Research Article

A Comparative Performance of U-net and Mask R-CNN for Lung Segmentation Across Public Chest X-ray Datasets

Andrew Mao

Sycamore High School, 7400 Cornell Rd, Montgomery, OH, 45242, USA

ABSTRACT

Lung segmentation is critical for detecting and monitoring respiratory conditions and abnormalities in the lungs, making it a useful diagnostic tool. This study compares the performance of two deep learning models, U-net convolutional neural network (U-net) and Mask Region-based Convolutional Neural Network (R-CNN), for segmenting lungs using publicly available datasets. The U-net model was trained and validated on the Montgomery dataset, while the Mask R-CNN model was evaluated after being pre-trained without fine-tuning. Both models were tested on both the Montgomery and Shenzhen datasets to assess generalizability. Mask R-CNN was found to have the best performance with a Dice Coefficient of 0.9302 and IoU of 0.8696 on the Shenzhen dataset. Although Mask R-CNN showed stronger performance on the unseen Shenzhen dataset, the comparison is limited since the two models are trained on different datasets. This study highlights strengths and limitations of each model and outlines future work to make the study more fair.

Keywords: Deep Learning; Convolutional Neural Networks; Lung Segmentation; Artificial Intelligence; Chest X-ray Images

INTRODUCTION

Lung segmentation is a fundamental step in medical image processing. Its main aim is to identify the outline of the lungs from 2D chest x-rays or chest radiographs (1). This process is critical for a broad range of applications, such as disease diagnosis (tuberculosis, COVID-19, pneumonia), automating clinical report generation, and tracking diseases over time. The segmentation of the lungs makes it easier to classify and detect discrepancies.

Historically, radiologists used manual segmentation, where radiologists would use software tools to draw outlines of the lungs in x-rays and CT scans (2). However, this process can be time-consuming and inaccurate, requiring advanced skills (2). Consequently, deep learning is now the leading solution for lung segmentation because of its ability to learn patterns from large datasets, removing the need of manual segmentation from doctors (2). The most common approaches to lung segmentation use convolutional neural networks with encoder-decoder architectures, region methods, and recently, dilated convolutions.

Corresponding author: Andrew Mao, E-mail: andrewmao268@gmail.com.

Received June 23, 2025**; Accepted** July 19, 2025
https://doi.org/10.70251/HYJR2348.347481

Among these methods, U-net convolutional neural network (U-net) is the primary architecture for lung segmentation due to its ability to extract rich features from images (3). U-net utilizes a symmetric encoder-decoder architecture with skip connections. Mask Region-based Convolutional Neural Network (R-CNN) is a region-based method that expands on Faster R-CNN for instance segmentation (4).

In this paper, the effectiveness of a U-net trained on a lung image dataset for lung segmentation is evaluated. This is compared to the performance of a Mask R-CNN model trained on a common image dataset to understand the importance of training on a specific dataset for lung segmentation performance.

LITERATURE REVIEW

U-net Based Lung Segmentation

U-net architectures and its variants have been used extensively for medical image segmentation over the span of many years. The U-net architecture, a convolutional neural network that was designed for biomedical image segmentation was first introduced by Ronneberger *et al*. in 2015 (5). Following that, Frid-Adar *et al*. in 2018 improved the U-net by utilizing an ImageNet pre-trained encoder, achieving Jaccard overlap scores of 96.1% for lungs on the JSRT dataset (1). This improved U-net model displayed better performance over the original U-net models. Keetha *et al*. in 2020 introduced the U-Det model which is a modified U-net with a bidirectional feature network achieving a Dice coefficient of 82.82% on the LUNA dataset, demonstrating success in lung nodule segmentation (2). Wu *et al*. in 2023 suggested changes to the U-net architecture causing the IoU score to increase from 87% to 92%, which indicates significant improvements in segmentation accuracy (3). Khaniki and Manthouri in 2024 made another modified U-net using a Convolutional Block Attention Module (CBAM), combining spatial, channel, and pixel attention mechanisms (6). As a result, this model led to improved segmentation performance.

Therefore, while U-net and its different variants have high accuracy in lung segmentation, it handles more complicated cases, such as severe lung deformations, inaccurately. In addition, many of these studies focus on specific datasets, which limit its generalizability of the models in different clinical scenarios.

