Automated Pneumothorax Segmentation in Frontal Chest X-rays using CheXNet as the Backbone for Two-stage U-Net: An application of Transfer Learning

Arman Haghanifar
arman.haghanifar@usask.ca
Dept. of Biomedical Engineering,
Univ. of Saskatchewan

Ian Stavness
ian.stavness@usask.ca
Dept. of Computer Science,
Univ. of Saskatchewan

Seok-Bum Ko
seokbum.ko@usask.ca
Dept. of Elec. & Comp. Engineering,
Univ. of Saskatchewan

ABSTRACT

Pneumothorax, or collapsed lung, happens when air is present in the pleural space between the lung and chest wall. It can be small, which does not need treatment, or large that causes even death not diagnosed and treated in time. Small pneumothoraces can be hard to detect and also time-consuming, which needs to be assessed via expert domain experts. Recently, deep learning has been providing significant assistance in detecting and segmenting pneumothoraces. In this study, we propose a two-stage training system to segment images with pneumothorax. The introduced system is built based on U-Net with Squeeze-and-Excitation Residual Networks (SeResNeXt) backbone for segmentation and CheXNet backbone for classification. Moreover, we utilize various techniques, including Stochastic Weight Averaging (SWA), aggressive data augmentation, and learning rate schedule. We use the chest x-ray dataset provided by the 2019 SIIM-ACR Pneumothorax Segmentation Challenge, including 12,047 training images and 3,205 testing images. Our proposed system achieves 0.8410 Dice Similarity Coefficient (DSC) on private test-set, being placed among the top 7% of models with a rank of 96 out of 1,475 teams.

KEYWORDS

Pneumothorax, Image Classification, Image Segmentation, CheXNet, SeResNeXt, U-Net, Transfer Learning

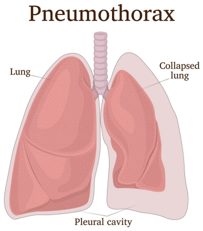


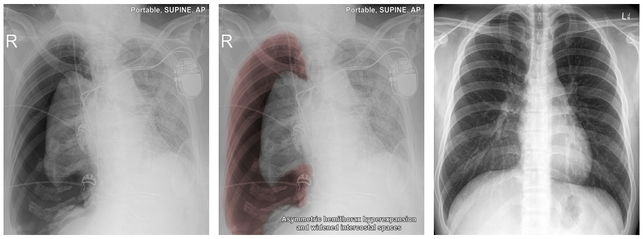
Figure 1: Illustration of pneumothorax or lung collapse

1 INTRODUCTION

Pneumothorax, also known as collapsed lung, happens when air comes through the pleural cavity between the lungs and the chest wall. Accumulation of air constantly enlarges the pleural space, resulting in compression of lung collapse and mediastinal structures. Typical symptoms of pneumothorax include sudden chest pain and shortness of breath [18]. A high-level illustration of pneumothorax is demonstrated in Fig. 1.

Pneumothorax is one of the most prevalent thoracic diseases, which happens because of excessive air accumulating in the pleural space between lung and chest wall. Pneumothorax can be caused by various reasons, such as other lung diseases, or in some cases from an accident or injury in the chest area. However, in other cases it may occur with no specific reason [14]. Pneumothoraces can also be classified as simple (no shift in mediastinal structures), tension (shift in mediastinal structures) or open (air passing through an open chest wound). Tension pneumothorax is a valve-liked fatal situation that results in air being trapped in the parietal and visceral pleura inside the chest. Pleural pressure increases as more gas are coming through the pleural space. The involved lung is then pushed inside, the hemidiaphragm is pushed down, mediastinal structures are displaced, and finally results in compression atelectasis involving the other lung. Traumatic and tension pneumothoraces are life-threatening and require immediate treatment. Therefore, early recognition of this condition is life-saving [3, 12].

Clinical detection of pneumothorax is complicated due to the variety of its symptoms and causes. Hence, a chest x-ray (CXR) is a common tool to help radiologists diagnosing pneumothoraces. However, detection can be difficult by visual inspection, specifically when its locations are atypical or the patient has heart or lung diseases at the same time [2]. Due to its different types, pneumothorax has various manifestations when diagnosed using CXRs. A frontal CXR is usually taken from the patient for pneumothorax detection. Fig. 2 demonstrates a visual comparison between normal x-ray and a case with pneumothorax present in the right lung 2.


R
Portable, SUPINE, AP
R
Portable, SUPINE, AP
Asymmetric hemithorax hyperexpansion and widened intercostal spaces
L

(a) (b) (c)

Figure 2: Comparison between normal and pneumothorax CXRs. (1) is a case with pneumothorax in the right lung, (2) is the annotated case, and (3) is a case with normal lungs

Pneumothoraces are different in terms of CXR manifestations and in many cases there might appear with subtle radiographic signs. Diagnosis accuracy of pneumothorax is thus highly dependent on the experience of the attending radiologists [8]. Besides, successful treatment is reliant on in-time review of the acquired CXR. In medical situations, a lack of well-trained radiologists is usual, and proper diagnosis and treatment of pneumothorax are often delayed, which can result in severe harm to patients, even death. Therefore, a Computer-Aided Diagnosis (CAD) tool is essential to assist examiners in the accurate detection of pneumothorax.

