CoRSAI: A System for Robust Interpretation of CT Scans of COVID-19 Patients Using Deep Learning

MANVEL AVETISIAN, ILYA BURENKO, KONSTANTIN EGOROV, VLADIMIR KOKH, and ALEKSANDR NESTEROV, Sberbank AI Laboratory
ALEKSANDR NIKOLAEV, Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies, Russia
ALEXANDER PONOMARCHUK and ELENA SOKOLOVA, Sberbank AI Laboratory
ALEX TUZHILIN, Sberbank AI Laboratory and New York University
DMITRY UMERENKOV, Sberbank AI Laboratory

Analysis of chest CT scans can be used in detecting parts of lungs that are affected by infectious diseases such as COVID-19. Determining the volume of lungs affected by lesions is essential for formulating treatment recommendations and prioritizing patients by severity of the disease. In this article we adopted an approach based on using an ensemble of deep convolutional neural networks for segmentation of slices of lung CT scans. Using our models, we are able to segment the lesions, evaluate patients' dynamics, estimate relative volume of lungs affected by lesions, and evaluate the lung damage stage. Our models were trained on data from different medical centers. We compared predictions of our models with those of six experienced radiologists, and our segmentation model outperformed most of them. On the task of classification of disease severity, our model outperformed all the radiologists.

CCS Concepts: • **Applied computing** → **Health care information systems**; • **Computing methodologies** → **Computer vision**; **Ensemble methods;**

Additional Key Words and Phrases: Convolutional neural network, deep learning, ensembling, COVID-19, segmentation, lesion detection

ACM Reference format:
Manvel Avetisian, Ilya Burenko, Konstantin Egorov, Vladimir Kokh, Aleksandr Nesterov, Aleksandr Nikolaev, Alexander Ponomarchuk, Elena Sokolova, Alex Tuzhilin, and Dmitry Umerenkov. 2021. CoRSAI: A System for Robust Interpretation of CT Scans of COVID-19 Patients Using Deep Learning. *ACM Trans. Manag. Inf. Syst.* 12, 4, Article 28 (September 2021), 16 pages.
https://doi.org/10.1145/3467471

Authors listed in alphabetical order.

Authors' addresses: M. Avetisian (corresponding author), Sberbank AI Laboratory, Oruzheynyy Pereulok, 41, Moscow, Russia; email: avetisian.m.s@sberbank.ru; I. Burenko (corresponding author), K. Egorov, V. Kokh, A. Nesterov, A. Ponomarchuk, E. Sokolova, and D. Umerenkov, Sberbank AI Laboratory, Moscow, Russia; emails: {burenko.i.m, egorov.k.ser, kokh.v.n, AINesterov, ponomarchuk.a.v, Sokolova.El.Vladimirov}@sberbank.ru, D.Umerenkov@gmail.com; A. Nikolaev, email: a.e.nikolaev@yandex.ru; A. Tuzhilin, Sberbank AI Laboratory, Russia and New York University, New York, NY 10003, United States, New York, USA; email: atuzhili@stern.nyu.edu.

2158-656X/2021/09-ART28 $15.00
https://doi.org/10.1145/3467471

1 INTRODUCTION

Coronavirus (COVID-19) has spread widely around the world since the beginning of 2020, and an extensive effort to combat the pandemic was launched that year. As a result of this effort, there have been several diagnostic tests developed in the medical community to detect COVID cases. One of the most prominent methods to confirm a COVID-19 infection is by conducting the **reverse transcriptional polymerase chain reaction (RT-PCR)** test, which has a lower sensitivity of 65% to 95%. Although useful and popular, the RT-PCR test has the problems of producing negative results even if the patient is infected and having to wait for the test results. Therefore, in some countries a chest **computed tomography (CT)** scan is widely used in clinical practice to detect typical changes in the pulmonary parenchyma associated with COVID-19 [6, 8, 24, 28] as a complement to the RT-PCR test, especially since CT is effective for early detection and diagnosis of COVID-19 [11, 18] and the results of CT scans can be analyzed immediately [1, 15]. Multifocal **ground-glass opacifications (GGOs)** are the most common finding of the CT scan, usually localized peripherally in both lungs, while a single ground-glass lesion can be common at an early stage of the disease [44]. Clinical manifestations of COVID-19 pneumonia and their severity correlate with the volume of lung damage, which can be assessed using visual or quantitative scale.

Although it is easy to assess the severity of lung damage using a visual scale, this is a subjective assessment that can vary substantially among radiologists [11]. Therefore, there exists a more objective classification of lung damage widely used in some countries, including Russia, consisting of the following five stages (referred in the article as CT classes): CT-0: absence of damage; CT-1: **pulmonary parenchymal involvement (PPI)** being ≤ 25%; CT-2: PPI being in the range of 25% to 50%; CT-3: PPI in the range of 50% to 75%; and CT-4: PPI ≥ 75% [28]. In the context of the current COVID-19 pandemic, radiologists in specialized departments need to process a very large number of CT images of subjects with suspected COVID-19, sometimes up to several hundred patients per day, which puts an incredible burden on them and also delays the COVID-19 detection event. Therefore, an automated system that can accurately detect the presence of COVID-19 and calculate the pathology of lung volume will significantly reduce the burden on the radiologist, help objectively assess the severity of the disease, make it possible to prioritize the radiologist work schedule, and provide better insights into the follow-up studies to assess the dynamics of the disease.

In this article, we present the CoRSAI system[1] that takes CT scans of COVID-19 patients and does the image classification and segmentation tasks using **Deep Learning (DL)**-based methods to find the affected areas, to determine the severity of the disease, and to track disease progression. The proposed system uses a novel ensemble of previously developed DL-based models that was architected specifically with the goal of detecting lung damage caused by COVID-19.

To test our system, we compared its performance with two existing DL-based baselines on two open datasets. As a result, our system outperformed these baselines. In addition, we also conducted a study in which we compared CoRSAI's performance with that of six radiologists having at least 3 years of practical experience across the following typical tasks:

- *Segmentation:* detection of the affected areas of the lungs
- *Patient's dynamics:* detection of positive response to the therapy or disease progression
- *Lesion share estimation:* assessment of the lung damage share (ratio of lesion volume to lung volume)
- *Classification:* identification of lung damage stage according to the CT class

[1] CoRSAI stands for *Ru*S*s*ian *CO*ronovirus *AI*-based detection system.

We performed these four experiments using 58 CT scans on 49 patients at a large Russian hospital and used the services of six experienced radiologists.