Mask R-CNN Lung Segmentation

Concurrently, to the development of U-net, He *et al*. in 2017 first showed Mask R-CNN, extending Faster R-CNN by adding a parallel mask prediction branch which allows for simultaneous object detection and instance segmentation with high accuracy (4). Kopelowitz and Engelhard in 2019 used Mask R-CNN for 3D lung segmentation, demonstrating Mask R-CNN's potential to handle three-dimensional medical imaging (7).

Podder *et al*. in 2021 applied Mask R-CNN for COVID-19 identification using chest x-ray images, displaying its ability to segment infected regions of the lungs (8). Cai *et al*. in 2020 used Mask R-CNN for lung nodule segmentation, which showed its effectiveness in delineating nodules in CT scans (9). This is extremely important for early cancer detection to properly understand the size and growth of lung nodules. In 2024, Doğan *et al*. made an enhanced Mask R-CNN architecture consisting of attention mechanisms and hybrid skip connections with the aim of improving segmentation accuracy for pulmonary embolism detection (10). Mask R-CNN models often require considerable computational resources and extensive training data. Additionally, its performance is sensitive to differences in image quality, making its generalizability weaker in different clinical settings.

Comparison Between the U-net and Mask R-CNN Models

Both U-net and Mask R-CNN architectures have been utilized comprehensively in lung segmentation functions. Each model has its own unique advantages. The U-net architecture, with pre-trained encoders and attention mechanisms, has high efficiency and accuracy, making it usable in many different clinical applications. The Mask R-CNN architecture is excellent in instance segmentation, providing detailed delineation of lung structures with a higher computational demand.

However, a significant gap in the current research is the lack of comparative studies evaluating the performance of U-net and Mask R-CNN across various datasets and clinical situations. Such comparisons are important to determine the most generalizable and effective models for lung segmentation tasks. Future research should focus on using these architectures under standardized conditions to determine their optimal use in clinical practice.

METHODS AND MATERIALS

Dataset Description

This study used two publicly available chest x-ray datasets for lung segmentation: the Shenzhen Hospital

X-ray set and the Montgomery County X-ray set (11).

The Montgomery dataset, collected by the U.S. National Library of Medicine, contains 138 posterior-anterior chest x-ray (CXR) images. These images were obtained from a tuberculosis screening program and include pixel-level annotations of the left and right lungs.

Image resolutions are either 4020x4892 or 4892x4020 pixels. The Montgomery dataset contains separate left and right mask images, which were merged together for training.

The Shenzhen dataset was released by the National Library of Medicine in conjunction with the Shenzhen No.3 People's Hospital. The dataset includes 662 annotated posteroanterior chest x-ray images, with an approximately balanced distribution of normal and tuberculosis-infected cases. 566 of the 662 PA CXR images have a corresponding manually segmented lung mask.

In this study, the Montgomery dataset was used exclusively for training and validation. The data was split into train, validation, and testing sets using scikit-learn's train_test_split function. The Shenzhen dataset was reserved entirely for external testing to measure cross-dataset generalization. All images and masks were resized to 256x256 pixels, normalized in a grayscale intensity in the [0,1] range.

The Montgomery dataset was selected to train the U-Net since it is one of the most common datasets encountered in lung segmentation literature and since it contains a limited number of images which makes it suitable for training under limited computational resources. The Shenzhen dataset was used for testing to evaluate the generalizability of the models. The two datasets have different image resolution, disease types, and patient demographics which may influence model performance. This is discussed further in the Discussion section.

Model Architectures

Two deep learning models—U-net, Mask R-CNN—were implemented using python scripts and open-source libraries, like Keras and Tensorflow. U-net was implemented in Google Colab, while Mask R-CNN was implemented locally in Visual Studio Code.

U-net was originally introduced in Ronneberger *et al*. in 2015 as a fully convolutional encoder-decoder architecture used specifically for biomedical image segmentation (5). As implemented in this study, U-net follows a symmetric encoder-decoder architecture. The encoder consists of four convolutional blocks, each containing two 3x3 convolutional layers followed by Batch Normalization and ReLU activation. Each block finishes with a 2x2 MaxPooling process that down sampled the spatial resolution. At the network's narrowing, a bridge block applies two additional convolutional layers to extract deep features. The decoder mirrors the encoder by using UpSampling2D layers for upscaling, combining skip connections from the corresponding encoder layers, and applying two more convolutional layers per level. A final 1x1 Conv2D layer with sigmoid activation outputs the lung mask. The entire architecture includes 23 Conv2D layers and maintains spatial alignment in consistent padding.

