Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Medical Image Analysis 71 (2021) 102054

Contents lists available at ScienceDirect

Medical Image Analysis

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

CT-Based COVID-19 triage: Deep multitask learning improves joint identification and severity quantification

Mikhail Goncharov[a,b,1], Maxim Pisov[a,1], Alexey Shevtsov[b,1], Boris Shirokikh[a,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 for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department, Russia
[d] Sklifosovsky Clinical and Research Institute for Emergency Medicine, Moscow, Russia

ARTICLE INFO

Article history:
Received 2 June 2020
Revised 21 March 2021
Accepted 26 March 2021
Available online 1 April 2021

Keywords:
COVID-19
Triage
Convolutional neural network
Chest computed tomography

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 patients with severe COVID-19, thus 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 could provide reasonable quality only for one of these setups. We employ a multitask approach to consolidate both triage approaches and propose a convolutional neural network to leverage all available labels within a single model. In contrast with the related multitask approaches, we show the benefit from applying the classification layers to the most spatially detailed feature map at the upper part of U-Net instead of the less detailed latent representation at the bottom. We train our model on approximately 1500 publicly available CT studies and test it on the holdout dataset that consists of 123 chest CT studies of patients drawn from the same healthcare system, specifically 32 COVID-19 and 30 bacterial pneumonia cases, 30 cases with cancerous nodules, and 31 healthy controls. The proposed multitask model outperforms the other approaches and achieves ROC AUC scores of 0.87 ± 0.01 vs. bacterial pneumonia, 0.93 ± 0.01 vs. cancerous nodules, and 0.97 ± 0.01 vs. healthy controls in *Identification* of COVID-19, and achieves 0.97 ± 0.01 Spearman Correlation in *Severity quantification*. We have released our code and shared the annotated lesions masks for 32 CT images of patients with COVID-19 from the test dataset.

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 (World Health Organization et al., 2020), chest imaging, including computed tomography (CT), is helpful in diagnosis and patient management (Bernheim et al., 2020; Akl et al., 2020; Rubin et al., 2020). Moreover, compared to RT-PCR, CT has higher sensitivity (98% compared to 71% at $p \leq .001$) for some cohorts Fang et al. (2020). Fleischner Society has addressed the role of thoracic imaging in COVID-19, providing recommendations intended to guide medical practitioners with one scenario including medical triage in moderate-to-severe clinical features and a high pretest probability of disease (Rubin et al., 2020). Radiology departments can respond to the pandemic by division into four areas (contaminated, semi-contaminated, buffer, and clean), strict disinfection and management criteria (Huang et al., 2020b). The International Society of Radiology surveyed current practices in managing patients with COVID-19 in 50 radiology departments representing 33 countries across all continents. In symptomatic patients with suspected

* Corresponding author.
E-mail address: m.belyaev@skoltech.ru (M. Belyaev).
[1] Equal contribution

https://doi.org/10.1016/j.media.2021.102054

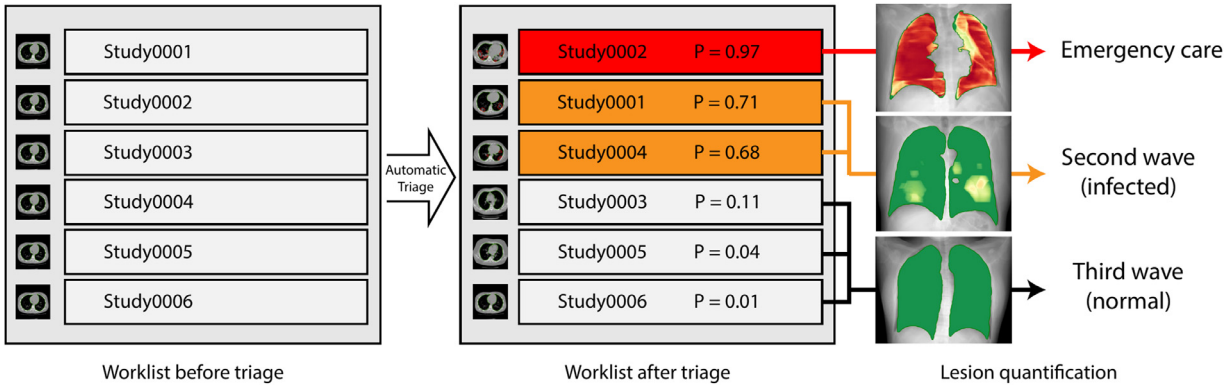

Fig. 1. A schematic representation of the automatic triage process. Left: the chronological order of the studies. Center: re-prioritized order to highlight findings requiring radiologist's attention (*P* denotes COVID-19 *Identification* probability). Right: accompanying algorithm-generated X-ray-like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

COVID-19, imaging was performed in 89% of cases, in 34% of cases - chest CT. Faster results than molecular tests (51%) and easy access (39%) were the main reasons for imaging use (Blažić et al., 2020)

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 Section 1.1.1. However, little research has been conducted to investigate automatic image analysis for triage, i.e. 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 a priority score for triage, e.g., output probability of a classifier or the total lesion volume extracted from a binary segmentation mask. However, these scores 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 adopted the following approach (Morozov et al., 2020c): the volume ratio of lesions in each lung is calculated separately and the maximal ratio is treated as the overall **severity score**. However, manual binary segmentation of the affected lung tissue is extremely time-consuming and may take several hours (Shan et al., 2020). For this reason, a visual semi-quantitative scale was implemented rather than a fully quantitative one. The original continuous score is split up into five categories: from CT-0 to CT-4 with a 25% step so that CT-0 corresponds to normal cohort and CT-4 - to 75%-100% of damaged lung tissue. Patients with CT-3 (severe pneumonia) are hospitalized, and CT-4 (critical pneumonia) are admitted to an intensive care unit. The scale is based on a visual evaluation of approximate lesion volume in both lungs (regardless of postoperative changes).

A retrospective study (Morozov et al., 2020b) analyzed the CT 0–4 scores and lethal outcomes in 13,003 COVID-19 patients. The chance of a lethal outcome increased from CT-0 to CT-4 by 38% on the average (95% CI 17.1–62.6%). Another retrospective analysis (Petrikov et al., 2020) found a significant correlation between an increase of CT grade and clinical condition deterioration ($r = 0.577$).

These two triage strategies, *Identification* and *Severity quantification*, are not 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-19 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.

