ai-corona: Radiologist-Assistant Deep Learning Framework for COVID-19 Diagnosis in Chest CT Scans

M. Yousefzadeh[1,2,3☯], P. Esfahanian[1,2☯], S. M. S. Movahed[4,3], S. Gorgin[1,5], D. Rahmati[1,6], A. Kiani[7], S. A. Nadji[9], S. Haseli[8], M. Hoseinyazdi[10], J. Roshandel[8], M. Bakhshayesh Karam[8], A. Abedini[8*], R. Lashgari[2*]

1 School of Computer Science, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran
2 Institute of Medical Science and Technology, Shahid Beheshti University, Tehran, Iran
3 Ibn-Sina Multidisciplinary Laboratory, Department of Physics, Shahid Beheshti University, Tehran, Iran
4 Department of Physics, Shahid Beheshti University, Tehran, Iran
5 Department of Electrical Engineering and Information Technology, Iranian Research Organization for Science and Technology (IROST), Tehran, Iran
6 Department of Computer Science and Engineering, Shahid Beheshti University, Tehran, Iran
7 Tracheal Diseases Research Center, National Research Institute of Tuberculosis and Lung Diseases (NRITLD), Shahid Beheshti University of Medical Sciences and Health Services, Tehran, Iran
8 Chronic Respiratory Diseases Research Center, National Research Institute of Tuberculosis and Lung Diseases (NRITLD), Shahid Beheshti University of Medical Sciences and Health Services, Tehran, Iran
9 Virology Research Center, National Research Institute of Tuberculosis and Lung Diseases (NRITLD), Shahid Beheshti University of Medical Sciences and Health Services, Tehran, Iran
10 Medical Imaging Research Center, Department of Radiology, Shiraz University of Medical Sciences, Shiraz, Iran

☯These authors contributed equally to this work.
*Corresponding authors rezalashgari@gmail.com

Abstract

Generation of medical assisting tools using recent artificial intelligence advances is beneficial for the medical workers in the global fight against COVID-19 outbreak. In this article we introduce *ai-corona*, a radiologist-assistant deep learning framework for COVID-19 infection diagnosis using the chest CT scans. Our framework incorporates an Efficient NetB3-based feature extractor. We employed three independent dataset in this work named: CC-CCII, MDH, and MosMedData; all includes 7184 scans from 5693 subjects which contained pneumonia, common pneumonia (CP), non-pneumonia, normal and COVID-19 classes. We evaluated *ai-corona* on test sets from the CC-CCII set and MDH cohort and the entirety of the MosMedData cohort, for which it gained AUC score of 0.997, 0.989, and 0.954, respectively. We further compared our framework's performance with other deep learning models developed on our employed data sets, as well as RT-PCR. Our results show that *ai-corona* outperforms all. Lastly, our framework's diagnosis capabilities was evaluated as assistant to several experts. We demonstrated an increase in both speed and accuracy of expert diagnosis when incorporating *ai-corona*'s assistance.

Introduction

Since the beginning of 2020, novel Coronavirus Disease 2019 (COVID-19) has widely spread globally and has taken countless lives. Patients infected with COVID-19 commonly display symptoms such as fever, cough, tiredness, breathing difficulties, and muscle ache [1–3].

Currently, the most common method of testing for COVID-19 is Real-Time Polymerase Chain Reaction (RT-PCR) to detect viral nucleotides from upper respiratory specimen obtained by nasopharyngeal, oropharyngeal, or nasal mid-turbinate swab [4]. It has been shown that RT-PCR has several drawbacks. Reports suggest that since oropharyngeal swabs tend to detect COVID-19 less frequently than nasopharyngeal swabs, RT-PCR tends to have a high false-negative rate. Furthermore, RT-PCR has demonstrated a decrease in sensitivity to below 70% due to a low viral nucleic acid load and inefficiencies in its detection. This might be caused by immature development of nucleic acid detection technology, variation in detection rate by using different gene region targets, or a low patient viral load [5]. Besides, the availability of test kits and expert personnel to take them is still suboptimal in some countries. Not to mention the extended time period for the test completion contributes to ruling out RT-PCR as a reliable early detection and screening method [7–9]. In contrast to RT-PCR, diagnosis from other measurements such as chest Computed Tomography (CT) and blood factors is shown to be an effective early detection and screening method with high sensitivity in both detection [10] and anticipation of the severity of the disease [6].

Chest CT scan of a COVID-19 infected patient reveals bilateral peripheral involvement in multiple lobes with areas of consolidation and ground-glass opacity that progresses to "crazy-paving" patterns as the disease develops [10]. Asymmetric bilateral subpleural patchy ground-glass opacities and consolidation with a peripheral or posterior distribution, mainly in middle and lower lobes, are described as the most common image finding of COVID-19 [11]. To elaborate more, additional common findings include interlobular septal thickening, air bronchogram, and crazy paving pattern in the intermediate stages of the disease [10]. The most common pattern in the advanced stage is subpleural parenchymal bands, fibrous stripes, and subpleural resolution. Nodules, cystic change, pleural effusion, pericardial effusion, lymphadenopathy, cavitation, CT halo sign, and pneumothorax are some of the uncommon but possible findings [10, 12]. Recent studies indicate that organizing pneumonia, which occurs in the course of viral infection, is pathologically responsible for the clinical and radiological manifestation of Coronavirus pneumonia [11].