Mask R-CNN is based on the architecture introduced by He *et al*. in 2017 (4). The architecture extends Faster R-CNN with a parallel mask prediction branch. It uses a ResNet50 backbone with a Feature Pyramid Network to extract multi-scale features. The Region Proposal Network generates object candidates, and for each region of interest, the network predicts bounding boxes, class labels, and segmentation masks. Due to the original datasets not containing bounding box annotations, bounding boxes were inferred directly from lung mask contours during preprocessing. Bounding box inference was implicitly handled by the Mask R-CNN regional proposal mechanism based on provided mask annotations.

Training Procedure

The U-net model was trained independently using only the Montgomery dataset. The data split was 70% for training, 10% for validation, and 20% for testing. The Shenzhen dataset was used in testing only. The Mask R-CNN model used a pre-trained COCO backbone, and further training was not done. Training was administered for up to 50 epochs with early stopping based on validation accuracy for the U-net model. The U-net model was optimized using the Adam optimizer with an initial learning rate of 0.0001 and a batch size of 4. U-net was trained using Binary Cross-Entropy loss, while Mask R-CNN was trained with its standard multi-task loss. A NVIDIA Tesla T4 GPU was used in Google Colab for all training and evaluation for the U-net model. For Mask R-CNN, a CPU was used for the loading of the pre-trained model and the evaluation on the Montgomery and Shenzhen datasets. Mask R-CNN was evaluated in Visual Studio Code and uses a modified Tensorflow 2.x version.

Evaluation

The models were evaluated on both the Montgomery and Shenzhen datasets using the following metrics: accuracy, mAP, IoU, and Dice coefficient. These metrics were used to comprehensively measure each model under different conditions. Accuracy is the measure of the

proportion of correctly classified pixels in the foreground and background over the total number of pixels. The formula for accuracy is given in equation 1 (12).

$$\text{Accurary} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad [1]$$

where TP is True Positive, TN is True Negative, FP is False Positive, and FN is False Negative.

The Intersection Over Union (IoU), also known as the Jaccard Index, measures the overlap between the predicted mask and the ground truth mask. The IoU ignores the true negatives and focuses on the overlap of the two regions. The formula for IoU is given in equation 2 (12).

$$\text{IoU} = \frac{TP}{(TP + FP + FN)} \quad [2]$$

The Dice coefficient is a similarity measure between the predicted and ground truth masks. The formula for the Dice coefficient is given in equation 3 (12).

$$\text{Dice coefficient} = \frac{(2 * TP)}{((2 * TP) + FP + FN)} \quad [3]$$

The Mean Average Precision (mAP) measures detection quality by averaging the precision over many IoU thresholds. In segmentation, it evaluates how well predicted masks match the ground truth for different strictness levels. The formula for mAP is given in equation 4 (13).

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^{N} \text{AP}_{t_i} \quad [4]$$

where N=number of IoU thresholds (usually 10), AP_{t_i}=average precision at IoU threshold t_i, and common thresholds: $t_i \in \{0.50, 0.55, 0.60, ..., 0.95\}$.

It is important to note that although mAP is commonly used in the literature, it is a less appropriate metric than the other metrics listed above for lung segmentation from U-net because it measures the accuracy of the bounding boxes, meaning it is suited for object segmentation with many objects.

RESULTS

In this section, the results of the experiments are presented. Table 1 provides a comparison of the performance of U-net and Mask R-CNN on the Montgomery dataset. U-net performs better than Mask.

R-CNN on this dataset over all the tested metrics. Table 2 provides the performance of both models on the Shenzhen dataset. Interestingly, on the Shenzhen dataset, Mask R-CNN outperforms U-net for all metrics aside from mAP. Therefore, it can be concluded that U-net performs best on the dataset it is trained on, however, Mask R-CNN generalizes better.

