the model's ability to generalize to new, unstructured images and perform effectively in real clinical environments, where images often come from different devices with a varying quality.

Learning Diverse Features: By learning features associated with different conditions, the model becomes more adept at distinguishing between multiple types of lung lesions and healthy states. This is crucial in clinical applications, where an accurate and rapid identification of multiple overlapping or similar pathologies is required.

Enhanced Clinical Applicability: For a segmentation model to be effective in real-world scenarios, it must handle diverse radiological images across a broad spectrum of conditions. Training and validating on varied databases ensure that the model can be reliably deployed in clinical settings, delivering consistent lung segmentation results even if the organ can feature one of the many possible pathologies.

5. Results

5.1 Performance Measures

The most common performance measures used for analysing the segmentation models' efficiency are Binary Accuracy, the Dice coefficient and the Jaccard coefficient.

The Binary Accuracy expressed in equation (1) measures how well a model classifies each pixel in an image, such as an X-ray with lungs as the primary subject and the background as the negative space:

$$\text{Binary Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

where TP (True Positives) are pixels correctly classified as lung, TN (True Negatives are pixels correctly classified as background (non-lung), FP (False Positives) are background pixels incorrectly classified as lung and FN (False Negatives) are lung pixels incorrectly classified as background.

Binary accuracy (BA) is a metric used to evaluate the performance of a binary classification model. It measures the proportion of correctly predicted cases out of the total cases.

The Dice coefficient is expressed in equation (2):

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

where A represents the ground truth (reference) segmentation - pixels in the lung segmentation mask, B is the set of pixels in the automatic segmentation obtained by the algorithm and $|A|\cap|B|$ is the number of common pixels (overlapping pixels) between the two segmentations.

The Dice coefficient is an overlap metric employed for evaluating the performance of image segmentation algorithms, especially in the field of medical image processing.

To compute the Jaccard coefficient based on the Dice coefficient, one can use equation (3), which provides the Jaccard index (JI) based on the Dice index:

$$\text{JI} = \frac{DICE}{2 - DICE} \quad (3)$$

The Jaccard coefficient is used to measure how well the predicted and actual lung regions overlap. If there are unique elements (noise) in the data set, they will not overly influence this coefficient, because the Jaccard index relies on the ratio of intersection over union, making it more robust to insignificant variations.

5.2 U-Net Performance in Different Training-testing Scenarios

Table 1 summarizes the results achieved by testing the U-Net models obtained under different scenarios.

Table 1. Assessment of the U-Net architecture for different scenarios

Scenario	Assessment metrics		
	BA (%)	DC (%)	JI (%)
Scenario 1 – no filters	85.45	74.65	59.55
Scenario 1 - with CLAHE and Gaussian filters	85.03	72.86	57.31
Scenario 1 - with CLAHE, Gaussian & Laplace filters	84.88	74.26	59.06
Scenario 2 – without filters, the Montgomery test set	94.03	89.70	81.32
Scenario 2 – without filters, the Shenzhen test set	93.67	89.32	80.70
Scenario 2 – with CLAHE and Gaussian filters, the Montgomery test set	94.05	89.80	81.49
Scenario 2 – with CLAHE and Gaussian filters, the Shenzhen test set	93.83	89.60	81.16
Scenario 2 – with CLAHE, Gaussian and Laplace filters, the Montgomery test set	93.68	89.20	80.51
Scenario 2 – with CLAHE, Gaussian and Laplace filters, the Shenzhen test set	93.62	89.23	80.55
Scenario 3a – without filters	97.07	96.13	92.55
Scenario 3b – without filters	95.87	99.16	98.33

A. Results Obtained for Scenario 1

Scenario 1 shows that training the model on datasets with normal chest X-rays or tuberculosis diagnoses does not yield a very strong performance (DC under 80%) when tested on a dataset containing images of viral pneumonia or large lung opacities.

By analysing Figure 2(a) one can observe that there are segmentation errors related to the shape and size of lung segmentation, and the model includes some areas that should not be part of the lungs. These differences suggest that the model needs improvement.

In Figure 2(c), the training loss (blue) and validation loss (orange) decrease significantly in the early epochs, indicating that the model is learning quickly. The convergence of the training and validation curves indicates that the model is not overfitting. Although the curves are generally smooth, minor fluctuations in validation loss, particularly between epochs 20 and 30, likely result from variations within the validation data itself.

By applying CLAHE and Gaussian filters (Figure 3), the generated mask is imprecise, containing artifacts and irregularities, particularly in areas where the lungs meet the rib cage. The boundaries are less sharp, with some misclassifications in areas where lung and non-lung regions are not clearly separated.

By applying all three filters, namely the CLAHE, Gaussian and Laplace filters, the segmentation results are those depicted in Figure 4.

The original image is a processed and filtered version of a chest X-ray, likely highlighting edges or certain structural details through edge detection or similar filtering techniques.

The lung segmentation performed by the employed model shows notable discrepancies from the radiologist's mask, with irregular boundaries, missing sections near the edges, under-segmented regions, and inaccuracies in the upper and lower lobes, resulting in a less smooth and well-defined output with false segmentations and visible gaps.