Deep learning is an area of Artificial Intelligence (AI) that has demonstrated tremendous capabilities in image feature extraction and has been recognized as a successful tool in medical imaging-based diagnosis, performing exceptionally with modalities such as X-Ray, Magnetic Resonance Imaging (MRI), and CT [13–16, 25]. Recently, the research of AI-assisted respiratory diagnosis, especially pneumonia, has gained a lot of attention. One of the well-established standards in this research is the comparison of AI with expert medical and radiology professionals. As a pioneering work in this filed, [17] introduced a radiologist-level deep learning framework trained and validated on the ChestX-ray8 dataset [18] for the detection of 14 abnormalities, including pneumonia, in chest X-Ray images, which was further developed to propose a deep learning framework with pneumonia detection capabilities equivalent to that of expert radiologists [19]. [20] introduced a novel dataset of chest X-Ray images annotated with 14 abnormalities (7 the same as ChestX-ray8) and a state-of-the-art deep learning framework. Lastly, [21] proposed a deep learning framework with a feature extractor based on AlexNet [22] to create a model capable of accurately diagnosing knee injuries from MRI scans and further showcases the positive impact of AI assistance in expert diagnosis.

In COVID-19 related research, [7] has reported a sensitivity of 0.59 for RT-PCR

test kit and 0.88 for CT-based diagnosis for patients with COVID-19 infection, and a radiologist sensitivity of 0.97 in diagnosing COVID-19 infected patients with RT-PCR confirmation. Furthermore, [23] introduces a deep learning framework with a 0.96 AUC score in the diagnosis of RT-PCR confirmed COVID-19 infected patients. Zhang *et al.* [24] proposed a model that on a dataset of 4154 subjects achieved an AUC score of 0.98 for diagnosing COVID-19 from two other classes; normal and CP (Common Pneumonia *i.e.* non COVID-19 viral and bacterial pneumonia). They further made their dataset, CC-CCII [24], publicly available. In addition, the model proposed by Jin *et al.* [26], developed on a dataset of 9025 subjects, which is an amalgamation of their own data and several other public dataset (*e.g.* LIDC–IDRI [27], Tianchi-Alibaba [28], MosMedData [29], and CC-CCII), gained an accuracy of 0.975 for diagnosing between COVID-19 and three other classes (non-pneumonia, non-viral community acquired pneumonia, Influenza-A/B), 0.921 for between COVID-19 and the CP and Normal classes on the CC-CCII dataset, and 0.933 for between COVID-19 from non-pneumonia on the MosMedData cohort. Further, this work manages to astoundingly diagnose between COVID-19 and influenza type-A, which is surprising given the small amount of influenza data in their study.

In this paper, we present *ai-corona*, a radiologist-level deep learning framework for COVID-19 diagnosis in chest CT scans. Our framework was developed on a set of 7184 lung CT scans from 5693 subjects, for which 2032 subjects are from the Masih Daneshvari Hospital (MDH) cohort and the rest belong to the CC-CCII set and MosMedData cohort. This data was gathered from three countries; China, Iran, and Russia. In this work, our framework diagnoses between COVID-19 and CP (common pneumonia), NCA (non COVID-19 abnormal), non-pneumonia, and normal classes. We evaluate and compare the performance of *ai-corona* with experts and RT-PCR in COVID-19 diagnosis and further compare our framework with AI models proposed by Zhang *et al.* [24] and Jin *et al.* [26]. Finally, we examine the impact of AI as assistance to expert diagnosis.

Materials and methods

Data

Three datasets were employed in this work; The MDH cohort, the CC-CCII set, and the MosMedData cohort. An overall summery of all the data employed in our work can be found in Table 1.

The first dataset was obtained by our group from patients hospitalized at the Masih Daneshvari Hospital (MDH) (Tehran, Iran). This cohort consists of 2121 lung CT scans from 2032 subjects annotated into 3 classes: (1) Normal; (2) Non-COVID Abnormal (NCA); and (3) COVID-19. Since differentiation between COVID-19 and Normal classes is easier than differentiating between COVID-19 and NCA (especially if there are similar imaging features), having the NCA class is very important. Using the search function of the hospital's PACS and by reviewing reports by two board-certified radiologists, we gathered a preliminary dataset of CT scans with a balanced distribution over all three classes.

All the participants in the MDH cohort gave written consent and our work has received the ethical license of IR.SBMU.NRITLD.REC.1399.024 from the Iranian National Committee for Ethics in Biomedical Research.

Cases in the Normal and NCA classes are from prior to the start of the Coronavirus global pandemic. A subset of the data in these two classes was randomly selected for testing. This portion was re-annotated by a different expert radiologist. Only the cases with consistent labels (*i.e.* same label as in the initial report) were retained in the test set. The MDH Normal and NCA cases that were not included in the test subset were

further divided randomly into a training subset and a tuning subset.