The results of this study show that our system outperformed the experienced radiologists for the segmentation and classification tasks on average. In all the cases, our system correctly determined the patients' dynamics. The results of the lesion share estimation are not directly usable due to a high degree of radiologists' subjectivity on this task. Correcting for this subjectivity bias allows our system to outperform all six radiologists on the classification task, three of them with statistical significance.

These results imply that our system can be used as a second-opinion tool that would help radiologists to deal with the coronavirus pandemic. In fact, our system has been favorably received by the medical community in Russia and has been successfully deployed in several hospitals in the country.

In this article, we make the following contributions:

- First, we propose an ensemble method specifically designed for the COVID detection problem for the CT scan data that we implemented as a part of the CoRSAI system.
- Second, we empirically compare CoRSAI with two existing baselines and demonstrate that our method outperforms these baselines on the public and on our proprietary CT scan data.
- Third, we conducted a study in which the CoRSAI system outperformed six experienced radiologists across various COVID detection tasks.

We give an overview of existing approaches to classification and segmentation of CT scans and chest X-ray studies in Section 2; in Section 3 we give a detailed description of datasets that we used for classification and segmentation tasks as well as for experiments with doctors; in Section 4 we describe models that we utilized, how we preprocess data, and how we combine individual models into an ensemble; Section 5 is devoted to experiments that we conducted and results we obtained; we give a conclusion and some final thoughts in Section 6.

2 RELATED WORK

Using **Convolutional Neural Networks (CNNs)** is a common practice for the task of image segmentation. Since its appearance in 2015, the U-Net architecture [35] and its modifications have been widely used for the medical segmentation tasks during the analysis of X-rays, CT scans, MRIs, and ultrasound signals for detecting pneumonia [33], breast cancer [40], stroke [5], liver tumor segmentation [23], prostate cancer [27], and many other medical problems [26].

Furthermore, there is a large body of work on applying CNNs to the task of nodule detection in the chest CT images [20], segmentation of the interstitial lung disease [2], chest organ segmentation [7, 36], and other related tasks [22].

There is a large body of recent work dedicated to the task of detecting COVID-19 lesions in lungs based on X-ray studies and CT scans. In particular, [43] and [29] focus on differentiating coronavirus-induced pneumonia from other pneumonia types and healthy controls. Both papers describe the experiments conducted on large samples of cases and produce comparable results with high levels of differentiation between these two types of pneumonias. In [39], a model was developed that classified whether a CT scan contains COVID-19 lesions or not, achieving ROC AUC of 0.959 based on the CT-level annotations. In [14], the authors describe a supervised and a semi-supervised approach to segmentation of lesions and lungs. In [41] a joint classification and segmentation model is built in order to achieve higher quality by extending an expensive segmentation dataset with a classification dataset. In [9] the same problem was solved by using contrastive learning to train a neural network that can later be adopted for the classification task.

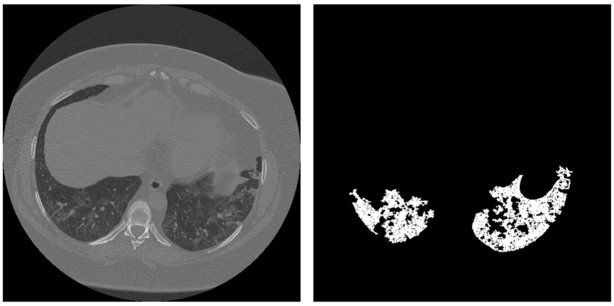

Fig. 1. An example of training set item: a CT scan and an annotation by radiologist.

Moreover, several publications, such as [4, 31], show that detecting coronavirus-induced lesions can be done using lightweight networks with a small number of parameters.

Ensemble learning is an old and well-studied subfield of machine and deep learning (see, e.g., [13], [46], [30]). Several works used ensembling to improve performance of individual models. To name a few, in [32] ensembling was used to improve performance of several classification tasks using chest X-ray scans. In [29] the authors utilized the ensemble-based approach to distinguish between COVID-19 and the commonly acquired pneumonia on the CT scan images. In [17] and [45] the authors used ensembling for the classification of CT slices. The former work uses two-stage transfer learning, where in the first stage weights of the convolutional part of networks were frozen and only classification heads were trained; in the second stage the whole pipeline was fine-tuned to achieve a high classification score while later relative majority voting was utilized to produce a classification result. There is no published work that utilizes ensemble learning of deep convolutional models for both the classification and segmentation tasks on CT studies of lungs affected by COVID.

In this article, we build on all this previous work of analyzing CT scan images by developing ensembles of the previously proposed neural networks that are specifically designed for the COVID-19-related problems. Following the terminology of [34], we conducted experiments with internal and external validation, i.e., experiments where the test data was either from the same distribution as train data or from a different distribution. Furthermore, we describe our clinical study involving several experienced radiologists on whom we test the quality of our ensemble-based model by comparing its performance with the performance of these radiologists across four coronavirus-related tasks.

3 DATA

In our work we utilize four different anonymized CT chest datasets that we use for training, validation, testing, lung segmentation, and experimentation purposes. We describe these datasets in the rest of this section.

Training and Validation. This dataset consists of 68 unique anonymized CT scans with slice thickness from 0.5 to 2.5 mm collected from several hospitals and performed for the patients having a COVID-19 diagnosis. It contains a collection of 18,383 original two-dimensional slices and 9,030 segmented two-dimensional slices with lesions that we used for training and validation purposes (see Table 1). Based on the radiologist reports provided with the CT scans, we selected CT scans that have increased lung opacity levels because of the ground-glass and consolidation findings (i.e., CT class distribution being CT-1 (44.6%), CT-2 (35.4%), CT-3 (15.4%), CT-4 (4.6%)). The CT scan series with lung window using a SeriesDescription DICOM tag like “Lung” level were selected for our research. Each of these series was segmented by the radiologists in a semi-automated fashion

Table 1. Summary for Segmentation Datasets

Dataset	Segmented Slices	Total
Train and Validation	9,030	18,383
Test	785	2,049

Table 2. Experimental Subsets with Labeling Types

#	**Subset Name**	**#CT Scans**	**Types of label**
First	F(20)	20	Segmentation CT classification
Second	S(18)	18	CT classification Dynamic classification Lesion share, %
Third	T(20)	20	CT classification Lesion share, %

using medical image viewer software for the segmentation purpose as follows. First, they used the grow region feature with the lower threshold set to −640 and the upper threshold set to −240. Second, radiologists fixed the results of automated segmentation by manually making corrections to the masks on each slice for the CT scan series by using the brush/erase feature. An example of a radiologist's annotation is depicted in Figure 1.