1.1. Related work

1.1.1. CT Analysis for COVID-19 identification and severity estimation

As briefly discussed above, we consider two problems: COVID-19 identification and severity quantification in chest CTs. In both cases, researchers usually calculate a continuous score of COVID-19 presence or severity, depending on their task. An overview of the existing 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 prepro-

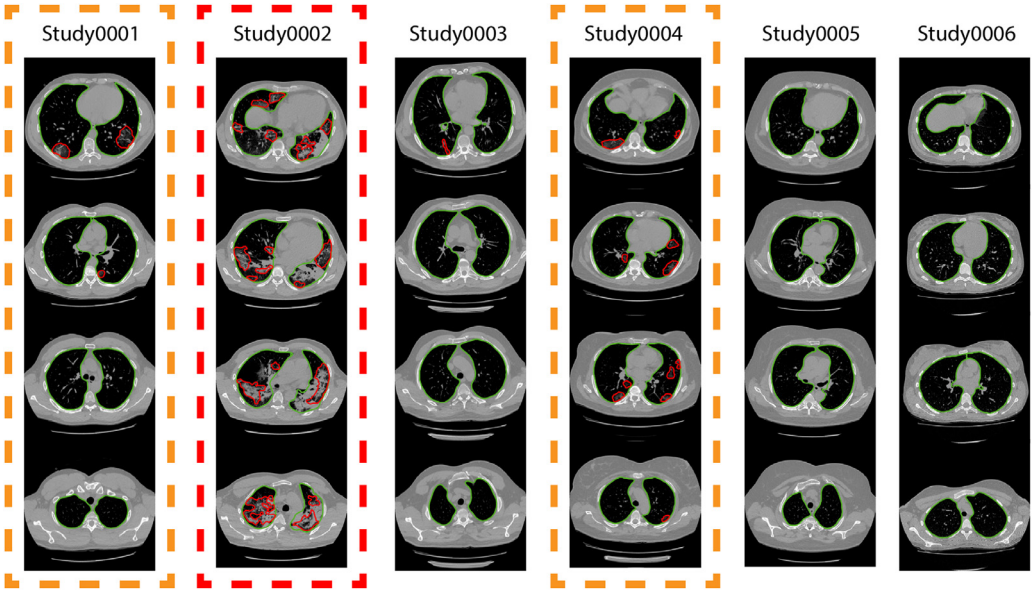

Fig. 2. An example of joint COVID-19 identification and severity estimation by the proposed method for several studies.

Table 1
Overview of continuous output indices proposed in previous works. The Type column denotes score type: COVID-19 *identification*, COVID-19 *severity* or *both*. Type of the Identification is given in brackets COVID vs. : P - Pneumonia, NP - non-Pneumonia, HC - Healthy controls, N - Nodules, C - Cancer. The Metric column contains reported ROC AUC values unless otherwise indicated. **Remarks**. 1. Accuracy because ROC AUC was not 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 does not provide a score, Dice score for the output masks is reported.

Paper	Ranking score description	Type	Metric
Bai et al. (2020)	Probabilities of 2.5D EfficientNet	Iden. (P)	0.95
Kang et al. (2020)	Probabilities of a NN for raidomics	Iden. (P)	Acc.[1] 0.96
Shi et al. (2020b)	Probabilities of RF for radiomics	Iden. (P)	0.94
Li et al. (2020a)	Probabilities of 2.5D ResNet-50	Iden. (P, NP)	0.96
Wang et al. (2020a)	Probabilities of a 3D Resnet-based NN	Iden. (HC, P)	0.97
Han et al. (2020)	Probabilities of a 3D CNN	Iden. (HC, P)	0.99
Jin et al. (2020b)	Probabilities of ResNet-50	Iden. (HC, P, N)	0.99
Jin et al. (2020a)	Custom aggregation of a 2D CNN predicitons	Iden. (HC, P)	0.97
Gozes et al. (2020a)	Fractions of affected slices (by 2D ResNet)	Iden. (HC, C)	0.99
Amine et al. (2020)	Probabilities of 3D U-Net (encoder part)	Iden. (HC, P)	0.97
Wang et al. (2020b)	Probabilities of a 3D CNN	Iden. (HC)	0.96
Chen et al. (2020)	2D Bounding boxes + post-processing	Iden. (other disease)	Acc.[1] 0.99
Gozes et al. (2020b)	A score based on 2D ResNet attention	Both (fever)	0.95[2]
Chaganti et al. (2020)	Affected lung percentage, a combined score	Sev.	Corr.[3] 0.95
Huang et al. (2020a)	Affected lung percentage by 2D U-Net	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] 0.60
Tang et al. (2020)	Random Forrest probabilities	Sev.	0.91

cessing step. We will skip the description of this step below for all works.

Binary classification Researchers usually treat the problem of identification 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) train 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 is proposed in (Gozes et al., 2020a), where the number of affected slices is used as the final output of the model. Also, this work employs Grad-cam (Selvaraju et al., 2017) to visualize network attention. A custom slice-level predictions aggregation is proposed in (Jin et al., 2020a) to filter out false positives.

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., 2020b). (Wang et al., 2020a) propose a two-headed architecture based on 3D ResNet. This approach is a way to obtain hierarchical classification as the first head is trained to classify CTs with and without pneumonia. In contrast, the second one aims 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 proposed in (Li et al., 2020a; Bai et al., 2020). Within this setup, all slices are processed by a 2D backbone (ResNet-50 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 severity score (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 is es-

timated in several ways including a non-trainable computer vision algorithm (Shen et al., 2020), 2D U-Net (Huang et al., 2020a), and 3D U-Net (Chaganti et al., 2020). Alternatively, an algorithm may predict the severity directly, e.g., with Random Forest based on a set of radiomics features (Tang et al., 2020) or a neural network.

Multitask approach 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) propose an original Grad-cam-based approach to calculate a single attention-based score. Though the authors mention both identification and severity quantification in the papers, they do not provide direct quality metrics for the latter. Amine et al. (2020) propose a multi-head architecture to solve both segmentation and classification problems in an end-to-end manner. They use a 2D U-Net backbone with an additional classification head after the encoder part, which takes a latent feature map from the bottom of U-Net as input. Even though they do not tackle the problem of severity identification, they demonstrate that solving two tasks jointly could benefit both. However, they report metrics only for classification and segmentation of 2D axial slices and do not propose an approach to applying their method to the whole 3D CT series.

1.1.2. Deep learning for triage

As mentioned above, we define triage as a process of ordering studies to be examined by a radiologist. There are two major scenarios where such an approach could 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) propose a deep learning-based algorithm to estimate the urgency of imaging findings on adult chest radiographs. The dataset includes 470388 studies annotated in an automated way via text report mining. The Inception v3 architecture (Szegedy et al., 2016) is 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 is 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, has been 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 is trained on 223109 screening mammograms. The model achieves 0.82 ROC AUC on the whole test population and demonstrates 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 is proposed in (Titano et al., 2018). A 3D adaption of ResNet-50 (Korolev et al., 2017) analyzes head CT images to predict critical findings. To train the model, the authors utilize 37236 studies; labels are also generated by text reports mining. The classifier's output probabilities serve as ranks for triage, and the system achieves ROC AUC 0.73-0.88. Stronger supervision is investigated in (Chang et al., 2018), where authors use 3D masks of all hemorrhage subtypes of 10159 non-contrast CT. The detection and quantification of 5 subtypes of hemorrhages are 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 is reported in (De Fauw et al., 2018) for optical coherence tomography (OCT). The authors employ a two-stage approach. First, 3D U-Net (Çiçek et al., 2016) is trained on 877 studies with dense 21-class segmentation masks. Then output maps for another 14884 cases are 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 highlight the need for triage systems of two types: for COVID-19 identification and severity quantification. We study existing approaches and demonstrate that a system trained for one task shows low performance in the other. *Second*, we have developed a multitask learning-based approach to create a single neural network which achieves top results in both triage tasks. In contrast to common multitask architectures, classification layers take the spatially detailed 3D feature map as input and return the single probability for the whole CT series. *Finally*, we provide 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 COVID-19 diagnosis 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 use (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 different populations.
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 than study-level prediction).

