CT-based COVID-19 Triage: Deep Multitask Learning Improves Joint Identification and Severity Quantification

Mikhail Goncharov[a,b,1], Maxim Pisov[a,b,1], Alexey Shevtsov[b,1], Boris Shirokikh[a,b,1], Anvar Kurmukov[b], Ivan Blokhin[c], Valeria Chernina[c], Alexander Solovev[d], Victor Gombolevskiy[c], Sergey Morozov[c], Mikhail Belyaev[a,*]

[a]*Skolkovo Institute of Science and Technology, Moscow, Russia*
[b]*Kharkevich Institute for Information Transmission Problems, Moscow, Russia*
[c]*Research and Practical Clinical Center of Diagnostics and Telemedicine Technologies, Department of Health Care of Moscow, Russia*
[d]*Sklifosovsky Clinical and Research Institute for Emergency Medicine, Moscow, Russia*

Abstract

The current COVID-19 pandemic overloads healthcare systems, including radiology departments. Though several deep learning approaches were developed to assist in CT analysis, nobody considered study triage directly as a computer science problem. We describe two basic setups: *Identification* of COVID-19 to prioritize studies of potentially infected patients to isolate them as early as possible; *Severity quantification* to highlight studies of severe patients and direct them to a hospital or provide emergency medical care. We formalize these tasks as binary classification and estimation of affected lung percentage. Though similar problems were well-studied separately, we show that existing methods provide reasonable quality only for one of these setups. To consolidate both triage approaches, we employ a multitask learning and propose a convolutional neural network to combine all available labels within a single model. We train our model on approximately 2000 publicly available CT studies and test it with a carefully designed set consisting of 33 COVID patients, 32 healthy patients, and 36 patients with other lung pathologies to emulate a typical patient flow in an out-patient hospital. The developed model achieved 0.951 ROC AUC for *Identification* of COVID-19 and 0.98 Spearman Correlation for *Severity quantification*. We release all the code and create a public leaderboard, where other community members can test their models on our dataset.

Keywords: COVID-19, Triage, Convolutional Neural Network, Chest Computed Tomography

*Corresponding author
Email address: m.belyaev@skoltech.ru (Mikhail Belyaev)
[1]Equal contribution

1. Introduction

During the first months of 2020, COVID-19 infection spread worldwide and affected millions of people Li et al. (2020b). Though a virus-specific reverse transcription-polymerase chain reaction (RT-PCR) testing remains the gold standard, CT is a valuable method in diagnosis and patient management Bernheim et al. (2020). The World Health Organization has split the diagnosis of coronavirus infection into two codes: *COVID-19, virus identified (U07.1)*, based on laboratory testing and *COVID-19, virus not identified (U07.2)*, based on clinical, epidemiological and radiological findings where laboratory confirmation is inconclusive or not available. These codes reflect that in an epidemic, diagnostic method speed is crucial in addition to its quality. Moreover, compared to RT-PCR, CT has not only a higher sensitivity (98% compared to 71% at $p < 0.001$) for some cohorts Fang et al. (2020) but also provides an opportunity to obtain results much faster than the RT-PCR analysis requiring a laboratory.

The pandemic dramatically increased the need for medical care and resulted in the overloading of healthcare systems Tanne et al. (2020). Many classification and segmentation algorithms were developed to assist radiologists in COVID-19 identification and severity quantification, see Sec. 1.1.1. However, little research has been conducted to investigate automatic image analysis for triage (or ranking) of CT studies. During an outbreak, many CT scans require rapid decision-making to sort patients into those who need care right now and those who will need scheduled care Mei et al. (2020). Therefore, the study list triage is relevant and may shorten the report turnaround time by increasing the priority of CT scans with suspected pathology for faster interpretation by a radiologist compared to other studies, see Fig. 1

The triage differs from other medical image analysis tasks, as in this case, automatic programs provide the first reading. The radiologist then becomes the second reading. Technically, many of the developed methods may provide an index for triage, e.g., output probability of a classifier or the total lesion volume

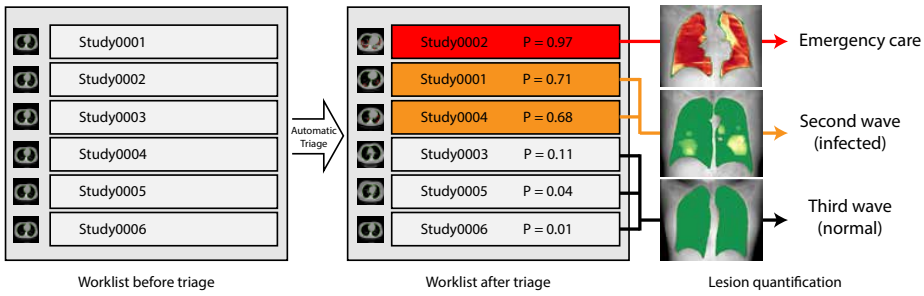


Figure 1: A schematic representation of the automatic triage process. Left: the chronological order of the studies. Center: re-prioritized order to highlight findings requiring radiologist's attention (P denotes COVID-19 *Identification* probability). Right: accompanying algorithm-generated X-ray-like series to assist the radiologist in fast decision making (color bar from green to red denotes *Severity* of local COVID-19-related changes).

extracted from a binary segmentation mask. However, these indices must be properly used. We assume that there are two different triage problems

1. *Identification*. The first challenging task is to identify studies of patients with COVID-19 and prioritize them so the physician can isolate potentially infected patients as early as possible Sverzellati et al. (2020).
2. *Severity quantification*. Second, within COVID-19 patients, a triage algorithm must prioritize those who will require emergency medical care Kherad et al. (2020).

Binary classification provides a direct way to formalize *Identification*, but the optimal computer science approach to estimate *Severity* is not as obvious. It was shown that human-based quantitative analysis of chest CT helps assess the clinical severity of COVID-19. Colombi et al. (2020) had quantified affected pulmonary tissue and established a high correlation between the healthy pulmonary tissue volume and the outcomes (transfer to an intensive care unit or death). The threshold value for the volume of healthy pulmonary tissue was 73%. This result and similar ones motivate clinical recommendations in several countries: COVID-19 patients need to be sorted based on quantitative evaluation of lung lesions. In particular, the Russian Federation operates the following approaches: the grading of COVID-19 pulmonary lesions in CT is performed with the visual semi-quantitative scale from CT1 to CT4 (less than 25 percent, 25-50 percent, 50-75 percent, more than 75 percent of lung volume affected, respectively) Morozov et al. (2020d). Patients with CT3 (severe pneumonia) are to be hospitalized, and CT4 (critical pneumonia) are to be admitted to an intensive care unit.

A visual semi-quantitative scale CT1-CT4 was implemented rather than a fully quantitative one because manual segmentation of the affected lung tissue with a threshold-based method is extremely time-consuming and may take several hours Shan et al. (2020). Thus, automated processing may provide an objective qualitative way to prioritize patients from CT3, CT4 groups.

These two triage strategies, *Identification* and *Severity quantification*, aren't mutually exclusive, and their priority may change depending on the patient population structure and current epidemiological situation.

- An outpatient hospital in an area with a small number of infected patients may rely on *Identification* solely.
- An infectious diseases hospital may use *Severity quantification* to predict the need for artificial pulmonary ventilation and intensive care units.
- Finally, an outpatient hospital during an outbreak needs both systems to identify and isolate COVID patients as well as quantify disease severity and route severe cases accordingly.

