2023 International Conference On Cyber Management And Engineering (CyMaEn) | 978-1-6654-9329-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/CyMaEn57228.2023.10050951

CXLSeg Dataset: Chest X-ray with Lung Segmentation

Wimukthi Nimalsiri
Department of Computer Science and Engineering
University of Moratuwa
Moratuwa, Sri Lanka
wimukthi.18@cse.mrt.ac.lk

Mahela Hennayake
Department of Computer Science and Engineering
University of Moratuwa
Moratuwa, Sri Lanka
mahela.18@cse.mrt.ac.lk

Kasun Rathnayake
Department of Computer Science and Engineering
University of Moratuwa
Moratuwa, Sri Lanka
kasunr.18@cse.mrt.ac.lk

Thanuja D. Ambegoda
Department of Computer Science and Engineering
University of Moratuwa
Moratuwa, Sri Lanka
thanujaa@uom.lk

Dulani Meedeniya
Department of Computer Science and Engineering
University of Moratuwa
Moratuwa, Sri Lanka
dulanim@cse.mrt.ac.lk

***Abstract*—With the advancement of robust deep learning techniques, a significant number of applications pertaining to biomedical research and clinical practice can be noticed within the computer vision domain. Radiologists use Chest X-Ray (CXR) images, prominent among medical imaging, to diagnose and treat diseases. Proper anatomical segmentation of CXR images is increasingly valuable as a critical image pre-processing step as well as in interpreting the deep learning models since it isolates the necessary area by extracting the region of interest. This paper proposes a CXR-segmented dataset based on the MIMIC-CXR dataset. A total of 243,324 frontal views of CXR images with segmented masks for each image are available in the dataset. A U-Net model with spatial attention (SA-UNet) architecture is utilized for segmenting the CXR images after a comparative analysis of different U-Net variants. The SA-UNet model achieved a 96.80% in dice similarity coefficient and 91.97% in IoU for lung segmentation. This study has shown that the visual feature extraction process is optimized using segmented CXR radiographs of CXLSeg instead of original MIMIC-CXR images by a significant margin.**

***Index Terms*—medical image segmentation, MIMIC-CXR, U-NET, SA-UNet, Shenzhen, Montgomery**

I. Introduction

Radiographs have become a major diagnostic support mechanism for many diseases and play an important role in predicting outcomes, as X-rays are easy to use and inexpensive. Chest X-ray (CXR) uses slight radiation to produce black-and-white images. The color of the different parts in the CXR will depend on the density of that body part. For instance, the air in the lungs has low density, indicated by the color black [1]. These images consist of the heart, lungs, and sternum, which are used by medical professionals to examine the chest area and treat the conditions like heart attack, lung diseases, and broken bones. Mainly there are three types of projections to collect the CXRs by passing an X-Ray beam over the chest [2]. The traversal of the X-Ray beam from back to front and front to back is known as Posterior-Anterior (PA) and Anterior-Posterior (AP), respectively. Thus, the frontal view is given by PA and AP. The side view of the chest is known as the lateral projection. At present, these images are used by Computer Scientists to build models that can analyze and predict the abnormalities in the images using modern machine-learning techniques [3].

With the improvements in deep learning techniques in the medical field, the requirement for CXR datasets has increased [4], [5]. Initially, there were smaller datasets, and currently, large datasets have been released to the public, which includes CXR images associated with labels or reports. As a solution for the limited CXR datasets, the MIMIC-CXR dataset was presented with 377,110 CXRs, and the associated reports [6]. These x-rays consist of a frontal and lateral view of each patient's chest. Since the models mainly focus on the lungs section, the image's background does not give any valuable information to the model. Because of that, we propose the CXLSeg dataset: Chest X-ray with Lung Segmentation, which uses semantic segmentation to remove unwanted regions of interest from the frontal views of the MIMIC-CXR dataset.