Deep Learning (DL) and computer vision techniques have been recently led to a remarkable breakthrough in medical image analysis. Deep learning networks use large amounts of data to automatically extract data features, making the training process fully automatic and resolving the need for manual feature engineering [4]. Among various deep learning algorithms, Convolutional Neural Networks (CNNs) are utilized to process images and have achieved great success in developing DL-based CAD systems. Deep neural networks have also been utilized for biomedical image segmentation in recent years.

However, a limited number of research studies have been conducted to develop DL-based methods for segmentation of pneumothorax from CXRs. The reason lies behind some major challenges in this particular area. First challenge is the large variation of pneumothoraces in terms of size, shape and location inside the lungs. Larger pneumothoraces are clearly easier to be detected, while smaller ones may also be hard to diagnose by the domain expert. Second is the lack of open datasets with sufficient confidently labeled data. There are currently a number of large CXR datasets, i.e. CheXpert [1] with approximately 225,000 images and PadChest [2] with more than 160,000 x-rays. These datasets mostly provide labels rather than annotated masks. To alleviate the shortage of annotated data, the Society for Imaging Informatics in Medicine (SIIM) created a pneumothorax segmentation challenge: *2019 SIIM-ACR Pneumothorax Segmentation Challenge*[3] providing a large set of annotated pneumothorax x-rays. Since then, there have been a number of research studies employing DL methods to improve classification and segmentation results on the dataset.

A very first approach is made by Jakhar *et al.* [7] to apply a U-Net with ResNet34 as the backbone on the public training set and achieved a dice similarity coefficient (DSC) of 0.8430 on the public test set. Further, Islam *et al.* [6] trained a SeResNeXt50-based U-Net using 80% of the public training set and got 0.6858 DSC score on the other 20%. Authors of [15] proposed a weakly supervised approach to deal with imperfect annotation problem, where they hit a DSC score of 0.7690 on the private test set. Wang *et al.* [16] proposed CheXLocNet as an ensemble of networks developed based on Mask R-CNN which achieved 0.82 DSC score on the private test set. The most recent effort is made by authors of [17] that resulted in a DSC score of 0.8883 on the private test set, placing them in the second best solution. They utilized a two-stage training network by employing various ensembles of pretrained networks as the backbone for the U-Net based on SeResNeXt and EfficientNet architectures. The proposed study discussed the possibility of using a pretrained network on similar type of images, i.e. CXRs, as the backbone of U-Net. A two-stage network is developed and optimized with different methods, and a good DSC score is achieved on the private test set. The rest of the paper is structured as follows: Dataset specs are explained in Section 2. Section 3 presents the details of the proposed classification and segmentation models. Experimental results and hyperparameter tuning is investigated in Section 4. Finally, discussion and conclusion are presented in Section 5 and Section 6, respectively.

2 MATERIAL

In this research study, CXR dataset of SIIM-ACR Pneumothorax Segmentation Challenge on Kaggle is used, as previously mentioned. Image data is in Digital Imaging and Communications in Medicine (DICOM) format, while bounding boxes are stored in a CSV file. DICOM is a format that includes metadata, such as patient sex and age, along with pixel data (image itself) attached to it. All CXR images are considered as frontal, with a view position of either Anterior-Posterior (AP) or Posterior-Anterior (PA). Patient names are anonymized and not visible in the dataset. Patient sex and age are also included in the metadata. All images are in a size of 1024 × 1024 pixels with capturing modality of either Computed Radiography (CR) or Digital Radiography (DR).

Annotations are in the form of image IDs accompanied by Run-Length Encoding (RLE) masks. RLE is a lossless data compression method in which runs of data are saved as a single data value and count. RLE form is used for images with pneumothorax mask values where pixel locations are measured from previous ends of the run. Thus, a function must be written to convert RLE-coded masks into mask images. On the other hand, images without pneumothorax have mask value of -1, which equals to a completely black image as the mask. Conversion from RLE to mask and vice versa is required to submit the model results to the leaderboard. A sample dataset image with its mask is shown in Fig. 3.

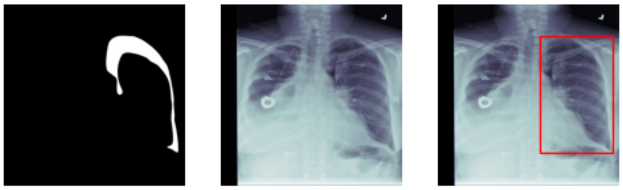


Figure 3: A sample x-ray from the dataset along with its pneumothorax mask

The CXR dataset consists of 12,954 binary masks along with 12,047 images in the training-set, and 3205 images in the test-set. Negative images may be from healthy chests or patients with other chest-related diseases, such as Pneumonia. In the training-set, there are more masks than the images. The reason is that some training images have multiple annotations, i.e., multiple Pneumothoraces. In the training-set, there are 9378 images from negative class (77.84%),

[1] https://stanfordmlgroup.github.io/competitions/chexpert/
[2] https://bimcv.cipf.es/bimcv-projects/padchest/
[3] https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation

while there are 2669 ones from positive class (22.16%). Hence, class imbalance is observed in the training-set.

In pneumothorax positive CXRs, 624 have multiple annotation masks and 2669 have only one mask. Among multi-mask images, the most number of masks per image is 10, and most multi-masks have only 2 masks per image. Worth mentioning that multiple masks in an image may overlap each other. Thus, to have one mask as the ground truth for each of the training-set images:

- In negative class images, a black mask is built and considered as the mask for that case.
- In single-masked images, the corresponding mask is the ground truth of the image
- In multi-masked images, all related masks are loaded and the union of them is computed to build the final mask as the ground truth.