This paper explores the automation of both *Identification* and *Severity quantification* intending to create a robust system for all scenarios, see Fig. 2.

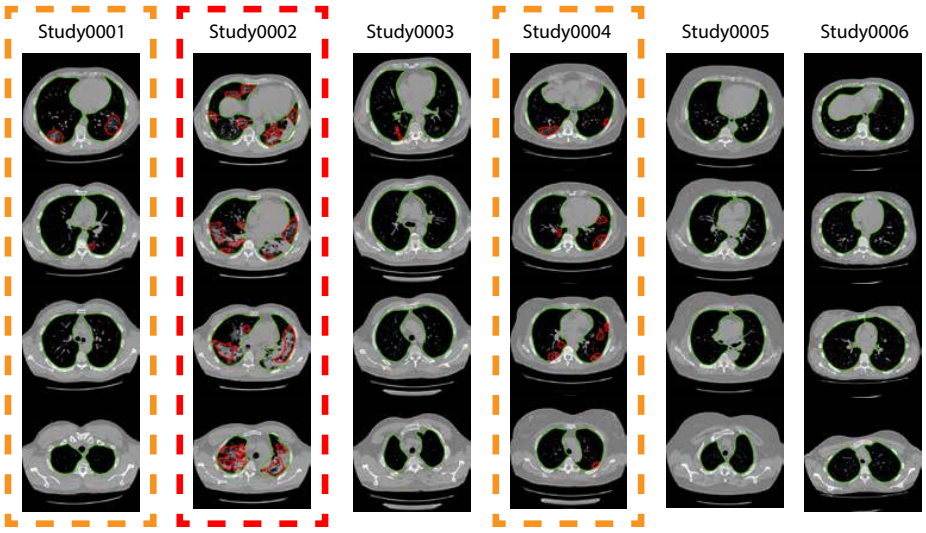


Figure 2: An example of joint COVID-19 identification and severity estimation by the proposed method for several studies.

1.1. Related work

1.1.1. CT analysis for COVID-19 Identification and Severity Estimation

As briefly discussed above, there are two major groups of machine learning-based approaches to analyze COVID-19 chest CTs: identification and severity quantification. In both cases, researchers usually calculate a continuous index of COVID-19 presence or severity (depending on their task). An overview of the existed indices can be found in Tab. 1. Below, we present only some of the existing CT-based algorithms, for a more comprehensive review, we refer to Shi et al. (2020a).

The majority of reviewed works use a pre-trained network for lung extraction or bounding box estimation as a necessary preprocessing step. We will skip the description of this step below for all works.

Identification. Researchers usually treat this problem as binary classification, e.g. COVID-19 versus all other studies. Likely, the most direct way to classify CT images with varying slice thicknesses is to train well established 2D convolutional neural networks. For example, authors of Jin et al. (2020b) trained ResNet-50 He et al. (2016a) to classify images using the obtained lung mask. An interesting and interpretable way to aggregate slice predictions into whole-study predictions was proposed in Gozes et al. (2020a), where the number of affected slices was used as the final output of the model. Also, this work employed Grad-cam Selvaraju et al. (2017) to visualize network attention. A custom slice-level prediction aggregation was proposed in Jin et al. (2020a) to filter out false positives.

Table 1: Overview of continuous output indices proposed in previous works. The Type column denotes index type: COVID-19 *identification*, COVID-19 *severity* or *both*. The Metric column contains reported ROC AUC values unless otherwise indicated. **Remarks**. 1. Accuracy because ROC AUC wasn't reported. 2. The metric was provided for the identification problem only. 3. Pearson correlation. 4. The average volume error, measured in cm^3. 5. The paper doesn't provide an index, Dice score for the output masks is reported.

Paper	Ranking index description	Type	Metric
Li et al. (2020a)	Probabilities of 2.5D ResNet50	Iden.	0.96
Bai et al. (2020)	Probabilities of 2.5D EfficientNet	Iden.	0.95
Wang et al. (2020)	Probabilities of a 3D Resnet-based NN	Iden.	0.973
Wang et al. (2020)	Probabilities of a 3D CNN	Iden.	0.959
Han et al. (2020)	Probabilities of a 3D CNN	Iden.	0.99
Jin et al. (2020b)	Probabilities of ResNet50	Iden.	0.991
Gozes et al. (2020a)	Fractions of affected slices (by 2D ResNet)	Iden.	0.996
Jin et al. (2020a)	Custom aggregation of a 2D CNN predicitons	Iden.	0.979
Chen et al. (2020)	2D Bounding boxes + post-processing	Iden.	Acc[1]. 98.8
Kang et al. (2020)	Probabilities of a NN for raidomics	Iden.	Acc[1]. 95.5
Shi et al. (2020b)	Probabilities of RF for radiomics	Iden.	0.942
Gozes et al. (2020b)	A score based on 2D ResNet attention	Both	0.948[2]
Chaganti et al. (2020)	Affected lung percentage, a combined score	Sev.	Corr.[3] 0.95
Huang et al. (2020)	Affected lung percentage by 2D Unet	Sev.	N/A
Shen et al. (2020)	Affected lung percentage by non trainable CV	Sev.	Corr.[3] 0.81
Shan et al. (2020)	Volume of segm. masks by a 3D CNN	Sev.	Vol.[4] 10.7
Fan et al. (2020)	Segmentation mask	Sev.	Dice[5] 0.597
Tang et al. (2020)	Random Forrest probabilities	Sev.	0.91

The need for a post-training aggregation of slice prediction can be avoided by using 3D convolutional networks, Han et al. (2020); Wang et al. (2020). Wang et al. (2020) proposed a two-headed architecture based on 3D-ResNet. This approach is a way to obtain hierarchical classification as the first head was trained to classify CTs with and without pneumonia. In contrast, the second one aimed to distinguish COVID-19 from other types of pneumonia. Alternatively, slice aggregation may be inserted into network architectures to obtain an end-to-end pipeline, as was proposed in Li et al. (2020a); Bai et al. (2020). Within this setup, all slices are processed by a 2D backbone (ResNet50 for Li et al. (2020a), EfficientNet Tan and Le (2019) for Bai et al. (2020)) while the final classification layers operate with a pooled version of feature maps from all slices.

Segmentation. The majority of papers for tackling severity estimation are segmentation based. For example, the total absolute volume of involved lung parenchyma can be used as a quantitative index Shan et al. (2020). Relative volume (i.e., normalized by the total lung volume) is a more robust approach taking into account the normal variation of lung sizes. Affected lung percentage was estimated in several ways including a non-trainable Computer Vision algorithm Shen et al. (2020), 2D Unet Huang et al. (2020), and 3D Unet Chaganti et al. (2020). Alternatively, an algorithm may predict the severity directly, e.g., with Random Forrest based on a set of radiomics features Tang et al. (2020) or a neural network.

As discussed above, many papers address either COVID-19 identification or severity estimation. However, little research has been conducted to study both tasks simultaneously. Gozes et al. (2020b) proposed an original Grad-cam-based approach to calculate a single attention-based score. Though the authors mentioned both identification and severity quantification in the papers, they do not provide direct quality metrics for the latter.

1.1.2. Deep learning for triage

As mentioned above, we consider triage to be the process of ordering studies to be examined by a radiologist. There are two major scenarios where such an approach can be useful:

- Studies with a high probability of dangerous findings must be prioritized. The most important example is triage within emergency departments, where minutes of acceleration may save lives Faita (2020), but it may be useful for other departments as well. For example, the study Annarumma et al. (2019) estimates the average reporting delay in chest radiographs as 11.2 days for critical imaging findings and 7.6 days for urgent imaging findings.