Figure 1 presents the predicted masks versus the ground truth masks for (a) the U-net and (b) Mask R-CNN. It is evident from the images that the generated masks closely match the ground truth masks for both architectures and both datasets. Likewise, this is reinforced by Figure 2, which shows the overlay of the predictions and ground truth masks. However, in Figure 2, it is evident that while the bulk of the mask is correctly predicted, the Mask R-CNN model misses the pixels along the outline of the lung. Additionally, in Figure 1 and 2 for the Shenzhen dataset in U-net, the example image was selected to reflect the median Dice coefficient similarity (≈ 0.91) across the Shenzhen dataset. Some boundary noise is typical of the model's average performance.

DISCUSSION

Based on the results, the U-net model performed better on the Montgomery dataset than Mask R-CNN. This is likely because the U-net model was trained on the Montgomery dataset. On the other hand, the Mask R-CNN model performed better on the unseen dataset, the Shenzhen dataset. As seen in Figure 1 and 2, Mask R-CNN produces more consistent masks than U-net. Therefore, it can be concluded that the U-net model has limited ability to generalize to unseen datasets, yet the Mask R-CNN model generalizes well.

Table 1. Segmentation performance on the Montgomery dataset

Model	Accuracy	mAP	IoU	Dice Coefficient
U-net	**0.9949**	**0.9947**	**0.9796**	**0.9896**
Mask R-CNN	0.9843	0.8100	0.8871	0.9401

Table 2. Segmentation performance on the Shenzhen dataset

Model	Accuracy	mAP	IoU	Dice Coefficient
U-net	0.9545	0.9120	0.8371	0.9073
Mask R-CNN	**0.9699**	0.7950	**0.8696**	**0.9302**

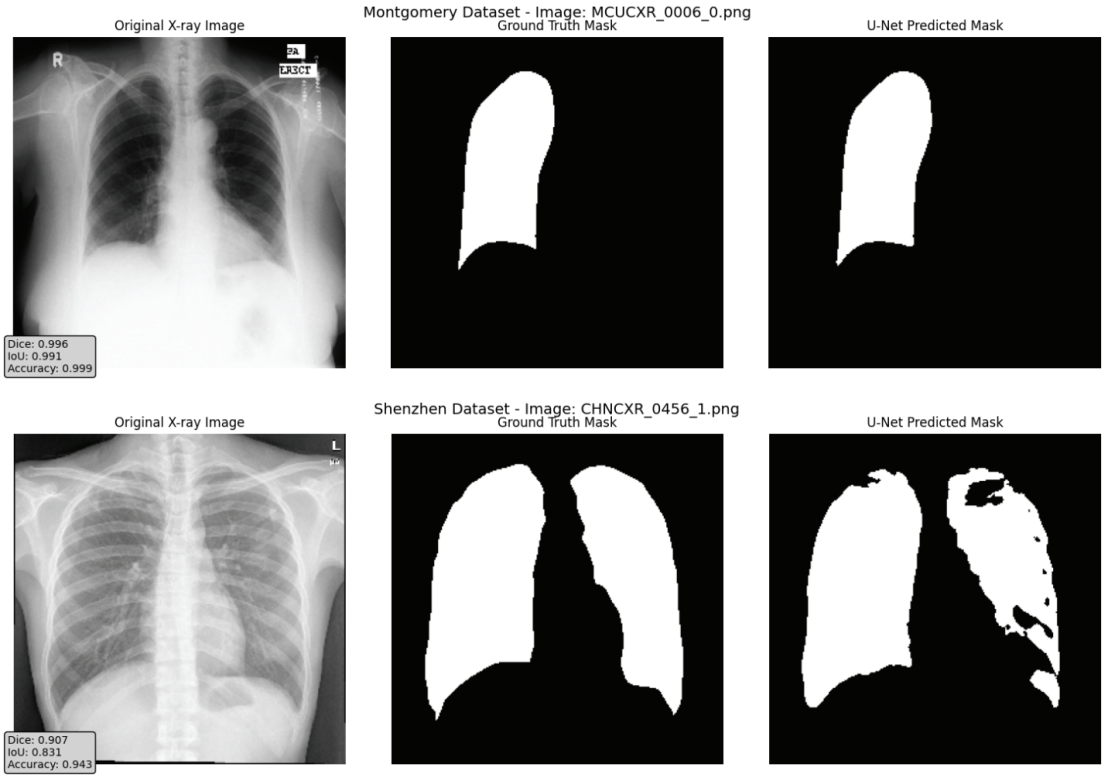

Figure 1a. U-net model's predicted masks versus the ground truth masks for the Montgomery and Shenzhen datasets.

