Computers and Electrical Engineering 78 (2019) 388–399

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

Deep neural network ensemble for pneumonia localization from a large-scale chest x-ray database ☆

Ilyas Sirazitdinov [a], Maksym Kholiavchenko [a], Tamerlan Mustafaev [b], Yuan Yixuan [c], Ramil Kuleev [a], Bulat Ibragimov [a,d,*]

[a] Innopolis University, Innopolis city, Russia
[b] Public Hospital #2, Department of Radiology, Kazan, Russia
[c] Department of Electronic Engineering, City University of Hong Kong, Hong Kong
[d] Stanford University, Department of Radiation Oncology, Palo Alto, USA

ARTICLE INFO

Article history:
Received 13 December 2018
Revised 17 June 2019
Accepted 5 August 2019
Available online 10 August 2019

Keywords:
Deep learning
Convolutional neural networks
Pneumonia detection
Chest x-ray
CXR, Pneumonia localization
Lung opacity detection

ABSTRACT

Pneumonia is a bacterial, viral, or fungal infection of one or both sides of the lungs that causes lung alveoli to fill up with fluid or pus, which is usually diagnosed with chest x-rays. This work investigates opportunities for applying machine learning solutions for automated detection and localization of pneumonia on chest x-ray images. We propose an ensemble of two convolutional neural networks, namely RetinaNet and Mask R-CNN for pneumonia detection and localization. We validated our solution on a recently released dataset of 26,684 images from Kaggle Pneumonia Detection Challenge and were score among the top 3% of submitted solutions. With 0.793 recall, we developed a reliable solution for automated pneumonia diagnosis and validated it on the largest clinical database publicity available to date. Some of the challenging cases were additionally examined by a team of physicians, who helped us to interpret the obtained results and confirm their practical applicability.

1. Introduction

Pneumonia is a common lung infection caused by bacteria, a virus or fungi. It kills thousands of people every year and the death rate from pneumonia had almost no improvement since antibiotics became widespread more than half a century ago [1].

Pneumonia usually appears as an area or areas of increased opacity [2] on chest x-ray images. However, diagnosing pneumonia is not a straightforward task since many patients come into the intensive care unit with pre-existing lung problems. Doctors try to decide if there is a change in the chest x-ray (CXR) image associated with pneumonia manifestation or other pre-existing condition. Moreover, patients often have multiple health problems such as pulmonary edema, bleeding, atelectasis, lung cancer or surgical interventions, which makes diagnosis even more complicated. The goal of this research is to facilitate diagnosing pneumonia by highlighting areas on chest x-ray image that look most like pneumonia thereby doctors can pay more attention to these areas.

☆ This paper is for CAEE special section SI-mip. Reviews processed and recommended for publication to the Editor-in-Chief by Guest Editor Dr. Li He.

* Corresponding author at: Innopolis University, Innopolis city, Russia.
E-mail address: b.ibragimov@innopolis.ru (B. Ibragimov).

https://doi.org/10.1016/j.compeleceng.2019.08.004
0045-7906/

Statistical analysis, machine learning and deep learning are remarkably powerful tools in computer-aided diagnosis (CAD) systems. They can be applied to tackle complex computer vision problems in the medical imaging domain, e.g., lungs segmentation [3,4], lungs pathologies classification [8] and etc. Recent advances in deep learning allowed to achieve human-level performance on a wide range of tasks and even slightly surpass it [6]. As next stage research, deep learning can also be applied to predict the results of treatment, e.g., cancer treatment [7]. Promising results in thorax diseases classification using a CXR modality are connected with large labeled datasets [5,6] and the deep learning based methods. Wang et al. presented a large labeled CXR dataset, proposed a baseline deep learning approach and used a class activation maps method for the coarse pathology localization [5]. P. Rajpurkar et al. [9] applied a deep densely connected convolutional neural network to predict one of fourteen possible pathologies using the same dataset. R. Abiyev compared supervised backpropagation neural networks with unsupervised competitive neural networks for diagnosing chest diseases [8]. However, all these works addressed multi-label image classification problem and used trained neural networks as an auxiliary tool for the coarse pathologies localization. In our work, we exclusively target pneumonia detection and localization problem and utilized the novel largest labeled pneumonia dataset (1353 in Chest14-xray [5] vs. 8964 in RSNA database [9]). The main contribution of the work are: 1) deep learning based predictions of pneumonia regions, 2) development of ensembling method that is superior to the existing approaches, 3) transparent results validation on the new largest publicly available pneumonia labeled dataset.

Doctors require a significant amount of time and proper qualifications in order to make a diagnosis. In this work, we present a model that is able to spot pneumonia on chest x-ray images and potentially be a great help for doctors. Our approach was ranked among the 3% of best-performing solutions on a public international competition RSNA Pneumonia Detection Challenge [9]. Given the high precision of our approach, it still remains fast and easy to deploy on most computing systems.

This paper is organized as follows. Section 2 presents a detailed review of the literature of the existing object detection methods. Section 3 presents the CXR database and the separation of the images into testing and training cases. Section 4 describes the details of the proposed pneumonia location method. Sections 5 and 6 summarizes the results of the application of the proposed methodology and the conclusions made from these results.

2. Background review

Object detection is a technology related to computer vision and image processing that deals with detecting instances of semantic objects of a certain class in digital images and videos. In the last decade, two main categories of methods had been used in classical visual recognition competitions (PASCAL VOC 2010 [10] and Microsoft COCO [11]). The first category of methods is based on handcrafted features, e.g., Viola-Jones algorithm based on Haar features, scale invariant feature transform (SIFT), a histogram of oriented gradients features (HOG) [12] usually classified by support vector machine (SVM). The main drawback of this category of methods is that handcrafted feature set cannot efficiently represent all images of interest and therefore cannot provide sufficient detection performance. The second category of methods overcomes the limitation of handcrafted features, and allow automated generation of suitable image features through a hierarchy of linear and non-linear operations – deep learning approach. Deep learning-based object detection was mainly driven by OverFeat [13] and region-based convolutional neural network (R-CNN) [14] models. R-CNN model combined ideas of handcrafted features methods to generate region proposals (selective search algorithm) with a convolutional neural network (CNN) which was used as a feature extractor. Obtained feature vector was classified with SVM with following greedy non-maximum suppression, which eliminates overlapping bounding box predictions. The R-CNN model formed a family of two-staged detectors: the first stage includes searching for regions containing target objects, the second stage includes object classification and possible refinement of bounding boxes of the proposed regions. Further evolution of the model was Fast R-CNN [20] that proposed region of interest (RoI) pooling layer that allowed to train the model almost end-to-end using selective search predictions to form RoI pooling layer. The RoI pooling also allowed to add a new bounding box regression branch which was responsible for adjusting regions proposals obtained from the selective search. The next iteration step Faster R-CNN [15] model introduced Region Proposal Network (RPN). RPN is an independent branch built on CNN's features to predict regions containing objects. This modification allowed to perform region of interest detection in a supervised manner since RPN is a trainable part of the network in contrast to selective search. The RPN works in a sliding window manner and predicts k regions for one sliding position and k anchor boxes. Anchor boxes are being used to represent a wide range of scales and aspect ratios of possible objects. A slight modification of Faster R-CNN (adding a new mask branch and transforming RoI pooling to RoI align layer) called Mask R-CNN allowed improving performance gain in a similar instance segmentation task [16]. The last improvements in object detection task are closely related to feature pyramid networks (FPN) [17]. The general architecture is presented in Fig. 1. The main idea of the network is independent multiple level predictions which help to capture objects of different size and scale. A similar idea of skip connections was successfully used in medical image segmentation task, e.g., in U-net model [18]. However, while U-net uses a concatenation of feature maps, FPN uses 1×1 convolution to equalize dimensionality of corresponding level feature maps and further summation, which made it more similar to residual blocks.