- The majority of studies do not contain critical findings. This is a common situation for screening programs, e.g., CT-based lung cancer screening Team (2011). In this scenario, triage systems aim to exclude studies with the smallest probability of important findings to reduce radiologists' workload.

Medical imaging may provide detailed information useful for automatic patient triage, as shown in several studies. Annarumma et al. (2019) developed a deep learning-based algorithm to estimate the urgency of imaging findings on adult chest radiographs. The dataset included 470388 studies annotated in an automated way via text report mining. The Inception v3 architecture Szegedy et al. (2016) was used to model clinical priority as ordinal data via solving several binary classification problems as proposed in Lin and Li (2012). The average reporting delay was reduced to 2.7 and 4.1 days for critical and urgent imaging findings correspondingly in a simulation on historical data.

A triage system for screening mammograms, another 2D image modality, was developed in Yala et al. (2019). The authors draw attention to reducing the radiologist's load by maximizing system recall. The underlying architecture is ResNet-18 He et al. (2016a), which was trained on 223109 screening mammograms. The model achieved 0.82 ROC AUC on the whole test population and demonstrated the capability to reduce workload by 20% while preserving the same level of diagnostic accuracy.

Prior research confirms that deep learning may assist in triage of more complex images such as volumetric CT. A deep learning-based system for rapid diagnosis of acute neurological conditions caused by stroke or traumatic brain injury was proposed in Titano et al. (2018). A 3D adaption of ResNet-50 Korolev et al. (2017) analyzed head CT images to predict critical findings. To

train the model, the authors utilized 37236 studies; labels were also generated by text reports mining. The classifier's output probabilities served as ranks for triage, and the system achieved ROC AUC 0.73-0.88. Stronger supervision was investigated in Chang et al. (2018), where authors used 3D masks of all hemorrhage subtypes of 10159 non-contrast CT. The detection and quantification of 5 subtypes of hemorrhages were based on a modified Mask R-CNN He et al. (2017) extended by pyramid pooling to map 3D input to 2D feature maps Lin et al. (2017). More detailed and informative labels combined with an accurately designed method provide reliable performance as ROC AUC varies from 0.85 to 0.99 depending on hemorrhage type and size. A similar finding was reported in De Fauw et al. (2018) for optical coherence tomography (OCT). The authors employed a two-stage approach. First, 3D-Unet Çiçek et al. (2016) was trained on 877 studies with dense 21-class segmentation masks. Then output maps for another 14884 cases were processed by a 3D version of DenseNet Huang et al. (2017) to identify urgent cases. The obtained combination of two networks provided excellent performance achieving 0.99 ROC AUC.

1.2. Contribution

First, we highlighted the need for triage systems of two types for COVID-19 identification and severity quantification. We studied existing approaches and demonstrated that a system trained for one task shows low performance in the other. *Second*, we developed a multitask learning-based approach to create a triage system for both types of problems, which achieves top results in both tasks. *Finally*, we provided a framework for reproducible comparison of various models (see the details below).

1.2.1. Reproducible research

A critical meta-review Wynants et al. (2020) of machine learning models for diagnosis of COVID-19 highlights low reliability and high risk of biased results for all 27 reviewed papers, mostly due to a non-representative selection of control patients and poor analysis of results including possible model overfitting. The authors used Wolff et al. (2019) PROBAST (Prediction model Risk Of Bias ASsessment Tool), a systematic approach to validate the performance of machine learning-based approaches in medicine and identified the following issues.

1. Poor patient structure of the validation set including several studies where control studies were sampled from a different population.
2. Unreliable annotation protocol where only one rater assessed each study without subsequent quality control or the model output influenced annotation.
3. Lack of comparison with other well-established methods for similar tasks.
4. Low reproducibility due to several factors such as unclear model description and incorrect validation approaches (e.g., slice-level prediction rather then study-level prediction).

The authors concluded the paper with a call to share data and code to develop an established system for validating and comparing different models collaboratively.

Though Wynants et al. (2020) is an early review and does not include many properly peer-reviewed papers mentioned above, we agree that current algorithmic research lacks reproducibility. Indeed, only one paper Fan et al. (2020) among 18 papers from section 1.1.1. We aim to follow best practices of reproducible research and address these issues in the following way

1. We select fully independent validation set where all COVID-19 positive and COVID-19 negative cases were retrieved collected from the same population and the same healthcare system, see details in Sec. 3.5
2. Two raters annotated the test data independently, and one experienced meta-rater validated the masks. If raters contours aren't aligned, the meta-rater requested annotation correction[2], also see Sec. 3.5.
3. We carefully selected several key ideas from the reviewed works and implemented them within the same pipeline as our method, see Sec. 2.
4. We publicly release all the code in order to provide the community a way to fully reproduce our experiments.

Finally, we used solely open datasets. The training was done on public CT studies and annotations. We have also annotated a set of publicly available COVID-19 images and will establish an online leaderboard to easily compare different models based on our annotations.

2. Method

We evaluated several baseline methods implementing core ideas from previous works:

- As a common preprocessing step for all networks, we use a CNN to segment lungs, see Sec. 2.1.1.
- A simple non trainable Computer Vision (CV) baseline, similar to Shen et al. (2020).
- 2D U-Net and 3D U-Net models trained on COVID-19 cases only, similar to Huang et al. (2020), Shan et al. (2020).
- 2D U-Net trained on COVID-19 cases plus a large external dataset with non-COVID-19 pathologies as suggested, e.g., in Jin et al. (2020a).
- A 2.5D model based on ResNet, similar to Li et al. (2020a).

[2]At the moment of submission, the second round of annotation correction hasn't been finalized, and not all discordant masks have been fixed. We believe that it only lowers our Dice scores as raters' quantification of severity match each other almost perfectly.

2.1. COVID-19 severity quantification

To triage patients by affected lung fraction we need to evaluate lungs' volumes and total lesion volume in each lung. For these purposes we solve two tasks: left and right lungs segmentation and lesions segmentation. Since we obtain lungs masks and lesions masks, we calculate the lesion volume to lung volume ratio for each lung and use the maximum of two ratios as a final score for ranking.

2.1.1. Lungs segmentation

We segment lungs in two steps: first we predict single binary mask for both lungs, then we split the obtained mask into separate left and right lungs' masks.

Binary segmentation is performed via fully-convolutional neural network in a standard fashion. Details of the architecture and training setup are given in Section 4.2.

On the second step voxels within the lungs are clustered using k-means algorithm with $k = 2$ and Euclidian distance as a metric between voxels. We treat the larger cluster as a right lung, and smaller as a left one.

2.1.2. Lesions segmentation

Threshold-based. As a dummy baseline for lesions segmentation we choose the thresholding-based method which exploits the knowledge that pathological tissues are more dense than healthy ones and therefore corresponding CT voxels have greater intensities in Hounsfield Units. The method consists of multiple steps. The first step implements thresholding: voxels with intensity value between $\mathrm{HU}_{\min}$ and $\mathrm{HU}_{\max}$ within the lung mask are assigned to positive class. At the second step we apply Gaussian blur with smoothing parameter σ to the resulting binary mask and reassign the positive class to voxels with values greater than 0.5. At the last step we remove 3D binary connected components with volumes less then $V_{\min}$. The hyperparameters $\mathrm{HU}_{\min}$, $\mathrm{HU}_{\max}$, σ and $V_{\min}$ are chosen via grid-search in order to maximize the average Dice score (see Section 5.1) between predicted and ground truth lesions masks for series from training dataset.