B. Results Obtained for Scenario 2

Scenario 2 shows that the training and evaluation of two separate models on the two types of images leads to mediocre results: a DC under 90% (Table 1) for lung segmentation either for lungs without lesions or for situations involving a diagnosis of viral pneumonia or tuberculosis.

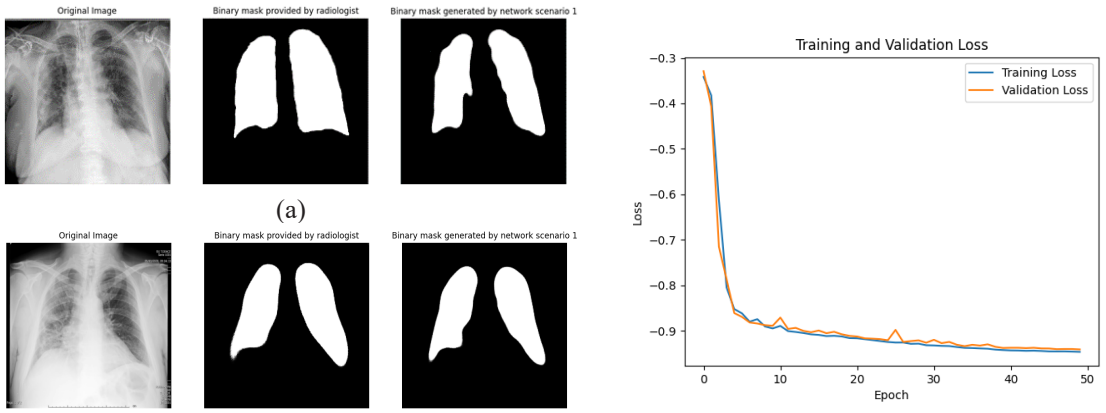


Figure 2. U-Net architecture segmentation results for scenario 1: (a) and (b) Two Chest X-rays - original image, binary mask provided by radiologist, the predicted mask; (c) Training and Validation Loss

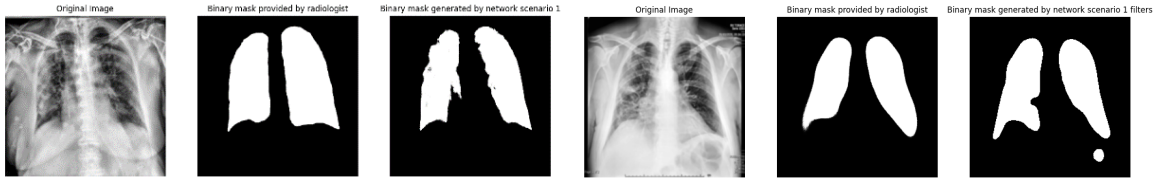


Figure 3. Two images containing the U-Net architecture segmentation results for scenario 1 with 2 filters (CLAHE and Gaussian)

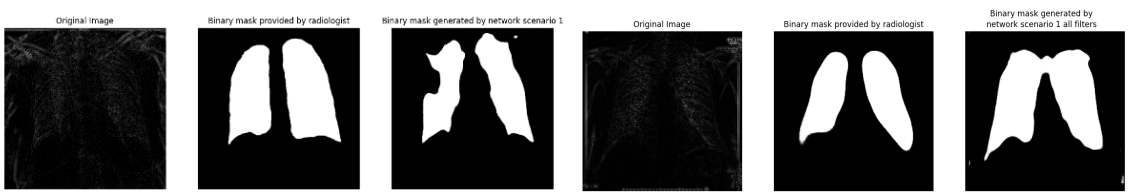


Figure 4. Two images containing the U-Net architecture segmentation results for scenario 1 with 3 filters

Figures 5 (a) and 5(b) show the masks obtained with a model trained in scenario 2; they feature a few imperfections, such as the inclusion of an additional region (in the lower right region).

The images on the right side display the binary mask produced by the model (network scenario 2). The model attempts to isolate the lungs in the X-ray image similarly to the radiologist's mask, but there are some notable discrepancies: the mask generated by network scenario 2 captures the general lung shape, but its edges are less smooth and precise than in the case of the mask provided by the radiologists. Some areas of the lungs appear to be over-segmented or under-segmented. In the lower region of the right lung (the left side of the binary mask), there is a misclassified region where a non-lung area is falsely included in the mask. This artifact indicates that the network is struggling to correctly differentiate lung tissue from other areas. The overall structure of the lungs is present, but parts of the lungs, particularly near the bottom, are either missing or not as cleanly segmented as they should be.

The plot in Figure 5(c) shows us that after an initial drop, both curves stabilize, with the training loss gradually decreasing to about -0.95, while the validation loss hovers around -0.90.

The close alignment between the two curves suggests that the model is not overfitting, as the training and validation losses follow similar trends and remain relatively close throughout the training process.

From around epoch 10 onwards, the validation loss plateaus with some slight fluctuations. This is a common which suggests that the model's performance on unseen data has stabilized. The validation loss has a minor upward trend in the later epochs (approximately between the epochs 40 and 50).

Further on, in scenario 2, CLAHE and Gaussian filters were applied, as shown in Figure 6.

