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.

Artificial Intelligence In Medicine 118 (2021) 102114

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

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

An explainable AI system for automated COVID-19 assessment and lesion categorization from CT-scans

Matteo Pennisi [a,1], Isaak Kavasidis [a,*,1], Concetto Spampinato [a], Vincenzo Schinina [b], Simone Palazzo [a], Federica Proietto Salanitri [a], Giovanni Bellitto [a], Francesco Rundo [c], Marco Aldinucci [d], Massimo Cristofaro [b], Paolo Campioni [b], Elisa Pianura [b], Federica Di Stefano [b], Ada Petrone [b], Fabrizio Albarello [b], Giuseppe Ippolito [b], Salvatore Cuzzocrea [e], Sabrina Conoci [e]

[a] DIEEI, University of Catania, Catania, Italy
[b] National Institute for infectious disease, "Lazzaro Spallanzani" Department, Rome, Italy
[c] STMicroelectronics - ADG Central R&D, Catania, Italy
[d] Department of Computer Science, University of Turin, Turin, Italy
[e] ChimBioFaram Department, University of Messina, Messina, Italy

ARTICLE INFO

Keywords:
COVID-19 detection
Lung segmentation
Deep learning

ABSTRACT

COVID-19 infection caused by SARS-CoV-2 pathogen has been 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 automatically identifying lung parenchyma and lobes. Next, we combine the segmentation network with classification networks for COVID-19 identification and lesion categorization. We compare the model's classification results with those obtained by three expert radiologists on a dataset of 166 CT scans. Results showed a sensitivity of 90.3% and a specificity of 93.5% for COVID-19 detection, at least on par with those yielded by the expert radiologists, and an average lesion categorization accuracy of about 84%. Moreover, a significant role is played by prior lung and lobe segmentation, that allowed us to enhance classification performance by over 6 percent points. The interpretation of the trained AI models 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 for making decisions and that proactively involves 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 155 million of people (with about 3.5 M 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

* Corresponding author.
E-mail address: kavasidis@dieei.unict.it (I. Kavasidis).
[1] Equal contribution

https://doi.org/10.1016/j.artmed.2021.102114
Received 25 September 2020; Received in revised form 6 May 2021; Accepted 12 May 2021
Available online 21 May 2021

widely used, this biomolecular method is time-consuming requiring up to several hours for being processed.

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 AI-based system to achieve both *COVID19 identification* and *lesion categorization* (ground glass, crazy paving and 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. Finally, 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.

To sum up, the main contributions of this paper are the following:

- We propose a novel lung-lobe segmentation network outperforming state-of-the-art models;
- We employ the segmentation network to drive a classification network that first identifies CT scans of COVID-19 patients, and, afterwards, automatically categorizes specific lesions;
- We then provide interpretation of the decisions made by the employed models and discover that, indeed, the proposed approach focuses on specific COVID-19 lesions for distinguishing whether a CT scan is related to positive 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. To the best of our knowledge, this is the first publicly available platform that offers COVID-19 diagnosis services based on CT scans with explainability capabilities. The free availability to the general public for such an important task, while the pandemic is still in full effect, is, in our opinion, an invaluable aid to the medical community.

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.

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.

The input to the model is a sequence of 3 consecutive slices – suitably resized to 224 × 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

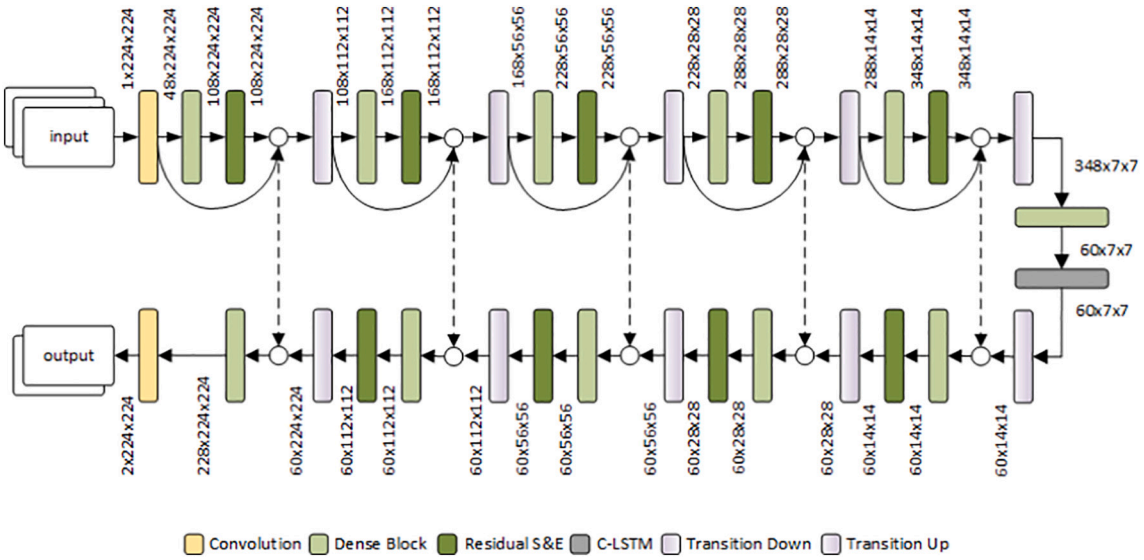

Fig. 1. The proposed segmentation architecture, consisting of a downsampling path (*top*) and an upsampling path (*bottom*), interconnected by skip connections and by the bottleneck layer.

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.

3.1.1. 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 in-formative 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 $C \times 1 \times 1$ 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.

