Deep Learning for Diagnosis and Segmentation of Pneumothorax: The Results on The Kaggle Competition and Validation Against Radiologists

Alexey Tolkachev, Ilyas Sirazitdinov, Maksym Kholiavchenko, Tamerlan Mustafaev, Bulat Ibragimov

***Abstract* — Pneumothorax is potentially a life-threatening disease that requires urgent diagnosis and treatment. The chest X­ray is the diagnostic modality of choice when pneumothorax is suspected. The computer-aided diagnosis of pneumothorax has received a dramatic boost in the last few years due to deep learning advances and the first public pneumothorax diagnosis competition with 15257 chest X-rays manually annotated by a team of 19 radiologists. This paper describes one of the top frameworks that participated in the competition. The framework investigates the benefits of combining the Unet convolutional neural network with various backbones, namely ResNet34, SE-ResNext50, SE-ResNext101, and DenseNet121. The paper presents a step-by-step instruction for the framework application, including data augmentation, and different pre- and post-processing steps. The performance of the framework was of 0.8574 measured in terms of the Dice coefficient. The second contribution of the paper is the comparison of the deep learning framework against three experienced radiologists on the pneumothorax detection and segmentation on challenging X-rays. We also evaluated how diagnostic confidence of radiologists affects the accuracy of the diagnosis and observed that the deep learning framework and radiologists find the same X-rays to be easy/difficult to analyze (p-value <1e4). Finally, the methodology of all top-performing teams from the competition leaderboard was analyzed to find the consistent methodological patterns of accurate pneumothorax detection and segmentation.**

***Index Terms* — pneumothorax, deep learning, convolutional neural network, chest X-rays, Kaggle competition, radiologist validation**

1. Introduction

A

BNORMAL presence of air in pleural space –pneumothorax – is a life-threatening condition that requires urgent medical attention. Pneumothorax is suspected in patients who exhibit chest pain, acute dyspnea had a previous history of pneumothorax incidence or/and other risk factors.

Bulat Ibragimov is the corresponding author; email: [bulat@di.ku.ku.](mailto:bulat@di.ku.ku)

Alexey Tolkachev was with Innopolis University, Innopolis, Russia, He is now with the Industry Center for Development and Implementation of Information Systems, Sochi, Russia (email: [aleksey.tolkachev@ocrv.ru](mailto:aleksey.tolkachev@ocrv.ru)).

Ilyas Sirazitdinov was with Innopolis University, Innopolis, Russia, He is now with Philips Research, Moscow, Russia (email: [ilyas.sirazitdinov@philips.com](mailto:ilyas.sirazitdinov@philips.com)).

The anterior-posterior chest X-ray is the main diagnostic image modality for confirming pneumothorax. Chest X-rays are essential for the detection of small air pockets, which usually exhibit inconclusive symptoms, and during treatment of tension pneumothorax, when there are doubts about the exact anatomical location of the disease. Visual inspection of X-rays, however, does not guarantee the correct diagnosis of pneumothorax and up to 14% of incidents remain undiagnosed at initial screening [1]. It has been reported that patients with misdiagnosed pneumothorax at early stages have a significant risk to develop tension pneumothorax and have longer hospitalization time. Automated screening of chest X-rays to identify patients with a high risk of pneumothorax and prioritizing them for a review by a physician can accelerate pneumothorax diagnosis and treatment.

Automated X-ray screening for pneumothorax can be decomposed into the detection of patients with air pockets in the lung fields and automated segmentation of air pockets because the urgency of pneumothorax treatment is related to the size of air pockets. The early attempts to automate pneumothorax diagnosis were based on the modeling the appearance of air pockets in X-ray images using Hough transforms [2], local intensity histograms and image edges to capture the visceral pleural edge [3], and texture information for the quantification of vascular markings in the lung [4]. The diagnostic accuracy of such algorithms remains relatively low as pre-defined appearance features cannot capture the rich variety of human lungs and air pockets. The development of deep learning opens a new pathway for automated diagnosis of pneumothorax. Deep learning algorithms can be trained to classify X-rays with and without pneumothorax using reference binary diagnostic labels [5]–[12]. Cicero et al. [12] were among the pioneers of pneumothorax diagnosis with deep learning, when they applied GoogleNet to detect five common lung pathologies using more than 35000 adult X-rays and achieved

Maksym Kholiavchenko is with Innopolis University, Innopolis, Russia (email: [m.kholyavchenko@innopolis.ru](mailto:m.kholyavchenko@innopolis.ru)).

Tamerlan Mustafaev is with the Public Hospital #2, Department of Radiology, Kazan, Russia, and Innopolis University, Innopolis, Russia. (email: [t.mustafaev@innopolis.ru](mailto:t.mustafaev@innopolis.ru)).

Bulat Ibragimov is with the University of Copenhagen, Department of Computer Science, Copenhagen, Denmark, and Innopolis University, Innopolis, Russia (e-mail: [bulat@di.ku.ku](mailto:bulat@di.ku.ku)).

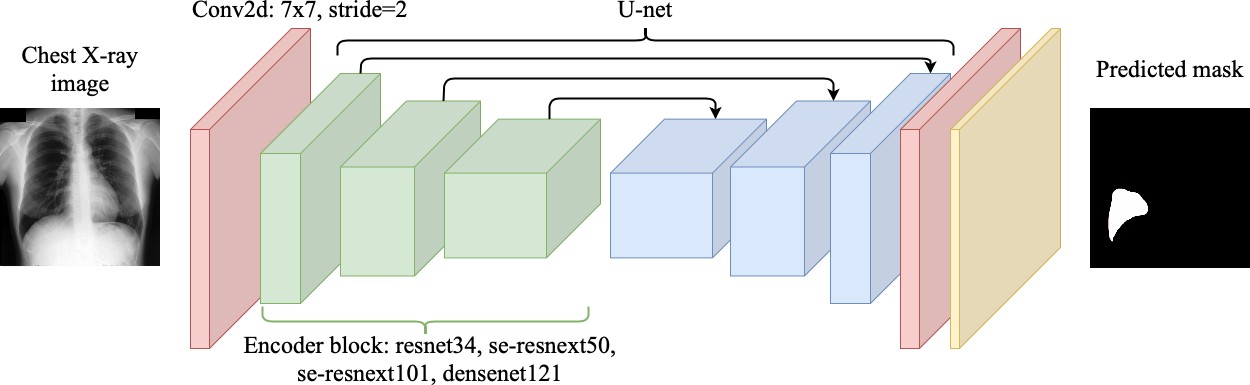


Fig. 1. The schematic illustration of the neural network architecture families utilized in the proposed framework for pneumothorax detection and segmentation.

pneumothorax detection accuracy of 0.86 measured in terms of the area under the receiving operator curve (AUC). Taylor et al.

[11] compared the performance of Xception, Inception, VGG, and ResNet neural network architectures and achieved 0.94 AUC for pneumothorax detection. Rajpurkar et al. [13] later demonstrated that the performance of ResNet is statistically comparable to the performance of radiologists on pneumothorax diagnosis from chest X-rays. Ouyang et al. [14] focused not only on automated pneumothorax diagnosis but also on the segmentation of individual pneumothorax pockets using Unet, LinkNet, and Tiramisu network architectures. Despite relatively accurate diagnosis results, their segmentation results were moderate with the accuracy of 0.67 measured in terms of the intersection of union. Park et al. [9] applied YOLO network to automatically recognize incidents of traumatic pneumothorax after percutaneous transthoracic needle biopsies and found out that the detection of such pneumothorax type is significantly less accurate than the detection of nontraumatic pneumothorax. Goosen et al. [5] compared the performance of the patch-based pneumothorax detection, where a neural network is iteratively applied to classify individual X-ray patches trying to predict if each patch contains pneumothorax, with the end-to-end pneumothorax detection, where a whole chest X-ray is analyzed in a single pass. The authors concluded that patch-based networks result in more accurate diagnoses but unable to segment individual pockets. The popularity and affordability of chest X-rays significantly facilitate the development of deep learning solutions. The median database in the mentioned-above studies contained more than 25000 cases, where 3%-10% were X-rays with pneumothorax. The reported pneumothorax detection accuracy can reach 0.96 AUC of ROC, which is comparable to the performance of human observers. However, Kitamura and Deible [6] have recently shown that the accuracy of pneumothorax detection drops to

0.59 AUC when deep learning solutions are trained and tested on databases from different hospitals in contrast to 0.9 AUC when the solutions are trained and tested on different parts of the same database. Such accuracy overestimation of published studies can be mitigated by independent algorithm evaluation

when the authors of automated methods have no access to test labels. The medical imaging community established the mechanism of public international competitions, where research groups submit their algorithms to be evaluated by the organizer on publicly inaccessible testing data. The Society for Imaging Informatics in Medicine (SIIM), the American College of Radiology (ACR) and the Society of Thoracic Radiology (STR) have recently addressed the need for objective evaluation of automated pneumothorax diagnosis and organized a public competition on pneumothorax detection and segmentation [15].