To have a better overview of the metadata, the distribution of training-set images in term of patient sex, patient age category, CXR view projection, and CXR modality is shown in Fig. 4 and Fig. 5.

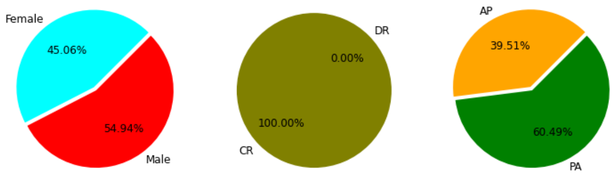


Figure 4: Distribution of dataset images in terms of (1) patient sex, (2) image modality, and (3) image view position

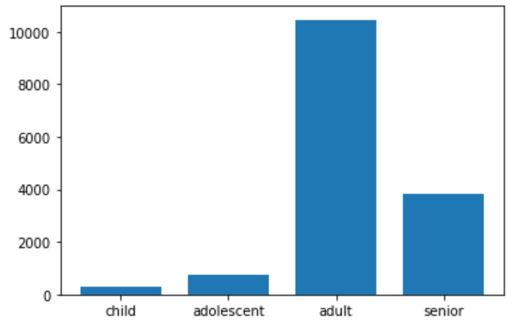


Figure 5: Distribution of images in terms of patient age categories

As seen in above figures, patients are divided into male and female classes with approximately equal ratio. All CXRs are taken with the modality of CR. Majority of CXRs are taken from PA view, which has higher image quality than AP view. More than 80% of patients are classified as adults, while youth patients are quite small. Age category guideline is taken from a CDC project [4].

Another point worth-investigating is the percentage of pneumothorax in dataset images where one or more pneumothoraces are present, which is illustrated in Fig. 6.

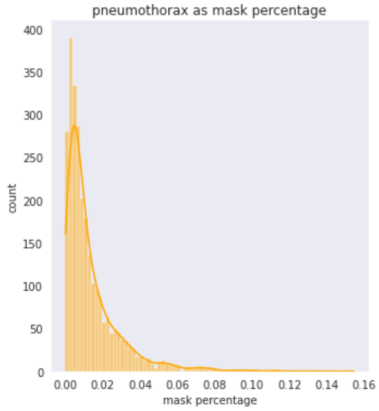


Figure 6: Distribution of mask percentage in positive cases

Segmentation models are in fact classifiers that classify each input pixel of an image into 0 or 1 classes. Thus, since most of the masks fall in negative class, the dataset is imbalance. To deal with the above-mentioned issue, we use class weights for binary cross-entropy loss. Class weight is set as 0.01 according to the mask percentage average as seen above.

3 METHODS

In the following section, we provide details about the proposed network pipeline, from image preprocessing to model development and post-processing. A high-level illustration of the methodology of our approach is in Fig. 7. Details of the design are discussed thereafter.

3.1 Preprocessing

Initially, images are in .dcm format with a size of 1024 × 1024 as provided by the challenge organizers. Pixel array is extracted and saved as a Numpy array along with metadata in CSV format. Then, to reduce the memory utilization, images are down-sized to 512 × 512 and also 256 × 256. Different histogram equalization algorithms are tested resulting in Contrast Limited Adaptive one (CLAHE) to give the best contrast enhancement. CLAHE is a famous enhancement algorithm used in many domains as the main preprocessing step. Besides, Adaptive Histogram Equalization (AHE) is also applied on the raw images. Main image along with AHE-applied and CLAHE-applied ones are concatenated to expand input dimensions to 3-channeled images. Before being fed to the network, a number of augmentation techniques are applied on the image to increase input variations and make the network more robust to unseen images in the test-sets.

3.2 Base Network: U-Net

U-Net is initially introduced by Ronneberger *et al.* [11] in 2015 for segmentation of biomedical images. U-Net is a u-shaped CNN which is an enhancement of Fully Convolutional Networks (FCN).