The MDH COVID-19 group scans for testing were taken in the early stages of the infection and included 119 lung CT scans from 109 patients hospitalized for more than three days. These scans were selected by the consensus of several metrics that indicate COVID-19 infection: (1) report by at least one radiologist on the scan; (2) confirmation of infection by two pulmonologists; (3) clinical presentation; and (4) RT-PCR report.

Furthermore, unlike other works that take a positive RT-PCT as the sole criterion to annotate a case with COVID-19 label, and since our evaluation includes comparing the diagnosis performance of *ai-corona* with experts and RT-PCT, we clearly could not use a dataset that was annotated solely based on RT-PCR test result. Our annotation strategy is, therefore, more comprehensive and incorporates additional available metadata.

The MDH COVID-19 training (1518 subjects, 1590 scans) and tuning (168 subjects, 174 scans) subsets were annotated using the aforementioned reports made by the two radiologists.

The CT scans in the MDH cohort contained between 21 to 46 slices acquired in axial orientation with a slice thickness between 8 and 10 mm, The histogram representation for the number of slices is indicated in Figure 9a, while Figure 9b and Figure 9c illustrate the age and sex distribution of the MDH cohort.

Moreover, as the NCA class of the MDH cohort includes many samples with non COVID-19 pneumonia, we can take this class as the equivalent of the CC-CCI set CP class for our model's training.

The second dataset employed in this work was the publicly available CC-CCII dataset [24]. After quality control (*e.g.* removing non-standard scans such as those with a small number of slices), this set contains 3953 CT scans from 2551 subjects. The scans in CC-CCII are annotated into three classes: Normal, Common Pneumonia (CP), and COVID-19. This CC-CCII dataset was randomly split into three subsets for: (1) training (2069 subjects, 3206 scans), (2) tuning (230 subjects, 352 scans), and (3) testing (252 subjects, 395 scans). The tuning subset was used for model checkpoint and selection the best overall model.

The third dataset, MosMedData cohort, is also publicly available and is comprised of 1110 CT scans from 1110 subjects. This dataset is annotated into two classes: Non-pneumonia and COVID-19. We used the entire MosMedData cohort for *external* testing, that is, testing on a dataset that has not been used for model training or tuning. To evaluate our model on this cohort, we take the prediction of COVID-19 class (for binary classification).

The public datasets LIDC–IDRI31 and Tianchi-Alibaba32 (which were used for the training of the model proposed by Jin *et al.* [26]) was not used in our framework's development, as these sets are for benign and malignant tumor diagnosis and they might introduce uncertainties to our framework.

For the RT-PCR evaluation set, 2672 subjects, each hospitalized more than three days, were tested 6419 times between February to October 2020. Respiratory samples including pharyngeal swabs/washing were obtained from the subjects. Nucleic acid was extracted from the samples using a QiaSymphony system (QIAGEN, Hilden, Germany) and SARS-CoV-2 RNA was detected using primer and probe sequences for screening and conformation on the basis of the sequence described by [30]. An RT-PCR diagnosis is considered correct when a patient has at least one positive test result.

Pre-Processing

For all the image slices, the top 0.5% of pixels with the highest values were selected and their values were clipped to the lowest one in the range. Then, the intensities were linearly transformed to the range [0, 255]. Since we utilize models pre-trained on the ImageNet dataset [31], an additional ImageNet normalization was also carried out.

Dataset	Classes	Training	Tuning	Test	#
MDH	Normal	467 (470)	51 (51)	120 (121)	628 (642)
	NCA	576 (578)	64 (64)	117 (117)	757 (759)
	COVID-19	475 (542)	53 (59)	109 (119)	637 (720)
CC-CCII	Normal	670 (844)	72 (94)	81 (105)	823 (1043)
	CP	665 (1131)	67 (127)	87 (147)	828 (1405)
	COVID-19	734 (1231)	82 (131)	84 (143)	900 (1505)
MosMedData	Non-Pneumonia	-	-	254 (254)	254 (254)
	COVID-19			856 (856)	856 (856)
#		3587 (4796)	398 (526)	1708 (1862)	5693 (7184)

Table 1. Number of subjects and (number of scans) for each class in the CC-CCII set, MDH cohort, and MosMedData cohort separated over training, tuning, and test.

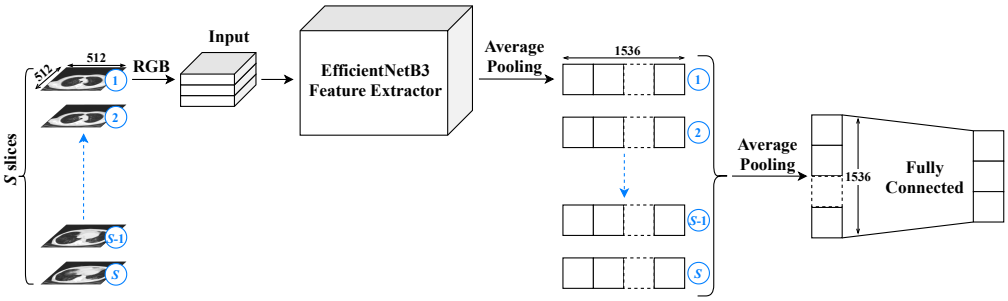


Fig 1. The schematic structure of ai-corona's deep learning model. The total number of utilized slices is labeled by S. Each selected slice is fed to the feature extractor block pipeline one by one so that we end up with S vectors, which is then transformed to a single vector via an average pooling function. Afterwards, the result is passed through a fully connected network to reach the three output neurons, corresponding to our three classes.

