

# An Explainable AI System for Automated COVID-19 Assessment and Lesion Categorization from CT-scans

Matteo Pennisi<sup>1,\*</sup>, Isaak Kavasidis<sup>1,\*</sup>, Concetto Spampinato<sup>1</sup>, Vincenzo Schininà<sup>2</sup>, Simone Palazzo<sup>1</sup>, Francesco Rundo<sup>3</sup>, Massimo Cristofaro<sup>2</sup>, Paolo Campioni<sup>2</sup>, Elisa Pianura<sup>2</sup>, Federica Di Stefano<sup>2</sup>, Ada Petrone<sup>2</sup>, Fabrizio Albarello<sup>2</sup>, Giuseppe Ippolito<sup>2</sup>, Salvatore Cuzzocrea<sup>4</sup>, Sabrina Conoci<sup>4</sup>

---

## Abstract

*COVID-19 infection caused by SARS-CoV-2 pathogen is a catastrophic pandemic outbreak all over the world with exponential increasing of confirmed cases and, unfortunately, deaths. In this work we propose an AI-powered pipeline, based on the deep-learning paradigm, for automated COVID-19 detection and lesion categorization from CT scans. We first propose a new segmentation module aimed at identifying automatically lung parenchyma and lobes. Next, we combined such segmentation network with classification networks for COVID-19 identification and lesion categorization. We compare the obtained classification results with those obtained by three expert radiologists on a dataset consisting of 162 CT scans. Results showed a sensitivity of 90% and a specificity of 93.5% for COVID-19 detection, outperforming those yielded by the expert radiologists, and an average lesion categorization accuracy of over 84%. Results also show that a significant role is played by prior lung and lobe segmentation that allowed us to enhance performance by over 20 percent points. The interpretation of the trained AI models, moreover, reveals that the most significant areas for supporting the decision on COVID-19 identification are consistent with the lesions clinically associated to the virus, i.e., crazy paving, consolidation and ground glass. This means that the artificial models are able to discriminate a positive patient from a negative one (both controls and patients with interstitial pneumonia tested negative to COVID) by evaluating the presence of those lesions into CT scans. Finally, the AI models are integrated into a user-friendly GUI to support AI explainability for radiologists, which is publicly available at <http://perceivelab.com/covid-ai>. The whole AI system is unique since, to the best of our knowledge, it is the first AI-based software, publicly available, that attempts to explain to radiologists what information is used by AI methods*

---

\*Equal contribution

\*\*Corresponding author

<sup>1</sup>DIEEI, University of Catania, Catania, Italy

<sup>2</sup>National Institute for infectious disease, "Lazzaro Spallanzani" Department, Rome, Italy

<sup>3</sup>STMicroelectronics - ADG Central R&D, Catania, Italy

<sup>4</sup>ChimBioFaram Department, University of Messina, Messina, Italy

*for making decision and that involves proactively them in the decision loop to further improve the COVID-19 understanding.*

---

## 1. Introduction

At the end of 2019 in Wuhan (China) several cases of an atypical pneumonia, particularly resistant to the traditional pharmacological treatments, were observed. In early 2020, the COVID-19 virus [1] has been identified as the responsible pathogen for the unusual pneumonia. From that time, COVID-19 has spread all around the world hitting, to date about 32 million of people (with about 1M deaths), stressing significantly healthcare systems in several countries. Since the beginning, it has been noted that 20% of infected subjects appear to progress to severe disease, including pneumonia and respiratory failure and in around 2% of cases death [2].

Currently, the standard diagnosis of COVID-19 is de facto based on a biomolecular test through Real-Time Polymerase Chain Reaction (RT-PCR) test [3, 4]. However, although widely used, this biomolecular method is time-consuming and appears to be not quite accurate suffering from a large number of false-negatives [5].

Recent studies have outlined the effectiveness of radiology imaging through chest X-ray and mainly Computed Tomography (CT) given the pulmonary involvement in subjects affected by the infection [5, 6]. Given the extension of the infection and the number of cases that daily emerge worldwide and that call for fast, robust and medically sustainable diagnosis, CT scan appears to be suitable for a robust-scale screening, given the higher resolution w.r.t. X-Ray. In this scenario, artificial intelligence may play a fundamental role to make the whole diagnosis process automatic, reducing, at the same time, the efforts required by radiologists for visual inspection [7].

In this paper, thus, we present an innovative artificial intelligent approach to achieve both COVID-19 identification and lesion categorization (ground glass, crazy and paving consolidation) that are instrumental to evaluate lung damages and the prognosis assessment. Our method relies only on radiological image data avoiding the use of additional clinical data in order to create AI models that are useful for large-scale and fast screening with all the subsequent benefits for a favorable outcome. More specifically, we propose an innovative automated pipeline consisting of 1) lung/lobe segmentation, 2) COVID-19 identification and interpretation and 3) lesion categorization. We tested the AI-empowered software pipeline on multiple CT scans, both publicly released and collected at the Spallanzani Institute in Italy, and showed that: 1) our segmentation networks is able to effectively extract lung parenchyma and lobes from CT scans, outperforming state of the art models; 2) the COVID-19 identification module yields better accuracy (as well as specificity and sensitivity) than expert radiologists. Furthermore, when attempting to interpret the decisions made by the proposed AI model, we found that it learned automatically, and without any

supervision, the CT scan features corresponding to the three most common lesions spotted in the COVID-19 pneumonia, i.e., consolidation, ground glass and crazy paving, demonstrating its reliability in supporting the diagnosis by using only radiological images. As an additional contribution, we integrate the tested AI models into a user-friendly GUI to support further AI explainability for radiologists, which is publicly available at <http://perceivelab.com/covid-ai>. The GUI processes entire CT scans and reports if the patient is likely to be affected by COVID-19, showing, at the same time, the scan slices that supported the decision.

## 2. Related Work

The COVID-19 epidemic caught the scientific community flat-footed and in response a high volume of research has been dedicated at all possible levels. In particular, since the beginning of the epidemic, AI models have been employed for disease spread monitoring [8, 9, 10], for disease progression [11] and prognosis [12], for predicting mental health ailments inflicted upon healthcare workers [13] and for drug repurposing [14, 15] and discovery [16].

However, the lion's share in employing AI models for the fight against COVID-19 belongs to the processing of X-rays and CT scans with the purpose of detecting the presence of COVID-19 or not. In fact, recent scientific literature has demonstrated the high discriminative and predictive capability of deep learning methods in the analysis of COVID-19 related radiological images [17, 18]. The key radiological techniques for COVID-19 induced pneumonia diagnosis and progression estimation are based on the analysis of CT and X-ray images of the chest, on which deep learning methodologies have been widely used with good results for segmentation, predictive analysis, and discrimination of patterns [19, 20, 21]. If, on one hand, X-Ray represents a cheaper and most effective solution for large scale screening of COVID-19 disease, on the other hand, its low resolution has led AI models to show lower accuracy compared to those obtained with CT data.