The authors conclude 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 COVID-19 algorithmic research lacks reproducibility. We aim to follow the best practices of reproducible research and address these issues in the following way.

1. We selected fully independent test dataset and retrieved all COVID-19 positive and COVID-19 negative cases from the

same population and the same healthcare system, see details in Section 3.5.
2. Two raters annotated the test data independently. If raters contours were not aligned, the meta-rater requested annotation correction, see Section 3.5.
3. We carefully selected several key ideas from the related works and implemented them within the same setup as our method, see Section 2.
4. We publicly released the code to share technical details of the compared architectures[2].

Finally, we use solely open CT images for training and testing. We also annotate and release the lesions masks for the COVID-19 positive cases from the test set, see details in the Section 3.5. Therefore, our experiments are reproducible as they rely on the open data.

2. Method

As discussed in Section 1 method should solve two tasks: identification of COVID-19 cases and ranking them in descending order of severity. Therefore, we organize Section 2 as follows.

- In Section 2.1 we describe lungs segmentation as a common preprocessing step for all methods.
- In Section 2.2 we tackle the severity quantification task. We describe methods which predict segmentation mask of lesions caused by COVID-19 and provide a severity score based on that.
- In Section 2.3 we discuss two straightforward baselines for the identification task. First is to use segmentation results and identify patients with non-empty lesions masks as COVID-19 positive. Second is to use separate neural network for classification of patients into COVID-19 positive or negative. However, as we show in Section 5 these methods yield poor identification quality, especially due to false positive alarms in patients with bacterial pneumonia.
- In Section 2.4 we propose a multitask model which achieves better COVID-19 identification results than the baselines. In particular, as we show in Section 5, this model successfully distinguishes between COVID-19 and bacterial pneumonia cases.
- In Section 2.5 we introduce quality metrics for both identification and severity quantification tasks to formalize the comparison of the methods.

2.1. Lungs segmentation

We segment lungs in two steps. First, we predict single binary mask for both lungs including pathological findings, e.g. ground-glass opacity, consolidation, nodules and pleural effusion. 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 ($k = 2$) with Euclidean distance as a metric between voxels. Then we treat resulting clusters as separate lungs.

2.2. COVID-19 Severity quantification

To quantify COVID-19 severity we solve COVID-19-specific lesions segmentation task. Using predicted lungs' and lesions' masks, we calculate the lesions' to lung's volume ratio for each lung and use the maximum of two ratios as a final severity score for triage, according to recommendations discussed in Section 1.

[2] https://github.com/neuro-ml/COVID-19-Triage

Threshold-based As a baseline for lesions segmentation, we choose a thresholding-based method. As pathological tissues are denser than healthy ones, corresponding CT voxels have greater intensities in Hounsfield Units. The method consists of three steps. The first step implements thresholding: voxels with intensity value between HU_{min} and HU_{max} within the lung mask are assigned to the 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. Finally, we remove 3D binary connected components with volumes smaller than V_{min}. The hyperparameters $HU_{min} = -700$, $HU_{max} = 300$, $\sigma = 4$ and $V_{min} = 0.1\%$ are chosen via a grid-search in order to maximize the average Dice score 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*. *2D U-Net* independently processes the axial slices of the input 3D series. *3D U-Net* processes 3D sub-patches of size $160 \times 160 \times 160$ and then stacks predictions for individual sub-patches to obtain prediction for the whole input 3D series. Thus, we do not need to downsample the input image under the GPU memory restrictions. 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).

2.3. COVID-19 Identification

We formalize COVID-19 identification task as a binary classification of 3D CT series. CT series of patients with verified COVID-19 are positive class. CT series of patients with other lung diseases, e.g. bacterial pneumonia, non-small cell lung cancer, etc., as well as normal patients are negative class.

Segmentation-based One possible approach is to base the decision rule on the segmentation results: classify a series as positive if the segmentation-based severity score exceeds some threshold. We show that this leads to a trade-off between severity quantification and identification qualities: models which yield the best ranking results perform worse in terms of classification, and vice versa. Moreover, despite some segmentation-based methods accurately classify COVID-19 positive and normal cases, all of them yields a significant number of false positives in patients with bacterial pneumonia (see Section 5.1).

ResNet-50 Another approach is to tackle the classification task separately from segmentation and explicitly predict the probability that a given series is COVID-19 positive. The advantage of this strategy is that we only need weak labels for model training, which are much more available than ground truth segmentations.

To assess the performance of this approach we follow the authors of (Li et al., 2020a; Bai et al., 2020) and train the ResNet-50 (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 via a pyramid max-pooling operation (He et al., 2014) along all the slices. The resulting vector is passed into two fully connected layers followed by sigmoid activation which predicts the final COVID-19 probability for the whole series. In our paper, we denote this architecture as *ResNet-50* (see other details in Section 4.4).

2.4. Multitask

Baselines for the identification task described in Section 2.3 do not perform well, as we show in Section 5. Therefore, we propose

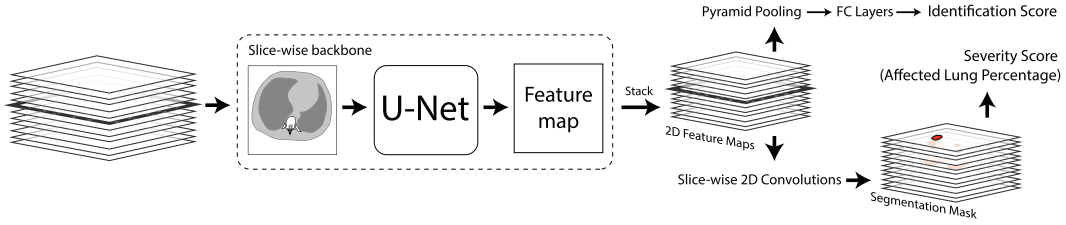

Fig. 3. Schematic representation of the *Multitask-Spatial-1* model. *Identification* score is the probability of being a COVID-19 positive series; *Severity* score is calculated using predicted lesions' mask and precomputed lungs' masks.

to solve the identification task simultaneously with the segmentation task via a single two-headed convolutional neural network.

The segmentation part of the architecture is slice-wise *2D U-Net* model. As earlier, its output is used for the evaluation of the severity score.

The classification head shares a common intermediate feature map (per slice) with the segmentation part. These feature maps are stacked and aggregated into a feature vector via a pyramid pooling layer (He et al., 2014). Finally, two fully connected layers followed by sigmoid activation transform the feature vector to the COVID-19 probability.