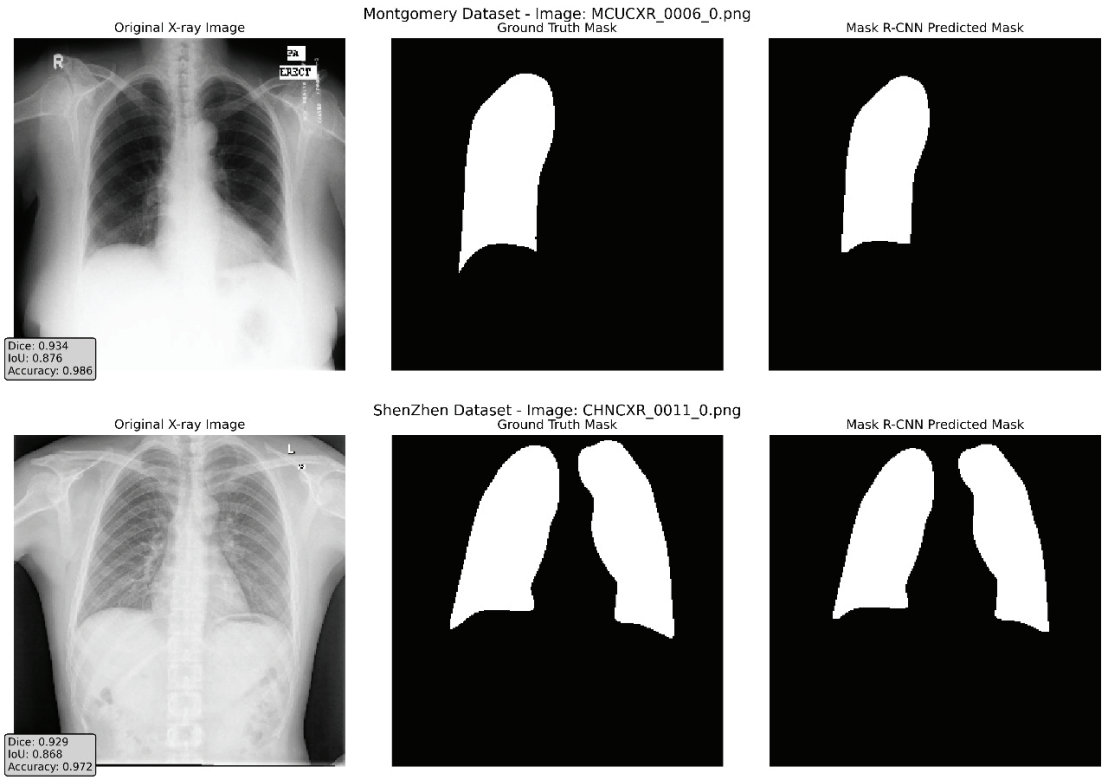

Figure 1b. Mask R-CNN model's predicted masks versus the ground truth masks for the Montgomery and Shenzhen datasets.