For the above reasons, CT scan has become the gold standard for investigation on lung diseases. In particular, deep learning, mainly in the form of Deep Convolutional Neural Networks (DCNN), has been largely applied to lung disease analysis from CT scans images, for evaluating progression in response to specific treatment (for instance immunotherapy, chemotherapy, radiotherapy) [22, 23], but also for interstitial lung pattern analysis [24, 25] and on segmentation and discrimination of lung pleural tissues and lymph-nodes [26, 27]. This latter aspect is particularly relevant for COVID-19 features and makes artificial intelligence an extremely powerful tool for supporting early diagnosis of COVID-19 and disease progression quantification. As a consequence, several recent works have reported using AI models for automated categorization of CT scans [21] and also on COVID-19 [28, 29, 30] but without being able to distinguish between the various types of COVID-19 lesions.

Thus, the main contributions of this paper w.r.t. the state of the art are the following ones:

- We propose a novel lung-lobe segmentation network outperforming state of the art models;
- We employ the segmentation network to drive a classification network in first identifying CT scans of COVID-19 patients, and, afterwards, in automatically categorizing specific lesions;
- We then provide interpretation of the decisions made by the employed models and discover that, indeed, those models focus on specific COVID-19 lesions for distinguishing whether a CT scan pertains COVID-19 patients or not;
- We finally integrate the whole AI pipeline into a web platform to ease use for radiologists, supporting them in their investigation on COVID-19 disease.

### 3. Explainable AI for COVID-19 data understanding

The proposed AI system aims at 1) extracting lung and lobes from chest CT data, 2) categorizing CT scans as either COVID-19 positive or COVID-19 negative; 3) identifying and localizing typical COVID-19 lung lesions (consolidation, crazy paving and ground glass); and 4) explaining eventually what CT slices it based its own decisions.

#### 3.1. AI Model for Lung Segmentation

Our lung-lobe segmentation model is based on the *Tiramisu* network [31], a fully-convolutional DenseNet [32] in a U-Net architecture [33]. The model consists in two data paths: the downsampling one, that aims at extracting features and the upsampling one that aims at generating the output images (masks). Skip connections (i.e., connections starting from a preceding layer in the network’s pipeline to another one found later bypassing intermediate layers) aim at propagating high-resolution details by sharing feature maps between the two paths.

In this work, our segmentation model follows the Tiramisu architecture, but with two main differences:

- Instead of processing each single scan individually, convolutional LSTMs [34] are employed at the network’s bottleneck layer to exploit the spatial axial correlation of consecutive scan slices.
- In the downsampling and upsampling paths, we add residual squeeze-and-excitation layers [35], in order to emphasize relevant features and improve the representational power of the model.

Before discussing the properties and advantages of the above modifications, we first introduce the overall architecture, shown in Fig. 1.

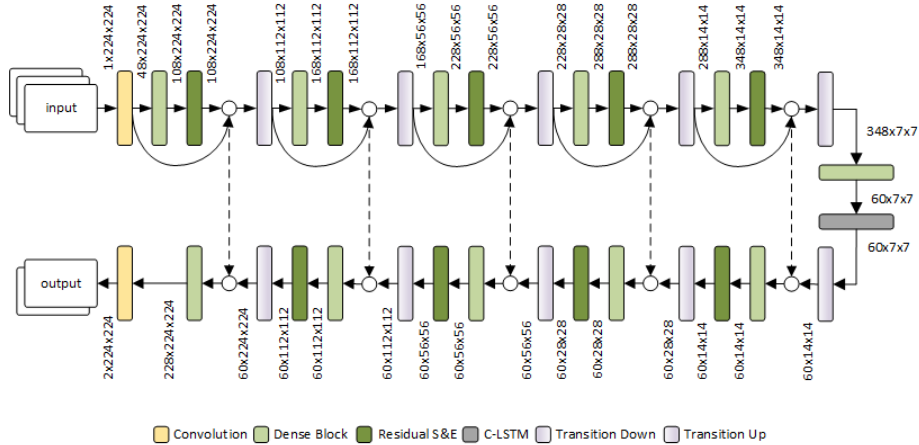


Figure 1: The proposed segmentation architecture, consisting of a downsampling path (*top*) and an upsampling path (*bottom*), interconnected by skip connections and by the bottleneck layer.

The input to the model is a sequence of 3 consecutive slices – suitably resized to  $224 \times 224$  – of a CT scan, which are processed individually and combined through a convolutional LSTM layer. Each slice is initially processed with a standard convolutional layer to expand the feature dimensions. The resulting feature maps then go through the downsampling path of the model (the encoder) consisting of five sequences of dense blocks, residual squeeze-and-excitation layers and transition-down layers based on max-pooling. In the encoder, the feature maps at the output of each residual squeeze-and-excitation layer are concatenated with the input features of the preceding dense block, in order to encourage feature reuse and improve their generalizability. At the end of the downsampling path, the *bottleneck* of the model consists of a dense block followed by a convolutional LSTM. The following upsampling path is symmetric to the downsampling one, but it features: 1) skip connections from the downsampling path for concatenating feature maps at the corresponding layers of the upsampling path; 2) transition-up layers implemented through transposed convolutions. Finally, a convolutional layer provides a 6-channel segmentation map, representing, respectively, the log-likelihoods of the lobes (5 channels, one for each lobe) and non-lung (1 channel) pixels.

In the following, we review the novel characteristics of the proposed architecture.

**Residual squeeze-and-excitation layers.** Explicitly modeling interdependencies between feature channels has demonstrated to enhance performance of deep architectures; squeeze-and-excitation layers [35] instead aim to select informative features and to suppress the less useful ones. In particular, a set of input features of size  $C \times H \times W$  is squeezed through average-pooling to a

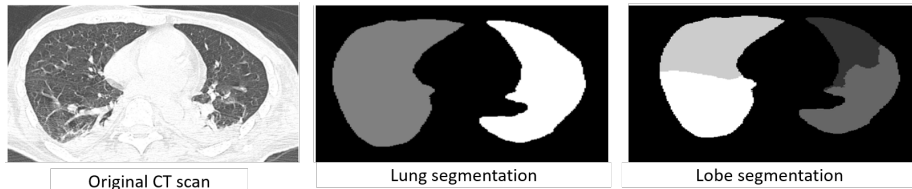


Figure 2: Example of lung and lobes segmentation.

$C \times 1 \times 1$  vector, representing global feature statistics. The “excitation” operator is a fully-connected non-linear layer that translates the squeezed vector into channel-specific weights that are applied to the corresponding input feature maps.

**Convolutional LSTM.** We adopt a recurrent architecture to process the output of the bottleneck layer, in order to exploit the spatial axial correlation between subsequent slices and enhance the final segmentation by integrating 3D information in the model. Convolutional LSTMs [34] are commonly used to capture spatio-temporal correlations in visual data (for example, in videos), by extending traditional LSTMs using convolutions in both the *input-to-state* and the *state-to-state* transitions. Employing recurrent convolutional layers allows the model to take into account the context of the currently-processed slice, while keeping the sequentiality and without the need to process the entire set of slices in a single step through channel-wise concatenation, which increases feature sizes and loses information on axial distance.

Fig. 2 shows an example of automated lung and lobe segmentation from a CT scan by employing the proposed segmentation network. The proposed segmentation network is first executed on the whole CT scan for segmenting only lung (and lobes); the segmented CT scan is then passed to the downstream classification modules for COVID-19 identification and lesion categorization.