In this paper, a dataset named CXLSeg is introduced, which is free of irrelevant background data around the CXR radiographs of MIMIC-CXR and provides a clear segmented image of the lung section. The dataset contains 243,324 frontal-view segmented images of the original CXR images taken from the MIMIC-CXR dataset. This study has proven that different Convolutional Neural Networks (CNNs) yielded better results with CXLSeg than the original MIMIC-CXR-JPG dataset, concluding the fact that the ability of visual feature extraction increases with segmented CXR images since those only contain lung lobes. This could aid researchers in

achieving better results during the computer vision pipeline. Additionally, our dataset includes corresponding masks for each segmented image, which can be utilized to train segmentation models to achieve better performance in deep learning applications. To the best of our knowledge, CXLSeg is the only segmented lung segmentation dataset with over 200,000 segmented images and masks and can be considered as a novel contribution. We highly believe this dataset will enhance the performance of the state-of-the-art computer vision-related models to a greater extent since this study has proven that for some of the models in Section VI.

II. Related Work

Segmented CXR images help the model to focus only on the lung lobes by removing the background areas. This enables the model to achieve better results. However, a limited number of datasets are available with segmented Chest X-Ray images [7]–[10]. Among them, segmented datasets with medical reports are rare to find. A set of sample CXR images and the corresponding segmented masks from different datasets are shown in Fig. 1. Additionally, a comparison of CXR image datasets with segmentation is included in Table I.

One of the popular CXR datasets is the Shenzhen Hospital dataset from China [7]. They captured X-Rays in 2012 using a Philips DR Digital Diagnost system. These images are in PNG format and with 3000×3000 pixels resolution as shown in Fig. 1(a), and patients' abnormalities are mentioned in a separate text file.

Another similar kind of dataset was introduced with the help of the Department of Health and Human Services in Maryland, USA. However, it consists of only 138 CXR frontal images [7]. All the images are either 4020x4892 or 4892x4020 pixels. The images in this dataset are captured using Eureka stationary X-ray machine. A text file is associated with this dataset, including the patient's details, such as gender, age, and lung condition. The specialty in this dataset is that they have made available the manual lung segmentations used for their lung segmentation algorithm. These images have larger lungs compared to the Shenzhen dataset, as shown in Fig. 1(b).

Another popular dataset for segmentation tasks is the Japanese Society of Radiological Technology Database (JSRT) [9], which was created in 1998. The resolution of the CXR images in this dataset is 2048x2048. JSRT is also facilitated with a text file mentioning the patient's abnormality. Sample CXR image from JSRT dataset is shown in Fig. 1(c).

V7-labs COVID-19 X-ray dataset is another widely used dataset for chest X-ray segmentation tasks [10]. Masks of the corresponding CXR images are also provided with this dataset. Initially, this dataset consists of 1602 normal, 4250 Pneumonia and 439 Covid-19 images. This dataset also contains frontal and lateral x-rays, but masks were not provided for lateral x-rays. Also, the resolution of the images varies in the dataset, where the largest image has a resolution of 5600x4700, and the smallest image has a resolution of 156x156. A sample image from the v7-labs dataset is shown in Fig. 1(d).

Using the MIMIC-CXR dataset [6], Li-Ching et al. [8] have published CXR segmentation images from the MIMIC-CXR dataset. They have used TernausNet, which is a U-Net architecture-based segmentation model, to segment images in the MIMIC-CXR dataset to create this new dataset. They have resized images to 512x512 before feeding them into the network. For this study, Li-Ching et al. have used randomly selected 4,091 CXRs from the MIMIC-CXR dataset containing 377,110 CXR images. After eliminating false recognitions by human examination, they produced only 1,141 segmented lung images.