Lung Segmentation. To build the model of the left and the right lung segmentation, we used a subset of the LIDC/IDRI database [3] from the Luna16 challenge [12]. This dataset contains 888 chest CT scans consisting of 227,301 normal two-dimensional slices and 194,805 segmented two-dimensional slices with the thickness levels ranging from 0.5 to 2.5 mm.

Testing. For the model testing, we used a subset of the MosMedData dataset [28] containing 50 anonymized CT scans that have been annotated by the radiology experts from the Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department. This testing dataset contained a collection of 2,049 original and 785 segmented two-dimensional slices across 50 anonymized chest CT scans with confirmed diagnosis. See the summary in Table 1.

Experimentation. Finally, we prepared an additional dataset for experimentation purposes in order to do the final comparison of the performance results of our model with that of the radiologists. This experimentation dataset consists of 58 anonymized chest CT scans of 49 patients. Since we used this dataset in four different experiments, radiologists applied different labeling methods for this dataset across these four cases. In particular, these four types of labels are designed for the tasks of segmentation, dynamic classification, CT classification, and lesion share of the chest CT scans that were described in the Introduction and will further be explained in Section 5. Furthermore, Table 2 summarizes the specifics of the Experimentation dataset that has been partitioned into three subsets of sizes 20, 18, and 20 corresponding to different experiments presented in Section 5.

4 METHODS

In this section we describe how we took the existing segmentation and classification models previously described in the literature and combined them in a unique fashion into our CoRSAI system using ensemble methods for the pneumonia-COVID detection problem.

4.1 COVID-19 Segmentation Models

We use the data described in Section 3 to train segmentation models that are able to localize COVID-19 lesions in lungs. We implemented the U-net with DPN-92 [10] and ResNet-21 [16] as encoders, FPN with the EfficientNet encoder [37], and a standalone ResNet-18 encoder. We first describe each of the networks in Sections 4.1.1 through 4.1.4 and then explain how we combined them into ensembles in Section 4.1.6.

4.1.1 DPN-92 U-Net. We followed [5] and trained the U-Net with a **Dual Path Network (DPN)** with 92 layers as an encoder with a lightweight decoder. Furthermore, we used the same learning rate, loss function, optimizer, and augmentations as described in [5], only having the number of training steps reduced from 20,000 to 2,500.

4.1.2 Resnet-21 U-Net. We trained the U-Net with ResNet-21 as an encoder [16]. We used the Adam optimizer with the initial learning rate 3×10^{-5} for the first 200 epochs and 1×10^{-5} until convergence. The weight decay was 1×10^{-4} for the whole training procedure. We used the batch size of 64 and the Dice measure as the loss function for this model [27].

4.1.3 FPN with EfficientNet Encoder. We also trained the Feature Pyramid Network model [25] with the EfficientNet-B0 [37] encoder from the open source repository [42]. We used the Adam optimizer with the flat learning rate of 3×10^{-3} until convergence. The weight decay was 1×10^{-8} for the whole training procedure. We used the batch size of 12 and binary cross-entropy as the loss function.

4.1.4 Standalone ResNet-18. We also trained the ResNet-18 as a standalone segmentation model by removing the pooling and the fully connected layers. We did it to diversify our ensemble by adding a different segmentation approach and examine whether plain convolution architecture like ResNet is able to extract features to handle the segmentation task. For the ResNet-18, we used the same hyperparameters and training regime as for ResNet-21, except the batch size was reduced to eight.

4.1.5 Preprocessing. Raw images from DICOM files are stored as 16-bit grayscale images. In order to make the learning process more stable and robust, we normalized input to the neural networks. For the DPN-92 U-Net and the FPN models we (see Algorithm 1)

- multiplied the value of each pixel in the DICOM image array by the rescale slope and added the rescale intercept—these two parameters are stored in DICOM file format;
- divided the result by the absolute value of the minimum pixel value in the image; and
- truncated the value to the range [-0.505, 0.505].

The data for the ResNet models was normalized to have zero mean and the unit standard deviation; i.e., we performed the following transformation for any input image x from the training, validation, and test[2] dataset:

$$x \leftarrow (x - \mu)/\sigma,$$

where the mean μ and the standard deviation σ for the normalization process were calculated on the training data from raw DICOM images.

4.1.6 Ensembling. To improve on the quality of the individual models, we experimented with various ensembling techniques for each model as well as across the models. The results for individual models and ensemble are summarized in Table 3.

[2] We emphasize that the test dataset was collected from different medical centers.

Table 3. Mean DSC and Standard Deviation for Segmentation Models on Test Dataset

	DPN	DPN-3D	FPN	ResNet-21	RN-21 + RN-18	Final
Individual model	0.565 ± 0.024	N/A	0.572 ± 0.016	0.508 ± 0.024	N/A	N/A
Ensemble	0.603[a]	0.613[b]	0.595[a]	0.601[c]	0.620[c]	0.643[d]

[a] Ensemble on 6 folds.

[b] Ensemble on 6 folds for each projection, 18 models total

[c] Ensemble of 6 models selected from 16 by ensemble result on test set.

[d] 3 ensembles merged with coefficients selected on test set.

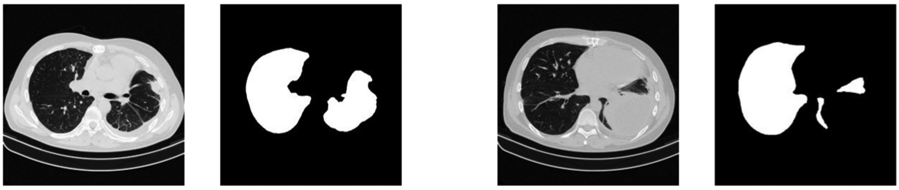

Fig. 2. Examples of the lung segmentation (Section 4.2).

ALGORITHM 1: Normalized input for DPN-92 and FPN networks

```
input  : Raw DICOM pixel array $x$
output : Normalized pixel array $x_{\text{norm}}$ for the DPN-92 and FPN
$b$ – rescale intercept, $k$ – rescale slope, $p_{\min} = |\min(x)|$;
for $\forall p \in x$ do
    $p \leftarrow \frac{p-b}{k}$;
    $p \leftarrow p * \frac{1}{p_{\min}}$;
    if $p < -0.505$ then
        $p \leftarrow -0.505$
    else
        if $p > 0.505$ then
            $p \leftarrow 0.505$
        end
    end
end
```

For the DPN and FPN architectures, we trained six models each, selecting a different validation set from the training set for each model. The models were ensembled by averaging their predictions for each architecture. We trained 16 models for the Resnet-21 U-net and two models for the standalone ResNet-18 using random subsets of the training data. For the ensemble model we take 5 of 16 ResNet-21 and the better of the two ResNet-18 models. We chose the best-performing tuple of ResNet-21 over 1,000 of $\binom{16}{5} = 4,368$ randomly generated choices. The models were ensembled by the unanimous vote of all the models in the positive class.