Following (Amine et al., 2020), the shared feature maps can be the outputs of the U-Net's encoder and have no explicit spatial structure in the axial plane. We refer to this approach as *Multitask-Latent*. In contrast, we argue that the identification task is connected to the segmentation task and the classification model can benefit from the spatial structure of the input features. Therefore, we propose to share the feature map from the very end of the U-Net architecture, as shown in Fig. 3. We refer to the resulting architecture as *Multitask-Spatial-1*. More generally, shared feature maps can be taken from the l-th upper level of the U-Net's decoder. Together they form a 3D spatial feature map, which is aligned with the input 3D series downsampled in the axial plane by a factor of 2^{l-1}. We denote this approach as *Multitask-Spatial-l*. Since *2D U-Net* architecture has 7 levels, l can vary from 1 to 7.

As a loss function we optimize a weighted combination of binary cross entropies for segmentation and classification (see other details in Section 4).

2.5. Metrics

To assess the quality of classification of patients into positive, i.e. infected by COVID-19, and negative, i.e. with other lung pathologies or normal, we use areas under the ROC-curves (ROC AUC) calculated on several subsamples of the test sample described in Section 3.5.

- The first subsample contains only COVID-19 positive and healthy subjects, while studies with other pathological findings are excluded (ROC AUC COVID-19 vs. Normal).
- The second subsample contains only patients infected by COVID-19 or bacterial pneumonia (ROC AUC COVID-19 vs. Bac. Pneum.).
- The third subsample contains COVID-19 positive patients and patients with lung nodules typical for non-small cell lung cancer (ROC AUC COVID-19 vs. Nodules).
- The last ROC AUC is calculated on the whole test sample (ROC AUC COVID-19 vs. All others).

ROC-curves are obtained by thresholding the predicted probabilities for ResNet-50 and multitask models, and by thresholding the predicted severity score for segmentation-based methods.

We evaluate the quality of ranking studies in order of descending COVID-19 severity on the test subsample, which contains only COVID-19 positive patients. As a quality metric, we use Spearman's rank correlation coefficient (Spearman's ρ) between the severity scores $\mathbf{y}^{\text{true}}$ calculated for ground truth segmentations and the predicted severity scores $\mathbf{y}^{\text{pred}}$. 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 descending order.

To evaluate the COVID-19 lesions segmentation quality we use Dice score coefficient between the predicted and the ground truth segmentation masks. Similar to Spearman's ρ, we evaluate the mean Dice score only for COVID-19 positive cases.

3. Data

We use several public datasets in our experiments:

- NSCLC-Radiomics and LUNA16 to create a robust lung segmentation model.
- Mosmed-1110, MedSeg-29 and NSCLC-Radiomics to train and validate all triage models.
- Mosmed-Test as a hold-out test set for the final evaluation of all models.

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., 2020c). According to this grading, the dataset contains 254 images without COVID-19 symptoms. The rest is split into 4 categories: CT1 (affected 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 the Center for Diagnostics and Telemedicine (CDT) independently conducted the second reading as a part of a total audit targeting all CT studies with suspected COVID-19.

Additionally, 50 CT scans were annotated with binary masks depicting regions of interest (ground-glass opacity 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 a web-site[4] that were 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 — lung cancerous nodules, atelectasis and pleural effusion — are included in the lung volume masks. Pleural effusion and cancerous nodules are also delineated separately, when present.

Automatic approaches for lungs segmentation often perform inconsistently for patients with diseased lungs, while it is usually the main case of interest. Thus, we use NSCLC-Radiomics to create robust for pathological cases lungs segmentation algorithm. Other pathologies, e.g. pneumothorax, that are not presented in NSCLC-Radiomics could also lead to poor performance of lungs segmentation. But the appearance of such pathologies among COVID-19 cases is extremely rare. For instance, it is less than 1% for pneumothorax (Zantah et al., 2020). Therefore, we ignore the possible presence of other pathology cases, while training and evaluating our algorithm.

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). We also exclude 7 cases with absent or completely broken lung masks and extremely noisy scans.

3.5. Mosmed-Test

We ensure the following properties of the test dataset:

- All cases are full CT series without missing slices and/or lacking metadata fields (e.g., information about original Hounsfield units).
- Data for all classes comes from the same population and the same healthcare system to avoid domain shifts within test data.

COVID-19 positive It is a subsample of Mosmed-20[5], 42 CT studies collected from 20 patients in an infectious diseases hospital during the second half of February 2020, at the beginning of the Russian outbreak. We remove 5 cases with artifacts related to patients' movements while scanning. The remaining 37 cases were independently assessed by two raters (radiologists with 2 and 5 years of experience) who have annotated regions of interest (ground-glass opacities and consolidation) via MedSeg[6] annotation tool for every of the 37 Mosmed-Test series. During the annotation process, 5 out of 37 images were identified to have no radiomic signs of COVID-19, so we remove these images from the list of COVID-19 positives. Then, we iteratively verify annotations based on two factors: Dice Score between two rates, and missing large connected components of the mask by one of the raters. The discrepancy between the two raters has been analyzed until the consensus is reached — 0.87 ± 0.17 Dice Score over 32 COVID-19 infected cases. We publicly release the final version of *COVID-19 positive* dataset including both images and annotated lesions masks.

Note, that the Mosmed-20 was collected at inpatient clinics, whereas Mosmed-1110 is a subset of Moscow out-patient clinics database created from two to six weeks later, which guarantees that studies are not duplicated.

Bacterial pneumonia We use 30 randomly selected cases from a dataset (Korb et al., 2021) with 75 chest CT studies with radiological signs of community-acquired bacterial pneumonia in 2019.

Lung nodules We use a subset of MoscowRadiology-CTLungCa-500[7], a public dataset containing 500 chests CT scans randomly selected from patients over 50 years of age. We selected 30 cases randomly among cases with radiologically verified lung nodules.

Normal controls The dataset with healthy patients consists of two parts: 5 Mosmed20 cases mentioned above without radiomic signs of COVID-19, and 26 cases from MoscowRadiology-CTLungCa-500 without lung nodules larger than 5mm and other lung pathologies.

4. Experiments

We design our experiments in order to objectively compare all the triage models described in Section 2. As shown in the Tab. 2, all the methods are trained on the same datasets and evaluated using the mean values and the standard deviations of the same quality metrics defined in Section 2.5 on the same hold-out test dataset described in Section 3.5. We believe, that the experimental design for training neural networks for triage described in Section 4.3 and 4.4 exclude overfitting. All computational experiments were conducted on Zhores supercomputer (Zacharov et al., 2019).

4.1. Preprocessing

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

In our COVID-19 identification and segmentation experiments we crop the input series to the bounding box of the lungs' mask predicted by our lungs segmentation network.

We further show (Section 5) that this preprocessing is sufficient for all the models. Despite the diversity of the training dataset, all the models successfully adapt to the test dataset.