TABLE I: Comparison of CXR segmentation datasets

Dataset	Format	Normal	Abnormal	Total
Shenzhen [7]	PNG	326	336	662
Montgomery [7]	PNG	80	58	138
JSRT [9]	PNG	154	93	247
V7-lab [10]	JPEG	1602	4689	6291
MIMIC-seg [8]	JPEG	-	-	1,141

III. MIMIC-CXR Dataset (Base Dataset)

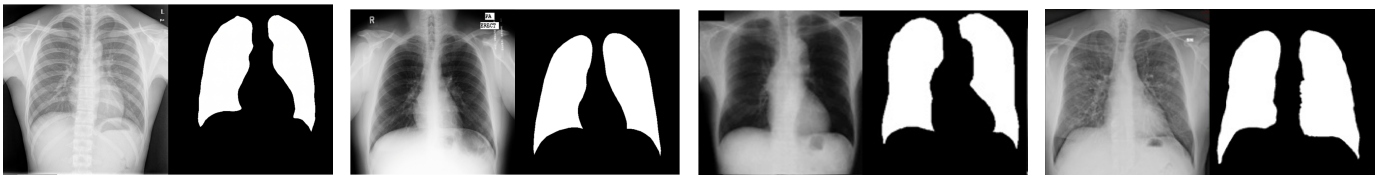
The MIMIC Chest X-ray (MIMIC-CXR) is a public dataset. It consists of 377,110 chest x-ray images and 227,827 studies associated with those CXRs [6]. It is available to the public in two versions.

- MIMIC-CXR DICOM version: Digital Imaging and Communications in Medicine, also known as DICOM, is used to transmit medical images. It can store multiple images in a single file and metadata in the header related to the image. The important fact about this data format is it ensures that the image is not separated from its data. In the MIMIC-CXR dataset, the authors have included some metadata for the images as well. The whole dataset is around 4.6 TB, which is larger than many public datasets.
- MIMIC-CXR JPG version: This is a compressed dataset that used the images from DICOM files in the original MIMIC-CXR dataset. The purpose of this dataset is to provide a lightweight version of the MIMIC-CXR for researchers since the DICOM version is too large. The jpg version of MIMIC-CXR is around 557.6 GB, much larger than the DICOM version.

In the MIMIC-CXR dataset, each study consists of a report with an examination, indication, technique, findings, and impression for each CXR, and metadata for each CXR is included in a separate compressed CSV file. The mapping from the image to the report and the metadata is done using study id, subject id, and dicom id [6].

- Study ID - Unique identifier for each report or study
- Subject ID - Unique identifier for a patient
- DICOM ID - Identifier for the DICOM file. Also, JPG file name is equal to the DICOM_id

The image file path can be formed using these identifiers as follows. Image file path - first two characters of subject_id followed by "p"/study_id/dicom_id.jpg.



(a) (b) (c) (d)

Fig. 1: Sample images from datasets (a) Shenzhen Hospital, (b) Montgomery County, (c) JSRT dataset and (d) V7-labs

Since the MIMIC-CXR is referred to around 350 studies, we were inspired to come up with this segmented dataset which is based on MIMIC-CXR.

IV. CXLSeg (Proposed Dataset)

Chest X-Ray Segmented (CXLSeg) consists of 243,324 frontal CXR images and the corresponding segmented masks for each CXR image. The original MIMIC-CXR dataset contains a total of 243,334 frontal images of both AP and PA views. Since there were ten studies with missing reports, the frontal images related to those missing studies were removed when creating the dataset. The segmented dataset has the same directory structure as the original MIMIC-CXR. Both the segmented image and the respective mask, as depicted in Fig. 2, are saved at the same location in JPEG format. The segmented images have the exact naming convention as the original images, and masks additionally contain a suffix as '-mask' at the end. The dataset comes with four compressed CSV files for users' convenience, as follows.