Figure 6 shows the results for a segmentation mask produced by the model (network scenario 2 with CLAHE and Gaussian filters). The model-generated mask accurately captures the general shape and boundaries of the lungs with minor curvature differences, particularly at the bottom edges, and is relatively smooth and clean with few artifacts; slight gaps are present at the lower right lung boundary but are minimal and unlikely to impact on clinical performance significantly.

By applying all three filters the results depicted in Figure 7 are obtained:

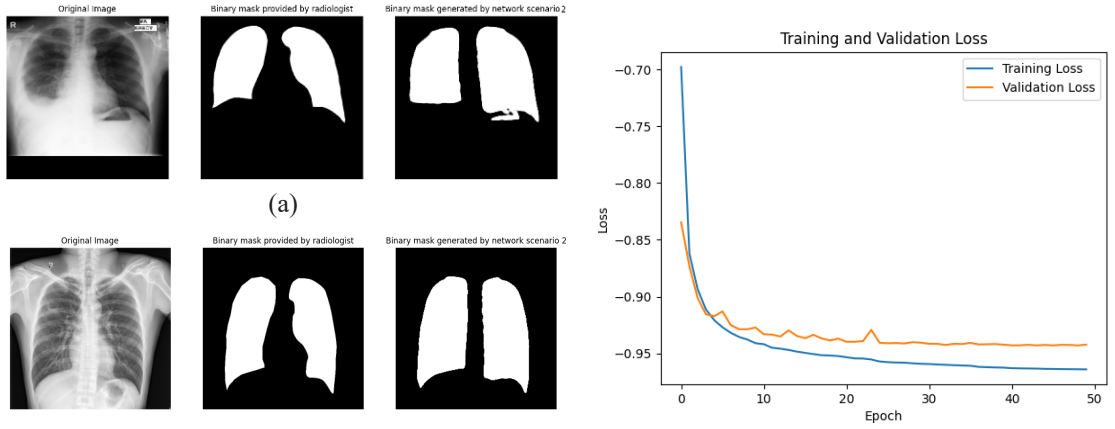


Figure 5. U-Net architecture segmentation results for scenario 2: (a) and (b) Two Chest X-rays - original image, binary mask provided by radiologist, the predicted mask; (c) Training and Validation Loss

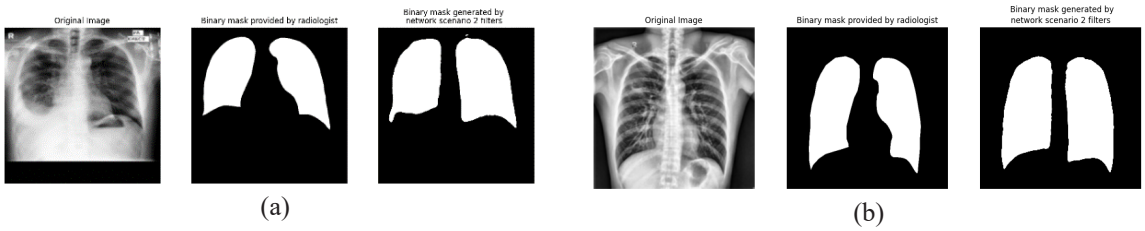


Figure 6. Two images containing the U-Net architecture segmentation results for scenario 2 with 2 filters (Clahe and Gaussian)

Figure 7 shows the binary mask generated by the model using scenario 2 with all filters applied. The model's attempt at segmentation has the following characteristics: the general shape of the lungs is well captured, but some inaccuracies are present, particularly near the lower edges of the lungs, and the contours of the lungs are less smooth in comparison with the radiologist's mask. Also, there are slight irregularities and gaps, especially in the lower and side areas, suggesting that the model struggles to perfectly capture the lung boundaries with all the filters applied, and there are small artifacts at the bottom of the right lung (the left side of the binary mask), where the network appears to have misclassified some regions.

For the test datasets the performance of the pre-processing strategies did not significantly increase either in Scenario 1 or Scenario 2 so in the next two scenarios they were no longer applied.

C. Results obtained for Scenario 3a

Scenario 3 includes the results of the training and evaluation across all three databases, with a DC of 96% (Table 1) which is in line with the state of the art results.

The model manages to segment the lungs reasonably well (Figures 8(a) and 8(b)), but there is room for improvement, particularly with regard to the contours and edge details. The model may benefit from fine-tuning or additional training to improve the segmentation accuracy.

Figure 8(c) shows that after approximately 5-10 epochs, the validation loss stabilizes, indicating that the model has reached an equilibrium point where there is no significant improvement in its ability to generalize on the validation dataset.