We also trained 12 additional DPN models in the sagittal and dorsal projections, an ensemble of six models for each projection, as described in [5], and calculated performance of the three-dimensional DPN ensemble.

To build the final ensemble used in the experiments, we modified the predictions of the best-performing ensemble (five Resnet-21 U-Net and one Resnet-18) with high-confidence predictions from the DPN-92 and FPN ensembles. The confidence thresholds for the final ensemble were tuned on the test set. Furthermore, we used the following scoring function for each pixel:

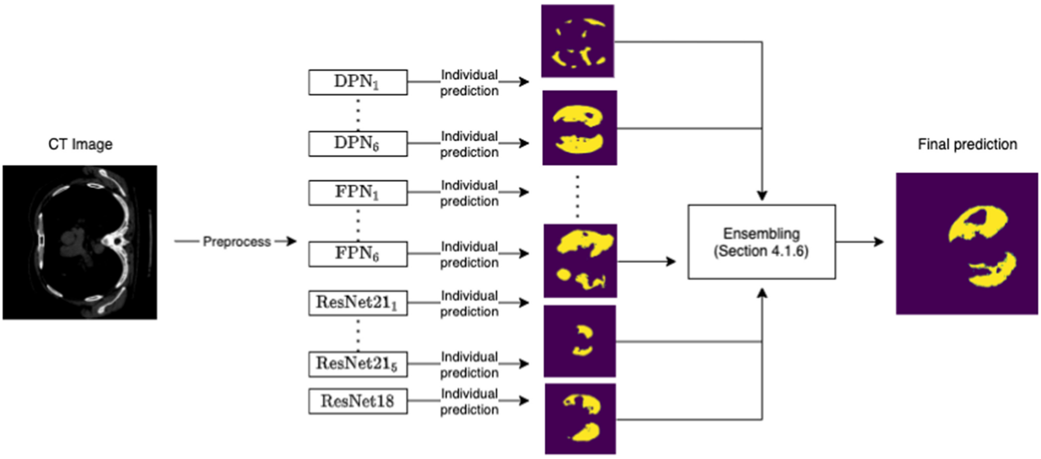

Fig. 3. Inference for the individual slice.

+1 for each:

- Predictions of all models in ResNet ensemble > 0.5
- Mean of DPN-92 U-Net models > 0.7
- Mean of FPN models > 0.85

−1 for each:

- Mean of DPN-92 U-Net models < 0.3
- Mean of FPN models < 0.15

Pixels with positive values were considered as positive class predictions. The whole pipeline of the inference is depicted in Figure 3.

4.2 Segmentation of Lungs

For the left and the right lung segmentation, we used FPN [25] with the lightweight encoder EfficientNet-B0 [37]. The final segmentation was the result of the ensemble of three separate two-dimensional networks for axial, coronal, and sagittal projections (see Figure 2). In order to give the model better spatial understanding, we added three-dimensional coordinates to the input as separate channels. We have also resized the inputs for all the projections to 128x128 pixels.

The lung segmentation dataset contains numerous mistakes. To deal with them, we used active learning as follows. First, we trained the ensemble on 50% of the data and evaluated the results on the other 50%. Second, the CT scans with the least Dice scores were manually reviewed and mask errors were excluded from the dataset. Then, we repeated this process with the other half of the dataset. Overall, we excluded 19 CT scans from the dataset as the result of this cleaning process. After this, we trained the final networks with the holdout 10% validation scheme. The resulting **Intersection-over-Union (IoU)** score for validation was 0.97, which is comparable to the human labeling quality of IoU = 0.96 [38].

5 EXPERIMENTS

To evaluate the efficiency of the proposed models, we conducted a study based on the retrospective data collected during the treatment process in a large clinic in Russia. In this study, 58 chest CT scans on 49 patients were selected (see Section 3) and were divided across the following four experiments described in the rest of this section.

Table 4. Mean DSC and Standard Deviation on the Segmentation Experiment

	Radiologist	Model	Number of Cases	p-Value
Radiologist 1	0.650(0.225)	**0.676(0.239)**	20	0.728
Radiologist 2	0.681(0.218)	**0.682(0.239)**	20	0.985
Radiologist 3	0.697(0.191)	**0.709(0.178)**	20	0.832
Radiologist 4	0.662(0.224)	**0.713(0.176)**	20	0.443
Radiologist 5	**0.699(0.202)**	0.686(0.235)	20	0.855
Radiologist 6	0.251(0.090)	**0.695(0.186)**	20	0.000
All radiologists	0.606(0.254)	**0.694(0.211)**	120	0.004

5.1 Segmentation

The goal of the first experiment was to compare the segmentation accuracy of pulmonary consolidation and the ground-glass opacity area on the CT images obtained by our segmentation model vis-a-vis the performance of experienced radiologists involved in our study. For this purpose, we used the experimental subset F(20), which was represented by 20 CT cases from 20 patients of varying severity and was described in Section 4 and presented in Table 2. These cases were manually segmented by six experienced practicing radiologists and by our model.

To measure the performance of an individual radiologist, we compared his or her results to the panel of the remaining five radiologists. Each pixel was considered to belong to the positive class if at least three radiologists marked it as positive. We used the mean **Dice similarity coefficient (DSC)** for all the CT scans as our measurement metric. Since the panel result is different for each radiologist, we calculated the metric for our segmentation model separately for each panel.

As Table 4 shows, our model outperforms five out of six radiologists and outperforms the average radiologist (represented by its last row) with statistical significance (p-value of 0.004).