3.1.2. 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 parenchyma lung segmentation (through the segmentation model presented in Section 3.1) a deep classification model analyzes slice by slice each segmented CT scan, and decides whether a single slice contains evidence of the COVID-19 disease. Note that slice-based COVID-19 classification is only the initial step towards the final prediction, which takes into account *all* per-slice predictions, and assigns the "positive" label in presence of a certain number of slices (10% of the total) that the model has identified as COVID-19 positive. Hence, COVID-19 assessment is actually carried out per patient, by combining per-slice predictions.

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.

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 by investigating output gradient with respect to feature map activations. More specifically, GradCAM generates class-discriminative localization map for any class c by first computing the gradient of the score for 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.:

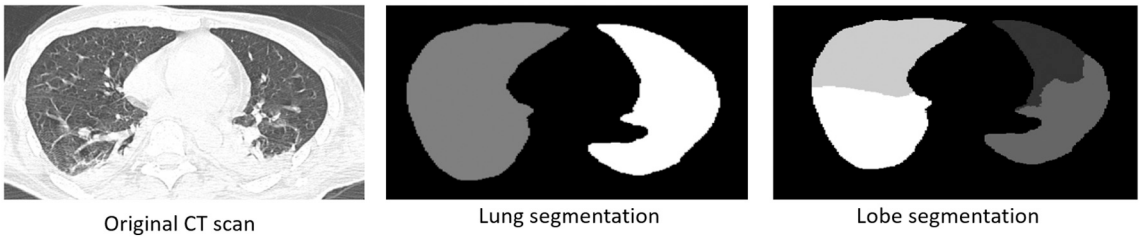

Fig. 2. Example of lung and lobes segmentation.

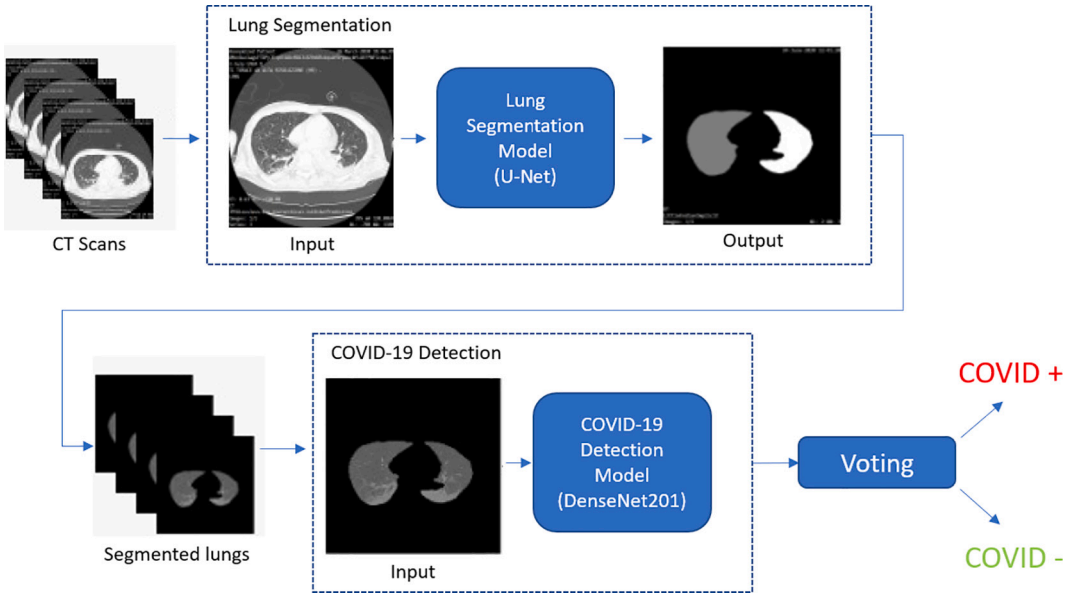

Fig. 3. Overview of the **COVID-19 detection approach** for CT scan classification as either COVID-19 positive or negative.

$$w_k^c = \sum_i \sum_j \frac{ay^c}{aA_{ij}^k} \tag{1}$$

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.:

VarGrad is a technique used in combination to GradGAM and consists in 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. 4.

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 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. 5). 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. 6).

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. 7).

The AI-empowered web system also integrates 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 and annotations

4.1.1. Data

Our dataset contains overall 166 CT scans: 72 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 of