### 3.2. Automated COVID-19 Diagnosis: CT classification

After parenchima lung segmentation (through the segmentation model presented in 3.1) a deep classification model analyzes slice by slice, each segmented CT scan, and decides whether a single slice contains some evidence of the COVID-19 disease. Afterwards, a voting method provides its final prediction according to all the per-slice decisions. At this stage, the system does not carry out any identification and localization of COVID-19 lesions, but it just identifies all slices where patterns of interest may be found and according to them, makes a guess on the presence or not of COVID-19 induced infection. An overview of this model is shown in Fig. 3: first the segmentation network, described in the previous section, identifies lung areas from CT scan, then a deep classifier (a DenseNet model in the 201 configuration [32]) processes the segmented lung areas to identify if the slice shows signs of COVID-19 virus.

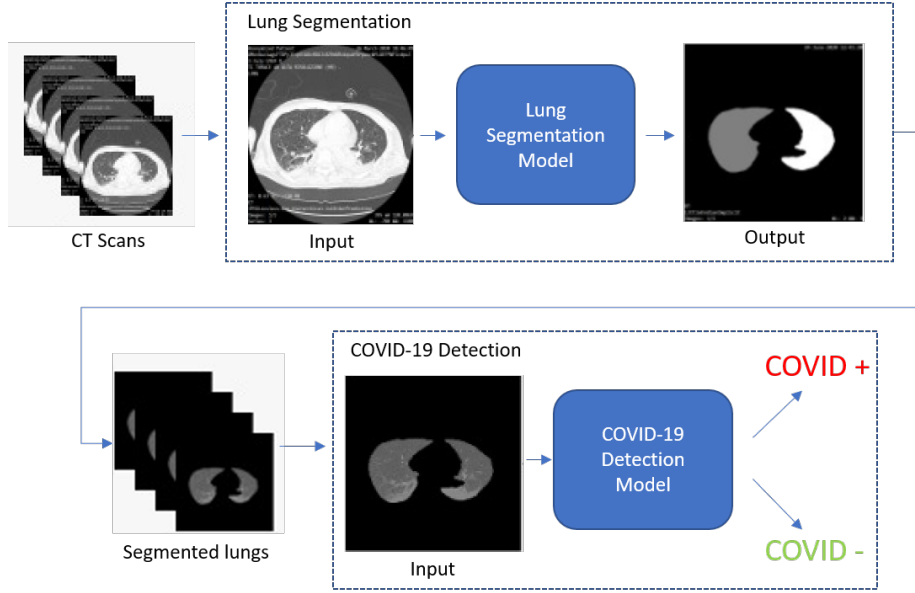


Figure 3: Overview of the **COVID-19 detection approach** for CT scan classification as either COVID-19 positive or negative.

Once the COVID-19 identification model is trained, we attempt to understand what features it employs to discriminate between positive and negative cases. Thus, to interpret the decisions made by the trained model we compute class-discriminative localization maps that attempt to provide visual explanations of the most significant input features for each class. To accomplish this we employ GradCAM [36] combined to VarGrad [37]. More specifically, GradCAM is a technique to produce such interpretability maps through by investigating output gradient with respect to feature map activations. More specifically, GradCAM generates class-discriminative localization map for any class  $c$ ,  $s^c$ , w.r.t feature activation maps  $A_k$  of a given convolutional layer. Such gradients are then global-average-pooled to obtain the activation importance weights  $w$ , i.e.:

$$w_k^c = \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad (1)$$

Afterwards, the saliency map  $S^c$ , that provides an overview of the activation importance for the class  $c$ , is computed through a weighted combination of activation maps, i.e.:

$$S^c = ReLU \left( \sum_k w_k^c A^k \right) \quad (2)$$

VarGrad is a technique used in combination to GradGAM and consists in

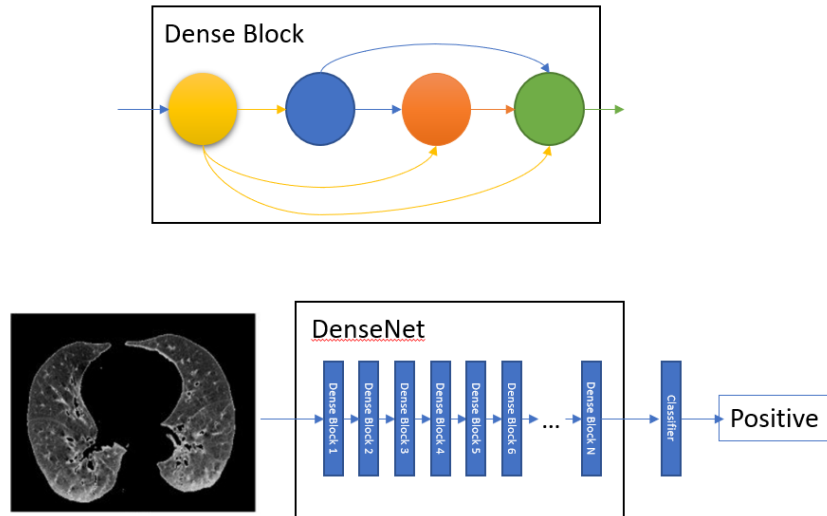


Figure 4: The DenseNet architecture. Convolutional processing layers are grouped in Dense Blocks (*top*). Features extracted in previous layers are concatenated and fed to all the next layers in the same Dense Block ensuring maximum information flow. Given that feature maps from previous layers are passed to the next layers, redundancy is avoided (i.e., later layers do not need to learn almost identical information from the immediately previous ones). In this way, each successive layer adds only a small number of feature maps, the so called *growth factor*, thus requiring fewer parameters to achieve state-of-the-art performance. Multiple Dense Blocks can be concatenated and form a deeper network (*bottom*).

performing multiple activation map estimates by adding, each time, Gaussian noise to the input data and then aggregating the estimates by computing the variance of the set.

### 3.3. COVID-19 lesion identification and categorization

An additional deep network activates only if the previous system identifies a COVID-19 positive CT scan. In that case, it works on the subset of slices identified as COVID-19 positives by the first AI system with the goal to localize and identify specific lesions (consolidation, crazy paving and ground glass). More specifically, the lesion identification system works on segmented lobes to seek COVID-19 specific patterns. The subsystem for lesion categorization employs the knowledge already learned by the COVID-19 detection module (shown in Fig. 3) and refines it for specific lesion categorization. An overview of the whole system is given in Fig. 5.