5.2 Patients' Dynamics

In the second experiment, we compared performance of the segmentation model and human performance in assessing patients' dynamics for the follow-up CT scans. For this purpose, we used the experimental subset S(18) (see Table 2) consisting of 18 CT scans on nine patients (2 CT scans per patient with different dates). The radiologists and our model independently estimated the percentage of lesions in the left and the right lung. Based on this information, one of the three classes for evaluating the patient dynamics was chosen by the radiologists and our model: a positive response to the therapy, disease progression, and a stable condition (for our model the change of $< 1\%$ was considered to be stable). In the case of one patient, the radiologists' assessments were tied 3 vs. 3 between the positive response and the stable condition. Therefore, we removed this case from the experiment since we were unable to determine the "ground truth" of this patient's dynamics. From the remaining eight cases, the radiologists unanimously agreed on the dynamics of the disease progression in seven cases, and in the remaining case they were split five to one in favor of the disease progression vs. the positive response. It turned out that our system correctly predicted the dynamics *in all eight cases.*

5.3 Lesion Share Estimation

In the third experiment we compared the performance of the segmentation model and human performance in assessing patients' lesion share, which is the base for identification of the lung damage stage. For that purpose we used radiologists' estimation of the lesion percentage in the left and the right lung made with the 5% increment. In total, we had 76 estimations from six radiologists for the right and the left lung for each of the 38 chest CT scans from experimental subsets F(18)

Table 5. Lesion Share Estimation Results of the Radiologists and Our Model

	Radiologist		Model		
	MAE	ME	MAE	ME	Cases
Radiologist 1	0.08	−0.06	0.13	−0.13	76
Radiologist 2	0.10	0.10	0.11	−0.11	76
Radiologist 3	0.06	−0.03	0.13	−0.13	76
Radiologist 4	0.07	0.03	0.12	−0.12	76
Radiologist 5	0.09	0.07	0.11	−0.11	76
Radiologist 6	0.12	−0.11	0.14	−0.14	76
All radiologists	0.09	0.00	0.13	−0.12	456

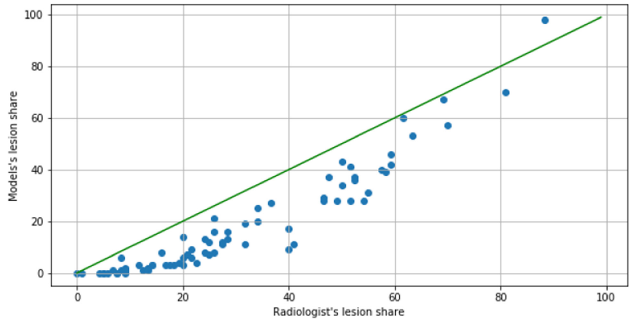

Fig. 4. Model's lesion share to radiologists' lesion share.

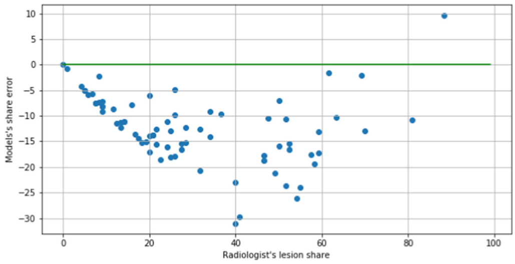

Fig. 5. Model's error to radiologists' lesion share.

and T(20). We performed the same estimation for lesion share using our segmentation models to calculate the COVID-19 lesions and lung volumes and divide them. To measure the performance of an individual radiologist in comparison to our system, we compared their results to the panel of the remaining five radiologists. As the ground truth, we took an average lesion share between the five remaining radiologists. We used the **mean absolute error (MAE)** for all the CT scans as our measurement metric. We also calculated the **mean error (ME)** to explore where there is a systemic component to the error.

As Table 5 shows, the radiologists' estimation is significantly subjective, with some radiologists biased to overestimation (Radiologist 2) or underestimation (Radiologist 6). Our model estimates the lesion share based on objective factors and is highly correlated with the mean estimation of six radiologists per each of 76 considered cases (Figure 4) while being biased to underestimation.

The subjectivity of the radiologists' estimate is correlated with lesion share. The lesion share range of 30 to 70 has the largest disagreement in estimation between the radiologists themselves (Figure 6) as between the model and mean radiologist estimation (Figure 5).

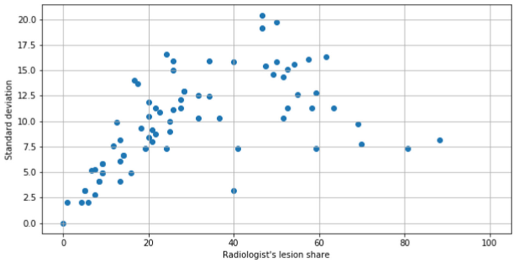

Fig. 6. Radiologists' lesion share standard deviation.

Table 6. CT Class Thresholds for Radiologists

	Split 1			Split 2			Full		
	CT-2	CT-3	CT-4	CT-2	CT-3	CT-4	CT-2	CT-3	CT-4
Radiologist 1	0.12	0.72	0.74	0.17	0.28	0.84	0.12	0.44	0.83
Radiologist 2	0.03	0.25	0.47	0.03	0.22	0.84	0.03	0.24	0.83
Radiologist 3	0.12	0.42	0.74	0.17	0.34	0.84	0.12	0.36	0.83
Radiologist 4	0.03	0.25	0.74	0.06	0.30	0.59	0.06	0.24	0.59
Radiologist 5	0.10	0.29	0.74	0.04	0.28	0.77	0.10	0.28	0.79
Radiologist 6	0.10	0.25	0.99	0.26	0.55	0.84	0.10	0.29	0.99
All radiologists	0.08	0.29	0.74	0.06	0.34	0.84	0.06	0.29	0.83

Based on this, we conclude that the doctors have systemic biases in their estimations of lesion volumes while using typical diagnostic tools. As our analysis shows, our system corrects these estimation biases and therefore is well suited for estimating lesion volumes.

5.4 Classification

In the fourth experiment we compared the performance of the radiologists and our segmentation model results on the CT classification task. We used classification accuracy as the metric for this experiment. For the classification task we used the segmentation model results. To estimate the CT class, our system calculated the maximum share of the lesions in the right or the left lung and then used precalculated thresholds to get the final classification result.

To correct for the radiologists' subjective bias, we fitted the thresholds for lesion share for each CT class to maximize prediction accuracy. We split the experimental dataset into two equal folds stratified by the CT classes retrieved form the hospital reports. For each radiologist we fitted the thresholds for the whole dataset and for each of the folds separately. As can be seen from Table 6, the thresholds vary greatly between radiologists even when fitted on all the data: 0.03 to 0.12 for CT1-CT2, 0.24 to 0.44 for CT2-CT3, and 0.59 to 0.99 for CT3-CT4. When fitting on separate folds, the individual thresholds become even more noisy due to the lower number of data points available.