We also opted to not perform any segmentation (*i.e.* patch extraction) in our pre-processing. This is due to the manual annotation of each dataset (like Jin *et al.* [26]) being time and resource consuming. On the other hand, using automated methods, such as image processing techniques and pre-trained segmentation deep learning models, would introduce further unwanted error and uncertainty to our data, and subsequently, to the model's inference.

Deep Learning Method

Inspired by [21], *ai-corona*'s deep learning model consists of two main blocks; a feature extractor and a classifier. The main challenge is mapping a 3-dimensional CT scan, which is a series of image slices, to a probability vector with a length equal to the number of classes. Another challenge is that all the scans not having the same number of slices and not all the slices being useful for diagnosis. To address this, we take the middle 50% image slices in each scan and denote the number of selected slices from each scan with S. We also experimented with other slice selection strategies (such as a portion larger than 50%, top/bottom 50%, *etc.*), from which none performed better.

As shown in Figure 1, the feature extractor block is a pipeline, receiving each slice with dimensions $512 \times 512 \times 3$ (3 represents the number of color channels, but with all templates being exactly the same as for each image) and outputting a vector of length 1536 through an average pooling function. After all the slices have passed through the feature extractor block, we end up with S vectors. After all the S slices have passed through the feature extractor block, another average pooling is applied to the results which yields a single vector of length 1536.

This pipeline manner ensures that our framework is independent of the number of slices in a CT scan, as we always end up with a single vector of length 1536 at the end of the feature extractor block. The pipeline receives any number of slices, extracts their features, and finally outputs a single vector of known length. Moreover, the use of only

a single feature extractor at a time significantly reduces the computational load of our framework, resulting in a much faster training and prediction time.

Convolutional Neural Networks (CNN) were used for the feature extraction block. We experimented with different CNN models, such as DenseNet, ResNet, Xception, and EfficientNetB0 through EfficientNetB5 [32–35], taking into account their accuracy and accuracy density on the ImageNet dataset [36]. All of these models were initialized with their respective pre-trained weights on the ImageNet dataset. At the end, the EfficientNetB5 model stripped of its last dense layers was chosen as the primary feature extractor for our deep learning framework. The vector output of the EfficientNetB3 feature extraction block is then passed through the classifier block, which contains yet another average pooling layer that is connected to the model's output neurons corresponding to the classes via a dense network of connections. *ai-corona* is implemented with Python 3.7 [37] and Keras 2.3 [38] framework and was trained on NVIDIA GeForce RTX 2080 Ti for 60 epochs in a total of three hours. The Pydicom [39] package was used to read the DICOM file of the cases.

Class Activation Maps

In order to generate the class activation map of an image slice, we computed a weighted average across the 1536 values of the feature vector using weights from the classification block to obtain a 10×10 image. The result map was then mapped to a color scheme, upsampled to 512×512 pixels, and overlaid with the original input image slice. Employing parameters from the classification block to weigh the feature vectors makes, more predictive features appear more bright. This leads to regions of the image slice that most influence the model's prediction to appear brighter. The class activation maps highlight which pixels in an image slice are important for the model's prediction [40].

Statistical Inference

In order to quantify the reliability of our findings and the performance of our results based on the model's detection of COVID-19 in chest CT scans, we provide a thorough comparison with expert practicing radiologists diagnosis. To achieve a more conservative discrimination strategy, we compute the following evaluation criteria ranging from sensitivity (true positive rate), specificity (true negative rate), F1-score, Cohen's kappa, and finally to AUC. Moreover, the confusion matrix for all the classes of each individual study is also calculated.

We set the presence of the underlying class with a positive label and the rest of the classes assigned by negative label. Incorporating error propagation and using the Bayesian statistics, we calculate the marginalized confidence region at 95% level for each computed quantity. The significance of diagnostic results is examined by computing the p-value statistics systematically. To achieve a conservative decision, the 3σ significance level is usually considered.

Since the radiologists' diagnosis is given by "Yes" or "No" statements for each class, it is necessary to convert the probability values computed by our model to binary values. Hence, we selected an operating point for distinguishing a given case among others and compute the true positive rate (sensitivity) versus false positive rate (1-specificity). This operating point was selected such that the model would yield a high specificity. To make more sense, as well as the other mentioned evaluation criteria, the Receiver Operating Characteristic (ROC) diagram is also estimated for our studies. All of our criteria were calculated using the scikit-learn [41] package.

Experts Evaluation

Our team of experts annotated cases in the CC-CCII test set and MDH test set, with "Yes" and "No" labels for each class. To prevent a loss in experts' diagnosis performance due to fatigue, they were asked to work on small time chunks. Their performance was then evaluated and recorded. Next, to evaluate the impact of AI assistance in the experts' performance, after an appropriate amount of time and shuffling the sets (to prevent any remembrance), the experts re-annotated the two sets for a second time, while this time having access to the output of the model. They incorporated the model's opinion for suspicious cases on their own authority. Their performance was evaluated and recorded again.