One stage detectors is a family of detectors which does not perform region proposal stage. They are applied over the dense sampling of object locations. Although this approach was sufficiently faster than the two-stage, the widespread models (YOLO [19]) had accuracy 10–40% lower than state-of-the-art two-stage methods. T.Lin et al. [20] identified a class imbalance problem in training one-stage detectors and proposed a new loss function (focal loss – formula 2) which is dynamically

$$CrossEntropy(p_t) = -log(p_t)(1)$$
$$FocalLoss(p_t) = -(1-p_t)^{\gamma} log(p_t)\,(2)$$

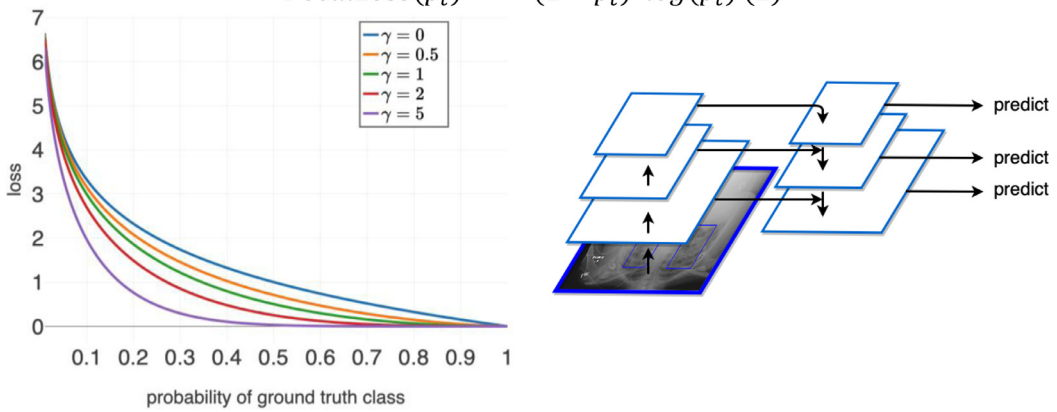

Fig. 1. Focal loss and FPN architecture utilized in the proposed models for pneumonia prediction.

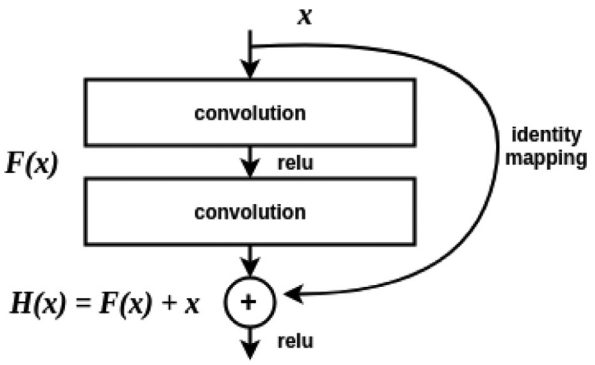

Fig. 2. The scheme of a residual block.

scaled cross-entropy loss:

$$CrossEntropy(p_t) = -log(p_t) \quad (1)$$

$$FocalLoss(p_t) = -(1-p_t)^{\gamma} log(p_t) \quad (2)$$

As it can be noticed from Fig. 1 focal loss and cross-entropy functions are different in $(1-p_t)^{\gamma}$ factor. This factor allows decreasing the relative loss for well-classified examples and giving more loss for hard examples.

The resulted model named RetinaNet is a one-stage detector that leverages FPN feature extractor with advanced ResNet-101 or ResNet-50 backbones and uses the focal loss to tackle the object – background imbalance problem. The model achieved top results, outperforming both one-staged and two-staged models in the COCO competition [20]. Detection of masses in mammograms using the RetinaNet was performed in [21], where the model achieved true positive rate similar to the state-of-the-art mass detection models. FPN (Fig. 1) is the encoder-decoder CNN with lateral connections. Each encoder block includes consecutive convolutional residual blocks and the max pooling block. Convolution is a filtering operation on two functions which in case of computer vision are image pixel values. Two-dimensional discretized convolution operation on image Im(p, q) with filter K(p, q) can be written as:

$$f(x) = \sum_{p}\sum_{q} Im(p,q) \times K(m - p,\ n - q) \quad (3)$$

The output of a convolution layer passes through a non-linearity function in order to avoid convergence of the neural network to a single layer perceptron model. Usually, the choice of non-linear functions is limited by the pool of functions such as hyperbolic tangent function, sigmoid, rectified linear unit function (ReLu) and its modifications. ReLu and its modification are considered to be more preferable since they provide better gradient pass and are computationally efficient. Feature maps obtained in convolutional layers normally is being reduced in dimension by max pooling or other pooling techniques.