To estimate the collective bias of the panel, we combine individual biases of each radiologist in the panel by fitting the optimal thresholds on all their estimations in the train split. As can be seen from Table 7, after applying correction for the panel bias, our system outperforms all the radiologists in the study, three of them with statistical significance (p-values of 0.01 or less).

5.5 Baseline Comparison

We have also compared the performance of our model with the existing COVID-19 detection baselines. It turns out that many existing models (e.g., Athanasios et al. [4], Qiu et al. [31], Wang et al.

Table 7. CT Classification Accuracy of the Radiologists and Our System

	Split 1 (29 Cases)		Split 2 (29 Cases)		Combined (58 Cases)		
	Radiologist	Model	Radiologist	Model	Radiologist	Model	p-Value
Radiologist 1	0.76	0.86	0.76	0.83	0.76	0.84	0.248
Radiologist 2	0.59	0.90	0.66	0.86	0.62	0.88	0.001
Radiologist 3	0.72	0.90	0.83	0.90	0.78	0.90	0.080
Radiologist 4	0.66	0.90	0.76	0.90	0.71	0.90	0.010
Radiologist 5	0.66	0.83	0.86	0.72	0.76	0.78	0.828
Radiologist 6	0.55	0.83	0.83	0.97	0.69	0.90	0.006
All radiologists	0.66	0.87	0.78	0.86	0.72	0.86	1.66×10^{-6}

[39], Zhang et al. [43]) were incomparable with our approach both for the segmentation and for the classification cases for the following reasons. First, some of the existing approaches did not provide code or data, while others used the classification and segmentation criteria that are different from our method, thus rendering them incomparable. For example, [39] used the binary (presence/absence of COVID-19) classification, whereas we deployed the CT-0/CT-1/CT-2/CT-3/CT-4 classification commonly used in Russia and some other countries. Similarly, we could not compare our approach to many segmentation methods due to lack of the code and data.

The segmentation baselines comparable with our approach are the Inf-Net model described in [14] and the MiniSeg model described in [31]. We compare our model with these baselines in the rest of this subsection using two testing datasets. The first dataset is the COVID-19 CT Segmentation Dataset[3] [19], referred hereafter as CovidCTSegmentation; this dataset is very small, containing only 100 CT slices from 60 patients. The second dataset is the MosMedData dataset mentioned earlier.

We compared the performance of the MiniSeg model available on GitHub[4] and CoRSAI on CovidCTSegmentation, training both models from scratch and using fivefold cross-validation. The resulting DSC score was 0.452 for MiniSeg and 0.744 for CoRSAI, showing excellent performance of our model even when presented with a minimal amount of training data. See predicted masks for MiniSeg and our model on Figure 7.

We also examined the following cases comparing CoRSAI, MiniSeg, and various baselines on the CovidCTSegmentation dataset using fivefold cross-validation as in [31]:

- Baseline results for (U-Net, Inf-Net, EfficientNet) as stated in [31]
- MiniSeg
- CoRSAI

The results are shown in Table 8, where the baseline DSC performance scores are taken directly from Qiu et al. [2020]. As Table 8 shows, the CoRSAI model outperformed the baselines in terms of the DSC metric.

Next we compared the Inf-Net, MiniSeg, and CoRSAI models on the MosMedData dataset. We took the initial Inf-Net model, as available on GitHub, including its architecture and the computed weights, and tested it vis-a-vis our model on the MosMedData dataset in terms of the Dice performance metric. It turned out that the "as-is" Inf-Net model produced only 0.195 Dice metric on MosMedData, which was significantly below our model, which had the value of 0.643 for the Dice metric. This inferior performance of Inf-Net was due to the fact that Inf-Net was trained on the very different dataset obtained for the Wuhan COVID-19 patients.

[3] Available at http://medicalsegmentation.com/covid19/.

[4] Available at https://github.com/yun-liu/MiniSeg.

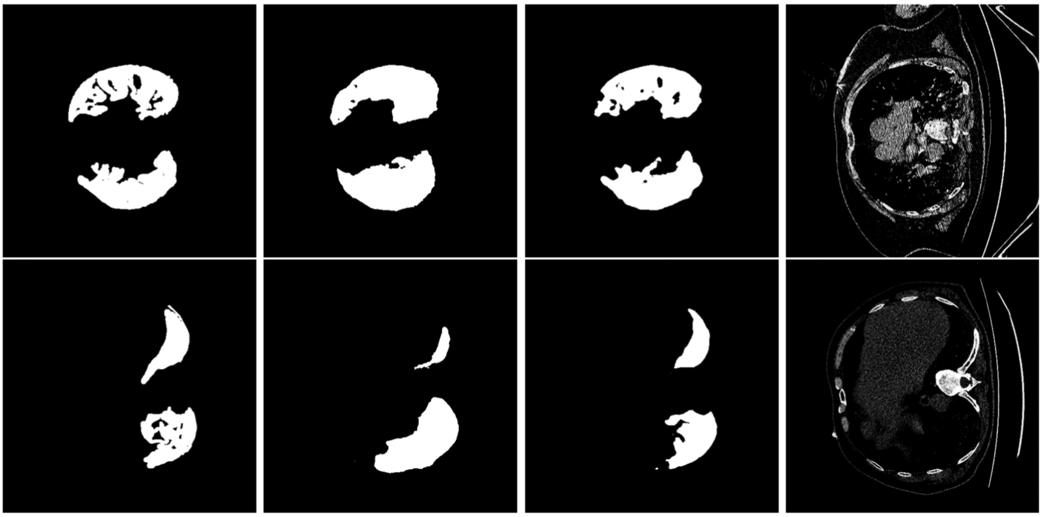

Fig. 7. Examples of the segmentation of a slice from CovidCTSegmentation dataset.
The leftmost image is the output of the MiniSeg model, next to the right is the output of the CoRSAI, next to the right is the ground truth, and the rightmost image is the input slice.

Table 8. Comparison on CovidCTSegmentation Dataset

Model	DSC Score
U-Net	0.684*
Inf-Net	0.744*
EfficientNet	0.705*
MiniSeg	0.759*
CoRSAI	**0.768**

*Results from [31].

Table 9. Comparison on MosMedData Dataset

Model	DSC Score
MiniSeg	0.032
Inf-Net	0.619
CoRSAI	**0.643**

To provide further comparison of the three models, we compared CoRSAI, MiniSeg, and Inf-Net on the MosMedData dataset using our private training dataset to retrain and fine-tune all the models using the same methods as we did for our model described in Section 4. The results are shown in Table 9, from which it is clear that CoRSAI outperformed Inf-Net and MiniSeg on the MosMedData dataset.