### 3.4. A Web-based Interface for Explaining AI decisions to Radiologists

In order to explain to radiologists, the decisions made by a “black-box” AI system, we integrated the inference pipeline for COVID-19 detection into a



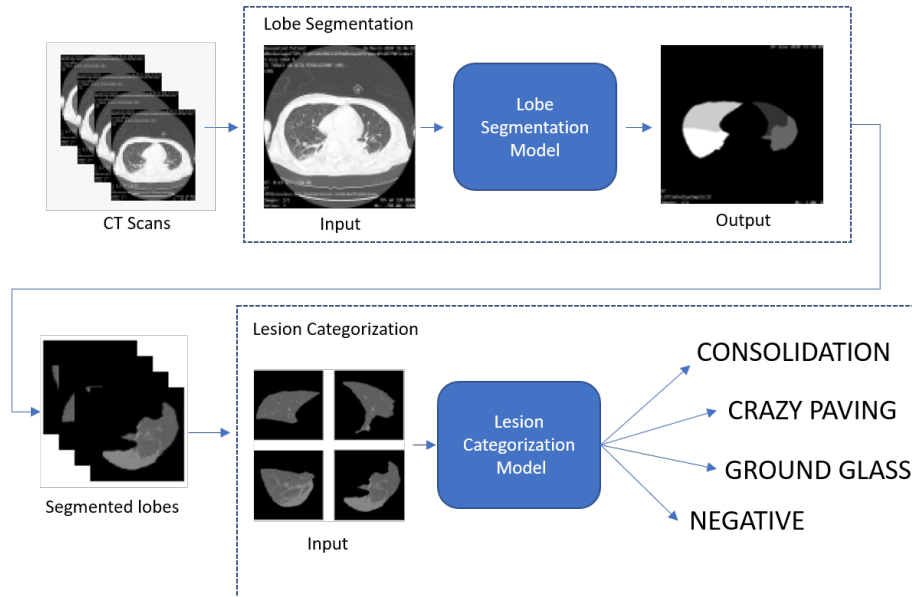


Figure 5: Overview of **COVID-19 lesion categorization** approach.

web-based application. The application was designed to streamline the whole inference process with just a few clicks and visualize the results with a variable grade of detail (Fig. 6). If the radiologists desire to see which CT slices were classified as positive or negative, they can click on “Show slices” where a detailed list of slices and their categorization is showed (Fig. 7).

Because the models may not achieve perfect accuracy, a single slice inspection screen is provided, where radiologists can inspect more closely the result of the classification. It also features a restricted set of image manipulation tools (move, contrast, zoom) for aiding the user to make a correct diagnosis (Fig. 8).

The AI-empowered web system integrates also a relevance feedback mechanism where radiologists can correct the predicted outputs, and the AI module exploits such a feedback to improve its future assessments. Indeed, both at the CT scan level and at the CT slice level, radiologists can correct models’ prediction. The AI methods will then use the correct labels to enhance their future assessments.

## 4. Results and Discussion

### 4.1. Dataset

Our dataset contains 72 CT scans of COVID-19 positive patients (positivity confirmed both by a molecular - reverse transcriptase–polymerase chain reaction for SARS-coronavirus RNA from nasopharyngeal aspirates - and an IgG or IgM antibody test) and 94 CT scans of COVID-19 negative subjects (35 patients with interstitial pneumonia but tested negative to COVID-19 and 59 controls).



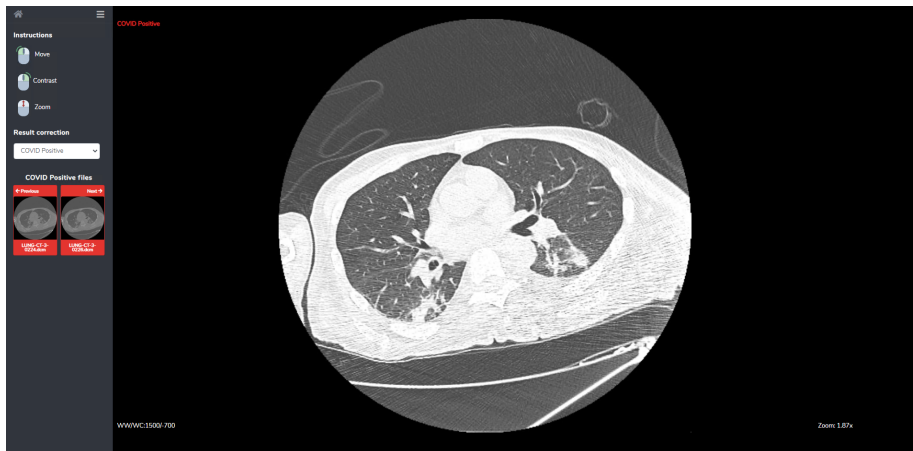


Figure 8: The slice inspection screen. In this screen the user can inspect each single slice and the AI models’ decisions.

CT scans were performed on a multi-detector row helical CT system scanner<sup>5</sup> using 120 kV pp, 250 mA, pitch of 1.375, gantry rotation time of 0,6 s and time of scan 5,7 s. The non-contrast scans were reconstructed with slice thicknesses of 0.625 mm and spacing of 0.625 mm with high-resolution lung algorithm. The images obtained on lung (window width, 1,000–1,500 H; level, -700 H) and mediastinal (window width, 350 H; level, 35–40 H) settings were reviewed on a picture archiving and communication system workstation<sup>6</sup>. CT scans of positive patients were also annotated by three expert radiologists (through consensus) who selected a subset of slices and annotated them with the type (Consolidation, Ground Glass and Crazy Paving) and the location (combinations of left/right/central and posterior/anterior) of the lesion. In total about 2,400 slices were annotated with COVID-19 lesions and about 3,000 slices of negative patients with no lesion. Tab. 1 provides an overview of all the CT scans and annotations in our dataset.

For training the lung/lobe segmentation model we adopted a combination of the LIDC [38], LTRC<sup>7</sup> and [39] datasets, for a total of 300 CT scans. Annotations on lung/lobe areas were done manually by three expert radiologists.

#### 4.2. Training Procedure

The COVID-19 detection network is a DenseNet201, which was used pre-trained on the ImageNet dataset [40]. The original classification layers in DenseNet201 were replaced by a 2-output linear layer for the COVID-19 positive/negative classification. Among the set of 166 CT scans, we used 95 scans

<sup>5</sup>Bright Speed, General Electric Medical Systems, Milwaukee, WI

<sup>6</sup>Impax ver. 6.6.0.145, AGFA Gevaert SpA, Mortsels, Belgium

<sup>7</sup><https://ltrcpublic.com/>

Dataset statistics	
<b>CT Data</b>	
CT Scans	166
COVID-19+	72
COVID-19-	94
<b>Annotations</b>	
Positive slices	2,390
Ground Glass	1,035
Crazy Paving	757
Consolidation	598
Negative slices	2,988

Table 1: CT Dataset for training and testing of the AI models.

(36 positives and 59 negatives) for training, 9 scans for validation (5 positives and 4 negatives) and 62 scans (31 positives and 31 negatives) for test. To compare the AI performance to the human one, the test set of 62 CT scans was provided to three expert radiologists for blind evaluation. Given the class imbalance in the training set, we used the weighted binary cross-entropy (defined in 3) as training loss and RT-PCR virology test as training/test labels.

The weighted binary cross-entropy loss for a sample classified as  $x$  with target label  $y$  is then calculated as:

$$WBCE = -w [y \cdot \log x + (1 - y) \cdot \log(1 - x)] \quad (3)$$

where  $w$  is defined as the ratio of the number negative samples to the total number of samples if the label is positive and vice versa. This way the loss results higher when misclassifying a sample that belongs to the less frequent class. It is important to highlight that splitting refers to the entire CT scan and not to the single slices: we made sure that full CT scans were not assigned in different splits to avoid any bias in the performance analysis. This is to avoid the deep models overfit the data by learning spurious information from each CT scan, thus invalidating the training procedure, thus enforcing robustness to the whole approach. Moreover, for the COVID-19 detection task, we operate at the CT level by processing and categorizing each single slice. To make a decision for the whole scan, we perform voting: if 10% of total slices is marked as positive then the whole exam is considered as a COVID-19 positive, otherwise as COVID-19 negative. The choice of the voting threshold was done empirically to maximize training performance.