Residual block (Fig. 2) consists of a chain of convolutional layers and uses identity mapping that adds the input of the block to the output: $H(x) = F(x) + x$, where x is input of the block, $F(x)$ is the mapping of stacked layers, $H(x)$ is the

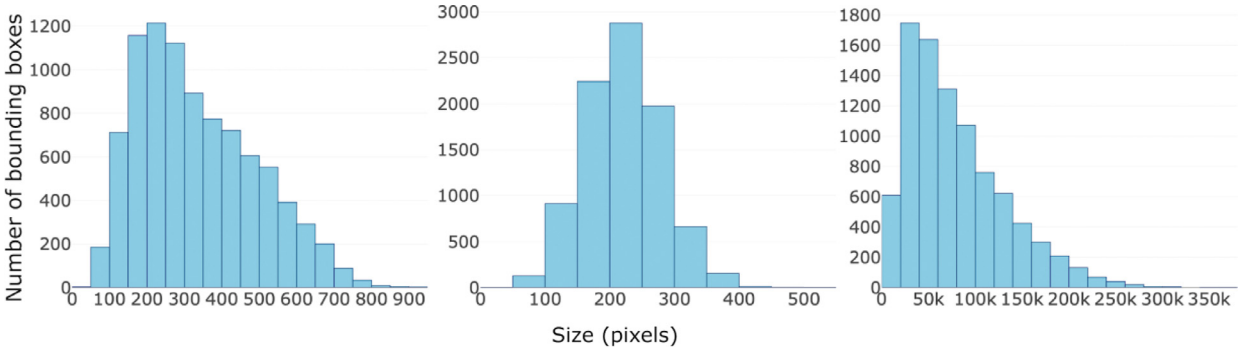

Fig. 3. Left to right: height, width, and a square of bounding boxes of the training set.

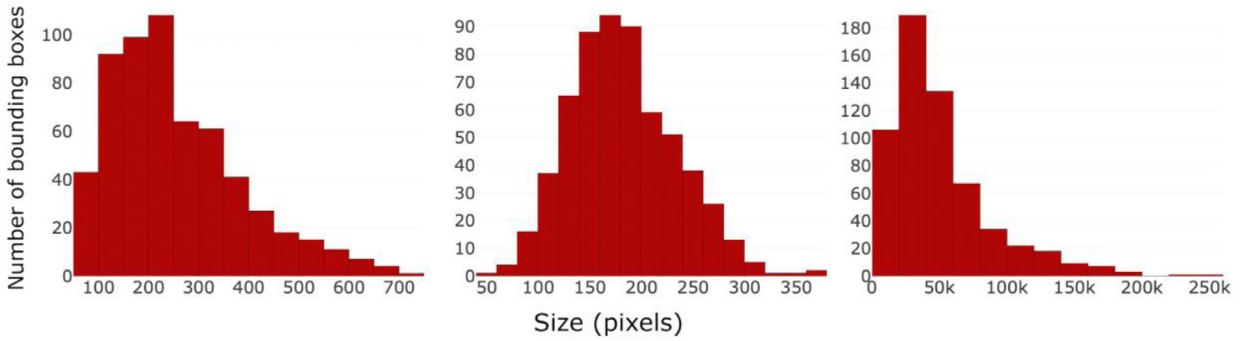

Fig. 4. Left to right: height, width, and a square of bounding boxes of the test set.

desired mapping of the whole block. The residual block allowed to provide a better gradient flow than in ordinarily stacked convolution layers, since the addition operation works as gradient splitter during the backward pass. As a result, it allowed to address a vanishing gradients problem and to create deeper networks. The residual blocks can augment most of the convolutional network architectures, and often used when the original architecture gets saturated.

A decoding block consists of the upsampling procedure with further encoder-decoder feature maps merging via lateral connections. Each decoder layer produces feature maps for the parallel classification and regression subnetworks that predict the probability of object presence for each predefined location (anchor) and the offset from each anchor box to the nearest ground truth object.

In medical imaging object detection solves a localization problem, e.g., pathologies localization, nucleus detection, outliers detection, and etc. Z. Xue et al. [22] used Circle Hough Transform and Viola-Jones algorithm to detect buttons on chest x-ray images. The Faster R-CNN-based approach was applied for the detection of glomeruli in multi-stained whole slide images (WSIs) of human renal tissue sections [23]. The annual Kaggle Data Science Bowl 2018 proposed a task of automating nucleus detection which was successfully solved using complex segmentation pipeline, as well as via instance segmentation neural networks.

3. Data and feature descriptions

In our work, we used publicly available RSNA Pneumonia Detection Challenge dataset [9] which consists of 26,684 unique CXRs. There are three classes of labels: normal (29%), no lung opacity / not normal (40%), lung opacity (31%). All lung opacity images were manually labeled with rectangular bounding boxes that encompass pneumonia region. The images were separated into training (25,684) and testing (1000) parts during the first stage of the competition and training (2684) and testing (3000) during the second stage. In our experiments, the training images were divided into actual training (90%) and validation (10%) parts.

4. Methodology

In this work, we propose an innovative approach based on an ensemble of RetinaNet and Mask R-CNN. Pneumonia regions are different in size and ratio (Figs. 3 and 4) with median height equals to 304 pixels (29.6% of image height) and median width equals to 219 pixels (21.3% of image width). Pneumonia is manifested on a relatively small region of a chest x-ray, which represents a challenge for modern object detectors. To tackle this problem, we used the FPN principle in the backbone of both models, since FPN generates multi-scale feature maps with better quality information than the default

Table 1
Statistics of the size of pneumonia regions presented as the reference standard in the RSNA database of chest x-rays.

	Train set	Test set
Median width (px)	219	178
Median height (px)	304	225
Median square (px)	$67{,}1 \times 10^3$	$40{,}4 \times 10^3$

Table 2
The configuration of RetinaNet and Mask R-CNN networks that were ensembled for automated pneumonia detection and localization.

	RetinaNet	Mask R-CNN
Input size	512 × 512	512 × 512
Backbone model	ResNet-50	ResNet-101
Batch size	8	6
Training data	All	Pneumonia only
Loss (sum)	Focal loss, bounding box regression loss	RPN regression loss, RPN binary cross-entropy(bce), bounding box regression loss, bounding box bce
	RetinaNet	Mask R-CNN
Optimizer	Adam	Adam
Learning rate	0.0001	0.001
Learning rate scheduling	Reduce on plato	Reduce on plato
Epoch length	25% of the training data	25% of the training data

feature pyramid [17]. As it can be noticed from Fig. 1, the FPN architecture combines low-resolution, i.e. semantically strong features, with high-resolution, i.e. semantically weak features, via a top-down pathway and lateral connections.

As a base backbone model, we used residual networks because they reduced the impact of degradation problem and allowed to create deeper models in contrast to ordinarily stacked convolutional layers.