[4] https://www.cdc.gov/nchs/products/databriefs/db334.htm

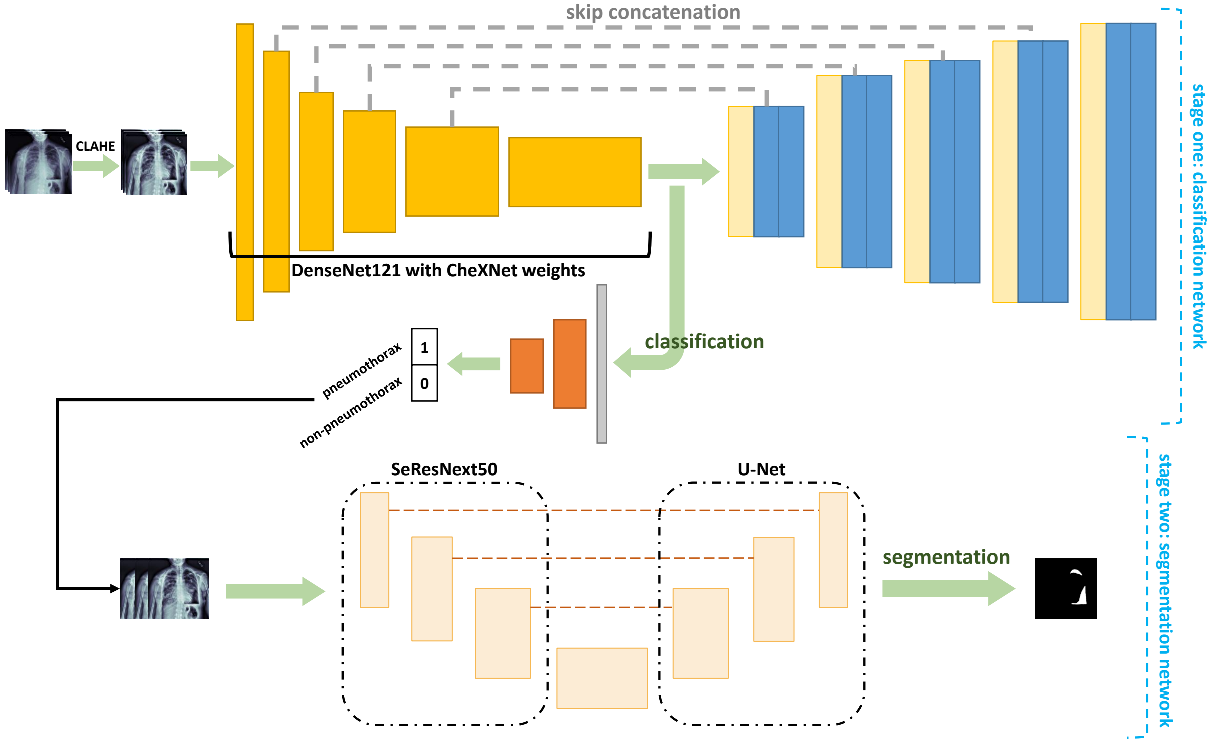

skip concatenation
CLAHE
DenseNet121 with CheXNet weights
classification
1
0
pneumothorax
non-pneumothorax
stage one: classification network
SeResNext50
U-Net
segmentation
stage two: segmentation network

Figure 7: The flow diagram of the proposed two-stage model. A U-net is used in the stage one with backbone set to DenseNet121 with weights coming from a pretrained network on chest x-rays. The output of the backbone is used along with flatten and dense layers to produce classification results. All images classified as having pneumothorax are then inserted into another U-Net with SeResNeXt50 as its backbone to produce the final segmented lesion as mask.

FCNs are introduced for semantic segmentation earlier in 2015 [9], and U-Net-based architectures hit better results when dealing with smaller datasets. Thus, U-Nets have been successfully applied for medical segmentation tasks.

By the development of several more complex U-net-based architectures, the initial one is now considered as the vanilla U-Net. Vanilla U-nets are created by adding a couple of convolutional layers as the encoder, followed by the same number of up-convolution layers as the decoder. The difference between U-Net and simple FCN is the concatenations between each convolutional layer in encoder, with its counterpart up-convolution layer in decoder. The baseline architecture for pneumothorax segmentation is shown in Fig. 8.

In each block, two convolution layers are followed by a max-pooling layer to reduce the input size. Activation function of convolution layer is set as Rectified Linear Unit (ReLU), which enhances the model performance by adding non-linearity to the network. The final layer is a convolutional layer with 1 × 1 kernel to merge the information of different channels.

Almost all of the previous works have been developed based on the U-Net as the main segmentation method. Hence, both stage one and stage two networks are designed followed by the popular U-Net architecture.

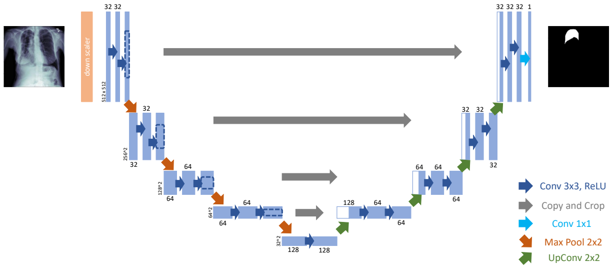


Figure 8: Architecture of a custom vanilla u-net for pneumothorax segmentation

3.3 Stage One: Classifier

A multi-task learning strategy is used to train the Stage one classifier network by using two output branches, one for pneumothorax detection and the other for pneumothorax segmentation. Worth mentioning that the segmentation branch is only used during Stage one model training and not at test time. Besides, same augmentation techniques are used both for classifier and segmenter networks.

Firstly, images are classified into two classes, pneumothorax and non-pneumothorax, based on the labels extracted from the RLE-coded masks. To develop the network, we used a pretrained network, trained on a large dataset of CXRs, as the backbone for U-Net: The CheXNet.

CheXNet is a robust model for lung disease detection based on chest x-ray images [10]. This model is trained on CXR-14, one of the largest publicly available dataset of chest x-rays from adult cases with 14 different diseases. And one of the disease labels is pneumothorax. CheXNet, which is claimed to have a radiologist-level diagnosis accuracy, is based on DenseNet architecture and has been trained on frontal CXRs.

The public dataset is divided into 80% dedicated for training and the other 80% for validation. Then, this procedure is repeated to perform a k-fold cross-validation method, where k is 5. Batch size of 32 is selected and training is continued for 100 epochs, while early stopping is present to stop the model from overfitting. Optimizer is set to Adam with default learning rate of 0.001.

Since our model is a binary classifier, binary cross-entropy or log loss is used as the loss function. Binary cross-entropy is a function to calculate the cross-entropy between the predicted classes and the true classes. It results in a prediction value between 0 and 1 for each image. Binary cross-entropy is formulated as below.

$$H_p(q) = \frac{-1}{N} \sum_{i=1}^{N} y_i \cdot log(p(y_i)) + (1 - y_i) \cdot log(1 - p(y_i))$$

where y is the label and $p(y)$ is the predicted probability of the image having pneumothorax or not.