This work presents one of the top-performing frameworks that participated in the pneumothorax detection and segmentation competition. There are three main contributions of the presented work. First, we present a step-by-step implementation of automated pneumothorax segmentation with the source code provided. The work capitalized on the advances in deep learning and implemented the U-net network with ResNet34, SE-ResNext50, SE-ResNext101, DenseNet121 backbones to segment pneumothorax pockets [16]. Second, we recruited three practicing radiologists to perform pneumothorax segmentation of a representative subset of chest X-rays with the aim to qualitatively assess the framework performance. The subset contained images with large air pockets to estimate the inter-observer variability and compare the reference labels to the annotations of the radiologists and deep learning performance. The subset contained X-rays significantly misclassified by the deep learning framework to check if such images represent a challenge for radiologists too. We also asked radiologists to assess how confident they were when diagnosing each patient. To conclude the experiment with human experts, we asked a radiologist to qualitatively assess the framework performance and the performance of his colleagues on challenging cases and suggest directions for future research. Finally, we studied the methodological components of all top-performing solutions to identify and summarize the key components of accurate pneumothorax segmentation. The overarching aims of this work are to provide the community with a step-by-step deep learning framework for pneumothorax diagnosis, systematically access the pneumothorax diagnostic

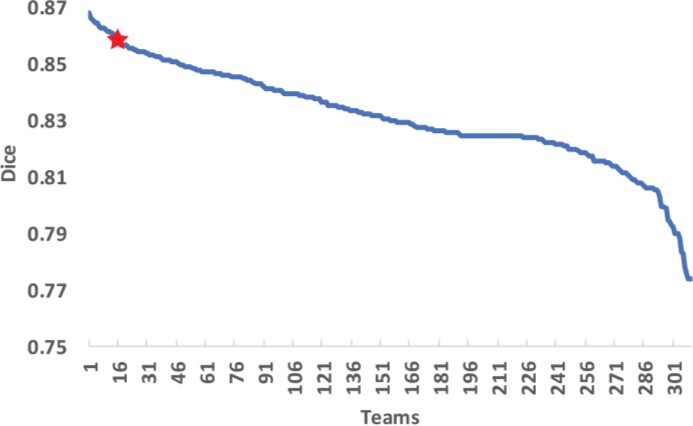


Fig. 2. The performance summary for the frameworks participated in the pneumothorax detection and segmentation competition. The plot corresponds to the final leaderboards computed using 99% of the testing X-rays, where all teams are sorted according to the average Dice coefficient of their frameworks. The teams outside the top 310 teams are not visualized to preserve the fine resolution of the plot. The proposed framework is shown with the red star superimposed over the plot.

performance of deep learning on the results of public Kaggle competitions, and critically compare the deep learning framework against radiologists on difficult-to-diagnose cases.

1. Methodology
2. *Database description*

Six radiologists from the SIIM and thirteen radiologists from STR participated in the preparation of the competition database. Most of the chest X-rays were sampled from the public ChestX-Ray8 database released previously by the National Institutes of Health (NIH). ChestX-Ray8 contains 108948 anterior-posterior and lateral chest X-rays labeled with 14 most-common thoracic disorders. The labeling of an X-ray was performed automatically by applying natural language processing (NLP) algorithms to parse the radiological report written for the corresponding patient [17]. Due to the imperfection of NLP, the resulting labels were not fully accurate and were updated for the pneumothorax competition. A database of 15302 chest X-rays with 5302, 5000, and 5000 X-rays initially labeled as pneumothorax, no finding, and other than pneumothorax abnormalities, respectively, was assembled [7]. Forty-five exams were excluded due to wrong fields of view leaving 15257 X-rays in the database. Six radiologists from the SIIM manually annotated six nonintersecting subsets of 1500-2000 X-rays covering 10902 X-rays. Each annotation consisted of any number of freeform contours encompassing each pneumothorax air pocket, including zero contours if the patients had no pneumothorax. The radiologists used a commercial toolbox from MD.ai for manual medical image annotation. The regional convolutional neural network (Mask R-CNN) was trained on the annotated 10902 X-rays and then applied to automatically annotate the remaining 4355 X-rays. The same six radiologists then corrected or approved the results of Mask R-CNN. The complete database of annotated 15257 X-rays was

next visually inspected by 12 radiologists from the STR, who were instructed to approve or disapprove the SIIM annotations. The SIIM and STR radiologists diagnosed 1490 X-rays differently. All the 1490 disputed X-rays were finally annotated by the thirteenth radiologist from STR who did not participate in the first two annotation rounds. Additional 902 chest X-rays were included into the competition database, but followed a different annotation protocol.

The chest X-rays in the database were non-isotropically rescaled to a unified size of 1024x1024 pixels. Around 72% of X-rays from the database were provided with reference annotations for training at the beginning of the competition. Around 8% of X-rays constituted the first testing database, for which reference annotations were released during the competition, while the remaining 20% of X-rays had no annotations released for the public and were used for evaluating the performance of the participating frameworks. The competition task was to detect and segment pneumothorax bounding boxes on testing chest X-rays. The result evaluation was performed using the mean Dice coefficient for all the images:

|  |  |
| --- | --- |
| 2|𝑆𝑟𝑒𝑓 ∩ 𝑆𝑎𝑢𝑡𝑜|  𝐷𝑖𝑐𝑒 =  |𝑆𝑟𝑒𝑓| ∪ |𝑆𝑎𝑢𝑡𝑜| | (1) |

where 𝑆𝑟𝑒𝑓 and 𝑆𝑎𝑢𝑡𝑜 are, respectively, the reference segmentation mask and automated segmentation mask submitted for evaluation. If both 𝑆𝑟𝑒𝑓 and 𝑆𝑎𝑢𝑡𝑜 masks were empty, the Dice coefficient was set to be equal 1. Such an extension of the Dice coefficient was utilized by the competition organizers. Annotations of test X-rays were not available for the public during the competition, and framework evaluation was automatically performed on the Kaggle platform [15]. A comprehensive description of the image curation and database creation was published by the organizers in the Journal of Digital Imaging [7].

1. *Deep neural network for pneumothorax detection and segmentation*

The state-of-the-art deep learning architectures were studied to best address the problem of pneumothorax detection and segmentation. The problems of similar nature, such as detection and segmentation of pedestrians on streets, tumors in the brain, human organs for radiotherapy planning, etc. are solved by deep convolutional neural networks (CNN) of the encoder-decoder type [16], [18]. In our work, we utilized the U-net architecture as it produces highly accurate medical image segmentation in various anatomical sites including chest [19], [20]. The U-net architecture consists of an encoder and decoder blocks augmented with skip-connections. The encoder block serves to extract meaningful features from the input image and downscale the feature maps so that local appearance features are progressively assembled into regional and global appearance features. The downscaling of the feature map size can be achieved using pooling layers or convolution layers with

>1 stride. The decoder block serves to upscale feature maps towards the input image resolution and effectively utilize fine

image information extracted by the encoder block by using skip connections. Each encoder and decoder blocks consist of interchangeable subblocks that generate and extract features. We implemented and tested ResNet34, SE-ResNext50, SE-ResNext101, DenseNet121 decoders that won several public competitions in the field of image analysis [21]–[24]. The ResNet34 and DenseNet121 utilize residual blocks that interconnect previous network layers with consecutive layers improving the gradient flow through the network during training and potentially reducing the risks of vanishing gradients. The squeeze-and-excitation component (SE-ResNet50, SE-ResNet101) utilized attention blocks and aggregation blocks to improve feature extraction. All convolution blocks were augmented with the batch normalization layers to improve network robustness and facilitate convergence. The final convolutional layer is followed by a sigmoid activation function to generate the pneumothorax segmentation results. The schematic illustration of the network is given in Figure 1. Please find the framework implementation in Python under the following link [https://github.com/dirtmaxim/super\_pneumothorax.](https://github.com/dirtmaxim/super_pneumothorax)

1. *Pneumothorax Detection and Segmentation*

1) *Preprocessing and Network Training*

All chest X-rays in the database 𝑰 ∈ 𝕀 were of the 1024x1024 pixels of uniform size. We isotropically downscaled all the X-rays to the 768x768 pixels, which significantly reduced the memory consumption during CNN training, while the quality

of the X-rays stayed relatively unchanged. The mean X-ray ̅𝑰

and standard deviation 𝝈 over all pixels in all the X-rays were computed and then used to normalize all X-rays to the same

Table I

The Parameters of the Presented Framework For Pneumothorax Segmentation. The Python Implementation of the Framework is Available Through the Following Link:

[https://github.com/dirtmaxim/super\_pneumothorax.](https://github.com/dirtmaxim/super_pneumothorax)

|  |  |
| --- | --- |
| Parameter | Value |
| Input image size | 768×768 |
| Batch size | 6 |
| Optimizer | Adam |
| Initial learning rate | 0.0001 |
| Learning rate scheduler | [15, 25, 35] |
| Scheduler coefficient | 0.1 |

the maximum factors of 0.15. A random combination of the corresponding rigid transformations was generated for every training image. Finally, a training image may be flipped in a horizontal direction with 0.5 probability. Such flipping is the simplest way of getting pneumothorax examples for the right and left lung fields from a single training image.