4.1. Training

Transfer learning from the models trained on a Microsoft COCO challenge was used as a weights initialization strategy. Transfer learning is a machine learning methodology that assumes usage of a model trained on one task, e.g., recognition of objects from natural images for another similar task, for which only a limited number of training samples are available. Although the COCO challenge images and labeled objects are different from the chest x-rays and labeled pneumonia regions, the transfer learning is expected to improve pneumonia detection, considering the fact that a similar type of weight initialization on the ImageNet dataset improved the chest x-ray classification problem [6]. The following set of data augmentations was used during training: vertical and horizontal flip, random degree rotation, random brightness, gamma transforms, random Gaussian noise and blur. Such augmentation can enrich the network training procedure and serve as an additional regularization and generalization strategy. Table 2 summarizes the network settings.

4.2. Adjusting networks for pneumonia prediction

Post-processing of object detectors outputs is a necessary step to eliminate overlapping bounding boxes obtained during the prediction stage. One of the common ways is a non-maximum suppression (NMS) algorithm. The NMS algorithm first picks the bounding box with the highest confidence among the same class predictions, next calculates the intersection over union (IoU) (4) with the remaining boxes, and finally suppresses predictions with IoU lower than a predefined suppression threshold. While high values of NMS threshold result in producing more overlapping predictions, close-to-zero threshold values result in non-overlapping bounding boxes:

$$IoU(A, B) = \frac{A \cap B}{A \cup B} \tag{4}$$

The second tunable parameter is the bounding box confidence that is the result of the sigmoid activation function. High values of the bounding box confidence result in a high precision score, whereas low values result in high recall score. In our work, we performed a grid search of the NMS suppression threshold and bounding box confidence level w.r.t. the mean average precision (mAP) metrics (5). mAP is a common metrics in object detection challenges, and it is usually used as the primary detection benchmark [10,11]. Thresholds is a set of values from 0.4 to 0.75 with a 0.05 step. A prediction and a ground truth label are considered to be true positive (TP) if the IoU between them is more or equal than the current value of a threshold, false positive (FP) if the IoU is less than a current threshold. We count a ground truth bounding box as false negative (FN) if none of IoU values is higher than the current threshold. mAP metrics do not count true negative (TN) image

Table 3
Individual model performance during the first stage of the competition.

	YOLO v1.	YOLO v3.	RetinaNet	Mask R-CNN
mAP	0.112	0.138	0.192	0.169

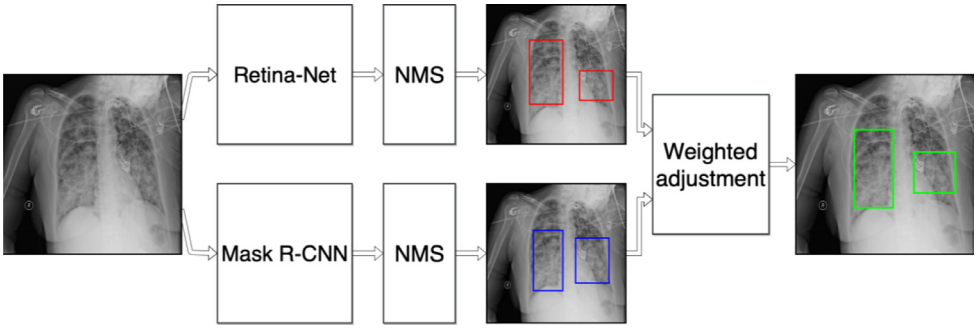

Fig. 5. A schematic illustration of ensembling of RetinaNet and Mask R-CNN.

predictions, because a non-pneumonia sample with zero predictions do not participate in averaging over all images:

$$mAP = \frac{1}{|images|}\sum_{img}\frac{1}{|thresholds|}\sum_{t}\frac{TP(t)}{TP(t) + FP(t) + FN(t)} \tag{5}$$

The grid search algorithm showed various optimal thresholds regarding the number of pneumonia and non-pneumonia samples on the validation set. To overcome the ambiguity, we performed the grid search algorithm multiple times on the subsamples of the validation set with different positive-negative ratios and further thresholds averaging.

4.3. Models ensembling

Supervised learning ensemble methods are based on the idea of using multiple learning algorithms in order to obtain better performance. While network ensembling for classification and segmentation tasks can be formulated using the weighted voting scheme, ensembling of object detection predictions is not straightforward. Indeed, an object detector produces many overlapping predictions with different confidence levels that can be merged with other object detector predictions. In the worst case scenario, if n and m are numbers of outputs of two object detection networks, the number of pairwise IoU calculations for simply merged predictions is $(n+m)^2$, while for individually merged predictions is (n^2+m^2). For instance, the number of RetinaNet outputs ($\sim 10^5$) leads to additional 2×10^{10} comparisons for the simply merged predictions. Another obstacle is the combining policy – ensemble participants usually have different confidence level scales that cannot be straightforwardly combined. In our work, we picked a «greedy» strategy of models ensembling to combat these problems. The first step of the strategy was to obtain the optimal individual models by selecting the optimal NMS and the confidence level threshold (Section 4.2.) that allowed to eliminate overlapping predictions and to achieve the highest mAP score for each model. To address the problem of confidence level scaling, all predictions were considered as positive if they have higher values than the classification threshold. We assumed all positive predictions to have equal weights during the bounding boxes merging step. It allowed performing confidence level scaling via grid search for optimal weighting strategy.

During the first stage of the competition, we examined four different object detection models to select the ones with the highest performance for the final ensembling. We selected three one-stage detectors (YOLO v1 [19], YOLO v3 [24], RetinaNet) since they were designed according to potentially complementary methodological principles: YOLO v1. uses ordinarily stacked convolutional layers, YOLO v3. is based on FPN, while RetinaNet is based on FPN and utilizes the focal loss. We also examined Mask R-CNN model that combined numerous advances of the two-stage detection: FPN, RPN, and RoI align. Table 3 summarizes the results of individual models. As it can be noticed from Table 3, RetinaNet and Mask R-CNN models showed better performance in comparison to both YOLO models. Therefore, RetinaNet and Mask R-CNN were picked for building an ensemble model.

In our work, we used RetinaNet as a primary more accurate model and Mask R-CNN as an auxiliary model for adjusting pneumonia regions (Fig. 5). The ensembling was performed in the following way: first, both models predicted pneumonia regions, then NMS with optimal thresholds was applied in the predicted regions. If some RetinaNet prediction overlapped some Mask R-CNN prediction with IoU $>$ 0.5, they were averaged with RetinaNet – Mask R-CNN weights ratio 3:1. The weight ratio was picked after the iterative grid search of weights ratio: 4:1, 3:1, 2:1 and vice versa on the validation set. If a predicted bounding box did not have a corresponding overlapping prediction from a different model, this bounding box was used in the ensemble model prediction without any changes.

Table 4
Averaged validation metrics of single models and ensemble model.