U-Net. The de facto standard approach for medical image segmentation is the U-Net model Ronneberger et al. (2015). We trained two U-Net-based architectures for lung parenchyma involvement segmentation which we refer to as *2D U-Net* and *3D U-Net*. The 2D U-Net independently processes the axial slices of the input 3D series, while the 3D U-Net processes the hole series in a single step. For each model we replace plain 2D and 3D convolutional layers with 2D and 3D residual convolutional blocks He et al. (2016b), correspondingly. Both models were trained using the standard binary cross entropy loss (see other details in Section 4.3).

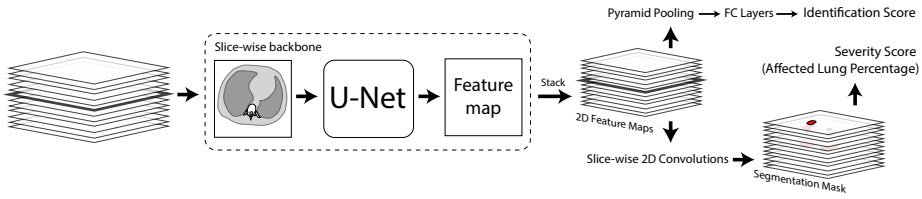


Figure 3: Schematic representation of the multitask model pipeline. *Identification* score is the probability of being a COVID-19 study; *Severity* score is calculated using predicted lesion masks and precomputed lung masks.

2.2. COVID-19 Identification

The triage problem statement implies identification of the studies with the pathology of interest. Therefore our method is supposed to be able to separate CT series with lesions caused by COVID-19 from series with other pathological findings and series of healthy subjects.

Segmentation-based. Possible approach is to base the decision rule on the segmentation results: classify a series as positive if the predicted fraction of lung parenchyma involvement exceeds some threshold. However, our experiments show that this approach leads to a trade-off between severity estimation and identification qualities: models which yield the best ranking and segmentation results perform worse in terms of classification, and vice versa (see Section 5.2).

ResNet50. Another approach is to tackle the classification task separately from segmentation and explicitly predict the probability that a given study contains lesions caused by COVID-19. The obvious advantage of this strategy is that we only need weak labels for model training, which are much more available than ground truth segmentation.

In order to assess the performance of this approach we followed the authors of Li et al. (2020a); Bai et al. (2020) and trained the ResNet50 He et al. (2016b) which takes a series of axial slices as input and independently extracts feature vectors for each slice. After that the feature vectors are combined by global max-pooling operation along all the slices. The resulting vector is fed into a fully connected layer followed by sigmoid activation which predicts the probability of lungs parenchyma involvement presence. This network is trained end-to-end using binary cross entropy loss between then final prediction and the weak label of the input series (see other details in Section 4.4).

2.3. Proposed multitask model

Finally, we propose to simultaneously solve both COVID-19 severity estimation and identification tasks via a single two-headed convolutional neural network. Its architecture is shown in Fig. 3. It starts with a U-Net based backbone consisting of 2D residual convolutional blocks, which slicewisely processes

the input 3D series and outputs the 3D spatial feature map of the same size. Then the feature map is separately processed by two heads.

The first head slicewisely applies single 2D residual convolutional block and 2D convolutional layer. Backbone followed by the first head forms exactly the *2D U-Net* architecture. The output of the first head is treated as a lesions segmentation probability map. As earlier, it is used for evaluation of the affected lung percentage.

The second head consists of pyramid pooling layer He et al. (2014), which transforms the 3D spatial feature map to the feature vector, followed by two fully connected layers, which transform the feature vector to the probability of COVID-19 lesions presence.

As a loss function we optimize a weighted combination of binary cross entropies for segmentation and classification (see other details in Section 4.5). Our experiments show that this multitask learning approach, when the both heads share the same features extracted by the backbone, yields the best results for ranking, segmentation and classification tasks.

3. Data

We used several public datasets in our experiments:

- LUNA16 and NSCLC to create a robust lung segmentation algorithm.
- Mosmed-1110, MedSeg-29 and NSCLC to train all the triage models.
- A subset from Mosmed-Lung-Cancer-500 and Mosmed-20 to assess the final performance of all the methods.

3.1. Mosmed-1110

1110 CT scans from Moscow outpatient clinics were collected from 1st of March, 2020 to 25th of April, 2020, within the framework of outpatient computed tomography centers in Moscow, Russia Morozov et al. (2020a).

Scans were performed on *Canon (Toshiba) Aquilion 64* units in with standard scanner protocols and, particularly 0.8 mm inter-slice distance. However, the public version of the dataset contains every 10th slice of the original study, so the effective inter-slice distance is 8mm.

The quantification of COVID-19 severity in CT was performed with the visual semi-quantitative scale adopted in the Russian Federation and Moscow in particular Morozov et al. (2020d). According to this grading the dataset contains 254 images without COVID-19 symptoms. The rest is split into 4 categories: CT1 (afected lung percentage 25% or below, 684 images), CT2 (from 25% to 50%, 125 images), CT3 (from 50% to 75%, 45 images), CT4 (75% and above, 2 images).

Radiologists performed an initial reading of CT scans in clinics, after which experts from the advisory department of Center for Diagnostics and Telemedicine (CDT) independently conducted the second reading as a part of total audit targeting all CT studies with suspected COVID-19.

Additionally, 50 CT scans were has been annotated with binary pixel masks depicting regions of interest (ground-glass opacities and consolidation)

3.2. MedSeg-29

MedSeg web-site[3] shares 2 publicly available datasets of annotated volumetric CT images. The first dataset consists of 9 volumetric CT scans from here[4] that were be converted from JPG to Nifti format.The annotations of this dataset include lung masks and COVID-19 masks segmented by a radiologist. The second dataset consists of 20 volumetric CT scans shared by Jun et al. (2020). The left and rights lungs, and infections are labeled by two radiologists and verified by an experienced radiologist.

3.3. NSCLC-Radiomics

NSCLC-Radiomics dataset Kiser et al. (2020); Aerts et al. (2015) represents a subset of The Cancer Imaging Archive NSCLC Radiomics collection Clark et al. (2013). It contains left and right lungs segmentations annotated on 3D thoracic CT series of 402 patients with diseased lungs. Pathologies — tumors, atelectasis, and effusion — are included in the lungs volumes masks. Pleural effusion is also delineated separately, when present. However, we used only united lungs binary masks in our experiments.

Automatic approaches for lungs segmentation often perform inconsistently for patients with diseased lungs, while it is usually the main case of interest. This dataset was prepared in order to train or benchmark the lungs segmentation algorithms which are aimed to be more robust for pathological cases.

3.4. LUNA16

LUNA16 Jacobs et al. (2016) is a public dataset for cancerous lung nodules segmentation. It includes 888 annotated 3D thoracic CT scans from the LIDC/IDRI database Armato III et al. (2011). Scans widely differ by scanner manufacturers (17 scanner models), slice thicknesses (from 0.6 to 5.0 mm), in-plane pixel resolution (from 0.461 to 0.977 mm), and other parameters. Annotations for every image contain binary masks for the left and right lungs, the trachea and main stem bronchi, and the cancerous nodules. The lung and trachea masks were originally obtained using an automatic algorithm van Rikxoort et al. (2009) and the lung nodules were annotated by 4 radiologists Armato III et al. (2011). During the preliminary experiments we excluded 7 cases with absent or completely broken lung masks and extremely noisy scans.