Our team of four experts incorporates two practicing academic senior radiologists with 15 years of experience each. In our study, they're referred to as *Senior Radiologist 1* and *Senior Radiologist 2*. Another expert is a practicing academic radiologists with 5 years of experience, which is referred to as *Junior Radiologist*. The last member is a senior radiology resident, referred as *Radiology Resident*. The team of experts were chosen such that a wide range of experience and background knowledge would be present for our studies, in order to make it more comprehensive.

Results

Training, Evaluation, and Testing Datasets

To develop *ai-corona*, we utilized data from three different sources: (1) the MDH cohort, (2) the publicly available CC-CCII dataset [24], and (3) the publicly available MosMedData cohort. The combined data were from multiple international sites and comprised of 7184 CT scans from 5693 subjects categorized into five classes: normal, CP, NCA, non-pneumonia, and COVID-19. For a better comparison of the diagnosis performance between RT-PCR and CT scans, the RT-PCR test records of 2672 patients in a 7-month period were gathered.

The MDH and the CC-CCII data were used for training, evaluation (tuning), and testing. The MosMedData was used entirely for testing. Overall, 5322 scans from 3985 subjects were used for training and tuning, and three sets were used for testing: (1) CC-CCII test set (105 normal, 147 CP, and 143 COVID-19 scans), (2) MDH test set (121 normal, 117 NCA, and 119 COVID-19 scans), and (3) the entire MosMedData cohort (254 non-pneumonia and 856 COVID-19 scans).

Taking into consideration the ground truth annotation of all the works involved, the CC-CCII test set was used to compare *ai-corona* with the models proposed by Zhang *et al.* [24], Jin *et al.* [26], and with expert radiologists. Furthermore, the MDH test set was used to compare *ai-corona* with the radiologists and RT-PCR. Lastly, The MosMedData cohort was used to compare *ai-corona* with the model proposed by Jin *et al.* [26].

RT-PCR Sensitivity

Since the truth annotation methodology described in section yields accurate labels, it was used to annotate a separate set for RT-PCR evaluation. This set is used to showcase the evolution of RT-PCR's sensitivity over a period of 7 months in Figure 2 (sensitivity of each day is calculated as the average sensitivity of a 15-day period centered around that day). RT-PCR's sensitivity oscillates in the range [0.351, 0.722]. The decrease in sensitivity to 0.351 on April 29^{th} is due to changing the specimen obtaining method to oropharygeal wash [42]. This changed later and nasopharyngeal and oropharyngeal swabs were used. The biggest value for RT-PCR's sensitivity in this evaluation is considered its best, denoted by *RT-PCR Best*.

Performance Evaluation and Comparison

Having three test sets, our framework's COVID-19 diagnosis performance for the CC-CCII test set, MDH test set, and the MosMedData cohort for all the studies is evaluated (an operating point was selected for each study). The confusion matrices for our evaluation results can be found in Figure 3. Moreover, for the COVID-19 class, ROC curves are showcased in Figure 4 and a more thorough look using the four metrics is depicted in Figure 5a and Figure 5b. At last, the complete numerical reports for this evaluation can be found in Table 3. Values denoted with "-" in the table correspond to a lack of report.

Figure 3a through Figure 3c show that *ai-corona* has performed better in all three classes (Normal, CP, COVID-19) compared to Zhang *et al.* [24] and Jin *et al.* [26] on the CC-CCII test set and achieves an AUC score of 0.997, sensitivity of 0.972, and specificity of 0.968 on the COVID-19 class. The confusion matrix in Figure 3d showcases our framework's performance on MDH test set for the three classes of Normal, NCA, and COVID-19. For this dataset, our framework gains scores of 0.989, 0.924, and 0.983 for AUC, sensitivity, and specificity, respectively. In addition, Figure 3e and Figure 3f showcase that our framework surpasses that of proposed by Jin *et al.* [26] on the MosMedData cohort with an AUC of 0.954. Although both have similar sensitivities in COVID-19 diagnosis, *ai-corona* outperforms Jin *et al.*'s model in non-pneumonia diagnosis with 83.07% accuracy, reporting fewer false positives.

The better diagnosis performance over the CC-CCII test set indicates that the task of diagnosing NCA from the other classes is indeed more difficult than diagnosing CP from the other classes. This due to all the different abnormalities present in the NCA class having their unique imaging features.

Comparison with Experts and RT-PCR

Figure 4a and Figure 5a showcase the COVID-19 diagnosis performance of *ai-corona* and its comparison with that of experts for the CC-CCII test set. As shown, our framework performs better in all cases (except for the specificity of Senior radiologist 1). Furthermore, Figure 4 and Figure 5b showcase the same comparison, but for MDH test set. This time, the framework performed similar to radiologists in specificity, but outperformed in the other metrics. In this comparison, 93.3% of COVID-19 cases in MDH test set (111 of 119) were diagnosed as infected by at least one expert. Out of the other 8 that were not, our framework managed to report one and RT-PCR reported three as infected. If RT-PCR was the only criteria for the truth annotation, the overall sensitivity of radiologists would improve to 97%, which would further confirm the findings in [7]. The complete reports for these two evaluations are in sections **a** and **b** of Table 3.