	Precision	Recall	F1-score
RetinaNet	0.293	0.267	0.279
Mask-RCNN	0.265	0.254	0.259
Ensemble model	0.288	0.284	0.286

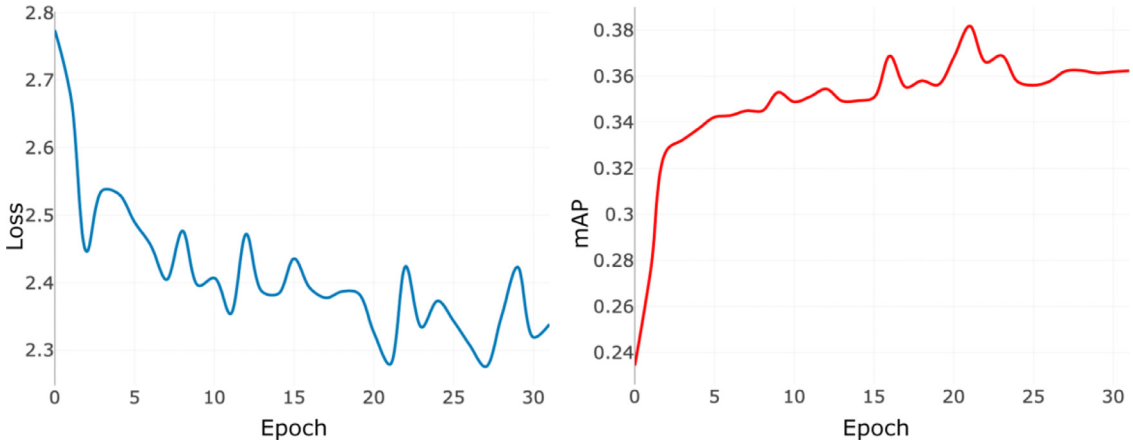

Fig. 6. RetinaNet learning curves. Blue line – training loss, red line – validation mAP. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4 shows precision, recall, and f1-score of the individual models and the ensemble model calculated on the validation set during the first stage of the competition; TP, FP, FN were calculated according to Eq. (5). As it can be noticed from Table 4, the ensemble model showed the highest recall that was higher than the results of individual models. As a result, the combined model had the highest f-1 score and showed relatively unbiased (w.r.t. specificity and sensitivity) predictions with almost equal precision and recall scores.

5. Results

RSNA Pneumonia Detection Challenge was announced in August 2018 and lasted for two months. A target metric of the competition was mAP (4). During the first stage of the competition, 25,684 images were provided for training and 1000 for testing. The second stage included the announcement of the previous stage testing labels and a new testing set of 3000 images. In our work, we used the official train-test splits in order to conduct transparent, comparable and repeatable experiments. The dataset was provided in DICOM data format and consisted of images in 1024 $\times$ 1024 resolution in 8-bit format and metadata, that was partially anonymized. The metadata includes the patient's gender, age and a view position which is either anterior-posterior (AP) or posterior-anterior (PA).

Our final model used RetinaNet and Mask R-CNN models implemented on a Keras framework. The training was performed on the Nvidia Tesla V100 for the RetinaNet (8 h) and Nvidia K80 for the Mask R-CNN (6 h).

Both stages results are presented in Table 5. The combined model showed the top 3% score during the second stage. Table 5 also presents a confusion matrix w.r.t. testing images. The TP result corresponds to a case where ground truth shows a presence of pneumonia and prediction result has at least one pneumonia region, the overlap between ground truth and predicted regions is not taken into account. The TN result corresponds to a case where ground truth either normal or non-pneumonia pathology and the prediction result indicates the absence of pneumonia. The FP result corresponds to a case where ground truth either normal or non-pneumonia pathology and the prediction result has at least one pneumonia region. The FN result corresponds to a case where ground truth shows the presence of pneumonia while the prediction result indicates the absence of pneumonia. The confusion matrix evaluates an object detector as a classifier. Resulting model's classification metrics are presented in Table 6.

6. Discussion

Training error and validation mAP of the RetinaNet model are presented in Fig. 6. As it can be noticed, training error is a decreasing function with periodic fluctuations. This behavior may be explained by the fact that one epoch used only 25% of data, so different data distribution might produce a higher error for epochs with difficult-to-analyze cases.

Table 5
Pneumonia prediction results for RetinaNet, Mask R-CNN and combined model measured in terms of mean average precision and predicted class matrix.

Object detection evaluation			
Stage / Model	RetinaNet	Mask R-CNN	Combined model
Stage 1	0.192	0.169	0.199
Stage 2	0.202	0.165	0.204
Classification evaluation (stage 1)			
		Actual class	
		Non-pneumonia	Pneumonia
Predicted class	Non-pneumonia	558	73
	Pneumonia	89	280

Validation mAP is a growing function which was used for final model selection – instead of picking the best validation score model we picked the last epoch model since it showed stable performance during the last four epochs. The same procedure was applied to the Mask R-CNN.

A significant difference between the validation and the testing scores is caused by the difference in mAP calculations (negative samples are not considered during validation) and different train-test region sizes distribution which can be seen from Figs. 3 and 4 and Table 1. As it was suggested by D. Poplavskiy [9] such difference might be explained by various labeling protocols: while the training regions were obtained from one radiologist, the testing data was obtained as the intersection from different doctors.

Fig. 7 shows the TP samples of the testing set. The first and third images contain two pneumonia regions in the right and in the left lung, respectively. Their predictions contain both regions in one bounding box, that reduce target mAP metrics. However, the practical importance of these predictions is high since they allow to localize pathology regions. Fig. 8 illustrates three cases of the individual and combined model performances. The first case (first row) shows AP improvement of the ensemble scheme where the final prediction contains more precise pneumonia localization in the right lung and Mask R-CNN corrected RetinaNet by adding missed pneumonia region in the left lung. The second case gives an example of performance improvement due to the selected weighted adjustment schema. The third case illustrates a situation when the combined model resulted in less accurate predictions compared to Mask R-CNN. Nevertheless, the ensemble model successfully localized left and right pneumonia regions in both cases.

Table 5 shows that RetinaNet outperformed Mask R-CNN in both stages, which may indicate certain superiority of the reformulated loss (focal loss) in tackling class imbalance problem of the pneumonia detection task. Besides, a comparison of similar one-stage detectors (Table 1) shows the dominance of models leveraging ideas of FPN and focal loss. The focal loss might help to overcome a class imbalance problem and to properly deal with hard negative examples that allowed performing better localization of pneumonia regions (Fig. 9).