The lesion categorization deep network is also a DenseNet201 model where classification layers were replaced by a 4-output linear layer (*ground glass*, *consolidation*, *crazy paving*, *negative*). The lesion categorization model processes

lobe segments (extracted by our segmentation model) with the goal to identify specific lesions. Our dataset contains 2,488 annotated slices; in each slice multiple lesion annotations with relative location (in lobes) are available. Thus, after segmenting lobes from these images we obtained 5,264 lobe images. We did the same on CT slices of negative patients (among the 2,950 available as shown in Tab. 1) and selected 5,264 lobe images without lesions. Thus, in total, the the entire set consisted of 10,528 images. We also discarded the images for which lobe segmentation produced small regions indicating a failure in the segmentation process. We used a fixed test split consisting of 195 images with consolidation, 354 with crazy paving, 314 with ground glass and 800 images with no lesion. The remaining images were split into training and validation sets with the ratio 80/20. Given the class imbalance in the training set, we employed weighted cross-entropy as training loss.

The weighted cross-entropy loss for a sample classified as  $x$  with target label  $y$  is calculated as:

$$WCE = -w \sum_C y \cdot \log(x) \quad (4)$$

where  $C$  is the set of all classes. The weight  $w$  for each class  $c$  is defined as:

$$w_c = \frac{N - N_c}{N} \quad (5)$$

where  $N$  is the total number of samples and  $N_c$  is the number of samples that have label  $c$ .

Since the model is the same as the COVID identification network, i.e., DenseNet201, we started from the network trained on the COVID-identification task and fine-tune it on the categorization task to limit overfitting given the small scale of our dataset.

For both the detection network and the lesion categorization network, we used the following hyperparameters: batch-size = 12, learning rate = 1e-04, ADAM back-propagation optimizer with beta values 0.9 and 0.999, eps = 1e-08 and weight decay = 0 and the back-propagation method was used to update the models' parameters during training. Detection and categorization networks were trained for 20 epochs. In both cases, performance are reported at the highest validation accuracy.

For lung/lobe segmentation, input images were normalized to zero mean and unitary standard deviation, with statistics computed on the employed dataset. In all the experiments for our segmentation model, input size was set to  $224 \times 224$ , initial learning rate to 0.0001, weight decay to 0.0001 and batch size to 2, with RMSProp as optimizer. When C-LSTMs were employed, recurrent states were initialized to zero and the size of the input sequences to the C-LSTM layers was set to 3. Each training was carried out for 50 epochs.

### 4.3. Performance Evaluation

In this section report the performance of the proposed model for lung/lobe segmentation, COVID-19 identification and lesion categorization.

#### 4.3.1. Lobe segmentation

Our segmentation model is based on the Tiramisu model [31] with the introduction of *squeeze-and-excitation* blocks and of a convolutional LSTM (either unidirectional or bidirectional) after the bottleneck layer. In order to understand the contribution of each module, we first performed ablation studies by testing the segmentation performance of our model using different architecture configurations:

- Baseline: the vanilla Tiramisu model described in [31];
- Res-SE: residual *squeeze-and-Excitation* module are integrated in each dense block of the Tiramisu architecture;
- C-LSTM: a unidirectional convolutional LSTM is added after the bottleneck layer of the Tiramisu architecture;
- Res-SE + C-LSTM: variant of the Tiramisu architecture that includes both residual *squeeze-and-Excitation* at each dense layer and a unidirectional convolutional LSTM after the bottleneck layer.

We also compared the performance against the U-Net architecture proposed in [39] that is largely adopted for lung/lobe segmentation.

All architectures were trained for 50 epochs by splitting the employed lung datasets into a training, validation and test splits using the 70/10/20 rule. Results in terms of Dice score coefficient (DSC) are given in Tab. 2. It has to be noted that unlike [39], we computed DSC on all frames, not only on the lung slices. The highest performance is obtained with the Res-SE + C-LSTM configuration, i.e., when adding *squeeze-and-excitation* and the unidirectional C-LSTM at the bottleneck layer of the Tiramisu architecture. This results in an accuracy improvement of over 4 percent points over the baseline. In particular, adding *squeeze-and-excitation* leads to a 2 percent point improvement over the baseline. Segmentation results are computed using data augmentation obtained by applying random affine transformations (rotation, translation, scaling and shearing) to input images. The segmentation network is then applied to our COVID-19 dataset for prior segmentation without any additional fine-tuning to demonstrate also its generalization capabilities.

#### 4.3.2. COVID-19 assessment

We compute results both for COVID-19 detection and lesion categorization and compare to those yielded by three experts with different degree of expertise:

1. Radiologist 1: a physician expert in thoracic radiology ( $\sim 30$  years of experience) with over 30,000 examined CT scans;
2. Radiologist 2: a physician expert in thoracic radiology ( $\sim 10$  years of experience) with over 9,000 examined CT scans;
3. Radiologist 3: a resident student in thoracic radiology ( $\sim 3$  years of experience) with about 2,000 examined CT scans.

Model	Lung segmentation	Lobe segmentation
Baseline Tiramisu [31]	$89.41 \pm 0.45$	$77.97 \pm 0.31$
Baseline + Res-SE	$91.78 \pm 0.52$	$80.12 \pm 0.28$
Baseline + C-LSTM	$91.49 \pm 0.57$	$79.47 \pm 0.38$
Baseline + Res-SE + C-LSTM	<b><math>94.01 \pm 0.52</math></b>	<b><math>83.05 \pm 0.27</math></b>

Table 2: Ablation studies of our segmentation network in terms of dice score. Best results are shown in bold. Note: we did not compute confidence intervals on these scores as they are obtained from a very large set of CT pixels.

	Sensitivity	C.I. (95%)
Radiologist 1	83.9%	[71.8% – 91.9%]
Radiologist 2	87.1%	[75.6% – 94.3%]
Radiologist 3	80.6%	[68.2% – 89.5%]
AI Model without lung segmentation	83.9%	[71.8% – 91.9%]
AI Model with lung segmentation	<b>90.3%</b>	<b>[79.5% – 96.5%]</b>

Table 3: Sensitivity (together with 95% confidence interval) comparison between manual readings of expert radiologists and the AI model for COVID-19 detection without lung segmentation and AI model with segmentation.

We also assess the role of prior segmentation on the performance. This means that in the pipelines showed in Figures 3 and 5 we removed the segmentation modules and performed classification using the whole CT slices using also information outside the lung areas. Results for COVID-19 detection are measured in terms of sensitivity and specificity and given in Tables 3 and 4.

Thus, the AI model using lung segmentation achieves the best performance outperforming expert radiologists in the COVID-19 assessment. Furthermore, performing lung segmentation improves by about 6 percent points both the sensitivity and the specificity, demonstrating its effectiveness. The important aspect to highlight is that expert radiologists during the annotation process did not have to segment lungs or lobes, showing the generalization capabilities of the proposed deep learning-based methods.