As a result of this extensive retraining, we improved the performance of Inf-Net from 0.195 to 0.619, which is in line with the performance results of the joint ResNet-21+ResNet-18 and DPN-3D ensembles presented in Table 5, such as DPN and FPN, but still inferior to the 0.643 of our model. The MiniSeg model achieved a DSC score of 0.032 after the same retraining. We need to mention that for the MiniSeg retraining we used all hyperparameters "as is" excluding the batch size. In our retraining we set the batch size to 24 to reduce the time needed for the training. The dataset that we used to retrain the MiniSeg model is much bigger than the original dataset used in [31] used for the training MiniSeg model. Since MiniSeg is an extremely small network with approximately 86,000 parameters, this might be a problem of catastrophic forgetting [21] and hence a reason for generalization performance on our data.

We maintain that the superior performance of our model is due to the careful deployment of ensembles both for the individual models (DPN, FPN, ResNet-21, etc.) and for combining of these individual models into one ensemble, as described in Section 4.1.6 and shown in Table 5.

6 CONCLUSION

In this article we described a novel ensemble of previously proposed deep convolutional neural networks specifically modified for the COVID-19-induced pneumonia segmentation task for the analysis of the chest CT scans and the corresponding CoRSAI system. Furthermore, we have created a segmentation-based classification model to categorize the severity level of the disease on those CT scans. To test the performance of our models, we conducted an experiment that showed that our model outperformed most of the experienced radiologists in the segmentation and all the radiologists in the classification tasks. It also managed to predict the dynamics of the disease with 100% accuracy.

Our model has been favorably received by the medical community in Russia and has been recently deployed in several hospitals in the country.

In particular, CoRSAI is publicly available for doctors and anybody else who is interested in our system on the website https://ai.sberhealth.ru/covid19/. Moreover, it was deployed in 46 medical institutions in 25 different regions of the Russian Federation, and thousands of CT studies have been processed with its help since May 2020. Although some discrepancies were highlighted between radiologists and the model in lesion estimation, the system demonstrated good acceptance by the medical community. It was emphasized that the system has speeded up the lesion estimation time and improved its accuracy, which is particularly important for evaluating patients' dynamics. However, doctors have also noted certain limitations of our system, and fixing these limitations is a part of the future plans for working on CoRSAI, including problems of differentiation of the ground glass and consolidation, assessment of the type of pneumonia (bacterial or viral), and extending the list of lung pathologies for diagnosis (tuberculosis, emphysema, lung cancer, pneumothorax, etc).

Therefore, as a part of the future work, we plan to measure the performance of this deployed model in actual clinical settings in terms of how much it helps doctors treat coronavirus patients. We also plan to fine-tune and further improve the model based on this feedback.

REFERENCES

[1] Tao Ai, Zhenlu Yang, Hongyan Hou, Chenao Zhan, Chong Chen, Wenzhi Lv, Qian Tao, Ziyong Sun, and Liming Xia. 2020. Correlation of Chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: A report of 1014 cases. *Radiology* 296, 2 (2020), 32–40.

[2] Marios Anthimopoulos, Stergios Christodoulidis, Lukas Ebner, Andreas Christe, and Stavroula Mougiakakou. 2016. Lung pattern classification for interstitial lung diseases using a deep convolutional neural network. *IEEE Transactions on Medical Imaging* 35, 5 (2016), 1207–1216.

[3] Samuel G. Armato et al. 2015. The Cancer Imaging Archive. http://doi.org/10.7937/K9/TCIA.2015.LO9QL9SX

[4] Voulodimos Athanasios, Protopapadakis Eftychios Katsamenis, Katsamenis Iason, Doulamis Anastasios, and Doulamis Nikolaos. 2020. Deep learning models for COVID-19 infected area segmentation in CT images. *medRxiv* (2020). https://www.medrxiv.org/content/early/2020/05/19/2020.05.08.20094664.

[5] Manvel Avetisian, Vladimir Kohn, Alexander Tuzhilin, and Dmitry Umerenkov. 2019. Radiologist-level stroke classification on non-contrast CT scans with deep U-Net. In *Medical Image Computing and Computer Assisted Intervention.*

[6] Damiano Caruso, Marta Zerunian, Michela Polici, Francesco Pucciarelli, Tiziano Polidori, Carlotta Rucci, Gisella Guido, Benedetta Bracci, Chiara de Dominicis, and Andrea Laghi. 2020. Chest CT features of COVID-19 in Rome, Italy. *Radiology* 296, 2 (2020), 201237.

[7] Jean-Paul Charbonnier et al. 2017. Improving airway segmentation in computed tomography using leak detection with convolutional networks. *Medical Image Analysis* 36 (2017), 52–60.

[8] Rodrigo Caruso Chate, Eduardo Kaiser Ururahy Nunes Fonseca, Rodrigo Bastos Duarte Passos, Gustavo Borges da Silva Teles, Hamilton Shoji, and Gilberto Szarf. 2020. Presentation of pulmonary infection on CT in COVID-19: Initial experience in Brazil. *Jornal Brasileiro de Pneumologia* 46, 2 (2020), e20200121.

[9] Xiaocong Chen, Lina Yao, Tao Zhou, Jinming Dong, and Yu Zhang. 2020. Momentum contrastive learning for few-shot COVID-19 diagnosis from chest CT images. arXiv:2006.13276 https://arxiv.org/abs/2006.13276.

[10] Yunpeng Chen, Jianan Li, Huaxin Xiao, Xiaojie Jin, Shuicheng Yan, and Jiashi Feng. 2017. Dual path networks. *CoRR* abs/1707.01629 (2017). arXiv:1707.01629 http://arxiv.org/abs/1707.01629.

[11] Michael Chung et al. 2020. CT imaging features of 2019 novel coronavirus (2019-nCoV). *Radiology* 295, 1 (2020), 202–207.

[12] Colin Jacobs, Arnaud Arindra Adiyoso Setio, Alberto Traverso, and Bram van Ginneken. 2016. LUNA16 Dataset. https://luna16.grand-challenge.org/Home/.

[13] Thomas G. Dietterich. 2000. Ensemble methods in machine learning. In *Multiple Classifier Systems*. Springer, Berlin, 1–15.

[14] Deng-Ping Fan et al. 2020. Inf-Net: Automatic COVID-19 lung infection segmentation from CT images. *IEEE Transactions on Medical Imaging* 39, 8 (2020), 2626–2637.