Typical performance metrics to evaluate a classifier model are accuracy, precision, recall and F-score. Since we have a comparison between pneumothorax and non-pneumothorax (may include several other diseases), we need to consider both recall and precision scores for evaluation. Besides, class imbalance may result in false high accuracy, while most correctly classified images may come from negative class. Therefore, the acceptable metric for model evaluation is F2-score; as follows.

$$F2 - score = 5 \times \frac{precision \times recall}{(4 \times precision) + recall}$$

where precision is fraction of true positive samples to all positive predicted ones, recall is fraction of true positives to all positives in the dataset. F2-score takes both false positives and false negatives into account, and is more accurate than the accuracy score when dealing with class imbalance.

3.4 Stage two: Segmenter

To construct the segmentation network, similar U-Net is utilized with a different backbone, pretrained on the ImageNet dataset: Squeeze-and-Excitation-based Residual Network (SeResNeXt). A variety of networks are experimentally tested to select the best one. Segmentation is only applied on images which are supposed to have at least one pneumothorax, according to the classifier results. Hence, all images are expected to have masks with pneumothorax present as white pixels.

Similar method is approached in terms of data augmentation, dataset split into training-set and validation-set, and also 5-fold cross-validation. Class weights are used to force the model look for white pixels, since producing all-black masks for input images yields to a misunderstanding high accuracy scores. Batch size is set to 32, and the optimizer is Adam. This time, initial learning rate is selected by applying an optimal learning rate finder technique calculated using the approach introduced in [13]. During the training, an aggressive learning rate schedule is used to apply the change at the end of each epoch based on the value computed using a cosine annealing schedule. Cosine annealing schedule is computed as follows.

$$\alpha = \frac{\alpha_0}{2} \times (cos(\frac{\pi \times \frac{n}{N}}{N}) + 1)$$

where α_0 is the previous learning rate, n is current epoch, N is the total number of training epochs, and α is the computed learning rate for the next epoch.

For the segmentation task, two metric are considered. First, Intersection over Union (IoU), which is the area of overlap between the ground truth and the predicted segmentation, divided by the area of union between them. IoU is one of the standard metrics for segmentation model evaluation, and is computed as follows.

$$IoU(P_{true}, P_{predicted}) = \frac{P_{true} \cap P_{predicted}}{P_{true} \cup P_{predicted}}$$

Moreover, to be able to compare the results with the leaderboard, competition's official evaluation metric should also be considered. Sorenson-Dice coefficient (DSC) is an evaluation metric used to compare a predicted segmentation and its corresponding ground truth, pixel-wisely. DSC is computed as follows.

$$DSC(X, Y) = \frac{2 \times |X \cap Y|}{|X| + |Y|}$$

where X is the predicted set of pixels and Y is the ground truth.

Our segmentation model acts similar to a binary classifier described before; it classifies each pixel into two groups of 0 being black and 1 being white. Hence, binary cross-entropy itself can also be used as the loss function. However, dice loss is another useful function which utilizes the dice coefficient, and is obtained by subtracting dice coefficient from 1. To achieve optimal performance, a combination of binary cross-entropy and dice loss is thus employed. Experimentally tested, best combination is to add both loss functions with equal weights.

Note that in both network, dropout layers are also devised after each convolutional and deconvolutional blocks to prevent overfitting on the input data after a few epochs. This is very important when training models with high number of parameters, such as DenseNet or EfficientNet.

3.5 Post-processing

A useful detail about the metrics used for segmentation network is that predicted values for each pixel are replaced with 1 if greater than 0.5, or to 0 otherwise. This helps the segmenter understand the certainty in mask annotations. However, best threshold for using step function to cast values into 0 or 1 is different in each case. After the training, we compute the IoU metric in different thresholds between 0 to 1 to find the best one for achieving highest

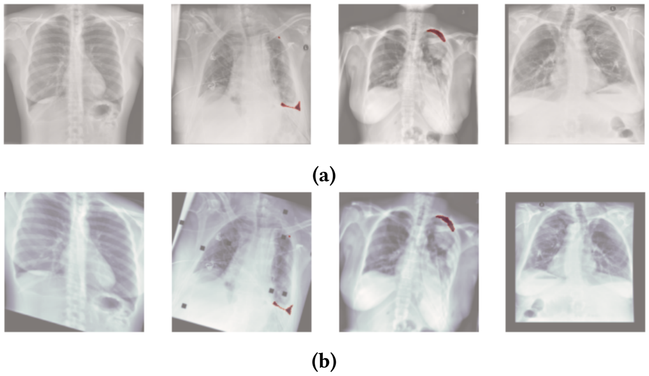
IoU from the model. The final threshold is replaced with initial one for testing the model with public and private test-sets.

4 RESULTS

Due to the memory limit of the system, original size of the images could not be used for training. Therefore, resized images to 512×512 and 256 × 256 are tested, where smaller images resulted in faster training time without a noticeable change in learning curves and metric scores. For the augmentation, Albumentations [5] library is used due to the fact that it is capable of applying same augmentation functions on the image and its mask at the same time for less time than Keras ImageDataGenerator class. A set of augmentation functions are selected:

- Coarse Dropout
- Horizontal Flip
- One of the followings randomly: Random Brightness or Random Contrast or Random Gamma
- One of the followings randomly: Grid Distortion or Optical Distortion or Elastic Transform
- Rotation
- Zoom in/out

Images are also down-scaled from $[0, 255]$ to $[0, 1]$. A random batch of images and their augmented outputs are shown in Fig. 9.