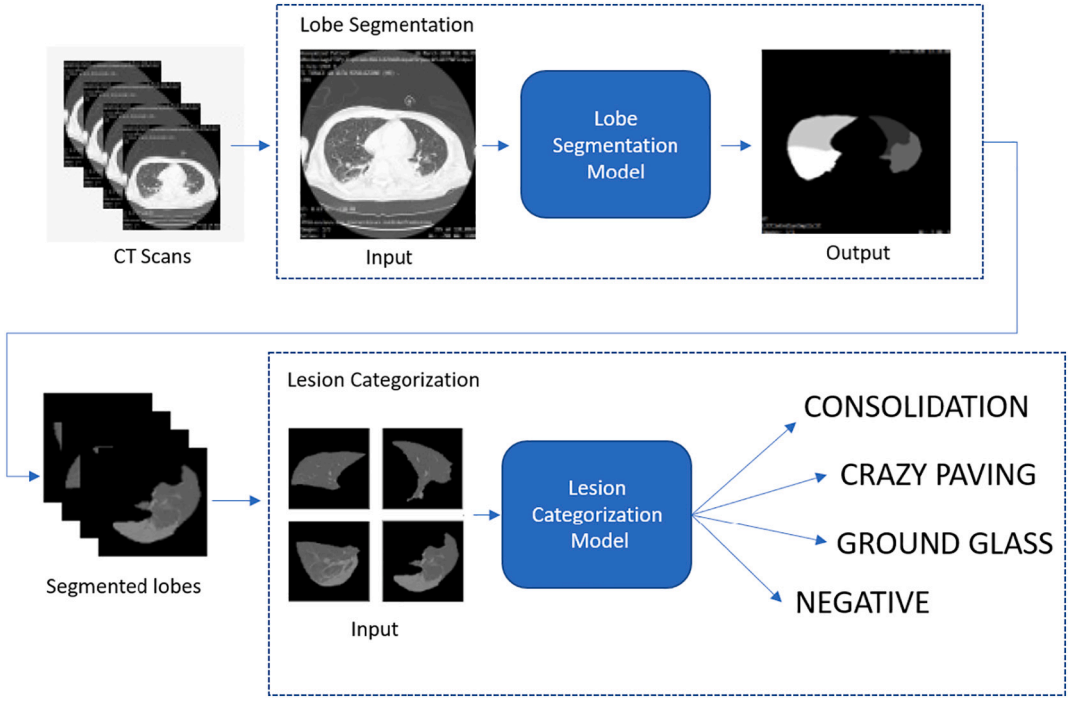

Fig. 4. Overview of **COVID-19 lesion categorization** approach.

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

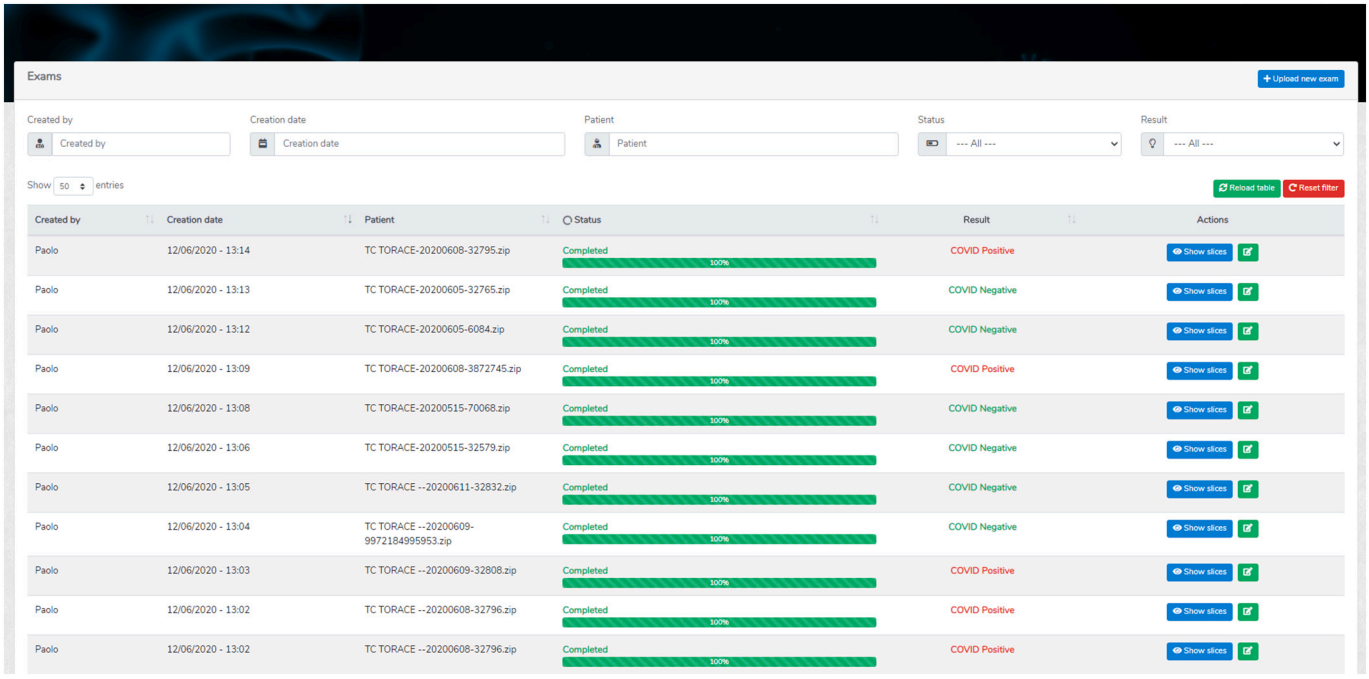

Fig. 5. The main page of the AI-empowered web GUI for explainable AI. The user is presented with a list of the CT scan classifications reporting the models' prediction.