[3] `https://medicalsegmentation.com/covid19/`
[4] `https://radiopaedia.org/articles/covid-19-3`

3.5. Hold-out test dataset

Mosmed-20. It is a dataset Morozov et al. (2020c) of 42 CT studies collected from 20 patients in a infectious diseases hospital during the second half of February 2020, at the beginning of Russian outbreak. We removed 4 cases with movement artifacts. The remaining 38 cases were split into healthy (5) and COVID-19 (27 CT1, 2 CT2, 3 CT3, 1 CT4) cases. As we see, at the beginning of the outbreak the majority of cases have mild severity and the resulted structure represents a typical out-patient clinic during the pandemic.

It also important to note that Mosmed-1100 was collected from a cloud PACS which connects all Moscow out-patient clinics. In-patient clinincs are not connected to this PACS. Finally, Mosmed-1100 collection were initiated 1-2 weeks after collection of Mosmed-20, so studies duplication is almost impossible.

Mosmed-Lung-Cancer-500. In a public dataset Morozov et al. (2020b) containing 500 chest CT scans randomly selected from patients over 50 years of age, 63 CT scans were found to have no annotated nodules. After the second reading, 36 patients with pathological conditions not corresponding to the lung nodules (segmental and lobar pneumonia, lung atelectasis) were found. The remaining 27 studies were without pathological changes in the lungs.

Covid-19 masks annotation protocol. Two raters (radiologists with 2 and 5 years of experience), independently tagged with binary pixel masks depicting regions of interest (ground-glass opacities and consolidation) on each slice using an internal contour-based annotation tool. Then the audit was conducted by a third rater (10 years of experience), and each discrepancy was analyzed until a consensus was reached.

4. Experiments

4.1. Preprocessing

In all our experiments we use the same preprocessing applied separately for each axial slice: rescaling to a pixel spacing of 2×2mm and intensity normalization to the $[0, 1]$ range.

Additionally, in our lesions segmentation and classification experiments we crop the input series to the bounding box of the lungs' mask predicted by our lungs segmentation network.

4.2. Lungs segmentation

For the lungs segmentation task we chose a basic U-Net Ronneberger et al. (2015) architecture with 2D convolutional layers, individually applied to each axial slice of an incoming series.

The model was trained on LUNA16 and NSCLC datasets for 16k batches of size 30. We used Adam Kingma and Ba (2014) optimizer with default parameters and an initial learning rate of 0.001, which is decreased to 0.0001 after 8k batches.

We assess the model's performance using 3-fold cross-validation and obtain a Dice score of 0.976 ± 0.023. The resulting network is used to delineate the lungs in all our subsequent experiments, as well as to obtain the ground truth affected lungs percentages on the test dataset described in Section 3.5.

4.3. Lesions segmentation

Global Thresholding. The hyperparameters that showed the best performance are the following: $\mathrm{HU}_{\min} = -700$, $\mathrm{HU}_{\max} = 300$, $\sigma = 4$ and $V_{\min} = 0.1\%$ of lungs' volume.

2D U-Net. We train the *2D U-Net* by using the 79 available 3D series with ground truth segmentations: 50 from Mosmed-1110 and 29 from MedSeg-29.

The network is trained for 30k batches of size 30 using Adam Kingma and Ba (2014) optimizer with default parameters and a learning rate of 0.0001.

2D U-Net+. Additionally, we train the same architecture on a larger dataset containing more normal cases by including both LUNA16 and NSCLC datasets (which do not contain cases with COVID-19) as well as all 254 normal cases from Mosmed-1110. We refer to it as *2D U-Net+*.

3D U-Net. Similarly, the *3D U-Net* architecture was trained on the same 79 series. We optimize the network via plain stochastic gradient descent for 10k batches of size 16 using a learning rate of 0.01.

4.4. ResNet50

We use the ResNet50 implementation from torchvision[5] and change the input channels from 3 to 1, because in our case the input images are grayscale. The network is trained on all available training data for 60k batches of size 5. Adam Kingma and Ba (2014) optimizer was used with default parameters and a learning rate of 0.0001.

4.5. Multitask learning

Our multitask architecture was trained for 60k batches of size 5 using Adam Kingma and Ba (2014) optimizer with default parameters and a learning rate of 0.0001.

Similarly to the *2D U-Net*, the segmentation head is trained on 79 images with available segmentation. The classification head, however, is trained on all available training data. Also, in order to compensate for loss imbalance, we multiply by 0.1 the loss from the classification head.

[5] `https://github.com/pytorch/vision`

5. Results

In this section we assess the performance of all the methods described in Section 2 on three tasks: COVID-19 identification, ranking by affected lung percentage and lesions segmentation.

5.1. Metrics

We use areas under the two ROC-curves (ROC AUC) to assess quality of classification of studies into positive, which have COVID-19 lesions, and negative, which may have other pathologies as well as have no critical findings. First ROC-curve is evaluated on the hole test dataset described in 3.5. The second one is evaluated on the subsample of this dataset, which contains only the studies of infected by COVID-19 or healthy subjects, while studies with other pathological findings are excluded. ROC-curves are obtained by thresholding the predicted affected lungs percentage for segmentation-based methods, and by thresholding the predicted probabilities for ResNet50 and multitask models.

We evaluate the quality of ranking studies by COVID-19 severity on the subsample of the test dataset, which contains only studies with COVID-19 lesions. As a quality metric we use Spearman's rank correlation coefficient (Spearman's ρ) between the ground truth affected lungs fractions $\mathbf{y}^{\text{true}}$ and the predicted fractions $\mathbf{y}^{\text{pred}}$. It is defined as

$$\rho(\mathbf{y}^{\text{true}}, \mathbf{y}^{\text{pred}}) = \frac{\text{cov}(\text{rg}(\mathbf{y}^{\text{true}}), \text{rg}(\mathbf{y}^{\text{pred}}))}{\sigma(\text{rg}(\mathbf{y}^{\text{true}})) \cdot \sigma(\text{rg}(\mathbf{y}^{\text{pred}}))},$$

where $\text{cov}(\cdot, \cdot)$ is a sample covariance, $\sigma(\cdot)$ is a sample standard deviation and $\text{rg}(\mathbf{y})$ is the vector of ranks, i.e. resulting indices of $\mathbf{y}$ elements after their sorting in the ascending order. We also report Spearman's ρ for ResNet50 model, calculated between the ground truth fractions and the predicted probabilities.

To evaluate the lungs parenchyma involvement segmentation quality we use standard Dice score coefficient between the predicted and the ground truth segmentation masks.

5.2. Segmentation-based methods

In this subsection we discuss the performance of four methods: threshold-based baseline, *3D U-Net*, *2D U-net* and *2D U-Net+*. The metrics values can be seen in Tab. 2.

The first method is HU *thresholding* with postprocessing. We expect two major weaknesses of this approach: False Positive (FP) predictions on the vessels and bronchi, and inability to distinguish COVID-19 related lesions from other pathologies. It is clearly seen from the extremely low ROC-AUC classification scores of 0.65 for COVID vs Healthy task and 0.57 for COVID vs All others task. One could also notice massive FP predictions even in healthy cases (see Fig. 4, D), we assume it is the main reason of poor classification results. Though, the method often provides a reasonable segmentation of the lesion area (Fig. 4, B).