(a)

(b)

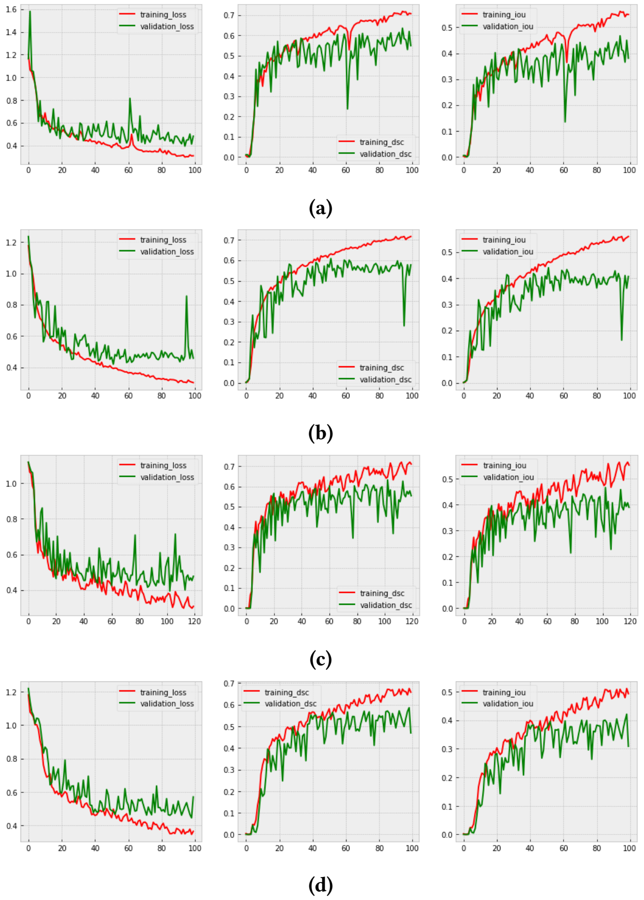
Figure 9: A random batch of images (a) before augmentation, and (b) after augmentation

For the classification task, various backbones are examined: plain CNN, DenseNet with pretrained ImageNet weights, DenseNet with pretrained weights from CheXNet, ResNet34 and ResNet50. Experiments are conducted in three circumstances; setting the whole backbone network as non-trainable, setting the network non-trainable except the last convolutional block, or setting the whole network to trainable. Making the network trainable increases training time into approximately twice and leads to 150 required epochs where the non-trainable and partly-trainable versions converge the best possible state in less than 70 epochs. Best f2-score is achieved from partly-trainable DenseNet with CheXNet weights, which is 0.9054. Table 1 is the confusion matrix of the proposed classifier network.

Table 1: Confusion matrix of the proposed stage-one classifier on 20% of the dataset

Classifier		**Predicted**	
		Non-pneumothorax	*Pneumothorax*
Actual	*Non-pneumothorax*	1786	88
	Pneumothorax-19	52	484

Segmentation network gets all positive samples from the classifier as the input. To select the backbone for U-Net of the segmenter, a number of networks are tested. Firstly, four experiments are conducted to tune two main hyperparameters; whether to freeze backbone and whether to use learning rate schedule or not. Trainings are done based on ResNet34 as the backbone for 100 or 120 epochs with batch size set to 32. Three plots are drawn, learning curves, DSC of training and validation sets, as well as IoU of these sets. Results are illustrated in Fig. 10.



(a)

(b)

(c)

(d)

Figure 10: Tuning main hyperparameters of segmentation network. (a) demonstrates training results of trainable resnet with default learning rate. (b) is non-trainable with default rate. (c) is trainable resent with learning rate schedule. And (d) is non-trainable backbone with learning rate schedule of cosine annealing.

[5] https://albumentations.ai

Trainable ResNet34 backbone with default learning rate is trained for more epochs to ensure an overfitting status is monitored. According to the plots, setting backbone as non-trainable along with learning rate schedule postpones overfitting, and hits the same best result as others.

After selecting best combination of hyperparameters, different networks are trained and compared. Fig 11 shows learning curves of these models on 20% of the public dataset.

As seen in plots above, SeResNeXt50 has most stable progress during training procedure and also hits slightly higher DSC for validation-set. Note that all the experiments are based on training the networks on 20% of the public dataset and showed from fold 0 of the 5-fold cross-validation process. After all folds are done, stochastic weight averaging is performed on top-3 models to create the best single segmentation network.

Finally, thresholding over different IoUs resulted in best threshold set to 0.69 to achieve higher IoU. Final model is submitted to the challenge, which obtained 0.8987 and 0.8410 in public and private leaderboards, respectively. Proposed system is placed in top 7% of all participating teams. A set of examples on the validation-set are shown in Fig. 12 to demonstrate model efficiency of localizing pneumothorax lesions.

5 DISCUSSION

Although our proposed approach is not placed in the top 1% of challenge leaderboard, it proves that the selected transfer learning and two-stage segmentation methods result in promising metric scores. Our solution made it to be among 100 top solutions among 1475 teams. Table 2 shows a comparison between the proposed model and some of the challenge participants publishing their works as a research article.