In Figure 4b, the sensitivity of RT-PCR based diagnosis and CT based diagnosis is compared. The figure shows that *RT-PCR Best* sensitivity of 0.722 is lower than every expert diagnosing via CT. The *RT-PCR Best* sensitivity is an upper bound. Because if instead of testing patient hospitalized for more than three days, every COVID-19 admitted patient was tested, RT-PCR's sensitivity would be much lower than 0.722.

Model as Expert Assistant

The goal of any AI assistant model is to improve the diagnosis performance of experts. For the evaluation, first, the radiologists annotate the test set. After an appropriate amount of time, the radiologist re-annotated the set for a second time while having the diagnosis of *ai-corona* for the entire set. The test set was also shuffled the second time to eliminate any remembrance of cases. Experts' diagnosis performance is depicted in Figure 5a, and Figure 5b. For the CC-CCII test set, all the experts (except the radiology resident) had an improvement in their sensitivity. A significant improvement in the other

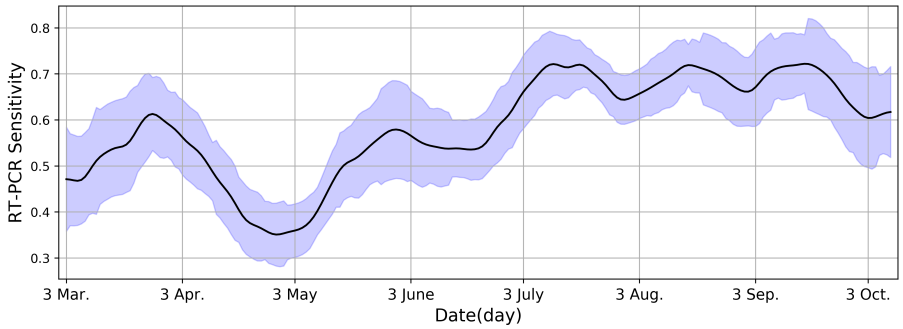


Fig 2. Fluctuations in RT-PCR sensitivity on a daily basis. The highest peek reaches 0.722, which is denoted as *RT-PCR Best*.

metrics is also seen for everyone (except Senior 1). For MDH test set, improvement in sensitivity can be seen for Senior 1 and Junior. Specificity only had an improvement in the Radiology Resident and remained unchanged for others. In every other evaluation criterion, the AI model had a positive impact on the experts' performance.

Interpretation of *ai-corona*

To ensure that *ai-ccorona* was learning the correct imaging features, class activation maps were generated Figure 6. This is done by following the methodology described in section . In the class activation map of a slice in a scan, more predictive areas (that hold the correct imaging features) appear brighter. Thus, the brightest areas of the class activation map correspond to regions that most influence the model's prediction.

Additional Evaluations

We made other evaluations as well, for which the complete details can be found in Supplementary Materials. First, over the MDH test set, performance of diagnosis between NCA and Normal classes was evaluated using the four metrics and was compared to the experts. Furthermore, all of the possible comparisons between every pair of classes were made to ensure the thoroughness and completeness of our evaluation. As an example, this extra study showcased that radiologists perform better in diagnosing NCA from Normal compared to the AI model.

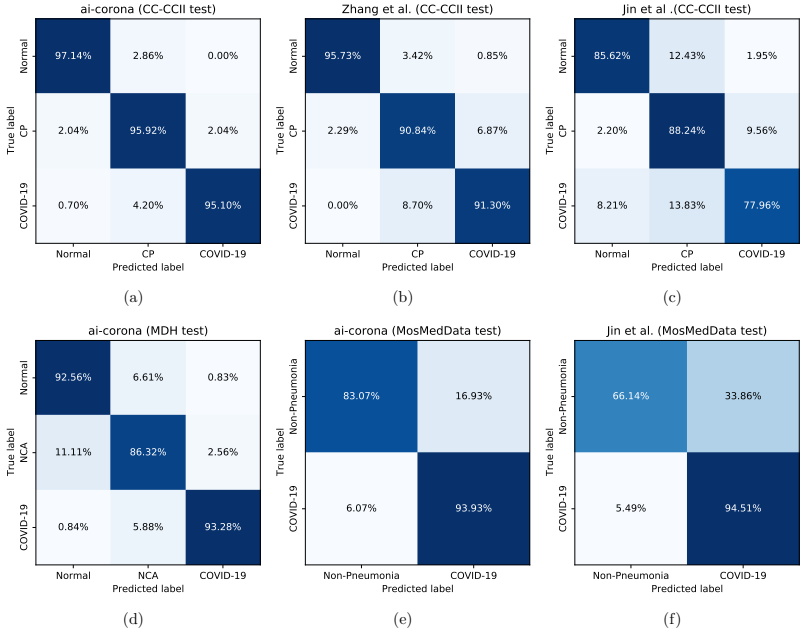
Lastly, it is important to note the speed at which different methodologies perform diagnosis. As shown in, RT-PCR is extremely slow. Moreover, our framework is faster than the best radiologist by 25 orders of magnitude. This is showcased in Table 2.

	ai-corona	Senior 1	Senior 2	Junior	Radiology resident
Diagnosis Time	12 *min.*	360 *min.*	300 *min.*	320 *min.*	400 *min.*

Table 2. Diagnosis time comparison for *ai-corona* and radiologists on the 357 case test set.