- CXLSeg-metadata.csv.gz - a compressed CSV file that provides metadata for the images, including Dicom ID, Subject ID, Study ID, medical report, patient orientation, view position, and an unknown date of the image accession time.
- CXLSeg-segmented.csv.gz - a compressed CSV file that provides Dicom ID, Subject ID, Study ID, image relative path, and labels generated by the CheXpert labeler.
- CXLSeg-split.csv.gz - a compressed CSV file that provides a standard train/validation/test data split of the ratio 80:10:10.
- CXLSeg-mask.csv.gz - a compressed CSV file that provides Dicom ID, Subject ID, Study ID, and mask file path. This can be utilized along with the original MIMIC-CXR-JPG dataset to train a segmentation model.

The dataset also includes additional medical reports derived from the original reports of the MIMIC-CXR dataset. Reports were extracted from the original reports after some preprocessing. Both the impression and findings sections were combined to make the report more verbose, and a comparison section was used if no other sections were available in the original report. The metadata CSV file contains the extracted reports with other necessary CXR image details, and it can be utilized along with the segmented CSV file for medical report generation purposes. To expand the mobility of the dataset, all the images and masks were resized to 224x224 and saved in JPG format, which drastically reduced the total size of the dataset to around 4.24 GB.

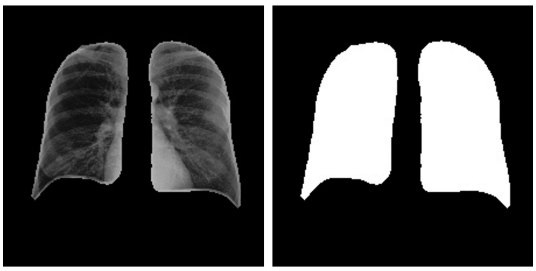


Fig. 2: Sample CXR Image and its mask from CXLSeg Dataset

Moreover, to enhance the usability of this dataset, we considered several aspects such as reduced dataset size, the same folder structure as the original dataset, and the same naming convention for the segmented images are prominent. The additional CSV files are designed in a way that helps the users to utilize them along with the original MIMIC-CXR dataset, which improves flexibility. All these changes help researchers to use this dataset more efficiently without deviating much from the original dataset yet achieving better results in terms of visual feature extraction.

V. Methodology

The MIMIC-CXR dataset is a large multi-class CXR image dataset. The segmentation model used to do the semantic segmentation needs to be trained with a comparatively larger dataset. However, most of the CXR segmentation datasets available to the public do not contain adequate images to train a segmentation model that can ultimately yield better results. Thus a combination of three popular datasets, namely the Shenzhen dataset, the Montgomery dataset, and the V7-labs COVID-19 dataset, has been used to assemble a comparatively larger dataset. Furthermore, data augmentation is used to reduce the class imbalance of the combined dataset. Traditional augmentation techniques like flipping (both horizontally and vertically), rotating, cropping, zooming and scaling, and adding gaussian noise by distortion of high-frequency features were used to generate new images and subsequently improve the training while reducing overfitting. Table II denotes the resulting dataset after minimizing the class imbalance, whereas Ori. denotes the original image count, and Aug. denotes the augmented image count. After data augmentation, image counts of 3940, 3619, 3987, and 3850 were available for

Tuberculosis, Covid-19, Pneumonia, and Normal, respectively, which resulted in 15396 total images.

TABLE II: Data augmentation for CXR segmentation

Dataset	**Tuberculosis**		**Covid-19**		**Pneumonia**		**Normal**	
	Ori.	***Aug.***	***Ori.***	***Aug.***	***Ori.***	***Aug.***	***Original***	***Aug.***
Shenzhen [7]	336	3360	-	-	-	-	326	326
Montgomery [7]	58	580	-	-	-	-	80	320
V7-labs [10]	-	-	517	3619	3987	3987	1602	3204

A U-Net model with spatial attention (SA-UNet) [11] is used for the image segmentation, and data preprocessing has been done to the original MIMIC-CXR dataset before the segmentation. The original image was resized to 224x224 before feeding into the model. CLAHE was used on the image since it improves visibility by amplifying the image contrast. It is an adaptive histogram equalization that operates on the image's small regions, hence increasing the accuracy of the generated masks. The images were normalized after dividing them by 255, which adjusted the values of the image into a standard scale of [0,1]. Morphological image dilation and erosion techniques were applied to the generated mask, and the segmented image was obtained after $bitwise\ AND$ operation of the original image and modified mask. Both the mask and the segmented image were resized to 224x224 before saving them.