Table 6 shows the comparison of the developed RetinaNet + Mask R-CNN ensemble model with state-of-the-art deep networks that are commonly used in CXR pathologies classification problem [25]. Both neural networks from the comparison were trained using the same data split and used weighted bce loss. We can notice that the proposed ensemble model outperformed classification networks in recall and f1-score.

Our additional experiment utilized RetinaNet architecture with a bce loss. Fig. 7 shows that focal loss ($\gamma = 2$) outperformed binary cross-entropy on the validation set that resulted to 0.192 mAP of focal loss in oppose to 0.173 mAP of bce during stage 1 and 0.204 mAP vs. 0.193 during stage 2.

One of the key ideas behind ensembling is combining low-biased uncorrelated models because algorithms with similar methodological principles will fail at similar samples and will not be able to correct each other. Low correlation or mutual independence can be achieved by bootstrap aggregation (bagging), random feature selection, a combination of algorithms

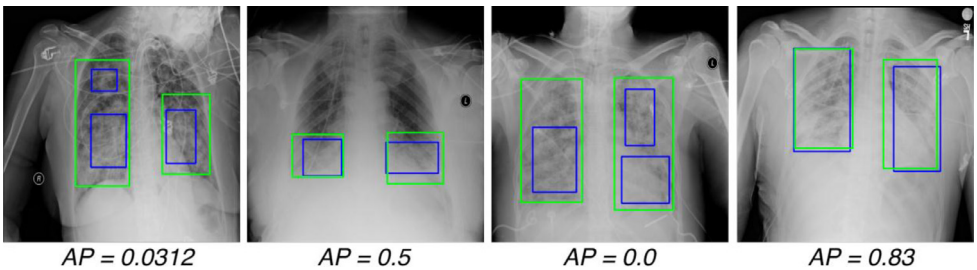

Fig. 7. TP predictions of the final mode with average precision (AP) per image. Blue boxes – ground truth regions, green boxes – predicted pneumonia regions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

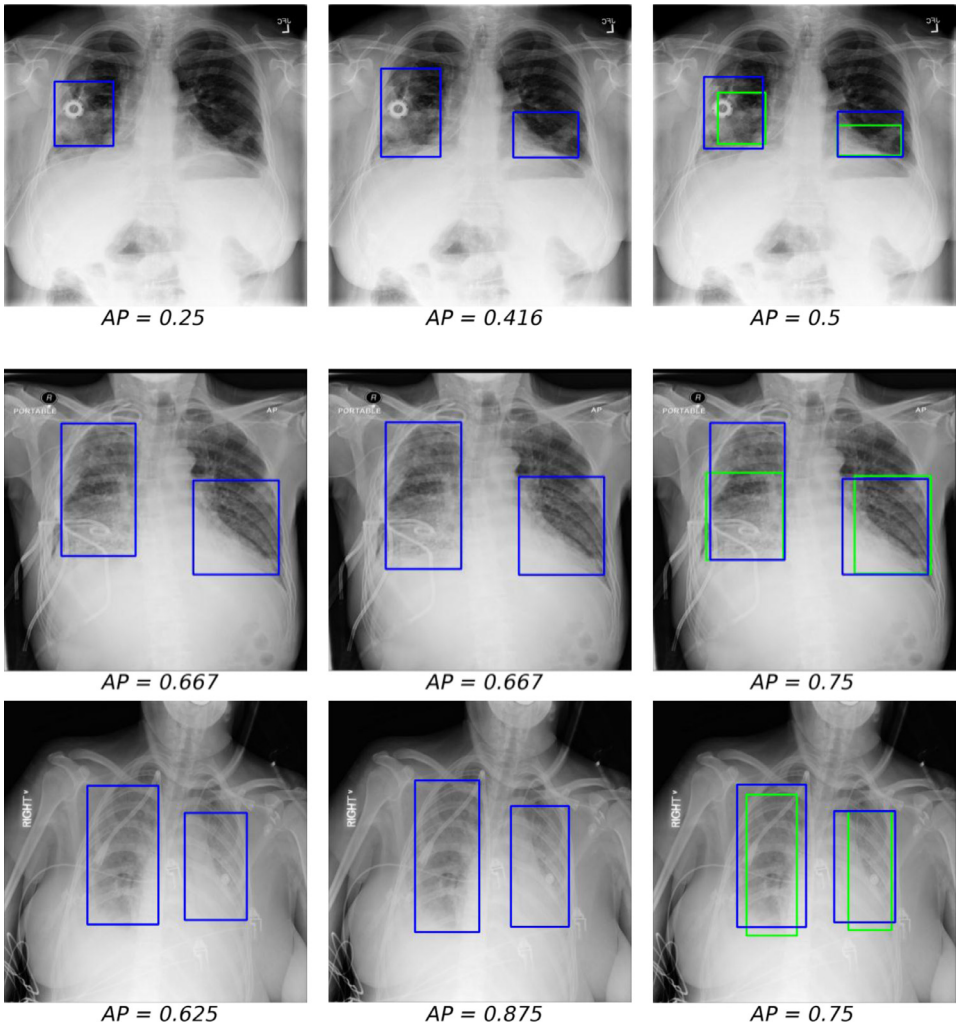

Fig. 8. From left to right: RetinaNet, Mask R-CNN, ensemble model predictions (blue) and the ground truth pneumonia regions (green). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

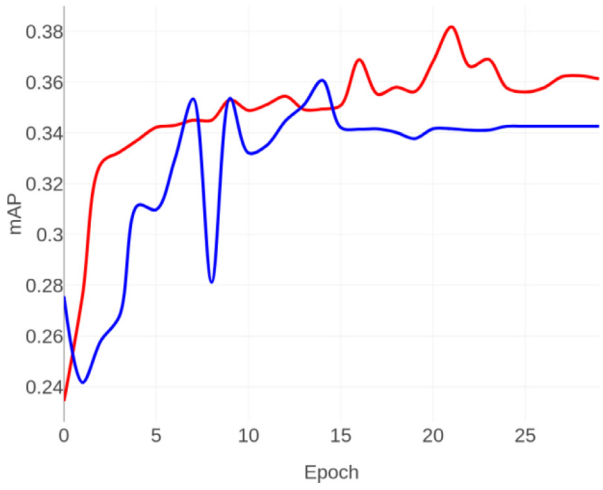

Fig. 9. mAP on the validation set RetinaNet with focal (red) and binary-cross entropy (blue) losses. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 6
Comparison of the utilized RetinaNet + Mask R-CNN networks against alternative state-of-the-art DenseNet-121and Resnet-50 networks trained for classification. The evaluation was performed using the images from the first stage of the RSNA pneumonia detection competition.