4.2. Lungs segmentation

For the lungs segmentation task we choose a basic U-Net (Ronneberger et al., 2015) architecture with 2D convolutional layers, individually apply to each axial slice of an incoming series. The model was trained on NSCLC-Radiomics and LUNA16 datasets for

[3] https://medicalsegmentation.com/covid19/
[4] https://radiopaedia.org/articles/covid-19-3
[5] https://mosmed.ai/en/datasets/ct_lungcancer_500/
[6] https://www.medseg.ai/
[7] https://mosmed.ai/en/datasets/ct_lungcancer_500/

Table 2
Training, validation and test data splits for all triage models. For each method, we give the optimized training objectives in the corresponding table cells for the training datasets. Every column of Mosmed-Test dataset represents the metrics which are calculated using the corresponding test subset. **Remarks**. 1. *pos. with mask*|*pos.* mean COVID-19 positive cases with or without lesions mask, correspondingly, and *neg.* means COVID-19 negative cases. 2. *DSC* means Dice Score. 3. *AUCs* means ROC AUC COVID-19 vs. All, vs. Normal, vs. Bac. Pneum. and vs. Nodules. 4. *Seg. BCE* and *class. BCE* means segmentation and classification Binary Cross-Entropy correspondingly. 5. ρ means Spearman's ρ. 6. Multitask-Latent, Multitask-Spatial-4, Multitask-Spatial-1.

	Training and validation datasets					Mosmed-test			
	Mosmed-1110			Medseg-29	NSCLC-Radiomics	COVID-19 pos.	Bac. Pneum.	Nodules	Normal
Ground truth[1]	pos. with mask	pos.	neg.	pos. with mask	neg.	pos. with mask	neg.	neg.	neg.
Num. of images	50	806	254	29	402	32	30	30	31
Thresholding	DSC[2]	-		DSC	-	AUCs[3], ρ, DSC	AUCs	AUCs	AUCs
2D U-Net, 3D U-Net	Seg. BCE[4]	-		Seg. BCE	-	AUCs, ρ, DSC	AUCs	AUCs	AUCs
2D U-Net+	Seg. BCE	-		Seg. BCE	Seg. BCE	AUCs, ρ^5, DSC	AUCs	AUCs	AUCs
ResNet-50	-	Class. BCE[4]		-	Class. BCE	AUCs	AUCs	AUCs	AUCs
Multitask models[6]	Seg. BCE	Class. BCE		Seg. BCE	Class. BCE	AUCs, ρ, DSC	AUCs	AUCs	AUCs

16k batches of size 30. We use Adam (Kingma and Ba, 2014) optimizer with default parameters and an initial learning rate of 0.001, which was decreased to 0.0001 after 8k batches.

We assess the model's performance using 3-fold cross-validation and additionally using MedSeg-29 dataset as hold-out set. Dice Score of cross-validation is 0.976 ± 0.023 for both LUNA16 and NSCLC-Radiomics datasets, and 0.962 ± 0.023 only on NSCLC-Radiomics dataset. The latter result confirms our model to be robust to the cases with pleural effusion. Dice Score on MedSeg-29 dataset is 0.976 ± 0.013, which shows the robustness of our model to the COVID-19 cases.

4.3. Lesions segmentation

We use all the available 79 images of COVID-19 positive patients with annotated lesions masks (50 images from Mosmed-1110 and 29 images from MedSeg-29) to train the threshold-based, *2D U-Net, 3D U-Net* models.

Additionally, we train the *2D U-Net*'s architecture on the same 79 cases along with 402 images from the NSCLC-Radiomics dataset. These 402 images were acquired long before the COVID-19 pandemic, therefore we assume that ground truth segmentations for them are zero masks. During training this model we resample series such that batches contain approximately equal numbers of COVID-19 positive and negative cases. We refer to this model as *2D U-Net+*.

2D U-Net and *2D U-Net+* were trained for 15k batches using Adam (Kingma and Ba, 2014) optimizer with default parameters and an initial learning rate of 0.0003. Each batch contains 5 series of axial slices. *3D U-Net* was optimized via plain stochastic gradient descent for 10k batches using a learning rate of 0.01. Each batch consists of 16 3D patches.

In order to estimate mean values and standard deviations of models' quality metrics defined in Section 2.5 each segmentation network was trained 3 times with different random seeds. Resulting networks were evaluated on the hold-out test dataset, described in Section 3.5.

4.4. Resnet-50 and multitask models

The remaining 806 positive images without ground truth segmentations and 254 negative images from the Mosmed-1110 and 402 negative images from NSCLC-Radiomics were split 5 times in a stratified manner into a training set and a validation set. Each of the 5 validation sets contains 30 random images.

For each split we train the *ResNet-50* and the classification heads of *Multitask-Latent, Multitask-Spatial-1* and *Multitask-Spatial-4* models on the defined training set, while segmentation heads of the multitask models were trained on the same 79 images, as *2D U-Net* (see Section 4.3).

For each network on each training epoch we evaluate the ROC AUC between the predicted COVID-19 probabilities and the ground truth labels on the validation set. We save the networks' weights which resulted in the highest validation ROC AUC during training. For all the multitask models as well as for *ResNet-50* top validation ROC AUCs exceeded 0.9 for all splits.

We train all networks for 30k batches using Adam (Kingma and Ba, 2014) optimizer with the default parameters and an initial learning rate of $3 \cdot 10^{-4}$ reduced to $1 \cdot 10^{-4}$ after 24k batches. Each batch contains 5 series of axial slices.

During training the multitask models we resample examples such that batches contained an approximately equal number of examples which were used to penalize either classification or segmentation head. However, we multiplied by 0.1 the loss for the classification head, because it resulted in better validation ROC AUCs.

For each of 5 splits, we evaluated each trained network on the hold-out test dataset described in Section 3.5. We report the resulting mean values and standard deviations of the quality metrics in Section 5.

5. Results

In this section we report and discuss the results of the experiments described in Section 4. In Tab. 3 we compare all the methods described in Section 2 using quality metrics defined in Section 2.5 and evaluated on the test dataset described in Section 3.5.

5.1. Segmentation-based methods

In this subsection we discuss the performance of four methods: the threshold-based baseline, *3D U-Net, 2D U-Net* and *2D U-Net+*.

We expect two major weaknesses of the threshold-based method: False Positive (FP) predictions on the vessels and bronchi, and inability to distinguish COVID-19 related lesions from other pathological findings. It is clearly seen from the extremely low ROC AUC scores (Tab. 3). One could also notice massive FP predictions even in healthy cases (Fig. 4, column B). However, the method often provides a reasonable segmentation of the lesion area (Fig. 4, column A).

Neural networks considerably outperform the threshold-based baseline in terms of any quality metric. We observe neither quantitative (Tab. 3) nor qualitative (Fig. 4) significant difference between *2D U-Net*'s and *3D U-Net*'s performances. They yield accurate severity scores within the COVID-19 positive population (Spearman's $\rho = 0.97$). However, severity scores quantified for the whole test dataset do not allow to accurately distinguish between COVID-19 positive cases and cases with other pathological findings (ROC AUC

Table 3
Quantitative comparison of all the methods discussed in Section 2. Trade-off between qualities of COVID-19 identification and ranking by severity is observed for segmentation-based methods. The proposed *Multitask-Spatial-1* model yields the best identification results. Results are given as $mean \pm std$.