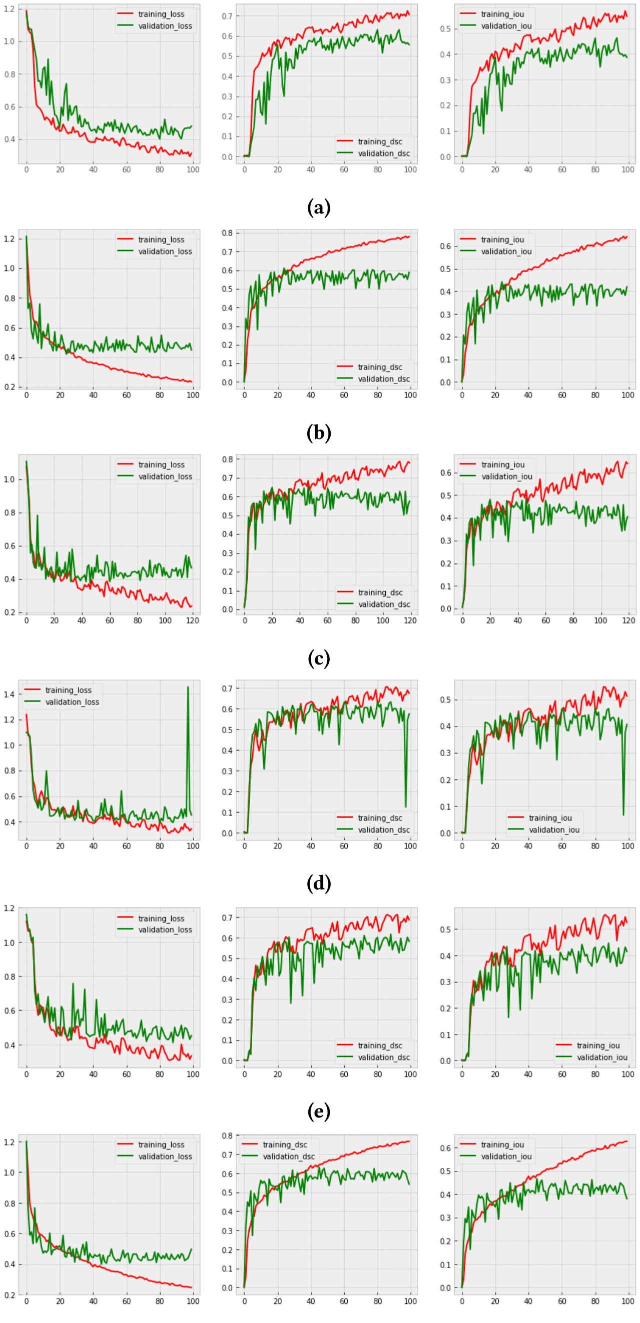
Worth mentioning that we have used single network as backbone for U-Nets, while most top solutions employed ensemble of different networks which have time-consuming training and inference process. No other study has reported to benefit from pretrained networks on chest x-rays, such as CheXNet used by the current study. Two-stage-solutions have resulted in higher DSC scores on the private test-set.

There are certain methods to improve model performance which are not investigated due to the time limit of this study:

- Label smoothing both for images and pixels, to hinder overfitting.
- Benefiting from metadata stored with images as network input to achieve a probable improvement in metric scores, which is not used by any other studies.
- Collecting pneumothorax CXRs from external sources to improve classifier network.
- Employ post-processing pixel-level techniques to boost the final score.

6 CONCLUSION

Deep learning has proved to promisingly classify and segment thoracic diseases in chest x-rays. In this paper, a two-stage U-Net-based system is introduced and compared with



(a)

(b)

(c)

(d)

(e)

(f)

Figure 11: Comparison of different networks as the backbone for segmentation U-Net-based model. (a) DenseNet121, (b) EfficientNet-B0, (c) EfficientNet-B7, (d) InceptionV3, (e) SeResNet34, and (f) SeResNeXt50. DSC plots are considered as the main source of comparison, due to the fact that leaderboard is sorted based upon DSC score on test-sets.

Table 2: Comparison between published models

Developed system	Network Type	DSC on Validation	DSC on Public Board	DSC on Private Board
Abedalla *et al.*[1]	Single	-	-	0.8356
Wang *et al.*[17]	Ensemble	-	0.9048	0.8665
Jakhar *et al.*[7]	Single	-	0.8430	-
Islam *et al.*[6]	Single	0.6858	-	-
Wang *et al.*[16]	Ensemble	-	-	0.82
Groza & Kuzin [5]	Ensemble	-	-	0.8614
Viniavskyi *et al.*[15]	Single	-	-	0.7690
Proposed method	Single	0.6450	0.8987	0.8410