Table 2: Quantitative comparison of all the methods discussed in Section 2. Trade-off between COVID-19 identification and severity estimation qualities is observered for segmentation-based methods. The multitask model yields the better results than models which separately tackle the segmentation and classification tasks.

	ROC-AUC (COVID-19 vs ·)		Spearman's ρ	Dice Score
	vs Healthy	vs All others *		
Thresholding	0.649	0.567	0.87	0.45
3D U-Net	0.865	0.772	0.94	**0.49**
2D U-Net	0.890	0.830	**0.98**	**0.50**
2D U-Net+	**0.996**	**0.957**	0.76	0.33
ResNet50	**0.966**	**0.934**	0.45	N/A
Proposed	**0.978**	**0.951**	**0.98**	**0.48**

* All others include healthy subjects and subjects with other pathologies.

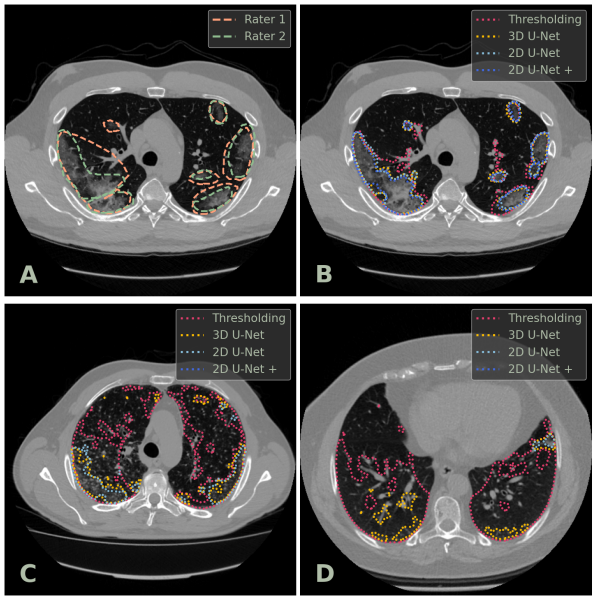


Figure 4: Contouring examples on axial CT slice of COVID case (A, B), other pathology case (C), and healthy case (D). We show predicted contours for four segmentation-based models with all corresponding details discussed in 5.2. For the COVID case we also include the ground truth masks from both Raters (A).

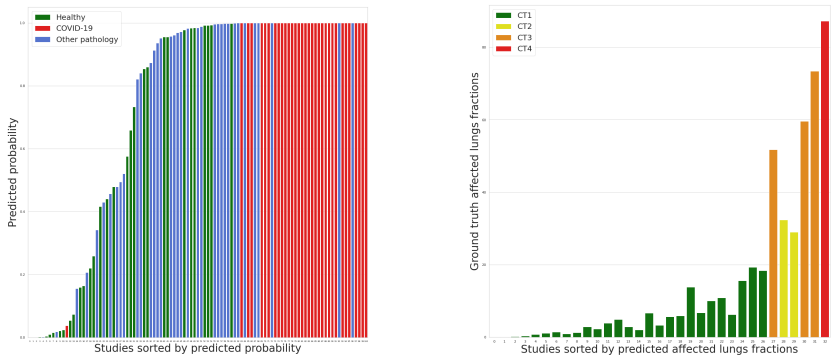


Figure 5: COVID-19 triage: joint identification (left) and severity quantification (rigth) of the test datatset via the proposed single multitask neural network.

3D U-Net considerably outperforms the thresholding baseline in terms of each quality metric. But it still makes a large number of mistakes, e.g. segments other pathologies and healthy tissue (Fig. 4, C and D respectively). We assume that the relatively poor performance of *3D U-Net* comes from the extreme variability of slice spacing in both training and test datasets (ranging from 0.3mm to 8mm).

Our experiments show that *2D U-Net* yields better ranking quality than *3D U-Net*, while their segmentation qualities are comparable. But again it has severe problems with distinguishing COVID cases from other pathologies (see an example on Fig. 4, C). *2D U-Net+* achieves almost perfect classification results (Tab. 2). On the other hand it yields lower segmentation quality than *2D U-Net*. We assume such a performance comes from becoming more conservative and missing small lesions on CT1 studies. As one can note, it gives accurate contours for COVID-19 studies with large areas of lung parenchyma involvement (Fig. 4, B) while mostly corrects FP predictions of the other models on not COVID-19 studies (Fig. 4, C and D).

Overall, *2D U-Net* provides promising COVID-19 severity quantification results, whereas *2D Unet+* demonstrates solid performance for COVID-19 identification at the expense of reduced ranking and segmentation qualities.

5.3. ResNet50

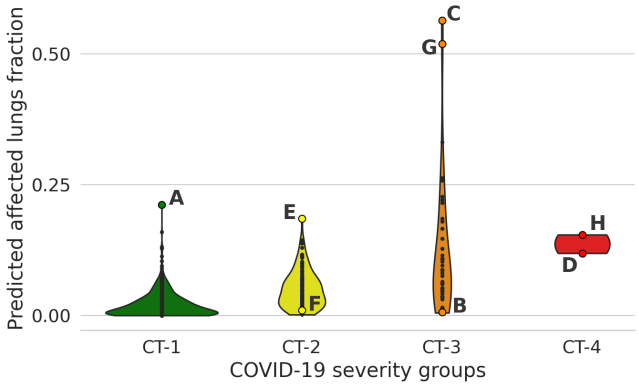
The ResNet50 model successfully copes with the classification of patients into COVID-19 and others. The possible pipeline that exploits this model can work as follows: at first step the ResNet50 identifies the COVID-19 studies, and at the second step they are segmented using the *U-Net 2D* model and ranked according to the predicted affected lungs percentages.

5.4. Multitask model

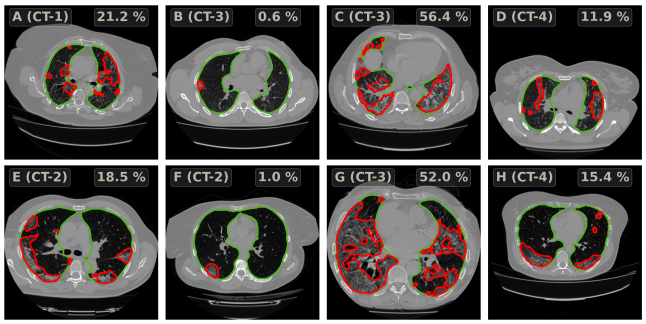
The multitask model described in Section 2.3 yields excellent results according to the metrics for all the tasks (see Tab. 2). We visualize its performance on

COVID-19 identification and ranking by severity tasks in Fig. 5. In the left picture it can be seen, that studies with other pathological findings are the most controversial in terms of model's prediction. The right picture demonstrates that model almost perfectly ranks COVID-19 patients from the most severe cases to the mild ones. The overall pipeline for triage, including preprocessing, lungs segmentation, and multitask inference takes 8s and 20s using nVidia V100 and GTX 980 GPUs respectively.

6. Discussion



(a) Predicted *Severity* for weakly annotated cases from the Mosmed-1110 dataset show inconsistency of visual subjective estimation. The expected ranges are [0; 25) for CT-1, [25; 50) for CT-2, [50; 75) for CT-3, [75; 100] for CT-4.



(b) A representative axial slice of various visually misclassified series from the Mosmed-1110 dataset. Each image contains a severity class (e.g. CT-1) as well as the predicted severity score (e.g. 21.2%).