	Precision	Recall	F1-score
Densenet-121	0.883	0.652	0.731
Resnet-50	0.855	0.569	0.683
RetinaNet + Mask RCNN	0.758	0.793	0.775

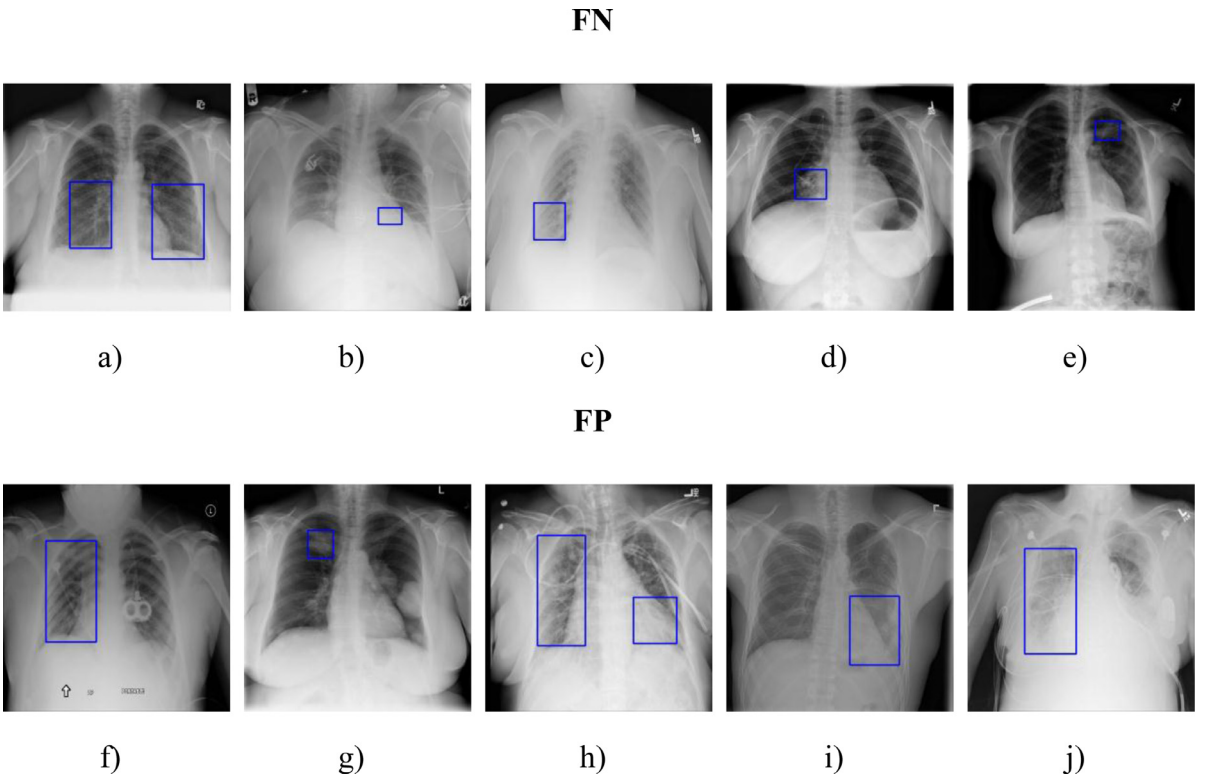

Fig. 10. Randomly selected cases from the database with false positive (FP) and false negative (FN) prediction results. These cases were visually inspected by a team of physicians to evaluate the human performance on challenging-for-machine cases.

from different families (e.g., linear with non-linear) and etc. Although the models in our study belong to the same family of deep CNNs based on the FPN principle, they utilize different techniques that address the class imbalance problem (i.e. two-stage approach for Mask R-CNN and the reformulated loss for RetinaNet). Such double regions checking might help to eliminate hard negative examples (i.e. non-pneumonia or another pathology lung opacity areas). As a result, the ensemble of the two networks exhibits superior results to the individual network performances.

The mAP that was used as a target competition metrics is a common choice for many object detection competitions. It allows evaluating how accurate predicted bounding boxes with the ground truth labels. However, we consider precision and recall is more informative metrics in medical image processing domain since it is more important to detect the presence of pathology, whereas precise detection of the pneumonia region is highly subjected to inter-observer variability and cannot be always considered fully representative. Moreover, the clinical adoption of such technology is expected to put recognition of healthy and diseased as the main priority, whereas physician will revise the diseased cases in order to validate the presence of pneumonia and annotate its exact location.

A team of four experienced radiologists was assembled to assess the performance of the proposed pneumonia detection network. The clinical visually examined 5 FP and 5 FN cases randomly selected from the 1000 testing x-rays from the first challenge stage (Fig. 10). For none of the examined cases, the physicians were able to univocally conclude the presence/absence of pneumonia. They emphasized that chest x-rays need to be accompanied with clinical disease picture, patient's history, demographic data and laboratory measurements for the definitive diagnosis. The physicians suspected the presence of pneumonia or pleural effusion for FN-1 (Fig. 10a), and pneumonia or tuberculosis or carcinosis for FN-3 pa-

tients - both have pneumonia according to the database; pneumonia or pleural edema for FP-1 (Fig. 10f) and pneumonia or fibrosis for FN-3 (Fig. 10c) patients – both do not have pneumonia according to the database. The physicians indicated a need for lateral lung x-rays for the FN-4 patient (Fig. 10d), where the pathology is obstructed with the lung root, and FP-3 patient (Fig. 10h) to eliminate/confirm multifocal pneumonia. Only for three cases out of ten the physicians reached certain consensus. They supported the network decision against the reference diagnosis from the database for FN-5 (Fig. 10e) and FP-4 patients (Fig. 10d), while agreed with the reference diagnosis for the FP-2 patient (Fig. 10b). They would have evaluated the FP-4 patient's response to antibiotics treatment in order to exclude pleural effusion. The FP-2 (Fig. 10g) patient is likely to have lung cyst or cancer due to the solid contour of the pathology in the x-ray. The pathology, however, overlaps with the clavicle, which may be the reason for pneumonia mis-prediction of the proposed network.

7. Conclusion

In our work, we presented an ensemble approach for pneumonia detection using the largest labeled dataset and pointed out the superiority of focal loss and object-detection approach in terms of classification metrics. After analyzing the performance of neural networks on pneumonia identification from chest x-rays and consulting with physicians, we identified several directions for future extension of the presented research. First, although medical imaging is the main information source for pneumonia identification, the comprehensive diagnosis should be based on manifestation on specific clinical symptoms, blood test, pulse oximetry, sputum test etc. Second, a frontal chest x-ray needs to be augmented with a lateral chest x-ray or/and computed tomography image to obtain a more detailed image of the lung field and ensure correct diagnosis of cases with uncertainties. Third, we consider meta information can be important in further investigations: while a patient's gender and age might be useful features for representing a prior pneumonia distribution, a view position might be a valuable feature for distinguishing opacity regions caused by pneumonia and fluid retention.