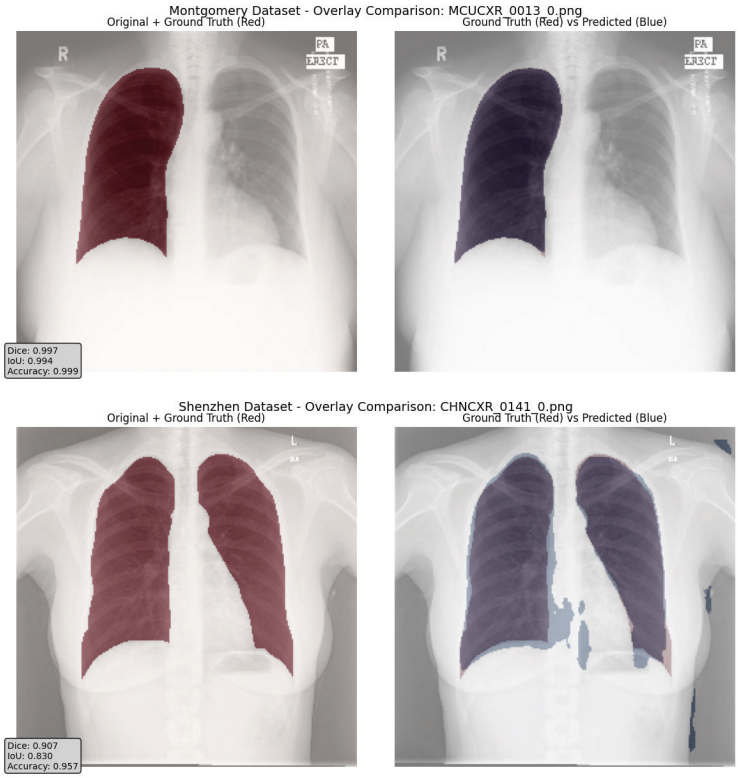

Figure 2a. U-net model's overlay of the ground truth mask and predicted mask for the Montgomery and Shenzhen datasets.

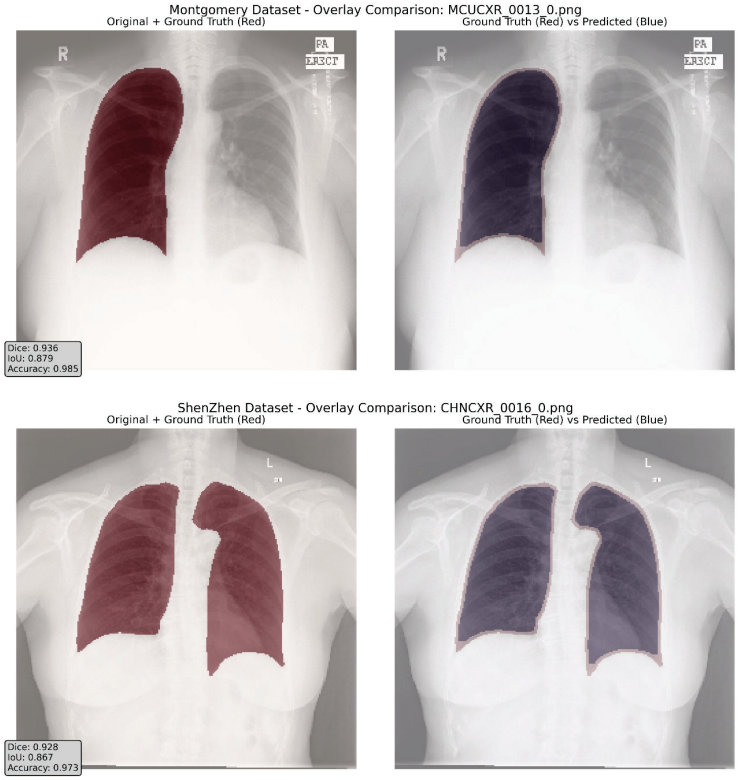

Figure 2b. Mask R-CNN model's overlay of the ground truth mask and predicted mask for the Montgomery and Shenzhen datasets.

Another factor that might impact the performance of the models is the differences in the image characteristics between the Montgomery and Shenzhen datasets. The Montgomery dataset contains higher resolution images than the Shenzhen dataset, and Shenzhen has a more diverse, larger population group from different geographic and healthcare contexts. This additional diversity is likely to have contributed to the reduced performance on U-Net for the Shenzhen dataset.

Although the mAP of the Mask R-CNN is lower than that of the U-net for the Shenzhen dataset, the mAP metric for U-net is artificial as discussed in the methods and materials section, making the performance of the Mask R-CNN look worse than it is.

The main limitation of the work is that the Mask R-CNN model was not trained on the Montgomery dataset and instead pre-trained using a COCO backbone. This creates a bias in the system due to different training datasets being used for the two models because U-net was optimized for the specific task, while Mask R-CNN was not. This means that proper comparison of the effectiveness of the models for lung segmentation cannot be made and rather shows that Mask R-CNN has a superior ability to generalize. Therefore, for future work, the Mask R-CNN model should be trained using the Montgomery dataset for better comparison with the U-net model. Additionally, to further strengthen the findings, more datasets can be tested.

CONCLUSION

This paper presented a deep learning-based system for lung segmentation from chest x-rays. The performance of two prominent models from the literature were compared: U-net and Mask R-CNN. The U-net model was trained on the Montgomery dataset, and the Mask R-CNN model was a pre-trained model with a COCO backbone. U-net and Mask R-CNN were tested on both the Montgomery and Shenzhen datasets. The Mask R-CNN model performed better than the U-net model for the majority of tested metrics over the Shenzhen dataset. However, it is important to emphasize that the two models were trained on different datasets, and future work should ensure that Mask R-CNN is trained on the Montgomery dataset for fair comparison. Nevertheless, this work highlights the promising performance of Mask R-CNN in medical image segmentation even when not specifically trained on medical images. Thus, this work paves the way for automated lung segmentation enabling easier diagnosis and disease tracking.