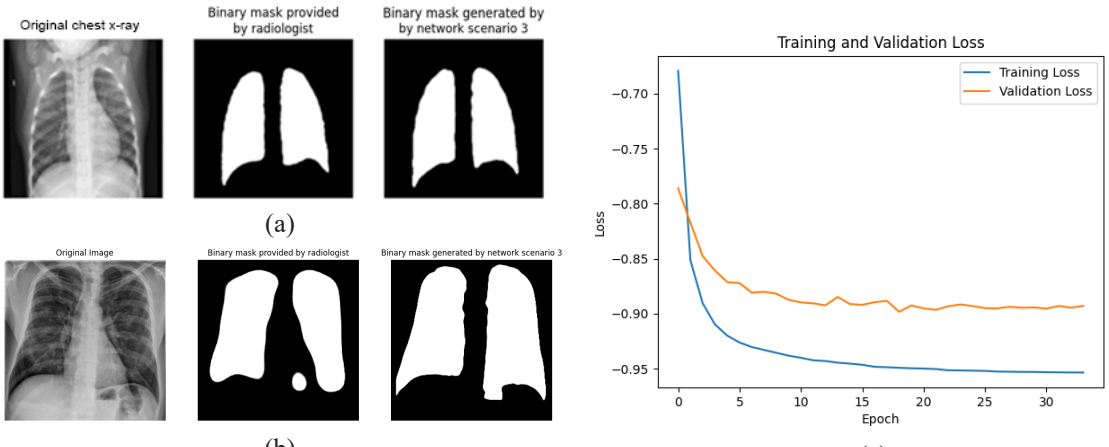
D. Results obtained for Scenario 3b

Scenario 3b includes the results obtained by fine-tunned model from scenario 3a. Fine-tunning lead to a higher segmentation accuracy (Table 1), with a DC of 99.16% (a 3% increase in DC in comparison with scenario 3a). The model size is unchanged, but its performance improves drastically, proving that the dataset labelling consistency is more critical than adding more parameters.

In Scenario 3b, the model was fine-tuned using pseudo-labels generated in Scenario 3a, while maintaining the original validation and test sets. This strategy aimed to reduce the inconsistencies

(a) (b)

Figure 7. Two images containing the U-Net architecture segmentation results for scenario 2 – with 3 filters



(a) (b) (c)

Figure 8. U-Net architecture segmentation results for scenario 3a: (a) and (b) Two Chest X-rays - original image, binary mask provided by radiologist, the predicted mask; (c) Training and Validation Loss

caused by varying annotation styles across datasets. By learning from unified pseudo-labels, the model adopted a more consistent segmentation approach, improving its alignment with general annotation patterns. As a result, it achieved higher Dice scores and reduced the uncertainty in ambiguous anatomical regions. These improvements reflect an enhanced generalization ability and robustness, confirming the effectiveness of pseudo-label-based fine-tuning for harmonizing the cross-dataset segmentation performance. The employed fine-tuning strategy in Scenario 3b inherently addresses the goal of ensemble learning by unifying diverse annotation styles into a single, consistent segmentation model. By leveraging pseudo-labels from the baseline model (Scenario 3a), this approach reconciles inter-dataset discrepancies without the computational cost of training multiple models. As a result, it achieves a robust and consistent performance, similar to what ensemble averaging would provide. Nonetheless, future work could consider further investigating model ensembles.

Figures 9 (a-c) depict well-segmented x-rays masks for different pathologies or normal chest X-rays. To that, the plot in Figure 9(d) shows that the model performance is good, with a robust generalization ability.

Also, in Figure 9(d) it can be observed that the training and validation losses stabilize at similar values, suggesting that the model is performing consistently and shows no obvious signs of over-training. No major modifications are required, but a continuous monitoring is recommended to maintain this performance.

Table 2 features the Dice coefficient, which was computed for each type of disease, considering scenario 3b. It can be observed that a good performance is obtained for pneumonia chest X-rays, while in the case of tuberculosis an improvement is needed. The weight of tuberculosis cases in the dataset is 4.37%, which, most probably, leads to poorer results. However, for the images in the Shenzhen dataset the results are better than those obtained in the previous scenarios.

Table 2. Assessment of the U-Net models form Scenario 3b for different diseases

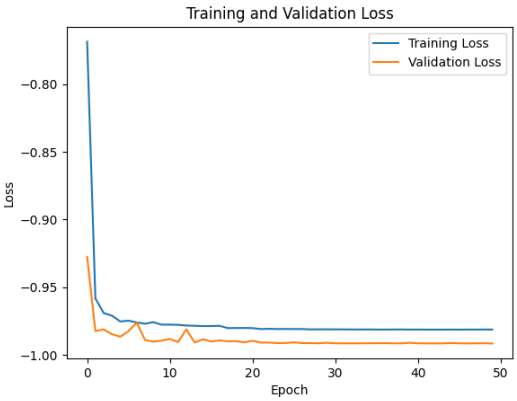
Disease	Dice Coefficient
Pneumonia	97.04%
Tuberculosis - the Montgomery dataset	82.98%
Tuberculosis – the Shenzhen dataset	91.51%

For a deeper understanding of how regularization affects model performance, this study evaluates the impact of different dropout rates on lung

(a) Chest X-rays original image – normal; binary mask provided by radiologist; predicted mask

(b) Chest X-rays original image –interstitial pneumonia; binary mask provided by radiologist; predicted mask

(c) Chest X-rays original image – tuberculosis; binary mask provided by radiologist; predicted mask



(d) Training and validation loss

Figure 9. U-Net architecture segmentation results for scenario 3b

segmentation performance. The U-Net models were trained with dropout rates of 0.1, 0.3, 0.5, and 0.7, and their segmentation accuracy was assessed using the Dice Coefficient (DC).