VI. Comparison of MIMIC-CXR vs. CXLSeg for visual feature extraction.

A classification study has been performed for both MIMIC-CXR and CXLSeg datasets to determine the accuracy improvement when using a segmented dataset for feature extraction. The training was carried out using different CNN architectures, namely ResNet, DenseNet, MobileNet, and Inception. The results of the study are shown in Table III. The accuracy of the trained model is the ability to correctly identify an X-ray image of a healthy and unhealthy person. The diseased people can be identified correctly using the sensitivity of a test, while healthy people can be identified using specificity. In a real clinical setting, identifying correct patients is essential to have a high level of correctness.

TABLE III: Classification result using MIMIC-CXR dataset and CXLSeg dataset. (Average scores of the classes Lung Lesion, Pneumonia, Atelectasis, and Pleural Effusion)

Model	**Avg. Acc(%)**		**Avg. Sensitivity(%)**		**Avg. Specificity(%)**	
	MIMIC	***CXLSeg***	***MIMIC***	***CXLSeg***	***MIMIC***	***CXLSeg***
ResNet50	66.21	67.30	63.55	64.11	75.29	76.82
MobileNetV2	60.33	60.98	59.26	60.38	63.86	64.55
DenseNet121	71.48	73.03	69.26	72.81	69.31	70.24
InceptionV3	67.25	69.12	64.79	66.27	68.85	69.13

Analyzing the results of Table III, the CXLSeg dataset yielded better results for each CNNs than the original MIMIC-CXR dataset. When the images are not segmented, the model extracts features, including the features in the background, which downgrade the overall results. In CXLSeg, since the images are segmented, the model will only extract features in the lung area to make decisions, thus achieving higher accuracy. Therefore, image segmentation and feature extraction-based tasks using CXLSeg will yield better results than the MIMIC-CXR dataset.

VII. Results and Discussion

Imbalance datasets are widely present since it is hard to create a dataset without class imbalance. The Shenzhen dataset, Montgomery dataset, and V7-labs dataset have class imbalances and need data augmentation before utilization. Dice Coefficient (DC) and Intersection over Union (IoU) scores are suitable for imbalanced datasets and are positively correlated. The DC is well-known for image segmentation and can act as a loss function. The DC (2) tends to have a better value when the similarity of the generated mask and the true mask is higher. In medical image segmentation tasks, cross-entropy loss, dice loss, or a combination of the two is used as the loss function. DC is calculated as shown in (2) where FP: False Positive, TP: True Positive, FN: False Negative, T: true pixels, and P: predicted pixels. The IoU (2) is similar to DC and has been used to evaluate the selected different U-Net variants.

$$DC(T,P) = \frac{2TP}{2TP + FP + FN} = \frac{2|T \cap P|}{|T| + |P|} \quad (1)$$