	ROC AUC (COVID-19 vs. ·)				Spearman's ρ	Dice Score
	vs. All others	vs. Normal	vs. Bac. Pneum.	vs. Nodules		
Thresholding	.51 ± 0.00	.68 ± 0.00	.46 ± 0.00	.45 ± 0.00	.92 ± 0.00	.42 ± 0.00
3D U-Net	.76 ± 0.02	.89 ± 0.02	.59 ± 0.01	.79 ± 0.03	**.97** ± 0.01	**.65** ± 0.00
2D U-Net	.78 ± 0.01	.93 ± 0.01	.62 ± 0.01	.79 ± 0.00	**.97** ± 0.00	.63 ± 0.00
2D U-Net+	.86 ± 0.01	**.98** ± 0.01	.68 ± 0.02	.91 ± 0.01	.80 ± 0.03	.59 ± 0.01
ResNet-50	.62 ± 0.19	.67 ± 0.21	.55 ± 0.13	.65 ± 0.22	N/A	N/A
Multitask-Latent	.79 ± 0.06	.84 ± 0.05	.73 ± 0.06	.80 ± 0.07	**.97** ± 0.00	.61 ± 0.02
Multitask-Spatial-4	.89 ± 0.03	.94 ± 0.03	.83 ± 0.05	.91 ± 0.03	**.98** ± 0.00	.61 ± 0.02
Multitask-Spatial-1	**.93** ± 0.01	**.97** ± 0.01	**.87** ± 0.01	**.93** ± 0.00	**.97** ± 0.01	.61 ± 0.02

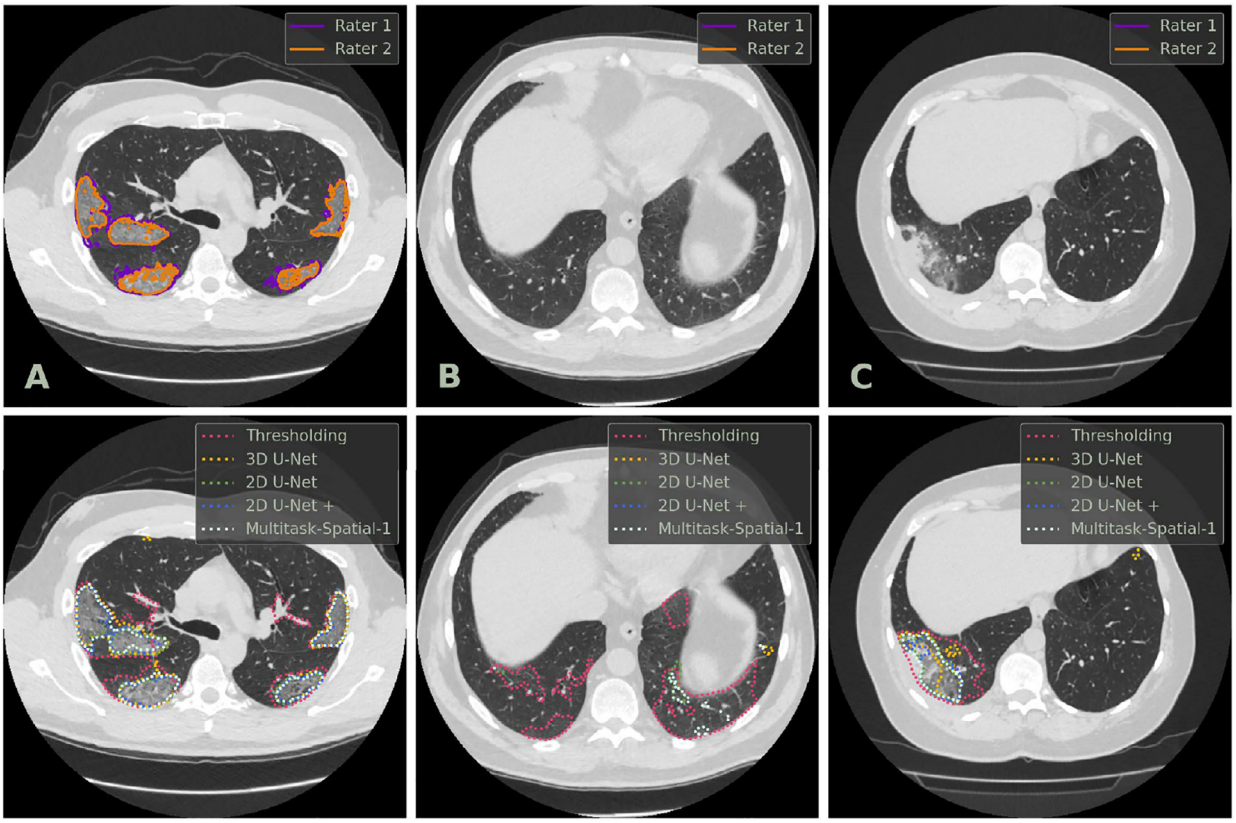

Fig. 4. Examples of axial CT slices from the test dataset along with ground truth annotations (first row) and predicted masks (second row) of COVID-19-specific lesions. *Column A*: COVID-19 positive case; *Column B*: normal case; *Column C*: case with bacterial pneumonia. Lesions' masks are represented by the contours of their borders for clarity.

COVID-19 vs. Bac. Pneum. ≈ 0.6, ROC AUC COVID-19 vs. Nodules $= 0.79$) due to FP segmentations (Fig. 4, columns B and C).

As one could expect, training on images with non-small cell lung cancer tumors from NSCLS-Radiomics dataset results in the enhancement of ROC AUC vs. Nodules (0.91 for *2D U-Net+* compared to 0.79 for *2D U-Net*). Interestingly, in this experiment we observe a degradation in terms of Spearman's ρ for ranking of COVID-19 positive cases (0.8 for *2D U-Net+* compared to 0.97 for 2D U-Net). We conclude that one should account for this trade-off and use an appropriate training setup depending on the task.

All the segmentation-based models perform poorly in terms of classification into COVID-19 and bacterial pneumonia (ROC AUC COVID-19 vs. Bac. Pneum. ≤ 0.7). This motivates to discuss the other methods.

5.2. Resnet-50

Despite that validation ROC AUCs for all the trained *ResNet-50* networks exceed 0.9, their performance on the test dataset is extremely unstable: ROC AUC COVID-19 vs. All varies from 0.43 to 0.85, see also high standard deviation values for all tasks in Tab. 3.

5.3. Multitask models

In this subsection we discuss the performance of *Multitask-Latent*, *Multitask-Spatial-4* and the proposed *Multitask-Spatial-1* models on identification, segmentation and severity quantification tasks in comparison to each other, *ResNet-50* and segmentation-based methods.