To estimate model uncertainty, Monte Carlo Dropout (using 30 forward passes per image) was applied, generating uncertainty maps which highlight the regions of inconsistent predictions. The results show that Scenario 3b exhibits a slightly lower mean uncertainty (0.003149) and standard deviation (0.000731) in comparison with Scenario 3 (mean uncertainty: 0.003471, standard deviation: 0.001214), indicating that the fine-tuned model provides more stable and confident segmentation predictions across the test set.

As shown in Figure 10, the obtained results indicate that:

- A low to moderate dropout rate (0.1–0.3) maintains a high segmentation accuracy and improves the generalization ability;
- Excessive dropout (>0.5) significantly degrades the model's performance, increasing the segmentation variability.

This analysis suggests that moderate dropout enhances the model's robustness, while excessive dropout negatively impacts on the segmentation quality. The integration of uncertainty estimation techniques can improve model reliability in clinical applications.

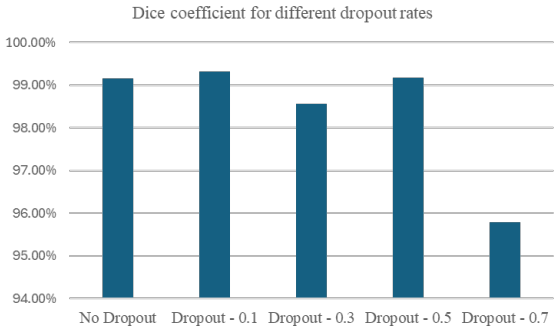
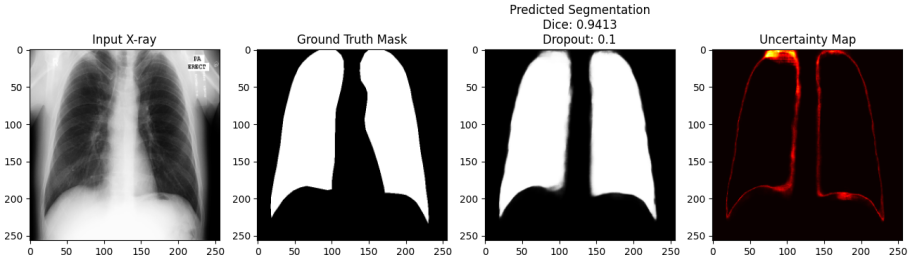
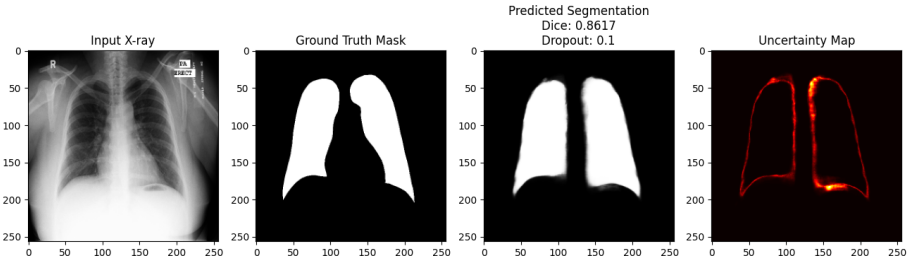


Figure 10. Analysis of the dropout effect on the model's performance

Figures 11 (a) and 11 (b) depict a segmentation analysis of a chest X-ray using the U-Net model trained with a Dropout of 0.1. In Figure 11(a), it can be observed that the model's predicted segmentation closely aligns with the ground truth mask, with only minor discrepancies along the lung boundaries. The lung fields are clearly delineated, and the overall contour accuracy is high, indicating a strong model performance in this case. In Figure 11 (b), it can be observed that the predicted segmentation deviates more noticeably from the ground truth, particularly around the lung contours. The uncertainty map highlights an elevated uncertainty along the boundaries of both lungs, especially in the lower and upper regions, suggesting that the model struggles to produce consistent predictions in areas where structural variations or a lower contrast are present.



(a)



(b)

Figure 11. Segmentation analysis, including the original chest X-ray, the ground truth mask, the predicted mask, and the uncertainty map

To gain insight into the regions where the model is less confident, the uncertainty maps generated through Monte Carlo Dropout were analysed, using 30 stochastic forward passes per image. These maps allowed to visualize and identify areas with a high prediction variability, typically located along lung boundaries or near regions featuring anatomical complexity or a low contrast.

Figures 11 (a) and (b) also highlight the regions of high uncertainty related to the segmentation. An increased uncertainty appears along the lung boundaries, particularly near the upper lung fields and costophrenic angles. This suggests the model is more confident in the central lung regions but less certain in areas with overlapping structures or a lower contrast.

To sum up, Scenario 1 revealed that CLAHE, Gaussian, and Laplacian filtering did not improve the segmentation performance. In fact, preprocessing resulted in comparable or slightly worse Dice scores (74.65% without filters vs. 72.86% with CLAHE and Gaussian filters). CLAHE filtering amplified noise in well-exposed images, while Gaussian filtering blurred fine lung structures, and Laplacian filtering overemphasized artifacts in diseased lungs. Since convolutional layers can inherently learn these transformations, external preprocessing was largely redundant. Instead, data augmentation or transfer learning may be more effective than fixed preprocessing pipelines.