As a backbone model for COVID-19 identification, we employed DenseNet201 since it yielded the best performance when compared to other state of the art models, as shown in Table 5. In all the tested cases, we used upstream segmentation through the model described in Sect. 3.1. Voting threshold was set to 10% on all cases.

In order to enhance trust in the devised AI models, we analyzed what features these methods employ for making the COVID-19 diagnosis decision. This is done by investigating which artificial neurons fire the most, and then projecting this information to the input images. To accomplish this we combined

	Specificity	C.I. (95%)
Radiologist 1	87.1%	[75.6% – 94.3%]
Radiologist 2	87.1%	[75.6% – 94.3%]
Radiologist 3	90.3%	[79.5% – 96.5%]
AI Model without lung segmentation	87.1%	[75.6% – 94.3%]
AI Model with lung segmentation	<b>93.5%</b>	<b>[83.5% – 98.5%]</b>

Table 4: Specificity (together with 95% confidence interval) comparison between manual readings of expert radiologists and the AI model for COVID-19 detection without lung segmentation and AI model with segmentation.

Model	Variant	Sensitivity (CI)	Specificity (CI)	Accuracy (CI)
AlexNet	–	71.0% (57.9–81.6)	90.3% (79.5–96.5)	80.7% (68.3–89.5)
ResNet	18	71.0% (57.9–81.6)	93.5% (83.5–98.5)	82.3% (70.1–90.7)
	34	80.7% (68.3–89.5)	90.3% (79.5–96.5)	85.5% (73.7–93.1)
	50	83.9% (71.9–91.9)	90.3% (79.5–96.5)	87.1% (75.6–94.3)
	101	77.4% (64.7–89.9)	87.1% (75.6–94.3)	82.3% (70.1–90.7)
	152	77.4% (64.7–89.9)	90.3% (79.5–96.5)	83.9% (71.9–91.9)
DenseNet	121	77.4% (64.7–89.9)	93.5% (83.5–98.5)	85.5% (73.7–93.1)
	169	67.9% (83.5–98.5)	93.5% (83.5–98.5)	81.4% (68.7–90.2)
	201	<b>90.3% (79.5–96.5)</b>	<b>93.5% (83.5–98.5)</b>	<b>91.9% (81.5–97.5)</b>
SqueezeNet	–	66.7% (54.5–78.9)	93.5% (83.5–98.5)	81.4% (68.7–90.2)
ResNeXt	–	77.4% (64.7–86.9)	90.3% (79.5–96.5)	83.9% (71.9–91.9)

Table 5: COVID-19 classification accuracy by several state of the art models. Values in parentheses indicate 95% confidence intervals (CI).



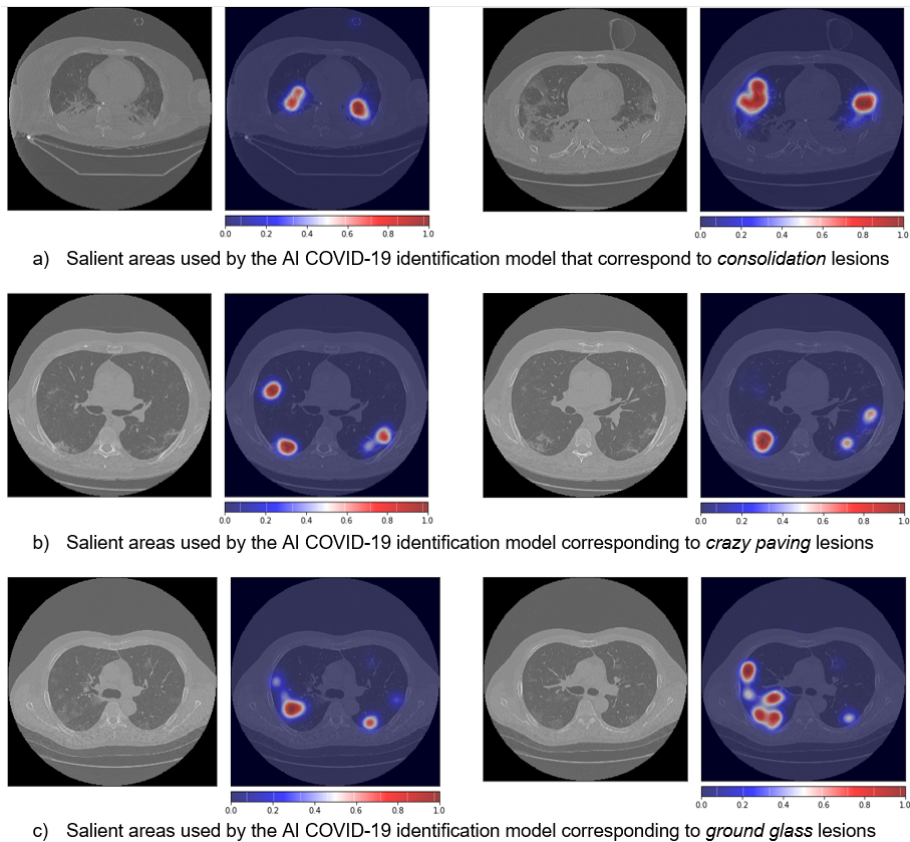


Figure 9: Lung salient areas identified automatically by the AI model for CT COVID-19 identification.

GradCAM [36] with VarGrad [37]<sup>8</sup> and, Fig. 9 shows some examples of the saliency maps generated by interpreting the proposed AI COVID-19 classification network. It is interesting to note that the most significant activation areas correspond to the three most common lesion types, i.e., ground glass, consolidation and crazy paving. This is remarkable as the model has indeed learned the COVID-19 peculiar patterns without any information on the type of lesions (to this end, we recall that for COVID-19 identification we only provide, at training times, the labels “positive” or “negative”, while no information on the type of lesions is given).

For COVID-19 lesion categorization we used mean (and per-class) classification accuracy over all lesion types and per lesion that are provided, respectively, in Table 6.

<sup>8</sup><https://captum.ai/>

	Model no_seg	Model w_seg
Consolidation	77.8% (69.9–84.1)	97.9% (93.6–99.8)
Ground glass	18.6% (14.1–24.1)	41.3% (35.1–47.7)
Crazy Paving	57.1% (49.4–64.4)	98.3% (94.8–99.8)
Negative	99.3% (98.6–99.7)	99.9% (99.5–100)
Average	63.2%	84.4%

Table 6: Per-class accuracy for lesion categorization between manual readings of expert radiologists and the AI model without lung segmentation and AI model with segmentation. Values in parentheses indicate 95% confidence intervals (CI).

Mean lesion categorization accuracy reaches, when operating at the lobe level, about 84% of performance. The lowest performance is obtained on ground glass, because ground glass opacities are specific CT findings that can appear also in normal patients with respiratory artifact. Operating at the level of single lobes yields a performance enhancement of over 21 percent points, and, also in this case, radiologists did not have to perform any lobe segmentation annotation, reducing significantly their efforts to build AI models. The most significant improvement when using lobe segmentation w.r.t. no segmentation is obtained *Crazy Paving* class, i.e., 98.3% against 57.1%.