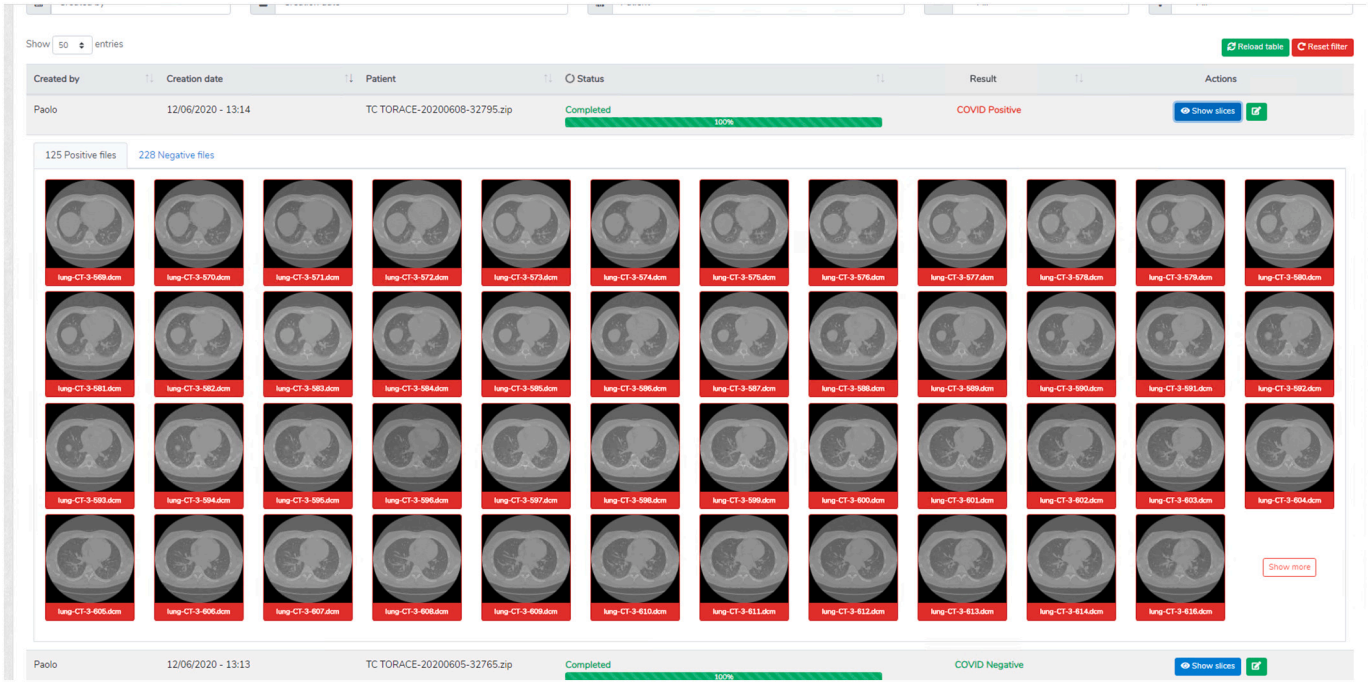

Fig. 6. The summarized classification result showing the CT slices that the neural network classified as positive or negative.

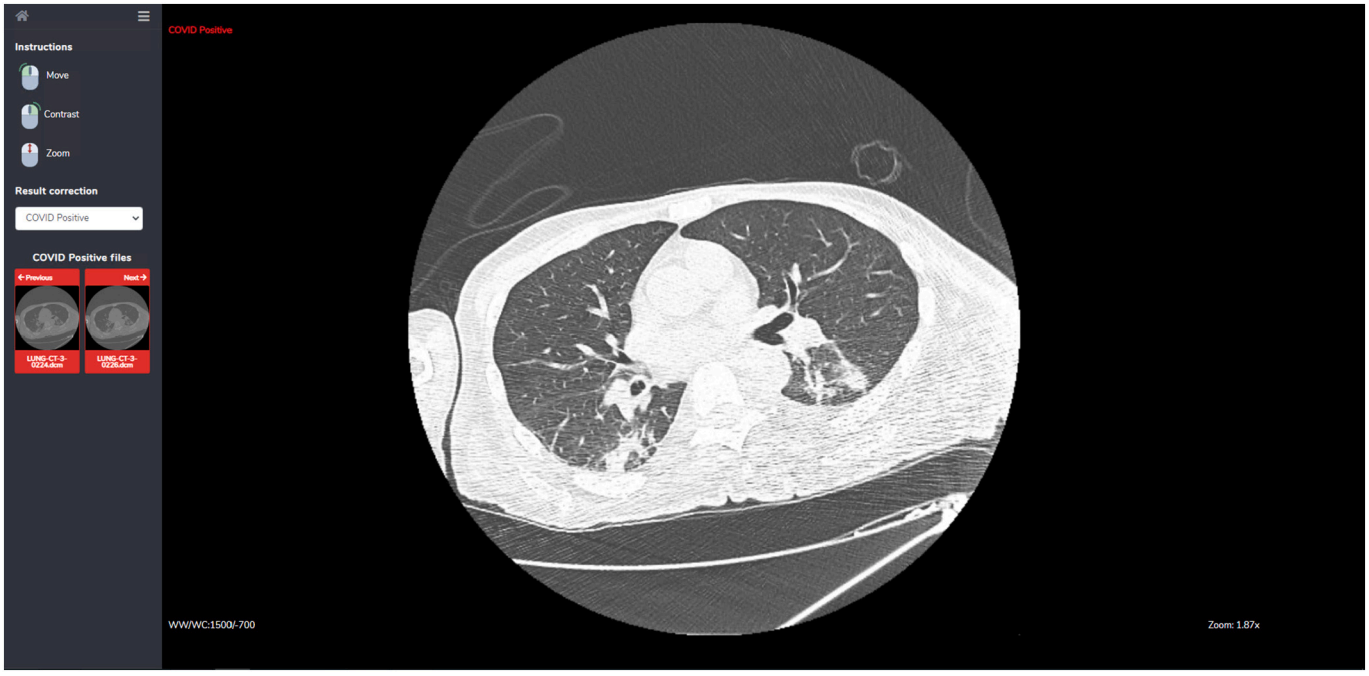

Fig. 7. The slice inspection screen. In this screen the user can inspect each single slice and the AI models' decisions.

COVID-19 negative subjects (35 patients with interstitial pneumonia but tested negative to COVID19 and 59 controls).

CT scans were performed on a multi-detector row helical CT system scanner[2]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, 1000–1500H; level, −700H) and mediastinal (window width, 350H; level, 35–40H) settings were reviewed on a picture archiving and communication system workstation.[3] For training the lung/lobe segmentation model we adopted a combination of the LIDC [38], LTRC[4] and [39] datasets, for a total of 300 CT scans.

[2] Bright Speed, General Electric Medical Systems, Milwaukee, WI

[3] Impax ver. 6.6.0.145, AGFA Gevaert SpA, Mortsel, Belgium

[4] https://ltrcpublic.com/

4.1.2. Annotations

We perform both COVID-19 identification and lesion categorization, thus the annotations are different according to the task. For COVID19 identification, ground truth consists of the results of the molecular and an IgG/IgM antibody test. Among the set of 166 CT scans, we used 95 scans (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.