Ground truth label	COVID-19 Prediction: Negative	COVID-19 Prediction: Positive
COVID-19	1	31
Bac. Pneum.	22	8
Nodules	26	4
Normal	31	0

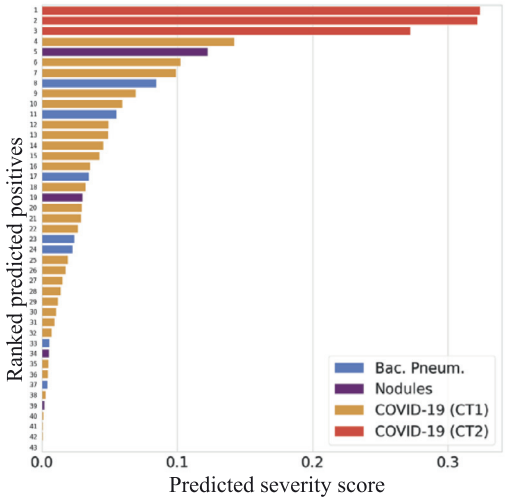

Fig. 5. COVID-19 triage: identification of COVID-19 positive patients (left) and ranking them in the descending order of severity (right) via the proposed single *Multitask-Spatial-1* model. In the right plot bars correspond to the ranked studies. Absolute values of the predicted affected lungs fractions are represented as bars' lengths along the *x*-axis. The bars' colors denote ground truth labels.

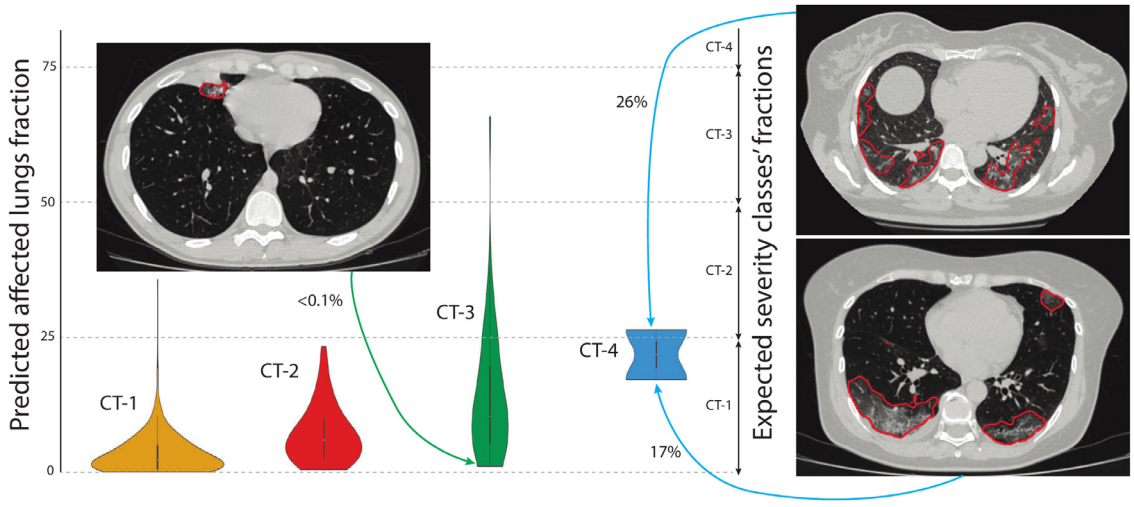

Fig. 6. The comparison of visual subjective estimation and automatic segmentation for weakly annotated cases from the Mosmed-1110 dataset. Each distribution corresponds to a set of cases with the same *Severity* group according to the radiologist's subjective judgment. The left y-axis shows the automatically estimated *Severity* by our method; the right one denotes expected *Severity* ranges that are [0; 25) for CT-1, [25; 50) for CT-2, [50; 75) for CT-3, [75; 100] for CT-4. The colored arrows denote the correspondence between some visually underestimated cases and their representative axial slices. Note the inconsistency of manual estimation.

As seen from mean values and standard deviations of ROC AUC scores in Tab. 3, *Multitask-Latent* model yields better and more stable identification results than *ResNet-50*. Both these models classify the latent representations of the input images. We show that sharing these features with the segmentation head, i.e. decoder of the U-Net architecture improves the classification quality. Moreover, one can see in Tab. 3 that this effect is enhanced by sharing the spatial feature maps from the upper levels of the U-Net's decoder. The proposed *Multitask-Spatial-1* architecture (see Fig. 3) with shallow segmentation and classification heads directly sharing the same spatial feature map shows the top classification results. Especially, it most accurately distinguish between COVID-19 and other lung diseases (ROC AUC COVID-19 vs. Bac. Pneum. = 0.87, ROC AUC COVID-19 vs. Nodules = 0.93).

As seen in Tab. 3 and Fig. 4 there is no significant difference in terms of segmentation and severity quantification qualities between the multitask models and the neural networks for single segmentation task.

Therefore, the single proposed *Multitask-Spatial-1* model can be applied for both triage problems: identification of COVID-19 patients followed by their ranking according to the severity. In Fig. 5 we visualize these two steps of triage pipeline for the test dataset, described in Section 3.5. One can see the several false positive alarms in cases with non-COVID-19 pathological findings. We discuss the possible ways to resolve them in Section 6. 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

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 showed that the proposed multitask architecture achieves top quality metrics for both tasks simultaneously. Finally, we have released the

code and used public data for training, so our results are fully reproducible.

Besides classification between COVID-19 and healthy patients we evaluate classification between COVID-19 and other lung abnormalities: bacterial pneumonia and cancerous nodules. As shown in Fig. 5, our method yields false positive alarms, mainly in patients with bacterial pneumonia (COVID-19 vs. bacterial pneumonia specificity $22/30 = 73.4\%$). However, we find this result promising, given the fact that we do not use any explicit training dataset with bacterial pneumonia patients. The proposed multitask model can be trained with the addition of bacterial or/and viral (not COVID-19) pneumonia cases, which can partially reduce the classification error. However, there is also an irreducible classification error in cases when radiomic features are not allow to distinguish between COVID-19 and non-COVID-19 pneumonia. Fortunately, in practice, usage of an automated triage system always implies second reading, so the model's false positives are assumed to be resolved by a radiologist, while the most controversial cases can be resolved by the RT-PCR testing. Thus, we conclude that the identification part of our triage system may be used as a highly sensitive first reading tool.

The role of the *Severity Quantification* part is more straightforward. As we mentioned in Section 1, radiologists perform the severity classification into groups from CT0 (no COVID-19 related lesions) and CT1 (up to 25% of lungs affected) to CT4 (more than 75%) in a visual semi-quantitative fashion. We believe that such estimation may be highly subjective and may contain severe discrepancies. To validate this assumption, we additionally analyzed Mosmed-1110, which includes not only 50 segmentation masks but also 1110 multiclass labels CT0-CT4. Within our experiments, we binarized these labels and effectively removed information about COVID-19 severity. We examined mask predictions for the remaining 1050 cases, excluding healthy patients (CT0 group) and grouped the predictions by these weak labels, as shown in Fig. 6. An expert radiologist validated analyzed the most extreme mismatches visualized in Fig. 6 and confirmed the correctness of our model's predictions. As we see, the severity of many studies was highly underestimated during the visual semi-quantitative analysis. This result implies that deep-learning-based medical image analysis algorithms, including the proposed method, are great intelligent radiologists' assistants in a fast and reliable estimation of time-consuming biomarkers such as COVID-19 severity.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

M. Belyaev is a founder and CEO of IRA Labs ltd, a medical image processing company. The company didn't support the study which was conducted on open-sourced datasets; the paper code is also public.