Despite the CT diagnosis of COVID-19 pneumonia seems an easy task for experienced radiologists, the results show that our system is able to outperform them providing more accurate decisions. Artificial intelligence (AI), in particular, is able to identify more accurately lung lesions, in particular the smaller and undefined ones (as those highlighted in Fig. 9) The identification elements increases the sensitivity and specificity of the method for the correct diagnosis. The results obtained both for COVID-19 identification and lesion categorization pave the way to further improvement by implementing an advanced COVID-19 CT/RX image-driven diagnostic pipeline interpretable and strongly robust to provide not only the diseases identification and differential diagnosis but also the risk of disease progression.

## 5. Conclusions

In this work we have presented an AI-based pipeline for automated lung segmentation, COVID-19 detection and COVID-19 lesion categorization from CT scans. Results showed a sensitivity of 90% and a specificity of 93.5% for COVID-19 detection and average lesion categorization accuracy of about 64%. Results also show that a significant role is played by prior lung and lobe segmentation that allowed us to enhance performance of about 6 percent points. The AI models are then integrated into an user-friendly GUI to support AI explainability for radiologists, which is publicly available at <http://perceiveai.com/covid-ai>. To the best of our knowledge, this is the first AI-based software, publicly available, that attempts to explain radiologists

what information is used by AI methods for making decision and that involve proactively in the loop to further improve the COVID-19 understanding. These results pave the way to further improvement to provide not only the diseases identification and differential diagnosis but also the risk of disease progression.

### **Acknowledgment**

We thank the “Covid 19 study group” from Spallanzani Hospital (Maria Alessandra Abbonizio, Chiara Agrati, Fabrizio Albarello, Gioia Amadei, Alessandra Amendola, Mario Antonini, Raffaella Barbaro, Barbara Bartolini, Martina Benigni, Nazario Bevilacqua, Licia Bordi, Veronica Bordoni, Marta Branca, Paolo Campioni, Maria Rosaria Capobianchi, Cinzia Caporale, Iliaria Caravella, Fabrizio Carletti, Concetta Castilletti, Roberta Chiappini, Carmine Ciaralli, Francesca Colavita, Angela Corpolongo, Massimo Cristofaro, Salvatore Curiale, Alessandra D’Abramo, Cristina Dantimi, Alessia De Angelis, Giada De Angelis, Rachele Di Lorenzo, Federica Di Stefano, Federica Ferraro, Lorena Fiorentini, Andrea Frustaci, Paola Gallì, Gabriele Garotto, Maria Letizia Giancola, Filippo Giansante, Emanuela Giombini, Maria Cristina Greci, Giuseppe Ippolito, Eleonora Lalle, Simone Lanini, Daniele Lapa, Luciana Lepore, Andrea Lucia, Franco Lufrani, Manuela Macchione, Alessandra Marani, Luisa Marchioni, Andrea Mariano, Maria Cristina Marini, Micaela Maritti, Giulia Matusali, Silvia Meschi, Francesco Messina Chiara Montaldo, Silvia Murachelli, Emanuele Nicastri, Roberto Noto, Claudia Palazzolo, Emanuele Pallini, Virgilio Passeri, Federico Pelliccioni, Antonella Petrecchia, Ada Petrone, Nicola Petrosillo, Elisa Pianura, Maria Pisciotta, Silvia Pittalis, Costanza Proietti, Vincenzo Puro, Gabriele Rinonapoli, Martina Rueca, Alessandra Sacchi, Francesco Sanasi, Carmen Santagata, Silvana Scarcia, Vincenzo Schininà, Paola Scognamiglio, Laura Scorzolini, Giulia Stazi, Francesco Vaia, Francesco Vairo, Maria Beatrice Valli) for the technical discussion and critical reading of this manuscript.

### **Regulation and Informed Consent**

All data and methods were carried out in accordance to the General Data Protection Regulation 2016/679. The experimental protocols were approved by the Ethics Committee of the National Institute for Infectious Diseases Lazzaro Spallanzani in Rome. All patients enrolled in the study were over 18 at the time of their participation in the experiment and signed informed consent.

### **Declarations of interest**

None.

## References

- [1] N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu, et al., A novel coronavirus from patients with pneumonia in china, 2019, *New England Journal of Medicine* (2020).
- [2] W. H. Organization, et al., Novel coronavirus (2019-ncov): situation report, 8 (2020).
- [3] P. Huang, T. Liu, L. Huang, H. Liu, M. Lei, W. Xu, X. Hu, J. Chen, B. Liu, Use of chest ct in combination with negative rt-pcr assay for the 2019 novel coronavirus but high clinical suspicion, *Radiology* 295 (1) (2020) 22–23.
- [4] M.-Y. Ng, E. Y. Lee, J. Yang, F. Yang, X. Li, H. Wang, M. M.-s. Lui, C. S.-Y. Lo, B. Leung, P.-L. Khong, et al., Imaging profile of the covid-19 infection: radiologic findings and literature review, *Radiology: Cardiothoracic Imaging* 2 (1) (2020) e200034.
- [5] H. Liu, F. Liu, J. Li, T. Zhang, D. Wang, W. Lan, Clinical and ct imaging features of the covid-19 pneumonia: Focus on pregnant women and children, *Journal of infection* (2020).
- [6] M. Chung, A. Bernheim, X. Mei, N. Zhang, M. Huang, X. Zeng, J. Cui, W. Xu, Y. Yang, Z. A. Fayad, et al., Ct imaging features of 2019 novel coronavirus (2019-ncov), *Radiology* 295 (1) (2020) 202–207.
- [7] F. Rundo, C. Spampinato, G. L. Banna, S. Conoci, Advanced deep learning embedded motion radiomics pipeline for predicting anti-pd-1/pd-l1 immunotherapy response in the treatment of bladder cancer: Preliminary results, *Electronics* 8 (10) (2019) 1134.
- [8] Z. Allam, D. S. Jones, On the coronavirus (covid-19) outbreak and the smart city network: universal data sharing standards coupled with artificial intelligence (ai) to benefit urban health monitoring and management, in: *Healthcare*, Vol. 8, Multidisciplinary Digital Publishing Institute, 2020, p. 46.
- [9] L. Lin, Z. Hou, Combat covid-19 with artificial intelligence and big data, *Journal of travel medicine* 27 (5) (2020) taaa080.
- [10] N. Zheng, S. Du, J. Wang, H. Zhang, W. Cui, Z. Kang, T. Yang, B. Lou, Y. Chi, H. Long, et al., Predicting covid-19 in china using hybrid ai model, *IEEE Transactions on Cybernetics* (2020).
- [11] X. Bai, C. Fang, Y. Zhou, S. Bai, Z. Liu, L. Xia, Q. Chen, Y. Xu, T. Xia, S. Gong, et al., Predicting covid-19 malignant progression with ai techniques (2020).
- [12] W. Liang, J. Yao, A. Chen, Q. Lv, M. Zanin, J. Liu, S. Wong, Y. Li, J. Lu, H. Liang, et al., Early triage of critically ill covid-19 patients using deep learning, *Nature communications* 11 (1) (2020) 1–7.