The described-above augmentations were randomly applied every time the CNN training procedure requested a new sample. The probability of applying these augmentations was the following: a) one augmentation among brightness, contrast, gamma, and CLAHE was randomly applied for each training sample; b) there was 0.5 probability of applying either additive Gaussian noise or random blur; c) the random shift and scale were applied for every sample, while flip was applied with 0.5 probability. The sample generator for CNN training first decided whether a specific augmentation should be applied for the next training sample and, if the decision was positive,

randomly generated the parameter for the augmentation from

intensity frame: 𝑰𝑘

← 𝑰𝒌−̅𝑰.

𝝈

the acceptable parameter range.

The performance of CNN is usually directly proportional to the size and representativeness of the training database. To enrich CNN without adding new chest X-rays from private sources, we augmented the already provided X-rays during CNN training. We utilized intensity appearance augmentations and augmentations based on rigid X-ray transformations. The intensity appearance augmentations included: a) brightness augmentation with the maximum factor of 0.2; b) contrast augmentation through rescaling the original intensities to 0-5 percent; c) gamma augmentation with the percent range from 70 to 130; d) contrast limited adaptive histogram equalization (CLAHE) with the upper threshold value for the contrast limit to be equal 0.1. For each training image, one augmentation from the list above was randomly selected and the augmentation parameters were then randomly sampled from the corresponding acceptable parameter range. An auxiliary intensity augmentation that included e) additive Gaussian noise with the predefined noise variance ranging from 10 to 50 or f) random blur with the maximal standard deviation and maximum kernel size equal to 5, was added with 0.5 probability. The rigid transformations included: a) random X­ray rotations of up to 20 degrees; b) image shifting by the maximal relative distance of 0.0225; and c) image scaling by

The stratified five-fold cross-validation protocol was

executed during the network training. The random fold splitting is shown to be suboptimal in the chest X-ray classification domain [25], as random splitting may result in a fold that does not encompass the rich variability of the target disease manifestations. For the stratified cross-validation, we first performed a k-means-based clustering of all training chest X-rays. This process separated X-rays into clusters according to four features, namely the number of pneumothorax air pockets, size of pockets, length of pocket contours, and their locations in the lung fields. All X-rays from a cluster were randomly separated into five subsets of a similar size and each subset was then included into individual folds for cross-validation. In other words, each fold contained the same proportion of X-rays from each cluster; therefore, all folds have a similar distribution of X-rays with different pneumothorax manifestations. Note that five-fold cross-validation was operated solely with the training samples from the competition. The labels for the testing samples were not available for the public and were only known to the organizers of the competition. The training was executed in a mixed-precision format, where most computations were performed with the half-precision training samples while retaining on the single-precision training samples was only

performed for critical parts of the network. The use of half-precision mode allowed not only to increase the batch size by a factor of 1.5 but also sufficiently speed-up the training. The encoder of the network was initially pre-trained on ImageNet using binary cross-entropy loss. After pre-training, the network was fine-tuned for 75 epochs using Dice loss and Adam optimizer with the initial learning rate of 3e-4, which was reduced by the factor of 10 on the 20th, 35th, and 50th epochs. The pre-trained encoder was incorporated into the segmentation network and then fine-tuned on the chest X-rays. The fine-tuning was performed for 25 epochs using the Lovasz-softmax loss [26] with the initial learning rate of 1e-5. The main parameters of the framework are summarized in Table I. The framework implementation can be found at GitHub under the following link

[https://github.com/dirtmaxim/super\_pneumothorax.](https://github.com/dirtmaxim/super_pneumothorax) We also utilized a public implementation of the segmentation backbones <https://github.com/qubvel/segmentation_models.pytorch> that ensures a straightforward integration of our framework into future research solutions. This public implementation has the backbones pre-trained on large-scale databases, which additionally speeds up the new solution development.

1. *Network Application and Postprocessing*

A three-step protocol was executed for generating CNN-based pneumothorax segmentations. A testing chest X-ray 𝑰 was first augmented with geometrical transformations for generating several slightly different examples of 𝑰 [27]. Adding such variability of the test sample helps to ensure that the X-ray will be viewed from different angles and is more likely to be labeled correctly. Second, a snapshot CNN ensemble segmented each example of 𝑰. The snapshot ensemble contains a collection of weights from the best-performing networks that were memorized every time the CNN training converged to a local optimum. Averaging the performance of several locally optimal CNNs is shown to produce more accurate classification results than the result of the single optimal CNN [28]. A collection of pneumothorax segmentation masks was generated for geometrically transformed 𝑰 examples. All the masks were returned to the original X-ray 𝑰 coordinate frame by applying the inverse of their augmentation transformation and then averaged into the final segmentation mask. As a result, the trained CNN produced the segmentation mask 𝑹 of the same size as the X-ray 𝑰. The array 𝑹 is thresholded according to parameter 𝒃 to form the binary segmentation array 𝑹𝒃, where

all the cells in 𝑹 with values < 𝒃 and ≥ 𝒃 are converted to 0

and 1, respectively. The binary array is then decomposed into connected components using 8-connectivity neighborhood. All the connected components smaller than 𝒄 pixels were removed because small pneumothorax candidates generated by the CNN are more likely to correspond to framework mis-segmentations than true pneumothorax air pockets. Both parameters 𝒃 and 𝒄 were automatically estimated during CNN training by maximizing the pneumothorax recognition performance on validation X-rays.

1. *Annotation by radiologists*

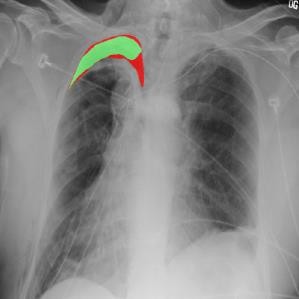
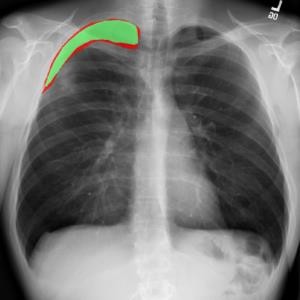
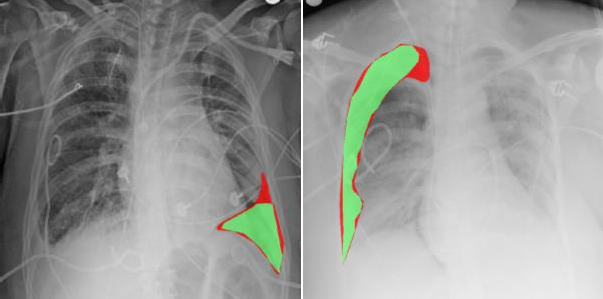
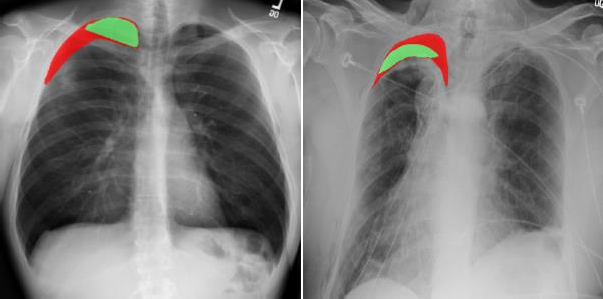
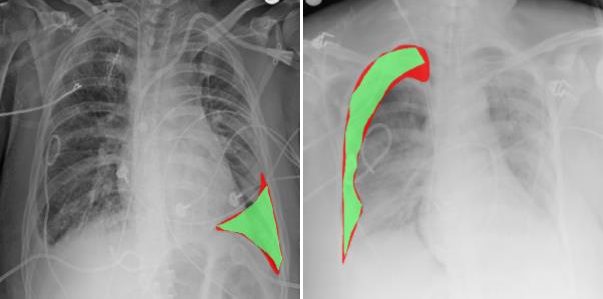
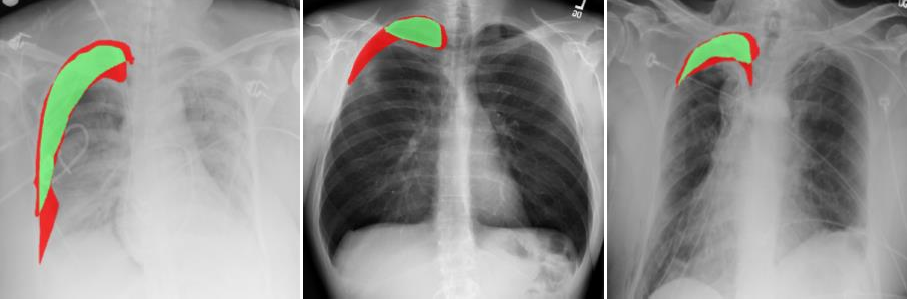
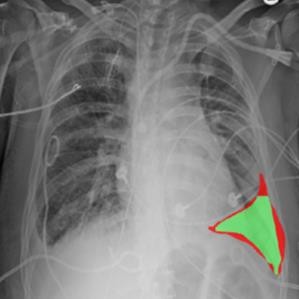
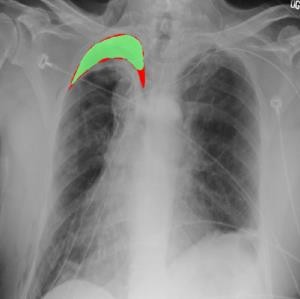
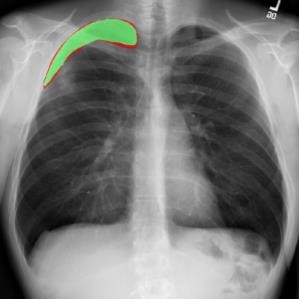
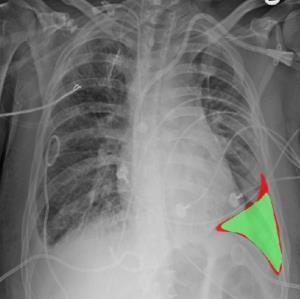
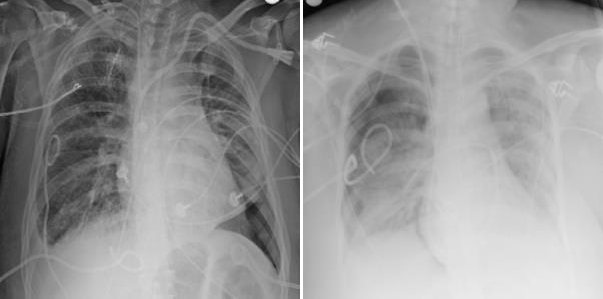
A confirmation subset of 100 chest X-rays was randomly sampled from the first testing database to compare the behavior of the deep learning framework against radiologist performance. The X-rays were sampled to include the following categories:

1. True positive (TP) examples. This category included chest X-rays with large air pockets, i.e. >10000 pixels that were correctly labeled by the CNN, i.e. Dice coefficient >0.9. The TP examples were needed to estimate the inter-observer variability when annotating large well-visible pneumothorax air pockets.
2. False positive (FP) examples. This category included chest X-rays, where the reference pneumothorax mask was empty, i.e. radiologists from SIIM and STR labeled the corresponding patient to be pneumothorax-free, while the CNN-generated mask contained at least one pneumothorax air pocket larger than b pixels.
3. False negative (FN) examples. This category included chest X-rays, where the reference mask was not empty, while the CNN-generated mask contained no pneumothorax air pocket larger than b pixels, i.e. the mask became empty after postprocessing.
4. True negative (TN) examples. This category included chest X-rays, where both the reference mask and post-processed CNN mask had no pixels labeled as pneumothorax.

The assembled confirmation subset of X-rays was given to three practicing radiologists to be manually labeled. The proportion of pneumothorax cases vs non-pneumothorax cases and the rules for selecting 100 X-rays were not revealed to the radiologists. The radiologists were asked to outline all pneumothorax air pockets on an X-ray using free-form contours in the supervise.ly annotation toolbox or label the X-ray as pneumothorax free. The radiologists were asked to ignore non-pneumothorax abnormalities potentially manifested on X-rays. The confirmation X-rays were first annotated by one radiologist to ensure the correctness of the protocol and then annotated by two other radiologists. All the radiologists were asked to additionally score their confidence when annotating each X­ray. The annotation results were evaluated against the reference masks following the same methodology as for the evaluation of the automated annotations. We finally asked a radiologist to give his expert opinion about the framework performance and typical human errors observed on the X-rays in the confirmation subset. The radiologist was armed with all available information about the confirmation X-rays including the reference pneumothorax annotations publicly released by the competition organizers, annotations by the three radiologists generated during the described-above experiment, and the results of the deep learning-based framework. All disputed cases from the confirmation subset were selected, i.e. cases where at least one radiologist or/and the framework disagreed with the reference diagnosis. The radiologist was asked to answer 14 questions and potentially provide a free-form opinion about each case. The questions targeted the following topics: a) agreement with the reference diagnosis; b) the observed pneumothorax properties that can make diagnostics challenging; c) alternative diagnosis options; d) image quality and patient positioning; e) the need for additional clinical or/and

|  |
| --- |
| **Original** |
| **Neural network** |
| **Radiologist #1** |
| **Radiologist #2** |
| **Radiologist #3** |

Fig. 3. The comparison of pneumothorax segmentation performed by deep learning framework, and three radiologists computed against the reference segmentations provided by the competition organizers. The green color denotes overlapping between a framework/radiologist segmentation and the reference segmentation, while the red color denotes areas of disagreement between segmentations.



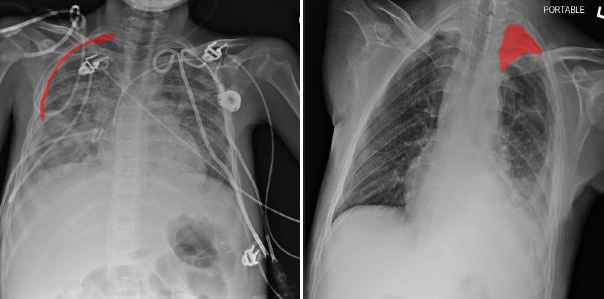
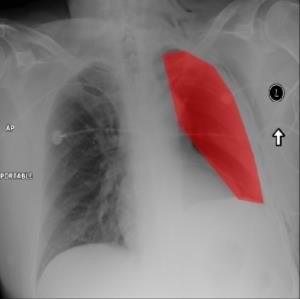
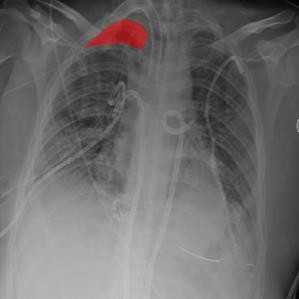
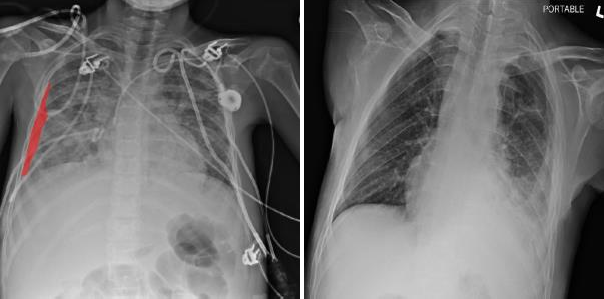
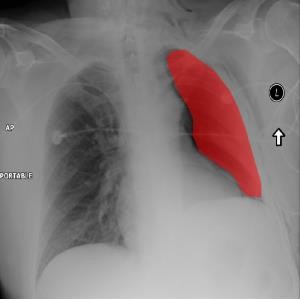
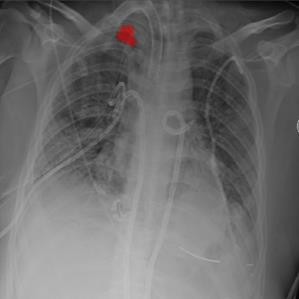
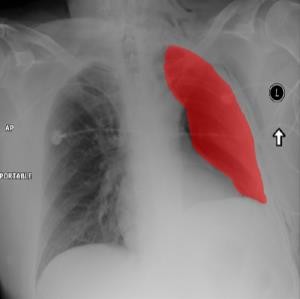
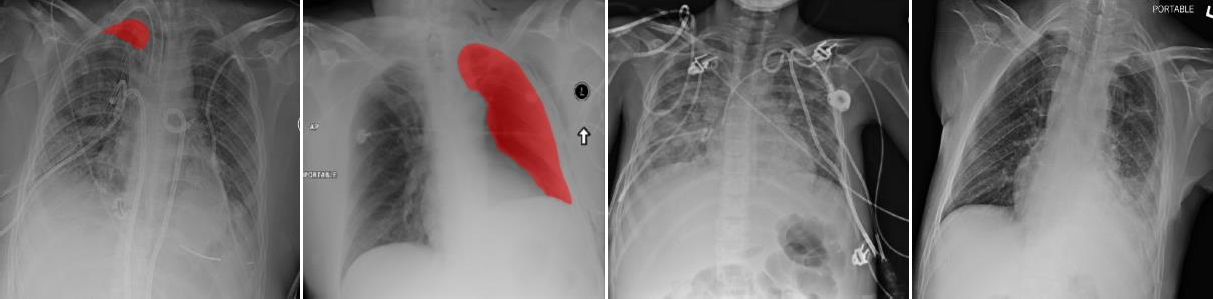
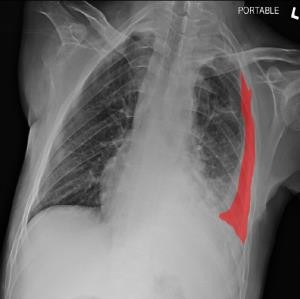
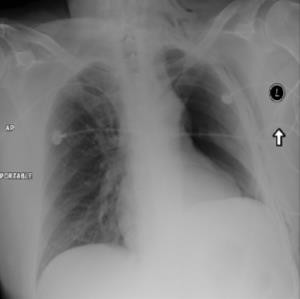
demographic data; and f) the significance of mistakes made by the algorithm. His answers were collected and summarized to qualitatively assess the framework performance on challenging cases.