Declaration of Competing Interest

The authors have no conflict of interest to declare. This work does not serve to promote any product of any company.

Acknowledgments

This research was supported by the Russian Science Foundation under Grant no. 18-71-10072. We also thank the team of radiologists from Public Hospital #2, Kazan, Russia for visual inspection of selected chest x-ray images, Yuriy Makarov and Vitaly Byrachonok for productive collaboration in designing and implementing models.

References

[1] https://www.thoracic.org/patients/patient-resources/resources/top-pneumonia-facts.pdf.
[2] Franquet T. Imaging of community-acquired pneumonia. J Thorac Imaging 2018;33(5):282–94.
[3] Shao Y, Gao Y, Guo Y, Shi Y, Yang X, Shen D. Hierarchical lung field segmentation with joint shape and appearance sparse learning. IEEE Trans Med Imaging 2014;33(9):1761–80.
[4] Ibragimov B, Likar B, Pernus F. A game-theoretic framework for landmark-based image segmentation. IEEE Trans Med Imaging 2012;31(9):1761–76.
[5] Wang X, Peng Y, Le Lu ZL, Bagheri M, Summers RM. Chestx-Ray8: hospital-scale chest x-Ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 2097–106.
[6] Rajpurkar, P., J. Irvin, K. Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, and Katie Shpanskaya. Chexnet: radiologist-Level pneumonia detection on chest x-Rays with deep learning. 2017. arXiv:1711.05225.
[7] Ibragimov B, Toesca D, Chang D, Yuan Y, Koong A, Xing L. Development of deep neural network for individualized hepatobiliary toxicity prediction after liver SBRT. Med Phys 2018;45(10):4763–74.
[8] Abiyev RH, Ma'aitah MKS. Deep convolutional neural networks for chest diseases detection. J Healthc Eng 2018:2018.
[9] RSNA pneumonia detection challenge | kaggle. [Accessed April 23], 2019. https://www.kaggle.com/c/rsna-pneumonia-detection-challenge.
[10] Everingham M, Gool LV, Williams CKI, Winn J, Zisserman A. The pascal visual object classes (Voc) challenge. Int J Comput Vis 2010;88(2):303–38.
[11] Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Lawrence Zitnick C. Microsoft Coco: common objects in context. In: European conference on computer vision. Springer; 2014. p. 740–55.
[12] Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: International conference on computer vision & pattern recognition (CVPR'05), 1. IEEE Computer Society; 2005. p. 886–93.
[13] Sermanet, P., D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: integrated Recognition, localization and detection using convolutional networks. 2013. arXiv:1312.6229.
[14] Girshick R. Fast R-Cnn. In: Proceedings of the IEEE international conference on computer vision; 2015. p. 1440–8.
[15] Ren S, He K, Girshick R, Sun J. Faster R-Cnn: towards real-time object detection with region proposal networks. In: Advances in neural information processing systems; 2015. p. 91–9.
[16] He K, Gkioxari G, Dollár P, Girshick R. Mask R-Cnn. In: Proceedings of the IEEE international conference on computer vision; 2017. p. 2961–9.
[17] Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on computer vision and pattern recognition; 2017. p. 2117–25.
[18] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer; 2015. p. 234–41.
[19] Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 779–88.
[20] Lin T-Y, Goyal P, Girshick R, He K, Dollár P. "Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision; 2017. p. 2980–8.
[21] Jung H, Kim B, Lee I, Yoo M, Lee J, Ham S, Woo O, Kang J. Detection of masses in mammograms using a one-stage object detector based on a deep convolutional neural network. PLoS One 2018;13(9):e0203355.

[22] Xue Z, Candemir S, Antani S, Rodney Long L, Jaeger S, Demner-Fushman D, Thoma GR. Foreign object detection in chest X-Rays. In: 2015 IEEE international conference on bioinformatics and biomedicine (BIBM). IEEE; 2015. p. 956–61.
[23] Kawazoe Y, Shimamoto K, Yamaguchi R, Shintani-Domoto Y, Uozaki H, Fukayama M, Ohe K. Faster R-CNN-Based glomerular detection in multistained human whole slide images. J Imaging 2018;4(7):91.
[24] Redmon, J., and A. Farhadi. Yolov3: an incremental improvement. 2018. arXiv:1804.02767.
[25] Qin C, Yao D, Shi Y, Song Z. Computer-Aided detection in chest radiography based on artificial Intelligence: a survey. Biomed Eng Online 2018;17(1):113.

Ilyas Sirazitdinov is a Researcher Scientist at the Artificial Intelligence Lab at the Innopolis University. He received his M.S. in Data Science from the Innopolis University in 2018. His research interests are medical image analysis, computer vision and digital signal processing.

Maksym Kholiavchenko is a Research Scientist at the Innopolis University and Lead Computer Vision Engineer at X5 Retail Group. From 2018 to 2019, he worked as Computer Vision Engineer in RoadAR. He received his B.S. in Computer Science from the Innopolis University in 2019. His research interests lie primarily in the area of Medical Image Analysis and Industrial Computer Vision.

Tamerlan Mustafaev is a physician at the Kazan Federal University Hospital. He was graduated at Kazan Medical University in 2018 and currently works in the radiology department. His research interests include medical radiology, especially MRI and PET, image processing and artificial intelligence.

Dr. Yixuan Yuan is an assistant professor in EE Department of City University of Hong Kong since April 2018. She was a Postdoctoral Fellow in the Department of Radiation Oncology, Stanford Cancer Center, Stanford University during 2017–2018. She received the Ph.D. degree in Electronic Engineering from the Chinese University of Hong Kong in 2016 with Hong Kong Postgraduate Fellowship.

Ramil Kuleev is head of Artificial Intelligence Lab in Innopolis University. He works in the computer vision domain since 2005 and received his Ph.D. degree in 2013. He has experience in scientific and industrial projects in different domains including medical image analysis, image segmentation and object detection.

Bulat Ibragimov received his Ph.D. degree in Electrical Engineering from the University of Ljubljana in 2014. Until 2018, he was a Postdoctoral Fellow in the Radiation Oncology Department at Stanford University. Currently, he is an Assistant Professor of Machine Learning at the University of Copenhagen and the Chief Investigator of the Artificial Intelligence Lab at Innopolis University.