In Scenario 2, training on one dataset and testing on two other datasets (the Montgomery and the Shenzhen datasets) resulted in slight performance variations (a Dice score of 89.70% vs. 89.32% without filters, and of 89.20% vs. 89.23% with all filters, respectively). The differences in contrast, exposure, and noise levels between datasets affected segmentation accuracy more than preprocessing did. Additionally, anatomical differences and varying lung opacity patterns in tuberculosis cases likely contributed to a weaker generalization ability. These findings suggest that dataset variability, rather than preprocessing techniques, plays a greater role in the segmentation performance.

In Scenario 3, training on all the three datasets yielded a Dice score of 96.13%, but fine-tuning in Scenario 3b improved it to 99.16%. This suggests that while initial training captured broad variations, fine-tuning enhanced lung boundary detection, particularly in lung disease cases. Dataset diversity, not preprocessing, proved to be the key driver of performance gains.

6. Discussion

A key finding of this study is that fine tuning with self-generated pseudo-labels significantly enhances a model's generalization ability across datasets. Unlike conventional approaches that improve model performance by increasing model complexity or adding parameters, the proposed method addresses labelling inconsistencies without modifying the model architecture. The 3% improvement in the Dice score (from 96.13% to 99.16%) demonstrates that harmonizing segmentation styles across datasets is as crucial - if not more so - as increasing model size. The obtained results suggest that ensuring consistency in training data can be a more effective strategy for improving segmentation accuracy than relying solely on larger, more complex networks.

With respect to model size, the one employed in this paper has 7.7 M parameters, while the ones in (Bombiński et al., 2024) and (Hryniewska-Guzik et al., 2024), have 9M (due to attention mechanisms) and 10 M parameters, respectively. At time same time, the DC obtained by employing this model is slightly better than the one presented in (Bombiński et al., 2024), as the standard U-Net achieved an average accuracy of 92% in metrics like DSC, while more complex architectures such as U-Net++ and hybrid models (e.g. with Attention Mechanisms) reached a DSC of over 94%.

The findings of this study demonstrate the potential of the U-Net architecture for lung segmentation in medical imaging, particularly in chest X-rays (CXR) featuring various pathologies such as tuberculosis, pneumonia (both viral and bacterial), and COVID-19. However, several important observations and limitations emerged from the evaluation of the model across different datasets and training-testing scenarios.

6.1 Generalization of the Model

One of the key objectives of this study was to assess the U-Net model's ability to generalize across datasets featuring diverse lung pathologies and different segmentation styles. The results across different training-testing scenarios suggest that the model can generalize reasonably well when trained and validated on multiple datasets. For instance, in Scenario 3b, where the model was trained

on a combined dataset (including the Darwin, Montgomery, and Shenzhen datasets) and fine-tuned, the model achieved a robust performance, with a Dice coefficient reaching 99.16%. This demonstrates that the U-Net architecture can handle a variety of lung conditions and achieve a high segmentation accuracy across different sources of data. The results are comparable to those obtained for larger and more complicated models.

Scenarios 1 and 2 highlighted the limitations of the model when trained on a dataset containing only normal and tuberculosis cases and then tested on datasets featuring viral and bacterial pneumonia. The model performance was significantly lower (a Dice coefficient under 90%) in comparison with other scenarios, illustrating the challenge of achieving generalization when training on a limited spectrum of lung diseases. This suggests that training on a more diverse dataset is crucial for creating a model capable of effectively handling real-world cases where multiple and overlapping pathologies are present.

6.2 Effect of Image Preprocessing on Segmentation Performance

Various image preprocessing techniques were applied to the employed datasets, such as CLAHE, Gaussian, and Laplace filters, to evaluate their impact on the segmentation performance. While the CLAHE filter significantly improved contrast in underexposed images, its benefits were marginal for images with an already good contrast. In some cases, excessive contrast enhancement introduced artifacts, making lung segmentation more challenging. And although Gaussian filtering effectively removed noise, excessive smoothing led to the loss of critical anatomical structures, such as small lesions or fine airway details. This was particularly problematic for images depicting lung diseases, where fine differences in lung opacities were crucial for segmentation. Further on, while Laplacian filtering helped enhance lung edges, it sometimes overemphasized certain structures, leading to false boundary detection. This was especially problematic in cases where lung opacity due to diseases (e.g. pneumonia) blurred the natural boundaries, causing the filter to highlight artifacts rather than meaningful structures.

Applying CLAHE, Gaussian filtering, and Laplacian filtering to each 256×256 image added an extra computational load. Although the resizing step reduced the dataset's complexity, the additional preprocessing operations slightly increased the overall data pipeline runtime.

Future work should focus on augmentation techniques such as rotation, flipping, or synthetic data generation to enhance model generalization ability rather than relying solely on preprocessing.

While the filters slightly improved the model's performance in some scenarios, the overall effect was marginal. Since convolutional layers inherently learn contrast and edge features, external preprocessing appears to be redundant when sufficient training data is available. These findings suggest that a more effective approach would be to focus on dataset diversity rather than on fixed preprocessing pipelines.