1. RESULTS

The pneumothorax detection and segmentation competition was launched on the 25th of June 2019 and closed on the 5th of September 2019. During the first stage of the competition that

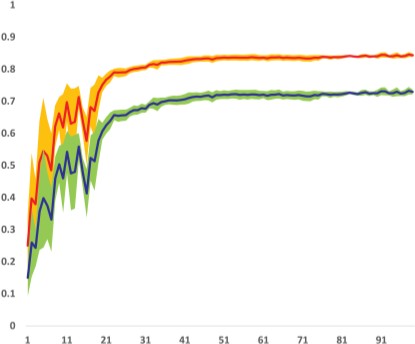
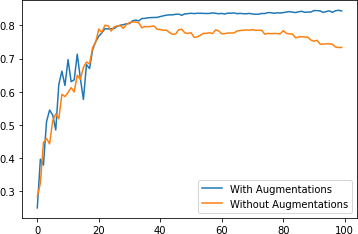
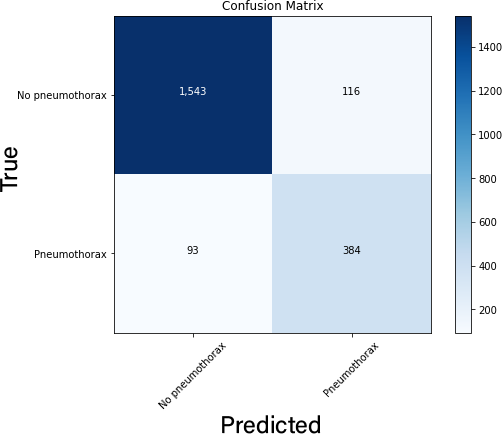
|  |
| --- |
| **Reference** |
| **Neural Network** |
| **Radiologist #1** |
| **Radiologist #2** |
| **Radiologist #3** |

Fig. 4. Segmentation of pneumothorax on four selected X-rays. The pneumothorax segmentation (red) superimposed over chest X-rays generated by the competition organizers (first row), i.e. reference segmentation, the proposed deep neural network-based framework (second row), and three physicians (third-fifth rows). The presented chest X-rays represented a significant challenge for the framework and physicians.



lasted until the 28th of August 2019, 11582 chest X-rays with reference annotations were released for algorithm training, and 1372 chest X-rays without reference annotations were released for algorithm testing. It was allowed to use any publicly available data for additional algorithm training, but the teams were asked to declare that they had used some additional data for training and provide references to the data sources. The

inclusion of training data from private chest X-ray collections was not allowed. Each participating team was asked to submit their algorithm performance on 1372 X-rays to the organizers for evaluation and their code before the end of the first stage. On the 29th of August 2019, the reference annotations for the first testing subset of 1372 X-rays, and new 3205 chest X-rays without reference annotations were released. All the

|  |  |  |
| --- | --- | --- |
| (a) | (b) | (c) |

Fig. 5. The framework performance on pneumothorax detection and segmentation form chest X-rays. (a) The framework performance measured in terms of Dice coefficient (red) and Intersection of Union (IoU) (blue) computed for pneumothorax pockets against validation chest X-rays. Framework training was performed on five stratified subsets of X-rays from the training data of the SIIM competition database. The shaded regions define the epoch-wise standard deviation across five subsets for Dice (yellow) and IoU (green). (b) The performance of the framework with and without training X-ray augmentation. (c) The confusion matrix of the framework computed over the validation dataset averaged against all five folds.

participants who passed the first stage were asked to apply their algorithms on the new test data until the 5th of September 2019. The organizers assembled the public algorithm leaderboard with the framework performances computed on a randomly sampled 1% of the second testing subset. The results on the remaining 99% of the second testing subset were released at the end of the competition and defined the final standing of the participants.

In total, 1475 teams participated in the competition. Some teams merged during the competition or did not make it to the second stage, resulting in 350 teams in the final leaderboard. The organizers disqualified teams who used multiple accounts to circumvent the restriction on the number of submission attempts or were reasonably suspected in other ways of cheating. During the framework development, we evaluated the performance of the framework with five-fold stratified cross-validation on the training part of the competition database (Figures 5 and 6). The proposed framework was ranked 16th in the final leaderboard and qualified to the top 1% of all 1475 participating teams. The obtained accuracy was 0.8574, while the top-performing team obtained an accuracy of 0.8679. The results from the final leaderboard for all the teams are presented

in Figure 2. Figure 5c summarizes the performance of the network in the confusion matrix form.

The performance of three radiologists was measured in terms of the Dice coefficient against the reference radiologists from SIIM and STR. The mean Dice coefficient was 0.63, 0.63, 0.64, and 0.55 for the chest X-rays from the confirmation subset for the first, second, and third radiologists, and the proposed framework, respectively. The summary of the radiologists' performance for TP, TN, FP, and FN subsets is presented in Table II. Figures 3 and 4 compare the pneumothorax segmentation performed by the competition organizers, CNN, and our radiologists for the selected X-rays. We evaluated how the confidence of the radiologists during the annotation of each X-ray agreed with the annotation accuracy. All the X-rays from the confirmation subset were separated into four categories depending on the number of confident radiologists. The overall confidence ranged from all the radiologists being confident to no radiologist being confident about the annotation of a particular X-ray. The mean Dice coefficient for the X-rays from each confidence subcategory was calculated and summarized in Table II. Table III summarizes the statistical comparison of the framework against each radiologist using the Z-test. For the

Table II

The Pneumothorax Segmentation Performance of the Presented Deep Learning Framework in Comparison to The Performance of Three Experienced Radiologists. The Table Presents the Results for a Representative Subsets of Chest X-Rays. Note That the Complexity of the Subset is Way Higher Than the Overall Complexity of the Testing Set of the Competition, As the Overall Framework Accuracy was 0.8574 Dice in Contrast to 0.5520 Dice for the Subset.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Dice coefficient | |  |
| Framework results | |  |  |  |
|  | Framework | Radiologist #1 | Radiologist #2 | Radiologist #3 |
| Correctly labeled pneumothorax X-ray | 0.9277 | 0.8509 | 0.8216 | 0.8746 |
| Mis-labeled pneumothorax X-ray | 0 | 0.0251 | 0.1501 | 0.1290 |
| Mis-labeled non-pneumothorax X-ray | 0 | 0.7500 | 0.5833 | 0.4167 |
| Correctly labeled non-pneumothorax X-ray | 1 | 0.9778 | 0.9333 | 1 |
| All X-rays | 0.5520 | 0.632 | 0.629 | 0.637 |

statistical comparison, all the scores were binarized, where a score for the framework/radiologist is considered positive if at least one pneumothorax air pocket is correctly identified. The positive values for confidence intervals indicate the superiority of the corresponding radiologist over the framework and vice versa.

1. Discussion

Pneumothorax is potentially a life-threatening disease that requires urgent diagnosis and treatment. Chest X-ray is the diagnostic modality of choice when the pneumothorax is suspected, as the X-ray does not only confirm the diagnosis but also allows estimating the exact location and size of pneumothorax. Pneumothorax air pockets do not have evident appearance features, so the development of automated pneumothorax detection algorithms with hand-crafted features has been limited [2]–[4]. The field has received a dramatic boost in recent years due to machine learning advances and the release of public databases [29]. The opportunity to objectively evaluate the performance of automated solutions has arisen due to the recent public SIIM-ACR Pneumothorax Segmentation competition [15].