Figure 6: Analysis of error in subjective visual estimation of *Severity*

We have highlighted two important scores: COVID-19 *Identification* and *Severity* and discussed their priorities in different clinical scenarios. We have

shown that these two scores aren't aligned well. Existing methods operate either with *Identification* or *Severity* and demonstrate deteriorated performance for the other task. We have presented a new method for joint estimation of COVID-19 *Identification* and *Severity* score and proved the proposed method achieves almost maximal scores for both tasks simultaneously. Thus, it can be used in real clinical settings for studies triage and assistance in fast decision making as we show by analyzing ranking order. Finally, we have released the code and used public data for training, so our results are fully reproducible.

As we mentioned in Section 1, radiologists perform the severity classification into groups CT0-CT4 in a visual semi-quantitative fashion. Mosmed-1100 contains a mixed set of labels: 50 binary masks and 1100 multiclass labels. Within our experiments, we binarized these labels and effectively removed weak *Severity* labels. We believe that this information is highly subjective and may contain severe discrepancies. We analyzed mask predictions for the remaining 1050 cases, excluding healthy patients (CT0 group), to validate our hypothesis. The predictions grouped by these weak labels, as shown in Fig. 6a. As we see, a significant number of studies were misclassified during the visual semi-quantitative analysis and subsequent second reading. Several studies with extreme mismatch are visualized in Fig. 6b. We believe that this result is the best evidence that deep-learning-based medical image analysis algorithms, including the proposed method, are great intelligent radiologists assistants in providing the best medical care.

References

Aerts, H., Velazquez, E.R., Leijenaar, R.T., Parmar, C., Grossmann, P., Cavalho, S., Bussink, J., Monshouwer, R., Haibe-Kains, B., Rietveld, D., et al., 2015. Data from nsclc-radiomics. The cancer imaging archive .

Annarumma, M., Withey, S.J., Bakewell, R.J., Pesce, E., Goh, V., Montana, G., 2019. Automated triaging of adult chest radiographs with deep artificial neural networks. Radiology 291, 196–202.

Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., et al., 2011. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. Medical physics 38, 915–931.

Bai, H.X., Wang, R., Xiong, Z., Hsieh, B., Chang, K., Halsey, K., Tran, T.M.L., Choi, J.W., Wang, D.C., Shi, L.B., et al., 2020. Ai augmentation of radiologist performance in distinguishing covid-19 from pneumonia of other etiology on chest ct. Radiology , 201491.

Bernheim, A., Mei, X., Huang, M., Yang, Y., Fayad, Z.A., Zhang, N., Diao, K., Lin, B., Zhu, X., Li, K., et al., 2020. Chest ct findings in coronavirus disease-19 (covid-19): relationship to duration of infection. Radiology , 200463.

Chaganti, S., Balachandran, A., Chabin, G., Cohen, S., Flohr, T., Georgescu, B., Grenier, P., Grbic, S., Liu, S., Mellot, F., et al., 2020. Quantification of tomographic patterns associated with covid-19 from chest ct. arXiv preprint arXiv:2004.01279 .

Chang, P.D., Kuoy, E., Grinband, J., Weinberg, B.D., Thompson, M., Homo, R., Chen, J., Abcede, H., Shafie, M., Sugrue, L., et al., 2018. Hybrid 3d/2d convolutional neural network for hemorrhage evaluation on head ct. American Journal of Neuroradiology 39, 1609–1616.

Chen, J., Wu, L., Zhang, J., Zhang, L., Gong, D., Zhao, Y., Hu, S., Wang, Y., Hu, X., Zheng, B., et al., 2020. Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography: a prospective study. medRxiv .

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3d u-net: learning dense volumetric segmentation from sparse annotation, in: International conference on medical image computing and computer-assisted intervention, Springer. pp. 424–432.

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., et al., 2013. The cancer imaging archive (tcia): maintaining and operating a public information repository. Journal of digital imaging 26, 1045–1057.

Colombi, D., Bodini, F.C., Petrini, M., Maffi, G., Morelli, N., Milanese, G., Silva, M., Sverzellati, N., Michieletti, E., 2020. Well-aerated lung on admitting chest ct to predict adverse outcome in covid-19 pneumonia. Radiology , 201433.

De Fauw, J., Ledsam, J.R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., ODonoghue, B., Visentin, D., et al., 2018. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature medicine 24, 1342–1350.

Faita, F., 2020. Deep learning in emergency medicine: Recent contributions and methodological challenges. Emergency Care Journal 16.

Fan, D.P., Zhou, T., Ji, G.P., Zhou, Y., Chen, G., Fu, H., Shen, J., Shao, L., 2020. Inf-net: Automatic covid-19 lung infection segmentation from ct scans. arXiv preprint arXiv:2004.14133 .

Fang, Y., Zhang, H., Xie, J., Lin, M., Ying, L., Pang, P., Ji, W., 2020. Sensitivity of chest ct for covid-19: comparison to rt-pcr. Radiology , 200432.

Gozes, O., Frid-Adar, M., Greenspan, H., Browning, P.D., Zhang, H., Ji, W., Bernheim, A., Siegel, E., 2020a. Rapid ai development cycle for the coronavirus (covid-19) pandemic: Initial results for automated detection & patient monitoring using deep learning ct image analysis. arXiv preprint arXiv:2003.05037 .

Gozes, O., Frid-Adar, M., Sagie, N., Zhang, H., Ji, W., Greenspan, H., 2020b. Coronavirus detection and analysis on chest ct with deep learning. arXiv preprint arXiv:2004.02640 .

Han, Z., Wei, B., Hong, Y., Li, T., Cong, J., Zhu, X., Wei, H., Zhang, W., 2020. Accurate screening of covid-19 using attention based deep 3d multiple instance learning. IEEE Transactions on Medical Imaging , 1–1.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn, in: Proceedings of the IEEE international conference on computer vision, pp. 2961–2969.

He, K., Zhang, X., Ren, S., Sun, J., 2014. Spatial pyramid pooling in deep convolutional networks for visual recognition. CoRR abs/1406.4729. URL: http://arxiv.org/abs/1406.4729, arXiv:1406.4729.

He, K., Zhang, X., Ren, S., Sun, J., 2016a. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

He, K., Zhang, X., Ren, S., Sun, J., 2016b. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700–4708.

Huang, L., Han, R., Ai, T., Yu, P., Kang, H., Tao, Q., Xia, L., 2020. Serial quantitative chest ct assessment of covid-19: deep-learning approach. Radiology: Cardiothoracic Imaging 2, e200075.

Jacobs, C., Setio, A.A.A., Traverso, A., van Ginneken, B., 2016. Lung nodule analysis 2016. URL: https://luna16.grand-challenge.org.

Jin, C., Chen, W., Cao, Y., Xu, Z., Zhang, X., Deng, L., Zheng, C., Zhou, J., Shi, H., Feng, J., 2020a. Development and evaluation of an ai system for covid-19 diagnosis. medRxiv .

Jin, S., Wang, B., Xu, H., Luo, C., Wei, L., Zhao, W., Hou, X., Ma, W., Xu, Z., Zheng, Z., et al., 2020b. Ai-assisted ct imaging analysis for covid-19 screening: Building and deploying a medical ai system in four weeks. medRxiv .