6.3 Performance Across Different Datasets

The model's performance varied across datasets, particularly with tuberculosis cases. In Table 2, it can be noticed that the Dice coefficient for tuberculosis cases was significantly lower (82.98% for the Montgomery dataset and 91.51% for the Shenzhen dataset) in comparison with other cases like normal lungs (98.48%) or pneumonia cases (97.04%). The lower model performance for tuberculosis cases is probably due to its underrepresentation among the training data. To improve model performance on underrepresented conditions, future work should consider applying data augmentation or oversampling techniques to balance the analysed dataset more effectively.

6.4 Potential for Clinical Application

Despite the challenges mentioned above, this study demonstrates that the U-Net can be a powerful tool for clinical applications in lung segmentation. The model's strong performance on diverse datasets, especially in Scenario 3b, shows promise for deploying such models in real-world settings where radiological images vary with regard to resolution, quality, pathology and manual labelling styles. Accurate segmentation is critical for assisting radiologists in diagnosing conditions such as COVID-19, pneumonia, and tuberculosis, especially when faced with high volumes of radiological images that need a rapid and consistent interpretation.

However, further improvements are needed before deploying the model in clinical practice.

Fine-tuning the model's architecture, addressing underrepresented diseases like tuberculosis, and validating the model on additional unseen datasets could enhance its robustness and reliability.

Several directions for future work arise from this study. Addressing the imbalance in the dataset, particularly for tuberculosis cases, could improve model performance across all disease types. Techniques such as synthetic data generation or oversampling of underrepresented classes may be necessary. Also, while this study focused on X-ray images, expanding the model to handle CT scans or incorporate multi-modal inputs (e.g. combining X-rays and CTs) could enhance its diagnostic capabilities.

As a future work direction, training an ensemble of models instead of a single network in Scenario 3b could offer multiple advantages. Ensemble-based pseudo-labelling may enhance the supervision level by averaging predictions across multiple models, thereby mitigating individual biases and improving the consistency of the fine-tuning process. Additionally, analysing prediction variability within the model ensemble would enable a more reliable pixel-wise uncertainty estimation, offering deeper insights into model confidence. Finally, ensemble learning could further improve a model's generalization ability, potentially leading to smoother and more robust segmentations, especially in anatomically complex or ambiguous regions.

Also, to improve consistency across diverse pathologies, a promising future direction is to train the model to focus on core anatomical structures of the lungs rather than on pathology-specific features, using augmentation techniques and adversarial training to promote robustness against disease-induced variability.

In conclusion, while the U-Net architecture shows strong potential for lung segmentation in various pathologies, further refinement is required to ensure a reliable performance across all disease types and in diverse clinical environments.

The key takeaways of the proposed approach are as follows: fine tuning with pseudo-labels improves model generalization (+3% DS) without increasing model complexity, dataset diversity is more impactful than preprocessing for achieving high segmentation accuracy and fixed preprocessing techniques (CLAHE, Gaussian, Laplacian) provided minimal improvements in some cases.

The findings of this study highlight the U-Net architecture's strong potential for lung segmentation across diverse chest X-ray (CXR) datasets, particularly in cases involving tuberculosis, pneumonia, and COVID-19. However, the challenges related to the model`s generalization ability and performance observed in various training-testing scenarios underscore the importance of comparing this approach with other strategies aimed at improving the segmentation quality through a better label generation.

Self-training using model-generated pseudo-labels could improve model performance on unlabelled data and enhance a model`s ability to generalize for unseen data, potentially raising the Dice coefficients above 90%. The obtained fine-tuning results on harmonized datasets (the Darwin, Montgomery, and Shenzhen datasets) suggest that domain adaptation and further fine-tuning can mitigate model performance discrepancies across diverse datasets.

Additionally, employing advanced data augmentation and synthetic data generation (e.g. GANs, diffusion models) (Sundaram & Hulkund, 2021) may address data variability. Although this was not the focus of this study, architectures such as Attention U-Net or transformer-based models (e.g. TransUNet) (Mahmud Auvy et al., 2024; Lin et al., 2022) can also be utilized for integrating clinical metadata or additional imaging modalities.

7. Conclusion

The obtained results challenge the common belief that improving segmentation models always requires larger models. Instead, this paper shows that fine-tuning on consistent pseudo-labels can significantly improve model performance results. The proposed approach is scalable, meaning it can be applied to new datasets without increasing the computational costs.

Future enhancements for a better model performance could include the implementation of data augmentation techniques (elastic deformations, adversarial augmentations) instead of static preprocessing, the use of transfer learning and domain adaptation to improve the model`s cross-dataset generalization ability and leveraging semi-supervised learning to reduce the dependence on large manually annotated datasets.

REFERENCES

Agrawal, T. & Choudhar, P. (2023) Segmentation and classification on chest radiography: a systematic survey. *The Visual Computer.* 39(3), 875-913. https://doi.org/10.1007/s00371-021-02352-7.

Anon. (n.d.) *COVID-19 xray dataset.* https://github.com/v7labs/covid-19-xray-dataset [Accessed 15th June 2024].

Azimi, H., Zhang, J., Xi, P. et al. (2022) Improving Classification Model Performance on Chest X-Rays through Lung Segmentation. [Preprint] https://arxiv.org/abs/2202.10971 [Accessed 23rd June 2024].