For lesion categorization, instead, 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 (left/right/central and posterior/anterior) of the lesion. In total, about 2400 slices were annotated with COVID-19 lesions and about 3000 slices of negative patients with no lesion. Table 1 provides an overview of all the CT scans and lesion annotations in our dataset. As for lung segmentation, annotations on lung/lobe areas were done manually by the same three expert radiologists who carried out lesion categorization.

4.2. Training procedure

4.2.1. COVID-19 identification model

The COVID-19 detection network is a DenseNet201, which was used pretrained 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. Given the class imbalance in the training set, we used the weighted binary cross-entropy (defined in Eq. (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:

$$WBCE = -w\,[y\,logx + (1-y)\,log(1-x)] \tag{3}$$

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 selected according to the best operating point in the ROC curve.

4.2.2. COVID-19 lesion categorization model

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 2488 annotated slices; in each slice multiple lesion annotations with relative location (in lobes) are available. Thus, after segmenting lobes from these images we obtained 5264 lobe images. We did the same on CT slices of negative patients (among the 2950 available as shown in Table 1) and selected 5264 lobe images without lesions. Thus, in total, 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:

$$WCE = -w\sum^{c} ylog(x) \tag{4}$$

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

$$w_c = \frac{N - Nc}{N} \tag{5}$$

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.

4.2.3. Lung/lobe segmentation model

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 × 224, initial learning rate to 0.0001, weight decay to 0.0001 and batch size to 2, with RMSProp as optimizer. When CLSTMs 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. All experiments have been executed using the HPC4AI infrastructure [41].

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;

Table 1
CT Dataset for training and testing the deep models.

	CT scans	Annotated slices			
		Ground glass	Crazy paving	Consolidation	Total
Positives	72	1035	757	598	2390
Negatives	94	–	–	–	2988

- 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 Table 2. It has to 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 diagnosis

We here report the results for COVID-19 diagnosis, i.e., classification between positive and negative cases. In this analysis, we compare model results to those yielded by three experts with different degree of expertise:

- Radiologist 1: a physician expert in thoracic radiology (~30 years of experience) with over 30,000 examined CT scans;
- Radiologist 2: a physician expert in thoracic radiology (~10 years of experience) with over 9000 examined CT scans;
- Radiologist 3: a resident student in thoracic radiology (~3 years of experience) with about 2000 examined CT scans.

It should be noted that the gold standard employed in the evaluation is provided by molecular and antibody tests, hence radiologists' assessments are not the reference for performance comparison.

We also assess the role of prior segmentation on the performance. This means that in the pipelines showed in Figs. 3 and 4 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, specificity and AUC, and are given in Tables 3, 4 and 5. Note that the AUC is a reliable metric in our scenario, since we explicitly defined the test set to be balanced among classes. More recent techniques [42] may be suitable when this assumption does not hold, as is often the case for new or rare diseases.

Our results show that the AI model with lung segmentation achieves higher performance than expert radiologists. However, given the relatively small scale of our dataset, statistical analysis carried out with the Chi-squared test does not show any significant difference between AI models and radiologists. Furthermore, performing lung segmentation improves by about 6 percent points both the sensitivity and the specificity, demonstrating its effectiveness.

In addition, we also measure how the sensitivity of the COVID-19 identification changes w.r.t. the level of disease severity. In particular, we categorize the 31 positive cases into three classes according to the percentage of the affected lung area: low severity (11 cases), medium severity (11 cases), high severity (9 cases). Results are reported in Table 6 that shows how our AI-based method seems to be yielding better assessment than the domain experts, especially at the beginning of the disease (low severity). This is important as an earlier disease detection may lead to a more favorable outcome. In case of high severity, two out of three radiologists showed difficulties in correctly identifying the COVID-19, mainly because when the affected lung area is significant, the typical COVID patterns are less visible. However, even in this case, our deep learning model was able to discriminate robustly COVID cases.

As a backbone model for COVID-19 identification, we employ

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 voxels.

Model	Lung segmentation	Lobe segmentation
Baseline Tiramisu [31]	89.41 ± 0.45	77.97 ± 0.31
Baseline + Res-SE	91.78 ± 0.52	80.12 ± 0.28
Baseline + C-LSTM	91.49 ± 0.57	79.47 ± 0.38
Baseline + Res-SE + C-LSTM	**94.01 ± 0.52**	**83.05 ± 0.27**

Table 3
Sensitivity (in percentage 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.

	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	**90.3**	**[79.5–96.5]**

Table 4
Specificity (in percentage 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.

	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	**93.5**	**[83.5–98.5]**

Table 5
AUC (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.

	AUC	C.I. (95%)
Radiologist 1	0.83	[0.72–0.93]
Radiologist 2	0.87	[0.78–0.96]
Radiologist 3	0.80	[0.69–0.91]
AI Model without lung segmentation	0.94	[0.87–1.00]
AI Model with lung segmentation	**0.95**	**[0.89–1.00]**

Table 6
Sensitivity (in percentage) changes w.r.t. disease severity. From the 31 test CTs for positive patients: 11 are with low severity, 11 with medium severity, and 9 with high severity. Values in parentheses indicate 95% confidence intervals (CI).

	Low severity	Medium severity	High severity
Radiologist 1	72.7 (50.6–88.5)	100.0 (90.9–70.6)	77.8 (54.7–92.6)
Radiologist 2	72.7 (50.6–88.5)	90.9 (70.6–100.0)	100.0 (81.5–100.0)
Radiologist 3	63.6 (42.3–81.3)	100 (90.9–70.6)	77.8 (54.7–92.6)
$Model_{wo}$ segmentation	72.7 (50.6–88.5)	90.9 (70.6–100.0)	88.9 (67.0–99.2)
$Model_{w}$ segmentation	81.8 (59.6–94.9)	90.9 (70.6–100.0)	100.0 (81.5–100.0)

Table 7
COVID-19 classification accuracy (in percentage) by several state of the art models. Values in parentheses indicate 95% confidence intervals (CI).