Jun, M., Cheng, G., Yixin, W., Xingle, A., Jiantao, G., Ziqi, Y., Minqing, Z., Xin, L., Xueyuan, D., Shucheng, C., Hao, W., Sen, M., Xiaoyu, Y., Ziwei, N., Chen, L., Lu, T., Yuntao, Z., Qiongjie, Z., Guoqiang, D., Jian, H., 2020. COVID-19 CT Lung and Infection Segmentation Dataset. URL: https://doi.org/10.5281/zenodo.3757476, doi:10.5281/zenodo.3757476.

Kang, H., Xia, L., Yan, F., Wan, Z., Shi, F., Yuan, H., Jiang, H., Wu, D., Sui, H., Zhang, C., et al., 2020. Diagnosis of coronavirus disease 2019 (covid-19) with structured latent multi-view representation learning. IEEE transactions on medical imaging .

Kherad, O., Moret, B.M., Fumeaux, T., 2020. Computed tomography (ct) utility for diagnosis and triage during covid-19 pandemic. Revue medicale suisse 16, 955.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

Kiser, K., Ahmed, S., Stieb, S., et al., 2020. Data from the thoracic volume and pleural effusion segmentations in diseased lungs for benchmarking chest ct processing pipelines [dataset]. The Cancer Imaging Archive .

Korolev, S., Safiullin, A., Belyaev, M., Dodonova, Y., 2017. Residual and plain convolutional neural networks for 3d brain mri classification, in: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), IEEE. pp. 835–838.

Li, L., Qin, L., Xu, Z., Yin, Y., Wang, X., Kong, B., Bai, J., Lu, Y., Fang, Z., Song, Q., et al., 2020a. Artificial intelligence distinguishes covid-19 from community acquired pneumonia on chest ct. Radiology , 200905.

Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K.S., Lau, E.H., Wong, J.Y., et al., 2020b. Early transmission dynamics in wuhan, china, of novel coronavirus–infected pneumonia. New England Journal of Medicine .

Lin, H.T., Li, L., 2012. Reduction from cost-sensitive ordinal ranking to weighted binary classification. Neural Computation 24, 1329–1367.

Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2117–2125.

Mei, X., Lee, H.C., Diao, K.y., Huang, M., Lin, B., Liu, C., Xie, Z., Ma, Y., Robson, P.M., Chung, M., et al., 2020. Artificial intelligence–enabled rapid diagnosis of patients with covid-19. Nature Medicine , 1–5.

Morozov, S., Andreychenko, A., Pavlov, N., Vladzymyrskyy, A., Ledikhova, N., Gombolevskiy, V., Blokhin, I., Gelezhe, P., Gonchar, A., Chernina, V.Y., 2020a. Mosmeddata: Chest ct scans with covid-19 related findings dataset. arXiv preprint arXiv:2005.06465 .

Morozov, S., Kulberg, N., Gombolevskiy, V., Ledikhova, N., Sokolina, E., Vladzymyrskyy, A., Bardin, A., 2020b. Mosmeddata: 500 lung cancer chest ct. URL: `https://storage.yandexcloud.net/selftest/For_Publication_v3.zip`.

Morozov, S., Protsenko, D., Smetanina, S., Ambrosi, O., Andreychenko, A., Balanyuk, E., Vladzymyrskyy, A., Gombolevskiy, V., Ledikhova, N., Lobanov, M., Pavlov, N., 2020c. Mosmeddata: Imaging studies of patients with covid-19 infection, 2020, v. 1.0. URL: `https://mosmed.ai/datasets/covid19`.

Morozov, S.P., Protsenko, D., Smetanina, S.e.a., 2020d. Imaging of coronavirus disease (covid-19): Organization, methodology, interpretation: Preprint no. cdt - 2020 - ii. version 2 of 17.04.2020.

van Rikxoort, E.M., de Hoop, B., Viergever, M.A., Prokop, M., van Ginneken, B., 2009. Automatic lung segmentation from thoracic computed tomography scans using a hybrid approach with error detection. Medical physics 36, 2934–2947.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer. pp. 234–241.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, pp. 618–626.

Shan, F., Gao, Y., Wang, J., Shi, W., Shi, N., Han, M., Xue, Z., Shi, Y., 2020. Lung infection quantification of covid-19 in ct images with deep learning. arXiv preprint arXiv:2003.04655 .

Shen, C., Yu, N., Cai, S., Zhou, J., Sheng, J., Liu, K., Zhou, H., Guo, Y., Niu, G., 2020. Quantitative computed tomography analysis for stratifying the severity of coronavirus disease 2019. Journal of Pharmaceutical Analysis .

Shi, F., Wang, J., Shi, J., Wu, Z., Wang, Q., Tang, Z., He, K., Shi, Y., Shen, D., 2020a. Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for covid-19. IEEE Reviews in Biomedical Engineering .

Shi, F., Xia, L., Shan, F., Wu, D., Wei, Y., Yuan, H., Jiang, H., Gao, Y., Sui, H., Shen, D., 2020b. Large-scale screening of covid-19 from community acquired pneumonia using infection size-aware classification. arXiv preprint arXiv:2003.09860 .

Sverzellati, N., Milanese, G., Milone, F., Balbi, M., Ledda, R.E., Silva, M., 2020. Integrated radiologic algorithm for covid-19 pandemic. J Thorac Imaging .

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818–2826.

Tan, M., Le, Q.V., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. arXiv preprint arXiv:1905.11946 .

Tang, Z., Zhao, W., Xie, X., Zhong, Z., Shi, F., Liu, J., Shen, D., 2020. Severity assessment of coronavirus disease 2019 (covid-19) using quantitative features from chest ct images. arXiv preprint arXiv:2003.11988 .

Tanne, J.H., Hayasaki, E., Zastrow, M., Pulla, P., Smith, P., Rada, A.G., 2020. Covid-19: how doctors and healthcare systems are tackling coronavirus worldwide. Bmj 368.

Team, N.L.S.T.R., 2011. The national lung screening trial: overview and study design. Radiology 258, 243–253.

Titano, J.J., Badgeley, M., Schefflein, J., Pain, M., Su, A., Cai, M., Swinburne, N., Zech, J., Kim, J., Bederson, J., et al., 2018. Automated deep-neural-network surveillance of cranial images for acute neurologic events. Nature medicine 24, 1337–1341.

Wang, J., Bao, Y., Wen, Y., Lu, H., Luo, H., Xiang, Y., Li, X., Liu, C., Qian, D., 2020. Prior-attention residual learning for more discriminative covid-19 screening in ct images. IEEE Transactions on Medical Imaging .

Wang, X., Deng, X., Fu, Q., Zhou, Q., Feng, J., Ma, H., Liu, W., Zheng, C., 2020. A weakly-supervised framework for covid-19 classification and lesion localization from chest ct. IEEE Transactions on Medical Imaging , 1–1.

Wolff, R.F., Moons, K.G., Riley, R.D., Whiting, P.F., Westwood, M., Collins, G.S., Reitsma, J.B., Kleijnen, J., Mallett, S., 2019. Probast: a tool to assess the risk of bias and applicability of prediction model studies. Annals of internal medicine 170, 51–58.

Wynants, L., Van Calster, B., Bonten, M.M., Collins, G.S., Debray, T.P., De Vos, M., Haller, M.C., Heinze, G., Moons, K.G., Riley, R.D., et al., 2020. Systematic review and critical appraisal of prediction models for diagnosis and prognosis of covid-19 infection. medRxiv .

Yala, A., Schuster, T., Miles, R., Barzilay, R., Lehman, C., 2019. A deep learning model to triage screening mammograms: a simulation study. Radiology 293, 38–46.