Bombiński, P., Szatkowski, P., Sobieski, B. et al. (2024) Underestimation of Lung Regions on Chest X-Ray Segmentation Masks Assessed by Comparison with Total Lung Volume Evaluated on Computed Tomography, [Preprint] https://arxiv.org/abs/2402.11510 [Accessed 2nd March 2025].

Danilov, V., Proutski, A., Kirpich, A. et al. (2022) Chest X-ray dataset for lung segmentation. https://data.mendeley.com/datasets/8gf9vpkhgy/1 [Accessed 23rd June 2024].

Gordaliza, P. M., Muñoz-Barrutia, A., Abella, M. et al. (2018) Unsupervised CT Lung Image Segmentation of a Mycobacterium Tuberculosis Infection Model. S*cientific Reports.* 8*,* art. no. 9802. https://doi.org/10.1038/s41598-018-28100-x.

Haralick, R. M. (1984) Digital Step Edges from Zero Crossing of Second Directional Derivatives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. PAMI-6(1), 58–68. https://doi.org/10.1109/TPAMI.1984.4767475.

Hryniewska-Guzik, W., Bilski, J., Chrostowski, B. et al. (2024) A Comparative Analysis of Deep Learning Models for Lung Segmentation on X-Ray Images. [Preprint] https://arxiv.org/abs/2404.06455 [Accessed 2nd March 2025].

Jaeger, S., Candemir, S., Antani, S. et al. (2014) Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative Imaging in Medicine and Surgery.* 4(6), 475-477. https://doi.org/10.3978/j.issn.2223-4292.2014.11.20.

Kim, M. & Lee, B-D. (2021) Automatic Lung Segmentation on Chest X-rays Using Self-Attention Deep Neural Network. *Sensors*. 21(2), 369. https://doi.org/10.3390/s21020369.

Lee, E. Y P., Ng, M.-Y. & Khong, P.-L. (2020) COVID-19 pneumonia: what has CT taught us? *The Lancet Infectious Diseases*. 20(4), 384-385. https://doi.org/10.1016/S1473-3099(20)30134-1.

Lei, Y., Tian, Y., Shan, H. et al. (2020) Shape and Margin-Aware Lung Nodule Classification in Low-dose CT Images via Soft Activation Mapping. *Medical Image Analysis*. 60, art. no. 101628. https://doi.org/10.1016/j.media.2019.101628.

Lin, A., Chen, B., Xu, J. et al. (2022) DS-TransUNet: Dual Swin Transformer U-Net for Medical Image Segmentation. *IEEE Transactions on Instrumentation and Measurement*. 71, art no. 4005615. https://doi.org/10.1109/TIM.2022.3178991.

Liu, W. Luo, J., Yang, Y. et al. (2022) Automatic lung segmentation in chest X-ray images using improved U-Net. *Scientific Reports*. 12, art. no. 8649. https://doi.org/10.1038/s41598-022-12743-y.

Mahmud Auvy, A. A., Zannah, R., Mahbub-E-Elahi et al. (2024), Semantic Segmentation with Attention Dense U-Net for Lung Extraction from X-ray Images. In: *2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT)*, *2-4 May 2024, Dhaka, Bangladesh*. Piscataway, New Jersey, USA, IEEE. pp. 658-663.

Marr, D. & Hildreth, E. (1980) Theory of Edge Detection. *Proceedings of the Royal Society of London*. *Series B, Biological Sciences*. 207(1167), 187–217.

Naqvi, A. J., Tauqeer, A, Bhatti, R. et al (2022) Improved Lung Segmentation Based on U-Net Architecture and Morphological Operations. [Preprint] https://arxiv.org/abs/2210.10545 [Accessed 14th February 2025].

Rahman, M. F., Tseng, T.-L. (B.), Pokojovy, M. et al. (2021) An automatic approach to lung region segmentation in chest x-ray images using adapted U-Net architecture. In: Bosmans, H., Zhao, W. & Yu, L. (eds.) *Proceedings of SPIE - Medical Imaging 2021: Physics of Medical Imaging (vol. 11595), 15-20 February 2021, Online Only, California, United States.* Bellingham, Washington, USA, SPIE. https://doi.org/10.1117/12.2581882.

Ronneberger, O., Fischer, P. & Brox, T. (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation. [Preprint] https://arxiv.org/abs/1505.04597 [Accessed 15th February 2025].

Selvan, R., & Dam, E. B., Detlefsen, N. S. et al. (2020) Lung Segmentation from Chest X-rays Using Variational Data Imputation. To be published in *IEEE Transactions on Medical Imaging.* [Preprint]. https://arxiv.org/abs/2005.10052 [Accessed 16th June 2024].

Shen, D., Wu, G. & Suk, H.-I. (2017) Deep Learning in Medical Image Analysis. *Annual Review of Biomedical Engineering*. 19, 221-248. https://doi.org/10.1146/annurev-bioeng-071516-044442.

Sundaram, S. & Hulkund, N. (2021) GAN-based Data Augmentation for Chest X-ray Classification. [Preprint] https://arxiv.org/abs/2107.02970. [Accessed 16th March 2025].

Zuiderveld, K. (1994) Contrast Limited Adaptive Histogram Equalization. In: Heckbert, P. S. (ed.) *Graphics Gems IV*. San Diego, CA, United States, Academic Press Professional, Inc., pp. 474-485.