	Sensitivity (95% CI)	Specificity (95% CI)	F1-score (95% CI)	Kappa (95% CI)	AUC (95% CI)
a (CC-CCII test set)					
ai-corona	0.972 (0.956, 0.988)	0.968 (0.954, 0.982)	0.970 (0.954, 0.986)	0.935 (0.909, 0.961)	0.997 (0.993, 0.999)
Zhang *et al.* [24]	0.949 -	0.911 -	-	-	0.980 (0.967, 0.990)
Jin *et al.* [26]	0.921 (0.918,0.926)	0.780 (0.770–0.789)	-	-	0.921 (0.918, 0.926)
Senior 1	0.895 (0.874, 0.916)	0.956 (0.945, 0.967)	0.908 (0.892, 0.924)	0.857 (0.837, 0.877)	-
Senior 1+AI	0.937 (0.919, 0.955)	0.948 (0.937, 0.959)	0.924 (0.91, 0.938)	0.88 (0.86, 0.9)	-
Senior 2	0.909 (0.888, 0.93)	0.917 (0.903, 0.931)	0.884 (0.868, 0.9)	0.816 (0.792, 0.84)	-
Senior 2+AI	0.965 (0.954, 0.976)	0.94 (0.927, 0.953)	0.932 (0.92, 0.944)	0.892 (0.876, 0.908)	-
Junior	0.636 (0.605, 0.667)	0.897 (0.882, 0.912)	0.7 (0.677, 0.723)	0.555 (0.523, 0.587)	-
Junior+AI	0.776 (0.75, 0.802)	0.913 (0.898, 0.928)	0.804 (0.782, 0.826)	0.7 (0.672, 0.728)	-
R. resident	0.839 (0.813, 0.865)	0.663 (0.639, 0.687)	0.69 (0.67, 0.71)	0.459 (0.426, 0.492)	-
R.res.+AI	0.853 (0.826, 0.88)	0.778 (0.758, 0.798)	0.76 (0.74, 0.78)	0.599 (0.568, 0.63)	-
b (MDH test set)					
ai-corona	0.924 (0.895, 0.953)	0.983 (0.961, 1.000)	0.953 (0.935, 0.971)	0.917 (0.887, 0.947)	0.989 (0.984, 0.994)
RT-PCR	0.722 (0.661, 0.783)	-	-	-	-
Senior 1	0.857 (0.833, 0.881)	0.979 (0.963, 0.995)	0.903 (0.886, 0.92)	0.858 (0.838, 0.878)	-
Senior 1+AI	0.908 (0.887, 0.929)	0.987 (0.975, 0.999)	0.939 (0.927, 0.951)	0.91 (0.892, 0.928)	-
Senior 2	0.899 (0.874, 0.924)	0.979 (0.965, 0.993)	0.926 (0.912, 0.94)	0.891 (0.868, 0.914)	-
Senior 2+AI	0.899 (0.877, 0.921)	0.992 (0.983, 1.0)	0.939 (0.928, 0.95)	0.91 (0.894, 0.926)	-
Junior	0.765 (0.738, 0.792)	0.992 (0.982, 1.0)	0.858 (0.838, 0.878)	0.8 (0.775, 0.825)	-
Junior+AI	0.857 (0.833, 0.881)	1.0 (1.0, 1.0)	0.923 (0.908, 0.938)	0.889 (0.869, 0.909)	-
R. resident	0.882 (0.858, 0.906)	0.92 (0.898, 0.942)	0.864 (0.846, 0.882)	0.794 (0.766, 0.822)	-
R.res.+AI	0.899 (0.877, 0.921)	0.966 (0.948, 0.984)	0.915 (0.901, 0.929)	0.873 (0.853, 0.893	-
c (MosMedData Cohort)					
ai-corona	0.939 (0.924, 0.954)	0.831 (0.802, 0.86)	-	-	0.954 (0.937, 0.971)
jin	0.945 (0.938, 0.951)	0.661 (0.636, 0.686)	-	-	0.933 (0.926, 0.938)