CRediT authorship contribution statement

Mikhail Goncharov: Writing - original draft, Formal analysis, Investigation. **Maxim Pisov:** Visualization, Writing - original draft, Investigation. **Alexey Shevtsov:** Investigation, Data curation, Software. **Boris Shirokikh:** Writing - original draft, Formal analysis, Investigation. **Anvar Kurmukov:** Software, Writing - review & editing. **Ivan Blokhin:** Writing - review & editing, Data curation. **Valeria Chernina:** Writing - original draft. **Alexander Solovev:** Data curation. **Victor Gombolevskiy:** Conceptualization, Validation. **Sergey Morozov:** Supervision. **Mikhail Belyaev:** Conceptualization, Methodology, Writing - original draft, Writing - review & editing.

Acknowledgment

We thank Dmitry Petrov (University of Massachusetts Amherst) for valuable comments and Tatiana Korb (Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies) for suggesting test CT studies.

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.

Akl, E.A., Blazic, I., Yaacoub, S., Frija, G., Chou, R., Appiah, J.A., Fatehi, M., Flor, N., Hitti, E., Jafri, H., et al., 2020. Use of chest imaging in the diagnosis and management of covid-19: a who rapid advice guide. Radiology 203173.

Amine, A., Modzelewski, R., Li, H., Ruan, S., 2020. Multi-task deep learning based ct imaging analysis for covid-19 pneumonia: classification and segmentation.. Comput. Biol. Med. 126.

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 (1), 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. Med. Phys. 38 (2), 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.

Blažić, I., Brkljačić, B., Frija, G., 2020. The use of imaging in covid-19-results of a global survey by the international society of radiology. Eur. Radiol. 1–9.

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

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 (9), 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. J. Digit. Imaging 26 (6), 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., O'Donoghue, B., Visentin, D., et al., 2018. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat. Med. 24 (9), 1342–1350.

Faita, F., 2020. Deep learning in emergency medicine: recent contributions and methodological challenges. Emergency Care Journal 16 (1).

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

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.. Rapid ai development cycle for the coronavirus (covid-19) pandemic: initial results for automated detection & patient monitoring using deep learning ct image analysis.

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

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 Trans. Med. 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. arXiv:1406.4729.

He, K., Zhang, X., Ren, S., Sun, J., 2016. 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., 2016. 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 (2), e200075.
Huang, Z., Zhao, S., Li, Z., Chen, W., Zhao, L., Deng, L., Song, B., 2020. The battle against coronavirus disease 2019 (covid-19): emergency management and infection control in a radiology department. Journal of the american college of radiology.
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., 2020. 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., 2020. 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. 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 Trans. Med. Imaging.
Kherad, O., Moret, B.M., Fumeaux, T., 2020. Computed tomography (ct) utility for diagnosis and triage during covid-19 pandemic. Rev. Med. Suisse 16 (692), 955.
Kingma, D.P., Ba, J.. Adam: a method for stochastic optimization.
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.
Korb, T., Chernina, V., Blokhin, I., Aleshina, O., Mokienko, O., Morozov, S., Gombolevskiy, V., 2021. Specificity of chest computed tomography in covid-19-associated pneumonia: a retrospective study (in russ.). Almanac of Clinical Medicine 49.
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., 2020. 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., 2020. Early transmission dynamics in wuhan, china, of novel coronavirus–infected pneumonia. N top N. Engl. J. Med..
Lin, H.-T., Li, L., 2012. Reduction from cost-sensitive ordinal ranking to weighted binary classification. Neural Comput. 24 (5), 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. Nat. Med. 1–5.
Morozov, S., Andreychenko, A., Pavlov, N., Vladzymyrskyy, A., Ledikhova, N., Gombolevskiy, V., Blokhin, I., Gelezhe, P., Gonchar, A., Chernina, V.Y., 2020. Mosmeddata: data set of 1110 chest ct scans performed during the covid-19 epidemic. Digital Diagnostics 1 (1), 49–59.
Morozov, S., Chernina, V., Blokhin, I., Gombolevskiy, V., 2020. Chest computed tomography for outcome prediction in laboratory-confirmed covid-19: a retrospective analysis of 38,051 cases. Digital Diagnostics 1 (1), 27–36.
Morozov, S.P., Protsenko, D., Smetanina, S. e. a., 2020c. Imaging of coronavirus disease (covid-19): Organization, methodology, interpretation: Preprint no. cdt - 2020 - ii. version 2 of 17.04.2020.
Petrikov, S., Popugaev, K., Barmina, T., Zabavskaya, O., Sharifullin, F., Kokov, L., 2020. Comparison of clinical data and computed tomography semiotics of the lungs in covid-19. Tuberculosis and Lung Diseases 98 (7).
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. Med. Phys. 36 (7), 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.
Rubin, G., Ryerson, C., Haramati, L., Sverzellati, N., Kanne, J., 2020. Others,'the role of chest imaging in patient management during the covid-19 pandemic: a multinational consensus statement from the fleischner society,'. Chest.
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.. Lung infection quantification of covid-19 in ct images with deep learning.
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. J. Pharm. Anal..
Shi, F., Wang, J., Shi, J., Wu, Z., Wang, Q., Tang, Z., He, K., Shi, Y., Shen, D., 2020. Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for covid-19. IEEE Rev. Biomed. Eng..
Shi, F., Xia, L., Shan, F., Wu, D., Wei, Y., Yuan, H., Jiang, H., Gao, Y., Sui, H., Shen, D.. Large-scale screening of covid-19 from community acquired pneumonia using infection size-aware classification.
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.. Efficientnet: rethinking model scaling for convolutional neural networks.
Tang, Z., Zhao, W., Xie, X., Zhong, Z., Shi, F., Liu, J., Shen, D.. Severity assessment of coronavirus disease 2019 (covid-19) using quantitative features from chest ct images.
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 (1), 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. Nat. Med. 24 (9), 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 Trans. Med. 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 Trans. Med. 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. Ann. Intern. Med. 170 (1), 51–58.
World Health Organization, et al., 2020. Clinical management of covid-19. who reference number: Who/2019-ncov/clinical/2020.5. 2020.[internet].
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 (1), 38–46.
Zacharov, I., Arslanov, R., Gunin, M., Stefonishin, D., Bykov, A., Pavlov, S., Panarin, O., Maliutin, A., Rykovanov, S., Fedorov, M., 2019. 'Zhores'–petaflops supercomputer for data-driven modeling, machine learning and artificial intelligence installed in skolkovo institute of science and technology. Open Engineering 9 (1), 512–520.
Zantah, M., Castillo, E.D., Townsend, R., Dikengil, F., Criner, G.J., 2020. Pneumothorax in covid-19 disease-incidence and clinical characteristics. Respir. Res. 21 (1), 1–9.