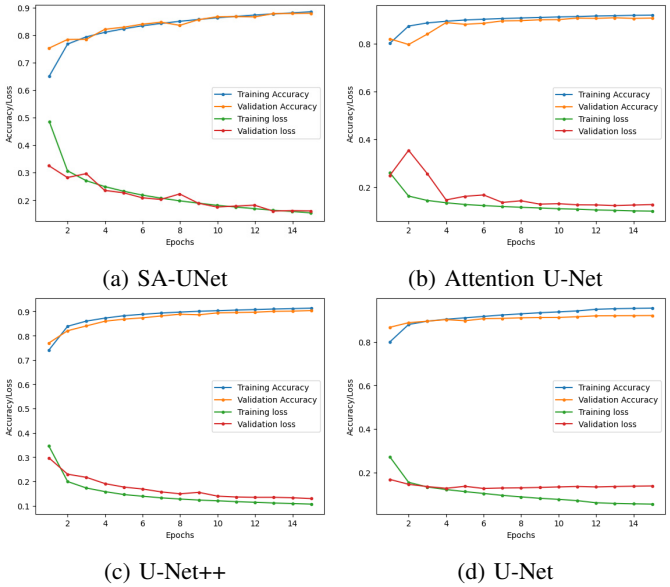
$$IoU(T,P) = \frac{TP}{TP + FP + FN} = \frac{|T \cap P|}{|T \cup P|} \quad (2)$$

Furthermore, a combination of dice loss and cross-entropy loss is used to calculate the loss. We selected four U-Net variants for the comparative analysis, namely Vanilla U-Net [12], U-Net++ [13], Attention U-Net [14], and SA-UNet [11] with a spatial attention mechanism.

TABLE IV: CXR segmentation results of U-Net variants

Model	**Loss(%)**	**DC(%)**	**IoU(%)**	**Recall(%)**	**Precision(%)**
U-Net [12]	16.92	92.88	86.75	93.68	94.68
U-Net++ [13]	13.02	94.88	90.28	95.42	96.06
Attention U-Net [14]	13.94	95.09	90.66	95.17	95.66
SA-UNet [11]	**12.59**	**96.80**	**91.97**	**96.05**	**96.23**

Table IV presents the image segmentation results using different variants of U-Net architecture. Bold text is used to highlight the best performance of each metric. Considering the table, it can be identified that the highest DC of 96.80% was achieved by the SA-UNet. Moreover, IoU, recall, and precision are also highest for SA-UNet. Regarding the loss, a minimum of 12.59% was achieved at the end of 15 epochs by SA-UNet. U-Net++ performance is relatively on par with Attention U-Net in most cases. In addition, Fig. 3 shows training and validation curves of accuracy and loss for the selected U-Net variants. The accuracy of training and validation curves increased along with each epoch, and the loss derived from the summation of dice loss and cross-entropy decreased with each epoch. Training and validation curves of the graphs portrayed in Fig. 3(a), Fig. 3(b), and Fig. 3(c) tend to be converging towards the latter part of the epochs. As shown in Fig. 3(b) and



(a) SA-UNet (b) Attention U-Net

(c) U-Net++ (d) U-Net

Fig. 3: Learning curves for the selected U-Net variants

Fig. 3(c), both U-Net++ and Attention U-Net have performed similarly despite of the first two epochs, where Attention U-Net shows low performance since there is a variation between the training and validation curves. Additionally, the gaps between the curves of U-Net tend to increase in the latter part of the epochs, as depicted in Fig. 3(d), resulting in overfitting. In comparison, SA-UNet architecture comprises a structured set of dropouts, which helps to perform better without overfitting [11]. Thus, Fig. 3(a) depicts a high level of convergence and better accuracy values compared to other models. Since SA-UNet achieved the best performance of all the models, it is used for the segmentation of the MIMIC-CXR images.

VIII. Conclusion

We present CXLSeg, a comprehensive dataset of lungs segmented in CXR images based on the MIMIC-CXR dataset. It comprises segmented images of frontal CXR images (both AP and PA) and their respective masks. A U-Net model with spatial attention (SA-UNet) architecture is used to segment the original images since it yielded high results after a comparative analysis of different UNet Variants. This study has shown that the proposed CXLSeg dataset resulted in better classification performance than the classification using the original MIMIC-CXR dataset. This is crucial in visual feature extraction, as using segmented images of CXLSeg helps the model focus more on the lung area and extract additional and more relevant features. Additionally, CXLSeg can be utilized as a lung segmentation dataset for biomedical image segmentation models using the masks provided. Furthermore, the dataset consists of a metadata CSV, a split CSV, and a mask CSV to increase the usability and convenience of the users. To the best of our knowledge, this is the first attempt to segment the entire dataset with 243,324 frontal CXR images and the corresponding segmented masks for each CXR image. Therefore, CXLSeg is the largest segmentation dataset available for Chest X-Rays. All the evaluation metrics used show that using CXLSeg can enhance the process of visual feature extraction. Furthermore, CXLSeg can be used for medical report generation since the visual feature extraction process is more effective when using the segmentation images.