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	**90.3 (79.5–96.5)**	**93.5 (83.5–98.5)**	**91.9 (81.5–97.5)**
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)

DenseNet201 since it yields the best performance when compared to other state of the art models, as shown in Table 7. In all tested cases, we use upstream segmentation through the model described in Section 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 GradCAM [36] with VarGrad [37][5] and, Fig. 8 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).

4.3.3. COVID-19 lesion categorization

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 8. Note that no comparison with radiologists is carried out in this case, since ground-truth labels on lesion types are provided by radiologists themselves, hence they are the reference used to evaluate model accuracy.

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 on the *Crazy Paving* class, i.e., 98.3% against 57.1%.

[5] https://captum.ai/

4.4. Discussion

Although COVID-19 diagnosis from CT scans may seem an easy task for experienced radiologists, our results show that this is not always the case: in this scenario, the approach we propose has demonstrated its capability to carry out the same task with an accuracy that is at least on par with, or even higher than, human experts, thus showing the potential impact that these techniques may have in supporting physicians in decision making. Artificial intelligence, in particular, is able to accurately identify not only if a CT scan belongs to a positive patient, but also the type of lung lesions, in particular the smaller and less defined ones (as those highlighted in Fig. 8). As shown, the combination of segmentation and classification techniques provides a significant improvement in the sensitivity and specificity of the proposed method.

Of course, although the results presented in this work are very promising in the direction of establishing a clinical practice that is supported by artificial intelligence models, there is still room for improvement. One of the limitations of our work is represented by the relatively low number of samples available for the experiments. In order to mitigate the impact of this issue, we carried out confidence level analysis to demonstrate the statistical significance of our results. Moreover, the employed dataset consists of images taken by the same CT scanner, not tested in multiple scanning settings. This could affect the generalization of the method on images taken by other CT scanner models; however, this issue can be tackled by domain adaptation techniques for the medical imaging domain, which is an active research topic [43,44,45].

Finally, one of the key features of our approach is the integration of explainability functionalities that may help physicians in understanding the reasons underlying a model's decision, increasing in turn, the trust that experts have in AI–enabled methods. Future developments in this regard should explore, in addition to model explainability, also *causability* features in order to evaluate the quality of the explanations provided [46,47].

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 90.3% and a specificity of 93.5% for COVID-19 detection and average lesion categorization accuracy of about 84%. Results also show that a significant role is played by prior lung and lobe segmentation, that allowed us to enhance diagnosis performance of about 6 percent points.

The AI models are then integrated into a user-friendly GUI to support AI explainability for radiologists, which is publicly available at http://perceivelab. 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 decisions and that proactively involves in the loop to further improve the COVID-19 understanding.

The results obtained both for COVID-19 identification and lesion categorization pave the way to further improvements, driven towards the implementation of an advanced COVID-19 CT/RX diagnostic pipeline, that is interpretable, robust and able to provide not only disease identification and differential diagnosis, but also the risk of disease progression.

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.