The competition gives us an unbiased estimation of the currently achievable pneumothorax segmentation accuracy. The exact algorithm details are not available for most of the participating submissions; however, all the teams who revealed their algorithms used CNNs to address the raised challenges. All the teams who qualified for the monetary awards or gold medal awards, i.e. 1-12 positions in the leaderboard, revealed their CNN architectures and implementation details. Two main strategies were selected by the top participants, namely to segment all pneumothorax pockets in one pass using an end-to-end network, or to first perform X-ray classification into healthy or pathological and then segment the X-rays labeled as

pathological. The first strategy was selected by seven out of the twelve teams, where two teams implemented a hybrid network with a designated classification path after the encoder of their segmentation CNNs. The remaining five teams decomposed the assignment into the classification and segmentation stages. After a segmentation mask was generated for a particular case, empty or almost empty segmentation masks were labeled as no-pneumothorax cases. There is no pattern in the results indicating that any of the two network composition strategies consistently resulted in more accurate results. The performance of a segmentation framework significantly depends on the selected CNN architecture. Among 12 top frameworks, UNet was utilized 10 times, DeepLabV3+ was utilized two times, while EncodingNet, FPNet, PANet, LinkNet, and EMANet were utilized one time each. Note that some participants combined several architectures in a segmentation ensemble. The UNet CNN was utilized by five top-performing frameworks, which suggests that the UNet may be an optimal architecture for pneumothorax segmentation. At the same time, the 19th, 21st, 31st, and 37th teams also used UNets but with less success. The next important property of a segmentation framework is the CNN backbone, i.e. its encoder part. The most commonly employed backbone was the residual neural network (ResNet) incorporated into 12 segmentation CNNs. Eight ResNets were augmented with the squeeze-and-excitation (SE-ResNet) modules that won the last ImageNet Large Scale Visual Recognition Challenge [23]. The remaining backbones were EfficientNet (employed three times), SENet (employed two times), and Dilated-ResNet (employed once). The third important property of a segmentation CNN is the loss function for network training. The pixel-wise binary cross-entropy (BCE) loss function was utilized by all the teams except the 3rd place team. As the BCE loss assigns the same value for both positive and negative samples, i.e. pneumothorax and non-

Table III

The Statistical Comparison of the Deep Learning Framework Performance Against Three Radiologists on a Selected Subset Of Challenging X-Rays. The Positive Value for Confidence Intervals Indicates That the Corresponding Radiologist is More Accurate in Detecting and Segmenting Pneumothorax Than the Framework for a Specified Category of X-Rays. The Difference is Significant for P-Values <0.016 Considering Bonferroni Correction.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Dice coef. difference, 95% confidence interval | p-value, two-tailed test | p-value, one-tailed test |
| Chest X-rays with reference pneumothorax label | | | |
| Radiologist #1 vs CNN framework | [0.022; 0.158] | <0.01 | <0.01 |
| Radiologist #2 vs CNN framework | [0.082; 0.258] | <0.01 | <0.01 |
| Radiologist #3 vs CNN framework | [0.083; 0.237] | <0.01 | <0.01 |
| Chest X-rays with reference no-pneumothorax label | | | |
| Radiologist #1 vs CNN framework | [-0.069; 0.069] | 1 | 0.5 |
| Radiologist #2 vs CNN framework | [0.158; 0.442] | <0.01 | <0.01 |
| Radiologist #3 vs CNN framework | [0.120; 0.430] | <0.01 | <0.01 |
|  | All chest X-rays |  |  |
| Radiologist #1 vs CNN framework | [0.049; 0.251] | <0.01 | <0.01 |
| Radiologist #2 vs CNN framework | [-0.023; 0.190] | 0.12 | 0.06 |
| Radiologist #3 vs CNN framework | [0.049; 0.251] | 0.02 | <0.01 |

pneumothorax X-ray pixels, a neural network trained with only BCE may incline towards correctly labeling predominant background pixels. To compensate for such BCE behavior, eight teams augmented BCE with the Dice or Lovasz loss functions. The Dice loss ignores correctly classified background pixels, and the network receives a positive reward for correctly classified pneumothorax pixels and a negative reward for classifying pneumothorax pixels as background or background pixels as pneumothorax.

Training data augmentation is currently considered an essential step for successful CNN training in the domain of computer-aided diagnosis. The appropriate data augmentation usually results in a several-percent improvement in the network performance, which is observed on various medical image modalities, anatomical regions, and diseases of interest [30]. The presented work and all top-performing frameworks augmented chest X-rays for network training. All the top teams used rigid transformations for data augmentation. A human observer would have no major difficulty analyzing a moderately rotated chest X-ray, whereas convolutional layers in CNNs are not rotation invariant, and unexpected rotations of X-rays may significantly compromise CNN performance. Augmenting X-rays with rotations partially compensates for the convolution layer rotation equivariance. X-ray flip has been commonly adopted as it generates examples of pneumothorax in the left lung field for every case with pneumothorax in the right lung field, and vice versa. Intensity augmentations were used less frequently suggesting that variability of local intensity patterns had been sufficiently covered in the database. The 1st, 4th, and 10th teams demonstrated the benefits of non-rigid deformations of X-rays for data augmentation. Although all top teams used various forms of data augmentation in their solutions, the benefits of data augmentation for pneumothorax detection and segmentation were not quantitatively assessed. We estimated how the framework performance changes when no data augmentation is used during the five-fold cross-validation experiment. The mean framework accuracy computed over all folds dropped from 0.84 to 0.81 measured in terms of Dice coefficient when augmentations were turned off. Summarizing the consistent patterns in the top frameworks, we conclude that the UNet with the SE-ResNet backbone trained using rigid data augmentations is the first configuration of choice for segmentation of pneumothorax in chest X-rays. Figure 7 gives a methodological summary of the top-performing networks.

Fifty top-performing frameworks resulted in Dice coefficient that lied in a narrow interval of [0.84-0.87]. Such accuracy saturation may indicate that the current methodological advances do not allow pushing automated pneumothorax diagnosis from X-rays to a higher accuracy level, or/and that X-rays do not always provide sufficient information for accurate diagnosis in around 10% of challenging cases. The second explanation is in agreement with the database creation report [7], where the authors stated that radiologists from the SIIM and STR disagreed on about 10% of cases. The disagreement could be in the diagnosis, where one group labeled an X-ray with pneumothorax, while the other did not, and segmentation. The

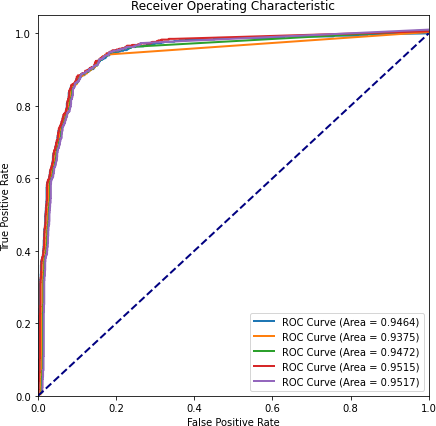


Fig. 6. The framework performance measured in terms of area under the receiver operating characteristic curve computed for pneumothorax pockets against validation chest X-rays. The results are reported for five stratified subsets of X-rays from the training data of the SIIM competition database.

fact that pneumothorax presence is sometimes not evident for expert radiologists indicates the complexity of the problem. We, therefore, cannot expect deep learning to perfectly detect and segment pneumothorax on all X-rays from the database. The observation that the highest achieved Dice coefficient of

0.87 is similar to the agreement level between radiologists from SIIM and STR does not necessarily mean that deep learning and human make similar mistakes. For example, in lung field segmentation, algorithms were able to, on average, surpass the inter-observer agreement. However, this achievement is due to the exceptional quality of automated segmentation on healthy lungs, whereas automated segmentation of pathological lungs is often unsatisfactory and does not reach the human level [20]. To better understand the framework performance and the errors it makes, we qualitatively assessed its results against radiologists using a representative confirmation subset of X-rays. The subset first contained X-rays with large pneumothorax air pockets that were accurately segmented by the CNN. The radiologists correctly detected all such pneumothorax air pockets; however, their segmentation did not perfectly match the reference standard. The observed mean Dice coefficient of 0.8490 is comparable to inter-observer variability for the segmentation of tumors and small poorly-visible organs [31]. Around 8% of 1372 testing X-rays were FN cases, where the framework produced empty segmentations for X-rays labeled as pneumothorax by the competition organizers. Approximately one-third of such X-rays were given to our radiologists for visual inspection. Their mean performance was of 0.1017 Dice, which suggests that the radiologists often agreed with the automated framework and not with the reference labels for FN cases. Such consistent disagreement with reference labels may be explained by a certain imperfection of the reference annotation. The framework

generated a non-empty pneumothorax mask for 13 chest X-rays from 1372 testing X-rays that were labeled as no-pneumothorax by the competition organizers. All of them were analyzed by the radiologists whose performance was of 0.5833 Dice. These results suggest that radiologists struggle to consistently label challenging no-pneumothorax subjects, but their results are still superior to the deep learning framework. It must be, however, noted that FPs correspond to 1% of the X-rays. Finally, our radiologists almost perfectly labeled TN X-rays from the confirmation database. The TN corresponds to more than 78% of the first testing database.

The summarized above results demonstrate a high level of agreement between our radiologists and the radiologists from the SIIM and STR on chest X-rays correctly labeled by the deep learning framework, i.e. TP and TN X-rays. We estimate the prevalence of such easy-to-analyze X-rays to be around 91.5% for the first testing database of the competition. The complexity of the confirmation subset was significantly higher than the complexity of the overall competition. This conclusion was derived from the dramatic gap between the overall framework accuracy of 0.8574 Dice and the framework accuracy of 0.5521 Dice on the confirmation subset. The radiologists’ mean performance on the confirmation subset was 0.6326 Dice indicating moderate superiority of the human against deep learning on challenging X-ray examples. The statistical comparison confirms this superiority with the p-values lower than the significance level 0.016 estimated after Bonferroni correction (Table 2). When analyzing subgroups with cases with positive and negative reference diagnoses, we observe that the difference becomes insignificant between the first radiologist and the framework for pneumothorax cases, and the second radiologist and the framework for no-pneumothorax cases. The low Dice coefficient for some X-rays annotated by human observers indicates that pneumothorax manifestations can be difficult to diagnose from X-rays without non-image data sources. Such data sources may include the clinical picture at the time of patient hospitalization. For example, a patient’s complaint about chest pain in the upper lobe of the left lung can be a decisive factor to correctly classify regions suspicious for pneumothorax at the corresponding location in the chest X-ray. Similarly, a record of traumatic percutaneous penetration into the lungs during a biopsy is a strong predictive feature for the pneumothorax diagnosis.