References

[1] "Segmentation and classification on chest radiography: a systematic survey," *The Visual Computer*, Jan. 2022. [Online]. Available: https://doi.org/10.1007/s00371-021-02352-7

[2] G. L. Jones, "Chest x-ray quality." [Online]. Available: https://www.radiologymasterclass.co.uk

[3] D. Meedeniya, H. Kumarasinghe, S. Kolonne, C. Fernando, I. De la Torre Díez, and G. Marques, "Chest x-ray analysis empowered with deep learning: A systematic review," *Applied Soft Computing*, vol. 126, p. 109319, 2022, doi: https://doi.org/10.1016/j.asoc.2022.109319.

[4] V. Wijerathna, H. Raveen, S. Abeygunawardhana, and T. D. Ambegoda, "Chest x-ray caption generation with chexnet," in *2022 Moratuwa Engineering Research Conference (MERCon)*, 2022, pp. 1–6, doi: https://doi.org/10.1109/MERCon55799.2022.9906263.

[5] H. Kumarasinghe, S. Kolonne, C. Fernando, and D. Meedeniya, "U-net based chest x-ray segmentation with ensemble classification for covid-19 and pneumonia," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 18, no. 07, p. pp. 161–175, 2022, doi: https://online-journals.org/index.php/i-joe/article/view/30807.

[6] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng, "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific data*, vol. 6, no. 317, pp. 1–8, 2019, doi: https://doi.org/10.1038/s41597-019-0322-0.

[7] S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wáng, P.-X. Lu, and G. Thoma, "Two public chest x-ray datasets for computer-aided screening of pulmonary diseases," *Quant. Imaging Med. Surg.*, vol. 4, no. 6, pp. 475–477, Dec. 2014.

[8] L.-C. Chen, P.-C. Kuo, R. Wang, J. Gichoya, and L. A. Celi, "Chest x-ray segmentation images based on MIMIC-CXR," 2022. [Online]. Available: https://physionet.org/content/lung-segment-mimic-cxr/1.0.0/

[9] J. Shiraishi, S. Katsuragawa, J. Ikezoe, T. Matsumoto, T. Kobayashi, K.-i. Komatsu, M. Matsui, H. Fujita, Y. Kodera, and K. Doi, "Development of a digital image database for chest radiographs with and without a lung nodule," *American Journal of Roentgenology*, vol. 174, no. 1, pp. 71–74, 2000, doi: https://doi.org/10.2214/ajr.174.1.1740071.

[10] X. He, X. Yang, S. Zhang, J. Zhao, Y. Zhang, E. Xing, and P. Xie, "Sample-efficient deep learning for covid-19 diagnosis based on ct scans," *medrxiv*, 2020.

[11] X. Fan, C. Yan, J. Fan, and N. Wang, "Improved u-net remote sensing classification algorithm fusing attention and multiscale features," *Remote Sensing*, vol. 14, no. 15, p. 3591, 2022, doi: https://doi.org/10.3390/rs14153591.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015, doi: https://doi.org/10.48550/arXiv.1505.04597.

[13] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: redesigning skip connections to exploit multiscale features in image segmentation," 2019, doi: https://doi.org/10.48550/arXiv.1912.05074.

[14] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention u-net: Learning where to look for the pancreas," 2018, doi: https://doi.org/10.48550/arXiv.1804.03999.