ACKNOWLEDGEMENTS

I would like to thank Dr. Kayla-Jade Butkow for their mentorship and support while working on this paper.

DECLARATION OF CONFLICT OF INTERESTS

The author declares that there are no conflicts of interest regarding the publication of this article.

REFERENCES

1. Frid-Adar M, Ben-Cohen A, Amer R and Greenspan H. Improving the Segmentation of Anatomical Structures in Chest Radiographs using U-Net with an ImageNet Pre-trained Encoder, *Lect Notes Comput Sci.* 2018; 11040: 159–168. *https://doi.org/10.48550/arXiv.1810.02113*
2. Keetha NV, Parisapogu S Annavarapu C. U-Det: A Modified U-Net Architecture with Bidirectional Feature Network for Lung Nodule Segmentation, *ArXiv* (Cornell University). 2020. *https://doi.org/10.48550/arxiv.2003.09293*
3. Wu Y. Lung Segmentation Using Modified U-Net, *ResearchGate*. 2023. *https://www.researchgate.net/publication/370553741_Lung_Segmentation_Using_Modified_U-Net*
4. He K, Gkioxari G, Dollar P Girshick R. Mask R-CNN, *IEEE Int Conf Comput Vis.* 2017; 2017: 2961–2969. *https://arxiv.org/abs/1703.06870*
5. Ronneberger O, Fischer P and Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation, *Int Conf Med Image Comput Comput Assist Interv.* 2015; 9351: 234–241. *https://arxiv.org/abs/1505.04597*
6. Khaniki M and Manthouri M. A Novel Approach to Chest X-ray Lung Segmentation Using U-net and Modified Convolutional Block Attention Module, *ArXiv* (Cornell University). 2024. *https://arxiv.org/abs/2404.14322*
7. Kopelowitz E and Engelhard G. Lung Nodules Detection and Segmentation Using 3D Mask-RCNN, *ArXiv* (Cornell University). 2019. *https://doi.org/10.48550/arxiv.1907.07676*
8. Podder S, Bhattacharjee S and Roy A. An Efficient Method of Detection of COVID-19 Using Mask R-CNN on Chest X-ray Images, *AIMS Biophys.* 2021; 8 (3): 281–290. *https://doi.org/10.3934/biophy.2021022*
9. Cai L, Long T, Dai Y and Huang Y. Mask R-CNN-Based Detection and Segmentation for Pulmonary Nodule 3D Visualization Diagnosis, *IEEE Access.* 2020; 8: 44400–44409. *https://doi.org/10.1109/access.2020.2976432*
10. Doğan K, Selçuk T and Alkan A. An Enhanced Mask R-CNN Approach for Pulmonary Embolism Detection and Segmentation, *Diagnostics.* 2024; 14 (11): 1102. *https://doi.org/10.3390/diagnostics14111102*
11. Jaeger S, Candemir S, Antani S, Wang YXJ, *et al*. Two

Public Chest X-ray Datasets for Computer-Aided Screening of Pulmonary Diseases, *Quant Imaging Med Surg.* 2014; 4 (6): 475–477. *https://doi.org/10.3978/j.issn.2223-4292.2014.11.20*

12. Mustapha B, Zhou Y, Nawel B, Chunyan S and Zhitao X. Optimized attention U-Net for enhanced lung and area of infection segmentation in chest X-Rays and CT scans, *Journal of Radiation Research and Applied Sciences.* 2025; 18 (3): 101650. *https://doi.org/10.1016/j.jrras.2025.101650*
13. Pathan RK, Lim WL, Lau SL, Ho CC, *et al*. Experimental Analysis of U-Net and Mask R-CNN for Segmentation of Synthetic Liquid Spray, in *Proc IEEE International Conference on Computing (ICOCO)*, Kota KInabalu, Malaysia. 14–16 Jan 2022: pp 237–242. *https://doi.org/10.1109/ICOCO56118.2022.10031951*