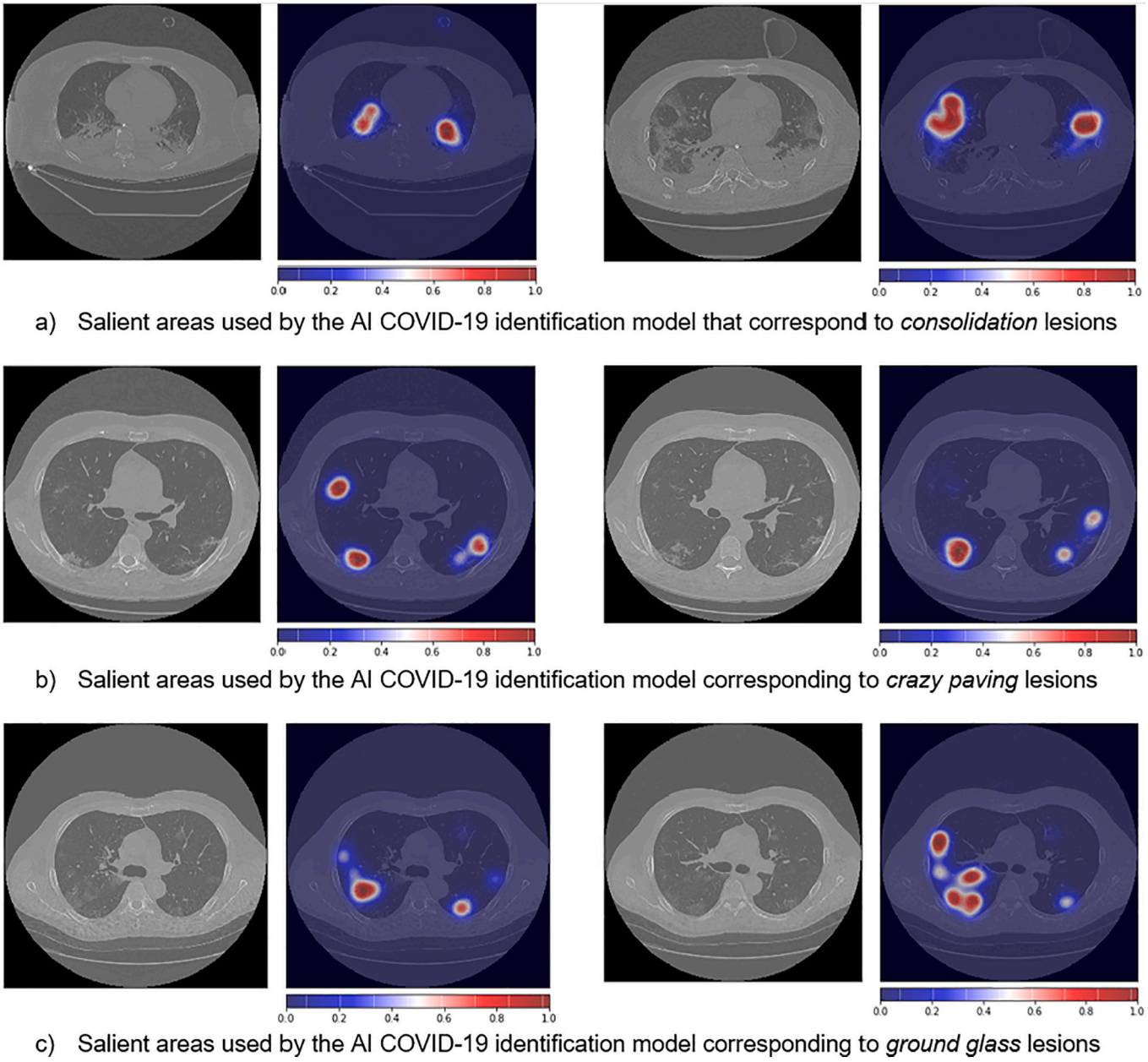

Fig. 8. Lung salient areas identified automatically by the AI model for CT COVID-19 identification.

Table 8
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).

	Model *no_segm*	Model *w_segm*
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%

Declaration of competing interest

None.

Acknowledgment

This work has been also partially supported by:

- The REHASTART project funded by Regione Sicilia (PO FESR 2014/2020 - Azione 1.1.5)
- The “Go for IT” project funded by the Conference of Italian University Rectors (CRUI)
- The DeepHealth project, funded under the European Union's Horizon 2020 framework, grant agreement No. 825111
- The HPC4AI project funded by Regione Piemonte (POR FESR 2014-20 - INFRA-P).

We also 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, Ilaria 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 Gall'ı, 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 Schinin'a, Paola Scognamiglio, Laura Scorzolini, Giulia Stazi, Francesco Vaia, Francesco Vairo, Maria Beatrice Valli) for the technical discussion and critical reading of this manuscript.

Finally, we would like to thank Antonino Bonanno for his contribution in the development of the website.