The diagnosis of pneumothorax is clinically based on both image and non-image information sources. A primary spontaneous pneumothorax is usually accompanied by chest pain and shortness of breath. A secondary pneumothorax occurs due to underlying lung diseases such as chronic obstructive pulmonary disease, tuberculosis, sarcoidosis, etc. A traumatic pneumothorax is the result of injury, stabbing, or a complication after penetrative interventions such as lung tissue biopsy. The clinical picture is often the key factor that swings the balance when information in X-rays is insufficient for a confident diagnosis. The absence of a clinical picture may have been the reason for a 10% disagreement between SIIM and STR radiologists during database creation and disagreement between our radiologists and reference labels. The radiologists were

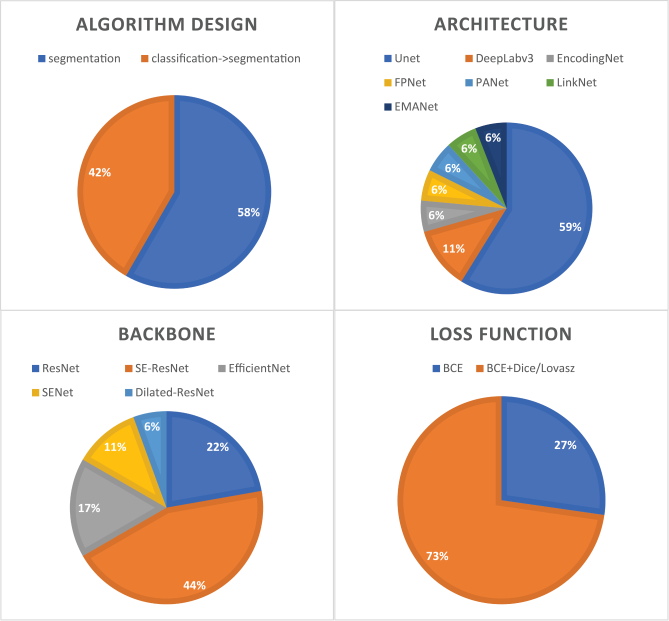


Fig. 7. A summary of top-performing frameworks for pneumothorax detection and segmentation participated in the SIIM completion. All the frameworks utilized deep neural networks. The summary is given in terms of the overall framework design, network architecture, backbone of the segmentation encoder, and loss function. Note that some frameworks combine different network architectures, backbones, and loss functions.

required to label X-rays even if they were not fully certain about the diagnosis. The problem of chest X-ray labeling uncertainty was raised by Stanford researchers, who published a database with 224316 X-rays, where each image did not only have diagnostic labels but also the level of diagnostic confidence. We asked all our radiologists to assess their diagnostic confidence for each X-ray from the confirmation subset. For 51% of X-rays, all three radiologists were confident about their diagnoses. For 33% and 15% of X-rays, two out of three and one out of three radiologists were confident about the diagnosis, respectively. No radiologist was confident about 1% of the confirmation X-rays. The different levels of confidence were reflected in the agreement between the radiologist and reference labels. The mean Dice coefficient was 0.81, 0.50, 0.34, and 0.19 for X-rays when three, two, one, and zero radiologists were confident, respectively. The performance of the framework was also lower with the deteriorating level of radiologist confidence (Pearson correlation coefficient = 0.85, 𝑝-value < 1𝑒−10, Table 1). These observations suggest that there are X-rays that are difficult-to-analyze for both human and deep learning. The diagnosis of patients with such ambiguous X-rays will require the inclusion of non-image information sources. Another interesting observation is that the performance of individual radiologists was almost the same in terms of the Dice coefficient, whereas their confidence ranged from being confident in 64% of diagnoses for the first radiologist to being confident in 94% of diagnoses for the third radiologist.

We asked one radiologist to go through the cases in the confirmation subset and give his clinical opinion about the diagnostic challenges he observed. For this experiment, only

the cases mislabeled by at least one of the three radiologists or/and framework were analyzed. The radiologist had access to the performance of all three radiologists and framework and the reference labels. All three radiologists agreed with the reference labels for 50% of cases, partially agreed for 38% of cases, and disagreed with the reference diagnosis for 12% of cases. Assuming that the reference diagnosis is always correct, the last radiologist considered 34% of the challenging pathological cases he inspected to represent uncommon manifestations of pneumothorax. For around 41% of the cases, the radiologist indicated that clavicles, ribs, and sometimes medical devices significantly obstructed the view so pneumothorax diagnosis became difficult. For some of such patients, the radiologist advised acquiring a lateral chest X-ray or a computed tomography (CT) image. He, however, acknowledged that the acquisition of additional images is not always possible if a patient is in an intensive care unit and cannot be transported. Intensive care patients are usually imaged in the horizontal positioning using a portable X-ray machine, which produces images of a lower quality than a stationary X-ray machine. Although it has not been specified explicitly by the competition organizers, the radiologist suggested that around 39% of X-rays from the confirmation subset were acquired using a portable machine. If a patient is in a vertical position, the pneumothorax pockets are most likely to be found in the upper parts of the lung, whereas the pneumothorax location for patients in a horizontal position is less predictable. The radiologist suggested that it may be beneficial to train neural networks to first recognize the positioning of a patient and then focus on pneumothorax detection and segmentation. Moreover, the horizontal patient positioning is an additional indicator of life-threatening conditions. The radiologist confirmed that some pneumothorax cases may be challenging to automatically diagnose due to the projective nature of X-ray imaging. The X-rays, where pneumothorax is obscured with bones, could be preprocessed with algorithms for bone suppression to improve pneumothorax visibility. A few algorithms of rib/clavicle suppression have been proposed but their performance remains insufficient and they sometimes suppress not only bones but also lung abnormalities [32]. For around 13% of X-rays, pneumothorax could have been confused with lung bullae. It has been acknowledged in the literature that distinguishing between pneumothorax and bullous lung diseases is sometimes non-trivial and can be achieved using dynamic ultrasound for detecting the lung sliding during inspiration [33] or CT for detecting the double-wall sign [34]. The radiologist stressed out that the most critical information for accurate diagnosis in the presence of uncertainty on X-rays comes from non-image data sources. Sudden shortness of breath, chest tightness, rapid heart rate, painful breathing, etc. could indicate a spontaneous pneumothorax. Laboratory tests for the detection of respiratory alkalosis and low level of blood oxygen would further confirm the pneumothorax. Finally, if a patient was in an intensive care unit, his disease history would guide the diagnosis and help to recognize acute pneumothorax cases. The experiments with radiologists give a better understanding of the performance of deep learning frameworks and highlight the directions for

future research. Easy-to-diagnose X-rays for deep learning turn to be easy-to-diagnose X-rays for human experts. This conclusion suggests that the framework is able to discover the appearance patterns associated with pneumothorax. At the same time, the framework performs slightly worse than humans, which indicates the rooms for further algorithmic improvements. It is possible that performance improvement will be achieved when chest X-rays will be first classified into images acquired in vertical and horizontal positioning. Significant performance improvement is likely to be achieved from the inclusion of non-image clinical factors. The radiologists consider such factors to be decisive for X-rays with uncertainties. The development of a multi-path neural network for the simultaneous analysis of image and non-image information sources is the direction for future research.

1. Conclusion

In this paper, we presented our deep learning framework for the detection and segmentation of pneumothorax on chest X-rays and its evaluation against radiologists. The presented model participated in a public competition and was ranked among the top-performing frameworks. We observed that deep learning is capable of achieving a relatively high diagnosis accuracy, is very much in agreement with human diagnostic performance, but still inferior to experienced radiologists in difficult-to-analyze cases.

ACKNOWLEDGMENT

This work was supported by the Russian Science Foundation (grant no. 18-71-10072). We thank two radiologists from the Public Hospital #2, Department of Radiology, Kazan for annotating the confirmation chest X-ray database.

References

1. L. Thomsen, O. Natho, U. Feigen, U. Schulz, and D. Kivelitz, “Value of Digital Radiography in Expiration in Detection of Pneumothorax,” *RöFo*

*- Fortschritte Auf Dem Geb. Röntgenstrahlen Bildgeb. Verfahr.*, vol. 186, no. 03, pp. 267–273, Mar. 2014, doi: 10.1055/s-0033-1350566.