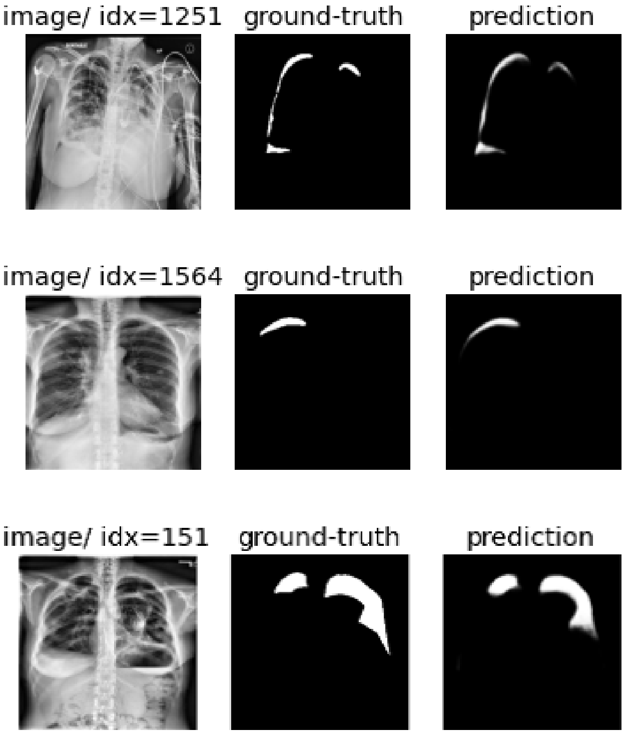


Figure 12: Sample predicted masks from validation-set

state-of-the-art solutions. CheXNet and SeResNeXt are selected experimentally as best backbone options for classification and segmentation tasks, respectively. Dataset images are provided by the 2019 SIIM-ACR Pneumothorax Challenge, and the proposed model achieved dice similarity scores of 0.8987 in its public test-set, and 0.8410 in the private leaderboard. Aggressive data augmentation, dropout layers, CLAHE image enhancement, thresholding over intersection-over-union metric, optimal learning rate finder using cyclical training, stochastic weight averaging, and learning rate schedule via cosine annealing are other improvement techniques used in the current approach.

REFERENCES

[1] Ayat Abedalla, Malak Abdullah, Mahmoud Al-Ayyoub, and Elhadj Benkhelifa. 2020. The 2ST-UNet for Pneumothorax Segmentation in Chest X-Rays using ResNet34 as a Backbone for U-Net. *arXiv preprint arXiv:2009.02805* (2020).

[2] Kuan-Yu Chen, Jih-Shuin Jerng, Wei-Yu Liao, Liang-Wen Ding, Lu-Cheng Kuo, Jann-Yuan Wang, and Pan-Chyr Yang. 2002. Pneumothorax in the ICU: patient outcomes and prognostic factors. *Chest* 122, 2 (2002), 678–683.

[3] Raiko Diaz and Daniel Heller. 2020. Barotrauma and mechanical ventilation. *Stat-Pearls, Treasure Island* (2020).

[4] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. 2016. *Deep learning*. Vol. 1. MIT press Cambridge.

[5] Vladimir GROZA and Artur KUZIN. 2020. Pneumothorax Segmentation with Effective Conditioned Post-Processing in Chest X-Ray. In *2020 IEEE 17th International Symposium on Biomedical Imaging Workshops (ISBI Workshops)*. IEEE, 1–4.

[6] Shariful Islam, Hasin Rehana, Sayed Asaduzzaman, Syed Mobassir Hossen, Rabby Hossain, Touhid Bhuiyan, Muhammad Shahin Uddin, and Nargis Akter. 2020. Automated Risk Prediction by Measuring Pneumothorax Size using Deep Learning. In *2020 IEEE Region 10 Symposium (TENSYMP)*. IEEE, 1747–1751.

[7] Karan Jakhar, Rohit Bajaj, and Ruchika Gupta. 2019. Pneumothorax Segmentation: Deep Learning Image Segmentation to predict Pneumothorax. *arXiv preprint arXiv:1912.07329* (2019).

[8] Brendan S Kelly, Louise A Rainford, Sarah P Darcy, Eoin C Kavanagh, and Rachel J Toomey. 2016. The development of expertise in radiology: in chest radiograph interpretation,"expert" search pattern may predate "expert" levels of diagnostic accuracy for pneumothorax identification. *Radiology* 280, 1 (2016), 252–260.

[9] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3431–3440.

[10] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. 2017. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225* (2017).

[11] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.

[12] Anita Sharma and Parul Jindal. 2008. Principles of diagnosis and management of traumatic pneumothorax. *Journal of Emergencies, Trauma, and Shock* 1, 1 (2008), 34–41. https://doi.org/10.4103/0974-2700.41789

[13] Leslie N Smith. 2017. Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 464–472.

[14] Charlie Strange. 1999. Pleural complications in the intensive care unit. *Clinics in chest medicine* 20, 2 (1999), 317–327.

[15] Ostap Viniavskyi, Mariia Dobko, and Oles Dobosevych. 2020. Weakly-Supervised Segmentation for Disease Localization in Chest X-Ray Images. In *International Conference on Artificial Intelligence in Medicine*. Springer, 249–259.

[16] Hongyu Wang, Hong Gu, Pan Qin, and Jia Wang. 2020. CheXLocNet: Automatic localization of pneumothorax in chest radiographs using deep convolutional neural networks. *PLoS One* 15, 11 (2020), e0242013.

[17] Xiyue Wang, Sen Yang, Jun Lan, Yuqi Fang, Jianhui He, Minghui Wang, Jing Zhang, and Xiao Han. 2020. Automatic Segmentation of Pneumothorax in Chest Radiographs Based on a Two-stage Deep Learning Method. *IEEE Transactions on Cognitive and Developmental Systems* (2020).

[18] Paul Zarogoulidis, Ioannis Kioumis, Georgia Pitsiou, Konstantinos Porpodis, Sofia Lampaki, Antonis Papaiwannou, Nikolaos Katsikogiannis, Bojan Zaric, Perin Branislav, Nevena Secen, et al. 2014. Pneumothorax: from definition to diagnosis and treatment. *Journal of thoracic disease* 6, Suppl 4 (2014), S372.