References

[1] Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A novel coronavirus from patients with pneumonia in China. N Engl J Med 2019;2020.
[2] W. H. Organization, et al. Novel coronavirus (2019-ncov): situation report8; 2020.
[3] Huang P, Liu T, Huang L, Liu H, Lei M, Xu W, et al. Use of chest ct in combination with negative rt-pcr assay for the 2019 novel coronavirus but high clinical suspicion. Radiology 2020;295(1):22–3.
[4] Ng M-Y, Lee EY, Yang J, Yang F, Li X, Wang H, et al. Imaging profile of the covid-19 infection: radiologic findings and literature review. Radiology 2020;2(1):e200034.
[5] Liu H, Liu F, Li J, Zhang T, Wang D, Lan W. Clinical and ct imaging features of the covid-19 pneumonia: focus on pregnant women and children. J Infect 2020;80(5):e7–13.
[6] Chung M, Bernheim A, Mei X, Zhang N, Huang M, Zeng X, et al. Ct imaging features of 2019 novel coronavirus (2019-ncov). Radiology 2020;295(1):202–7.
[7] Rundo F, Spampinato C, Banna GL, Conoci S. 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 2019;8(10):1134.
[8] Allam Z, Jones DS. 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, J Travel Med 27 (5) (2020) taaa080.
[10] Zheng N, Du S, Wang J, Zhang H, Cui W, Kang Z, et al. Predicting covid-19 in China using hybrid ai model. IEEE Trans Cybern 2020:2891–904. https://doi.org/10.1109/TCYB.2020.2990162.
[11] Bai X, Fang C, Zhou Y, Bai S, Liu Z, Xia L, et al. Predicting covid-19 malignant progression with ai techniques. 2020.
[12] Liang W, Yao J, Chen A, Lv Q, Zanin M, Liu J, et al. Early triage of critically ill covid-19 patients using deep learning. Nat Commun 2020;11(1):1–7.
[13] Cosic K, Popovic S, Sarlija M, Kesedzic I, Jovanovic T. Artificial intelligence in prediction of mental health disorders induced by the covid-19 pandemic among health care workers. Croat Med J 2020;61(3):279.
[14] Mohanty S, Rashid MHA, Mridul M, Mohanty C, Swayamsiddha S. Application of artificial intelligence in covid-19 drug repurposing. Diabetes Metab Syndr Clin Res Rev 2020;14(5):1027–31.
[15] Ke Y-Y, Peng T-T, Yeh T-K, Huang W-Z, Chang S-E, Wu S-H, et al. Artificial intelligence approach fighting covid-19 with repurposing drugs. Biom J 2020;43(4):355–62.
[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] Brunese L, Mercaldo F, Reginelli A, Santone A. Explainable deep learning for pulmonary disease and coronavirus covid-19 detection from x-rays. Comput Methods Prog Biomed 2020;196:105608.
[18] Huang L, Han R, Ai T, Yu P, Kang H, Tao Q, et al. Serial quantitative chest ct assessment of covid-19: deep-learning approach. Radiology 2020;2(2):e200075.
[19] Nardelli P, Jimenez-Carretero D, Bermejo-Pelaez D, Washko GR, Rahaghi FN, Ledesma-Carbayo MJ, et al. Pulmonary artery– vein classification in ct images using deep learning. IEEE Trans Med Imaging 2018;37(11):2428–40.
[20] Navab N, Hornegger J, Wells WM, Frangi A. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part IIIVol. 9351. Springer; 2015.
[21] Mei X, Lee H-C, Diao K-Y, Huang M, Lin B, Liu C, et al. Artificial intelligence–enabled rapid diagnosis of patients with covid-19. Nat Med 2020:1–5.
[22] Setio AAA, Ciompi F, Litjens G, Gerke P, Jacobs C, Van Riel SJ, et al. Pulmonary nodule detection in ct images: false positive reduction using multi-view convolutional networks. IEEE Trans Med Imaging 2016;35(5):1160–9.
[23] Cha KH, Hadjiiski L, Chan H-P, Weizer AZ, Alva A, Cohan RH, et al. Bladder cancer treatment response assessment in ct using radiomics with deep-learning. Sci Rep 2017;7(1):1–12.
[24] Bermejo-Pelaez D, Ash SY, Washko GR, Estepar RSJ, Ledesma-Carbayo MJ. Classification of interstitial lung abnormality patterns with an ensemble of deep convolutional neural networks. Sci Rep 2020;10(1):1–15.
[25] Gao M, Bagci U, Lu L, Wu A, Buty M, Shin H-C, et al. Holistic classification of ct attenuation patterns for interstitial lung diseases via deep convolutional neural networks. Comput Methods Biomech Biomed Engin 2018;6(1):1–6.
[26] Moltz JH, Bornemann L, Kuhnigk J-M, Dicken V, Peitgen E, Meier S, et al. Advanced segmentation techniques for lung nodules, liver metastases, and enlarged lymph nodes in ct scans. IEEE J Sel Top Sign Proces 2009;3(1):122–34.
[27] Shin H-C, Roth HR, Gao M, Lu L, Xu Z, Nogues I, et al. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging 2016;35(5):1285–98.
[28] Li L, Qin L, Xu Z, Yin Y, Wang X, Kong B, et al. Artificial intelligence distinguishes covid-19 from community acquired pneumonia on chest ct. Radiology 2020 [Published online 2020 Mar 19].
[29] Shi F, Wang J, Shi J, Wu Z, Wang Q, Tang Z, et al. Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for covid-19. IEEE Rev Biomed Eng 2021;14:4–15.
[30] Bai HX, Wang R, Xiong Z, Hsieh B, Chang K, Halsey K, et al. Ai augmentation of radiologist performance in distinguishing covid-19 from pneumonia of other etiology on chest ct. Radiology 2020:201491 [Published online 2020 Apr 27].
[31] S. Jegou, M. Drozdzal, D. Vazquez, A. Romero, Y. Bengio, The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation, in: CVPRW 2017, IEEE, 2017, pp. 1175–1183.
[32] Huang G, Liu Z, Maaten L Van Der, Weinberger KQ. Densely connected convolutional networks. In: CVPR. vol. 1; 2017. p. 3.
[33] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer; 2015. p. 234–41.
[34] Xingjian S, Chen Z, Wang H, Yeung D-Y, Wong W-K, Woo W-C. Convolutional lstm network: a machine learning approach for precipitation nowcasting. NIPS; 2015.
[35] Hu J, Shen L, Sun G. Squeeze-and-excitation networks, arXiv preprint arXiv:170901507 7. 2017.
[36] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: visual explanations from deep networks via gradient-based localization. In: 2017 IEEE international conference on computer vision (ICCV); 2017. p. 618–26.
[37] Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity checks for saliency maps. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, editors. Advances in neural information processing systems 31. Curran Associates, Inc; 2018. p. 9505–15 [URL], http://papers.nips.cc/paper/8160-sanity-checks-for-saliency-maps.pdf.
[38] Armato S, McLennan G, Bidaut L, McNitt-Gray M, Meyer C, Reeves A, et al. The lung image database consortium, (lidc) and image database resource initiative (idri):: a completed reference database of lung nodules on ct scans. Med Phys 2011;38(2):915–31. https://doi.org/10.1118/1.3528204.
[39] Hofmanninger J, Prayer F, Pan J, Rohrich S, Prosch H, Langs G. Automatic lung segmentation in routine imaging is a data diversity problem, not a methodology problem, arXiv preprint arXiv:200111767. 2020.
[40] Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. Imagenet: a largescale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, Ieee; 2009. p. 248–55.
[41] Aldinucci M, Rabellino S, Pironti M, Spiga F, Viviani P, Drocco M, et al. HPC4AI, an AI-on-demand federated platform Endeavour. Ischia, Italy: ACM Computing Frontiers; 2018. https://doi.org/10.1145/3203217.3205340 [URL], https://iris.unito.it/retrieve/handle/2318/1765596/689772/2018_hpc4ai_ACM_CF.pdf.
[42] Carrington AM, Fieguth PW, Qazi H, Holzinger A, Chen HH, Mayr F, et al. A new concordant partial auc and partial c statistic for imbalanced data in the evaluation of machine learning algorithms. BMC Med Inform Decis Mak 2020;20(1):1–12.
[43] Perone CS, Ballester P, Barros RC, Cohen-Adad J. Unsupervised domain adaptation for medical imaging segmentation with self-ensembling. NeuroImage 2019;194:1–11.
[44] Mahmood F, Chen R, Durr NJ. Unsupervised reverse domain adaptation for synthetic medical images via adversarial training. IEEE Trans Med Imaging 2018;37(12):2572–81.
[45] Madani A, Moradi M, Karargyris A, Syeda-Mahmood T. Semisupervised learning with generative adversarial networks for chest x-ray classification with ability of

data domain adaptation. In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). IEEE; 2018. p. 1038–42.
[46] Holzinger A. From machine learning to explainable ai. In: 2018 world symposium on digital intelligence for systems and machines (DISA). IEEE; 2018. p. 55–66.
[47] Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. Wiley Interdiscip Rev 2019;9 (4):e1312.