1. S. Sanada, K. Doi, and H. MacMahon, “Image feature analysis and computer-aided diagnosis in digital radiography: automated detection of pneumothorax in chest images,” *Med. Phys.*, vol. 19, no. 5, pp. 1153–1160, Oct. 1992, doi: 10.1118/1.596790.
2. O. Geva, G. Zimmerman-Moreno, S. Lieberman, E. Konen, and H. Greenspan, “Pneumothorax detection in chest radiographs using local and global texture signatures,” in *Medical Imaging 2015: Computer-Aided Diagnosis*, Mar. 2015, vol. 9414, p. 94141P, doi: 10.1117/12.2083128.
3. Y.-H. Chan, Y.-Z. Zeng, H.-C. Wu, M.-C. Wu, and H.-M. Sun, “Effective Pneumothorax Detection for Chest X-Ray Images Using Local Binary Pattern and Support Vector Machine,” *J. Healthc. Eng.*, vol. 2018, p. 2908517, 2018, doi: 10.1155/2018/2908517.
4. A. Gooßen *et al.*, “Deep Learning for Pneumothorax Detection and Localization in Chest Radiographs,” *ArXiv*, 2019.
5. G. Kitamura and C. Deible, “Retraining an open-source pneumothorax detecting machine learning algorithm for improved performance to medical images,” *Clin. Imaging*, vol. 61, pp. 15–19, May 2020, doi: 10.1016/j.clinimag.2020.01.008.
6. R. W. Filice *et al.*, “Crowdsourcing pneumothorax annotations using machine learning annotations on the NIH chest X-ray dataset,” *J. Digit. Imaging*, Nov. 2019, doi: 10.1007/s10278-019-00299-9.
7. M. Annarumma, S. J. Withey, R. J. Bakewell, E. Pesce, V. Goh, and G. Montana, “Automated Triaging of Adult Chest Radiographs with Deep Artificial Neural Networks,” *Radiology*, vol. 291, no. 1, pp. 196–202, Jan. 2019, doi: 10.1148/radiol.2018180921.
8. S. Park *et al.*, “Application of deep learning-based computer-aided detection system: detecting pneumothorax on chest radiograph after biopsy,” *Eur. Radiol.*, vol. 29, no. 10, pp. 5341–5348, Oct. 2019, doi: 10.1007/s00330-019-06130-x.
9. E. J. Hwang *et al.*, “Development and Validation of a Deep Learning-Based Automated Detection Algorithm for Major Thoracic Diseases on Chest Radiographs,” *JAMA Netw. Open*, vol. 2, no. 3, p. e191095, 01 2019, doi: 10.1001/jamanetworkopen.2019.1095.
10. A. G. Taylor, C. Mielke, and J. Mongan, “Automated detection of moderate and large pneumothorax on frontal chest X-rays using deep convolutional neural networks: A retrospective study,” *PLOS Med.*, vol. 15, no. 11, p. e1002697, Nov. 2018, doi: 10.1371/journal.pmed.1002697.
11. M. Cicero *et al.*, “Training and Validating a Deep Convolutional Neural Network for Computer-Aided Detection and Classification of Abnormalities on Frontal Chest Radiographs,” *Invest. Radiol.*, vol. 52, no. 5, pp. 281–287, 2017, doi: 10.1097/RLI.0000000000000341.
12. P. Rajpurkar *et al.*, “Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists,” *PLoS Med.*, vol. 15, no. 11, p. e1002686, 2018, doi: 10.1371/journal.pmed.1002686.
13. X. Ouyang *et al.*, “Weakly Supervised Segmentation Framework with Uncertainty: A Study on Pneumothorax Segmentation in Chest X-ray,” in *Medical Image Computing and Computer Assisted Intervention –MICCAI 2019*, Cham, 2019, pp. 613–621, doi: 10.1007/978-3-030-32226-7\_68.
14. “SIIM-ACR Pneumothorax Segmentation.” https://kaggle.com/c/siim-acr-pneumothorax-segmentation (accessed Mar. 18, 2020).
15. O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” May 2015, Accessed: Oct. 29, 2016. [Online]. Available: [http://arxiv.org/abs/1505.04597.](http://arxiv.org/abs/1505.04597)
16. L. Oakden-Rayner, “Exploring Large-scale Public Medical Image Datasets,” *Acad. Radiol.*, vol. 27, no. 1, pp. 106–112, Jan. 2020, doi: 10.1016/j.acra.2019.10.006.
17. A. Chaurasia and E. Culurciello, “LinkNet: Exploiting Encoder Representations for Efficient Semantic Segmentation,” *2017 IEEE Vis. Commun. Image Process. VCIP*, pp. 1–4, Dec. 2017, doi: 10.1109/VCIP.2017.8305148.
18. A. A. Novikov, D. Lenis, D. Major, J. Hladůvka, M. Wimmer, and K. Bühler, “Fully Convolutional Architectures for Multiclass Segmentation in Chest Radiographs,” *IEEE Trans. Med. Imaging*, vol. 37, no. 8, pp. 1865–1876, Aug. 2018, doi: 10.1109/TMI.2018.2806086.
19. M. Kholiavchenko *et al.*, “Contour-aware multi-label chest X-ray organ segmentation,” *Int. J. Comput. Assist. Radiol. Surg.*, vol. 15, no. 3, pp. 425–436, Mar. 2020, doi: 10.1007/s11548-019-02115-9.
20. K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.
21. G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely Connected Convolutional Networks,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 2261–2269, doi: 10.1109/CVPR.2017.243.
22. J. Hu, L. Shen, and G. Sun, “Squeeze-and-Excitation Networks,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 7132–7141, doi: 10.1109/CVPR.2018.00745.
23. T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature Pyramid Networks for Object Detection,” *ArXiv161203144 Cs*, Apr. 2017, Accessed: Apr. 28, 2020. [Online]. Available: [http://arxiv.org/abs/1612.03144.](http://arxiv.org/abs/1612.03144)
24. I. M. Baltruschat, H. Nickisch, M. Grass, T. Knopp, and A. Saalbach, “Comparison of Deep Learning Approaches for Multi-Label Chest X-Ray Classification,” *Sci. Rep.*, vol. 9, no. 1, Art. no. 1, Apr. 2019, doi: 10.1038/s41598-019-42294-8.
25. M. Berman, A. Rannen Triki, and M. B. Blaschko, “The Lovász-Softmax Loss: A Tractable Surrogate for the Optimization of the Intersection-Over-Union Measure in Neural Networks,” 2018, pp. 4413–4421, Accessed: Apr. 27, 2020. [Online]. Available: <http://openaccess.thecvf.com/content_cvpr_2018/html/Berman_The_Lo> vaSz-Softmax\_Loss\_CVPR\_2018\_paper.html.
26. K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” *ArXiv14091556 Cs*, Apr. 2015, Accessed: Apr. 28, 2020. [Online]. Available: [http://arxiv.org/abs/1409.1556.](http://arxiv.org/abs/1409.1556)
27. G. Huang, Y. Li, G. Pleiss, Z. Liu, J. E. Hopcroft, and K. Q. Weinberger, “Snapshot Ensembles: Train 1, get M for free,” *ArXiv170400109 Cs*, Mar.

2017, Accessed: Apr. 27, 2020. [Online]. Available: [http://arxiv.org/abs/1704.00109.](http://arxiv.org/abs/1704.00109)

1. X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, “ChestX-ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases,” 2017, pp. 2097–2106, Accessed: Mar. 24, 2020. [Online]. Available: <http://openaccess.thecvf.com/content_cvpr_2017/html/Wang_ChestX->ray8\_Hospital-Scale\_Chest\_CVPR\_2017\_paper.html.
2. F. Isensee, J. Petersen, S. A. A. Kohl, P. F. Jäger, and K. H. Maier-Hein, “nnU-Net: Breaking the Spell on Successful Medical Image Segmentation,” *ArXiv190408128 Cs*, Apr. 2019, Accessed: Mar. 25, 2020. [Online]. Available: [http://arxiv.org/abs/1904.08128.](http://arxiv.org/abs/1904.08128)
3. S. K. Vinod, M. G. Jameson, M. Min, and L. C. Holloway, “Uncertainties in volume delineation in radiation oncology: A systematic review and recommendations for future studies,” *Radiother. Oncol. J. Eur. Soc. Ther. Radiol. Oncol.*, vol. 121, no. 2, pp. 169–179, 2016, doi: 10.1016/j.radonc.2016.09.009.
4. M. Gusarev, R. Kuleev, A. Khan, A. Ramirez Rivera, and A. M. Khattak, “Deep learning models for bone suppression in chest radiographs,” in *2017 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, Aug. 2017, pp. 1–7, doi: 10.1109/CIBCB.2017.8058543.
5. C. Gelabert and M. Nelson, “Bleb Point: Mimicker of Pneumothorax in Bullous Lung Disease,” *West. J. Emerg. Med.*, vol. 16, no. 3, pp. 447–449, May 2015, doi: 10.5811/westjem.2015.3.24809.
6. B. Aramini, C. Ruggiero, A. Stefani, and U. Morandi, “Giant bulla or pneumothorax: How to distinguish,” *Int. J. Surg. Case Rep.*, vol. 62, pp. 21–23, Aug. 2019, doi: 10.1016/j.ijscr.2019.08.003.