- [13] K. Ćosić, S. Popović, M. Šarlija, I. Kesedžić, T. Jovanovic, Artificial intelligence in prediction of mental health disorders induced by the covid-19 pandemic among health care workers, *Croatian Medical Journal* 61 (3) (2020) 279.
- [14] S. Mohanty, M. H. A. Rashid, M. Mridul, C. Mohanty, S. Swayamsiddha, Application of artificial intelligence in covid-19 drug repurposing, *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* (2020).
- [15] Y.-Y. Ke, T.-T. Peng, T.-K. Yeh, W.-Z. Huang, S.-E. Chang, S.-H. Wu, H.-C. Hung, T.-A. Hsu, S.-J. Lee, J.-S. Song, et al., Artificial intelligence approach fighting covid-19 with repurposing drugs, *Biomedical Journal* (2020).
- [16] P. Richardson, I. Griffin, C. Tucker, D. Smith, O. Oechsle, A. Phelan, J. Stebbing, Baricitinib as potential treatment for 2019-ncov acute respiratory disease, *Lancet (London, England)* 395 (10223) (2020) e30.
- [17] L. Brunese, F. Mercaldo, A. Reginelli, A. Santone, Explainable deep learning for pulmonary disease and coronavirus covid-19 detection from x-rays, *Computer Methods and Programs in Biomedicine* 196 (2020) 105608.
- [18] L. Huang, R. Han, T. Ai, P. Yu, H. Kang, Q. Tao, L. Xia, Serial quantitative chest ct assessment of covid-19: Deep-learning approach, *Radiology: Cardiothoracic Imaging* 2 (2) (2020) e200075.
- [19] P. Nardelli, D. Jimenez-Carretero, D. Bermejo-Pelaez, G. R. Washko, F. N. Rahaghi, M. J. Ledesma-Carbayo, R. S. J. Estépar, Pulmonary artery–vein classification in ct images using deep learning, *IEEE transactions on medical imaging* 37 (11) (2018) 2428–2440.
- [20] N. Navab, J. Hornegger, W. M. Wells, A. Frangi, *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III*, Vol. 9351, Springer, 2015.
- [21] X. Mei, H.-C. Lee, K.-y. Diao, M. Huang, B. Lin, C. Liu, Z. Xie, Y. Ma, P. M. Robson, M. Chung, et al., Artificial intelligence–enabled rapid diagnosis of patients with covid-19, *Nature Medicine* (2020) 1–5.
- [22] A. A. A. Setio, F. Ciompi, G. Litjens, P. Gerke, C. Jacobs, S. J. Van Riel, M. M. W. Wille, M. Naqibullah, C. I. Sánchez, B. van Ginneken, Pulmonary nodule detection in ct images: false positive reduction using multi-view convolutional networks, *IEEE transactions on medical imaging* 35 (5) (2016) 1160–1169.
- [23] K. H. Cha, L. Hadjiiski, H.-P. Chan, A. Z. Weizer, A. Alva, R. H. Cohan, E. M. Caoili, C. Paramagul, R. K. Samala, Bladder cancer treatment response assessment in ct using radiomics with deep-learning, *Scientific reports* 7 (1) (2017) 1–12.

- [24] D. Bermejo-Peláez, S. Y. Ash, G. R. Washko, R. S. J. Estépar, M. J. Ledesma-Carbayo, Classification of interstitial lung abnormality patterns with an ensemble of deep convolutional neural networks, *Scientific reports* 10 (1) (2020) 1–15.
- [25] M. Gao, U. Bagci, L. Lu, A. Wu, M. Buty, H.-C. Shin, H. Roth, G. Z. Papadakis, A. Depeursinge, R. M. Summers, et al., Holistic classification of ct attenuation patterns for interstitial lung diseases via deep convolutional neural networks, *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* 6 (1) (2018) 1–6.
- [26] J. H. Moltz, L. Bornemann, J.-M. Kuhnigk, V. Dicken, E. Peitgen, S. Meier, H. Bolte, M. Fabel, H.-C. Bauknecht, M. Hittinger, et al., Advanced segmentation techniques for lung nodules, liver metastases, and enlarged lymph nodes in ct scans, *IEEE Journal of selected topics in signal processing* 3 (1) (2009) 122–134.
- [27] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, R. M. Summers, Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning, *IEEE transactions on medical imaging* 35 (5) (2016) 1285–1298.
- [28] L. Li, L. Qin, Z. Xu, Y. Yin, X. Wang, B. Kong, J. Bai, Y. Lu, Z. Fang, Q. Song, et al., Artificial intelligence distinguishes covid-19 from community acquired pneumonia on chest ct, *Radiology* (2020).
- [29] F. Shi, J. Wang, J. Shi, Z. Wu, Q. Wang, Z. Tang, K. He, Y. Shi, D. Shen, Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for covid-19, *IEEE reviews in biomedical engineering* (2020).
- [30] H. X. Bai, R. Wang, Z. Xiong, B. Hsieh, K. Chang, K. Halsey, T. M. L. Tran, J. W. Choi, D.-C. Wang, L.-B. Shi, et al., Ai augmentation of radiologist performance in distinguishing covid-19 from pneumonia of other etiology on chest ct, *Radiology* (2020) 201491.
- [31] S. Jégou, M. Drozdal, D. Vazquez, A. Romero, Y. Bengio, The one hundred layers tiramisú: Fully convolutional densenets for semantic segmentation, in: *CVPRW 2017, IEEE, 2017*, pp. 1175–1183.
- [32] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks., in: *CVPR, Vol. 1, 2017*, p. 3.
- [33] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: *International Conference on Medical image computing and computer-assisted intervention, Springer, 2015*, pp. 234–241.

- [34] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, W.-c. Woo, Convolutional lstm network: A machine learning approach for precipitation nowcasting, in: NIPS, 2015.
- [35] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, arXiv preprint arXiv:1709.01507 7 (2017).
- [36] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618–626.
- [37] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, B. Kim, Sanity checks for saliency maps, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems 31, Curran Associates, Inc., 2018, pp. 9505–9515.  
URL <http://papers.nips.cc/paper/8160-sanity-checks-for-saliency-maps.pdf>
- [38] S. Armato, G. McLennan, L. Bidaut, M. McNitt-Gray, C. Meyer, A. Reeves, H. MacMahon, R. Engelmann, R. Roberts, A. Starkey, P. Caligiuri, D. Aberle, M. Brown, R. Pais, D. Qing, P. Batra, C. Jude, I. Petkovska, A. Biancardi, B. Zhao, C. Henschke, D. Yankelevitz, D. Max, A. Farooqi, E. Hoffman, E. van Beek, A. Smith, E. Kazerooni, P. Bland, G. Laderach, G. Gladish, R. Munden, L. Quint, L. Schwartz, B. Sundaram, L. Dodd, C. Fenimore, D. Gur, N. Petrick, J. Freymann, J. Kirby, B. Hughes, A. Castele, S. Gupte, M. Sallam, M. Heath, M. Kuhn, E. Dharaiya, R. Burns, D. Fryd, M. Salganicoff, V. Anand, U. Shreter, S. Vastagh, B. Croft, L. Clarke, 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 (2) (2011) 915–931. doi:10.1118/1.3528204.
- [39] J. Hofmanninger, F. Prayer, J. Pan, S. Rohrich, H. Prosch, G. Langs, Automatic lung segmentation in routine imaging is a data diversity problem, not a methodology problem, arXiv preprint arXiv:2001.11767 (2020).
- [40] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.