[15] Yicheng Fang, Huangqi Zhang, Jicheng Xie, Minjie Lin, Lingjun Ying, Peipei Pang, and Wenbin Ji. 2020. Sensitivity of chest CT for COVID-19: Comparison to RT-PCR. *Radiology* (2020), 200432.

[16] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*. 770–778.

[17] Jose Francisco Hernandez Santa Cruz. 2021. An ensemble approach for multi-stage transfer learning models for COVID-19 detection from chest CT scans. *Intelligence-Based Medicine* 5 (2021), 100027. https://doi.org/10.1016/j.ibmed.2021.100027

[18] Chaolin Huang et al. 2020. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet* 395, 10223 (2020), 497–506.

[19] H. B. Jenssen. 2020. Covid-19 ct segmentation dataset. http://medicalsegmentation.com/covid19/.

[20] Kai-Lung Hua, Che-Hao Hsu, Shintami Chusnul Hidayati, Wen-Huang Cheng, and Yu-Jen Chen. 2015. Computer-aided classification of lung nodules on computed tomography images via deep learning technique. *OncoTargets and Therapy* 8 (2015), 2015–2022.

[21] Ronald Kemker, Marc McClure, Angelina Abitino, Tyler Hayes, and Christopher Kanan. 2018. Measuring catastrophic forgetting in neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 1 (Apr. 2018). https://ojs.aaai.org/index.php/AAAI/article/view/11651.

[22] Sang Min Lee et al. 2019. Deep learning applications in chest radiography and computed tomography: Current state of the art. *Journal of Thoracic Imaging* 34, 2 (2019), 75–85.

[23] Xiaomeng Li et al. [n.d.]. H-DenseUNet: Hybrid densely connected unet for liver and tumor segmentation from CT volumes. *IEEE Transactions on Medical Imaging* 37, 12 ([n. d.]), 2663–2674.

[24] Yan Li and Liming Xia. 2020. Coronavirus disease 2019 (COVID-19): Role of Chest CT in diagnosis and management. *American Journal of Roentgenology* 214, 6 (2020), 1280–1286.

[25] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*. 936–944.

[26] Geert Litjens et al. 2017. A survey on deep learning in medical image analysis. *Medical Image Analysis* 42 (2017), 60–88.

[27] F. Milletari, N. Navab, and S. Ahmadi. 2016. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 4th International Conference on 3D Vision (3DV'16)*. 565–571.

[28] Sergey Morozov et al. 2020. MosMedData: Chest CT Scans with COVID-19 related findings. *medRxiv* (2020). https://mosmed.ai/datasets/covid19/_1110.

[29] Xi Ouyang et al. 2020. Dual-sampling attention network for diagnosis of COVID-19 from community acquired pneumonia. *IEEE Transactions on Medical Imaging* 39, 8 (2020), 2595–2605. doi:10.1109/TMI.2020.2995508.

[30] Christian S. Perone, Pedro Ballester, Rodrigo C. Barros, and Julien Cohen-Adad. 2019. Unsupervised domain adaptation for medical imaging segmentation with self-ensembling. *NeuroImage* 194 (2019), 1–11. https://doi.org/10.1016/j.neuroimage.2019.03.026

[31] Yu Qiu, Yun Liu, and Jing Xu. 2020. MiniSeg: An Extremely Minimum Network for Efficient COVID-19 Segmentation. https://arxiv.org/abs/2004.09750.

[32] Sivaramakrishnan Rajaraman et al. 2020. Iteratively pruned deep learning ensembles for COVID-19 detection in chest X-rays. (2020).

[33] Pranav Rajpurkar et al. 2017. CheXNet: Radiologist-Level Pneumonia Detection on Xhest X-rays with Deep Learning. https://arxiv.org/abs/1711.05225.

[34] Michael Roberts, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, et al. 2021. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence* 3 (2021), 199–217. https://www.nature.com/articles/s42256-021-00307-0.

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI'15) (LNCS, Vol. 9351)*. Springer, 234–241.

[36] Brahim Ait Skourt, Abdelhamid El Hassani, and Aicha Majda. 2018. Lung CT image segmentation using deep neural networks. *Procedia Computer Science* 127 (2018), 109–113.

[37] Mingxing Tan and Quoc V. Le. 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. *CoRR* abs/1905.11946 (2019). arXiv:1905.11946 http://arxiv.org/abs/1905.11946.

[38] Eva M. van Rikxoort, Bartjan de Hoop, Max A. Viergever, Mathias Prokop, and Bram van Ginneken. 2009. Automatic lung segmentation from thoracic computed tomography scans using a hybrid approach with error detection. *Medical Physics* 4236, 10 (2009), 2934–2947.

[39] Xinggang Wang et al. 2020. A weakly-supervised framework for COVID-19 classification and lesion localization from chest CT. *IEEE Transactions on Medical Imaging* 8 (2020). https://doi.org/10.1109/TMI.2020.2995965

[40] Nan Wu et al. 2020. Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE Transactions on Medical Imaging* 39, 4 (Apr. 2020), 1184–1194.

[41] Yu-Huan Wu, Shang-Hua Gao, Jie Mei, Jun Xu, Deng-Ping Fan, Rong-Guo Zhang, and Ming-Ming Cheng. 2020. JCS: An explainable COVID-19 diagnosis system by joint classification and segmentation. arXiv:2004.07054 https://arxiv.org/abs/2004.07054.

[42] Pavel Yakubovskiy. 2020. Segmentation Models Pytorch. https://github.com/qubvel/segmentation_models.pytorch.

[43] Kang Zhang et al. 2020. Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography. *Cell* 181, 6 (2020), 1423–1433.

[44] Shuchang Zhou, Yujin Wang, Tingting Zhu, and Liming Xia. 2020. CT features of coronavirus disease 2019 (COVID-19) pneumonia in 62 patients in Wuhan, China. *American Journal of Roentgenology* 214, 6 (2020), 1287–1294.

[45] Tao Zhou, Huiling Lu, Zaoli Yang, Shi Qiu, Bingqiang Huo, and Yali Dong. 2021. The ensemble deep learning model for novel COVID-19 on CT images. *Applied Soft Computing* 98 (2021), 106885. https://doi.org/10.1016/j.asoc.2020.106885

[46] Zhi-Hua Zhou, Jianxin Wu, and Wei Tang. 2002. Ensembling neural networks: Many could be better than all. *Artificial Intelligence* 137, 1 (2002), 239–263. https://doi.org/10.1016/S0004-3702(02)00190-X

Received September 2020; revised March 2021; accepted May 2021