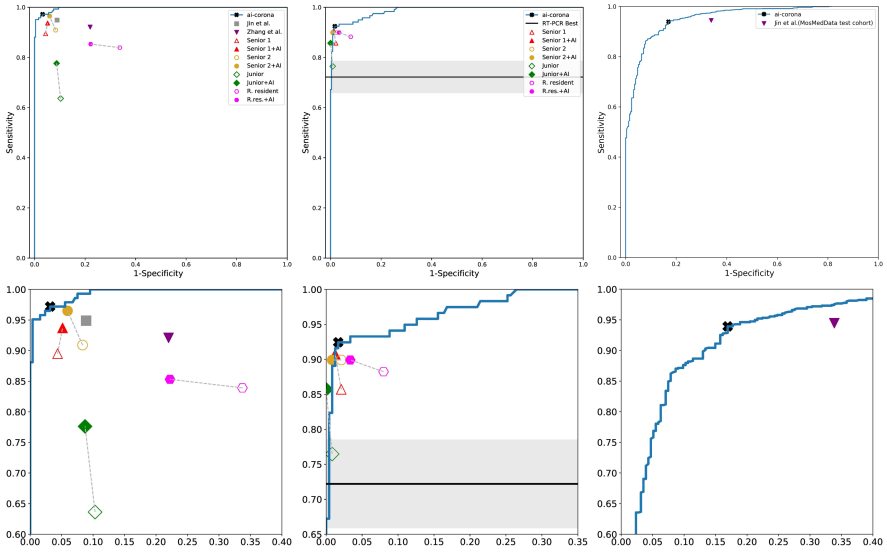
Table 3. Evaluation results of all the studies with a 95% confidence interval using the metrics sensitivity, specificity, F1-score, Kappa, and AUC. A "-"" value corresponds to a lack of data. Reports in the sections **a**, **b**, and **c** are for the CC-CCI test set, the MDH test set, and the MosMedData Cohort, respectively.



(a) (b) (c)

(d) (e) (f)

Fig 3. Top row left-to-right: Confusion matrices for *ai-corona*, the model proposed by Zhang *et al.* [24], and the model proposed by Jin *et al.* [26] for the CC-CCI test set, respectively. Bottom row left and middle: Confusion matrices for *ai-corona* on the MosMedData cohort, respectively. Bottom row right: Confusion matrix for the model proposed by Jin *et al.* [26] for the MosMedData cohort.



(a) CC-CCII test (b) MDH test (c) MosMedData cohort

Fig 4. ROC curve diagrams for *ai-corona* on the (a) CC-CCII test set (b) MDH test set (c) MosMedData cohort. Diagrams in the bottom row correspond to a zoom-in of their respective curves. Hollow shapes represent an expert un-aided by AI, where filled shapes represent expert with AI assistance. As RT-PCR sensitivity was not available, its sensitivity is shown as a solid line in Figure 4b.

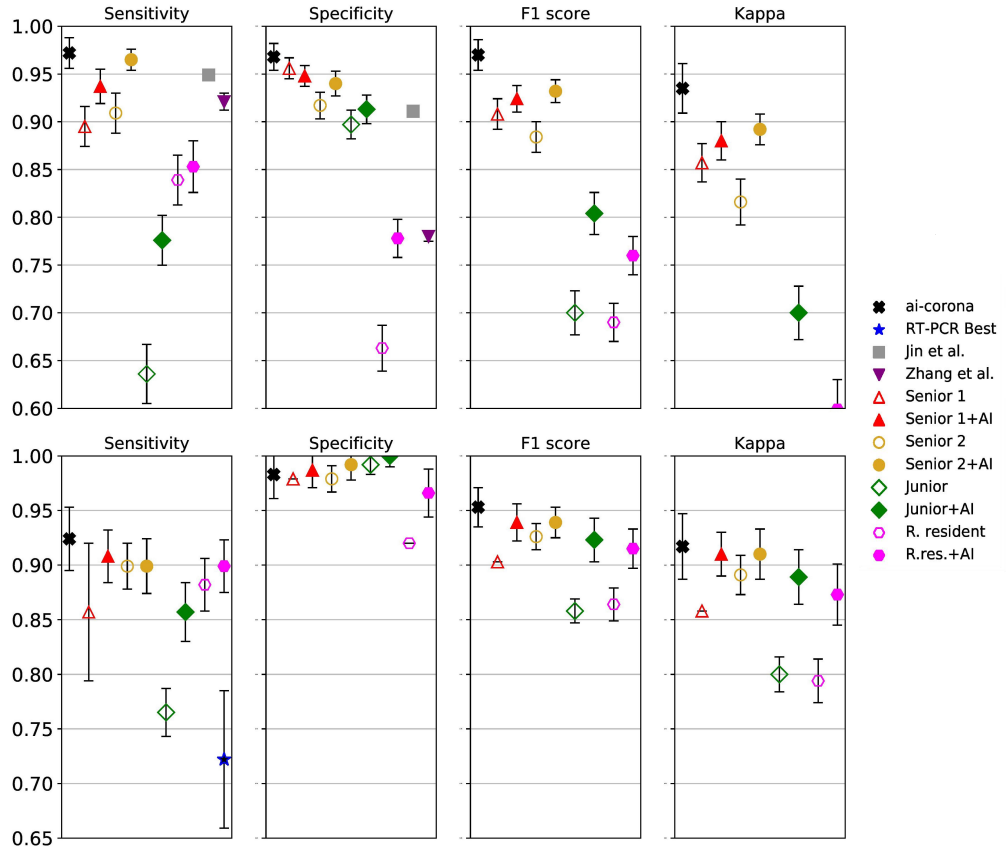


Fig 5. Detailed comparison of all the studies using our evaluation metrics for above: CC-CCII test set, below: MDH test set. Hollow shapes represent an expert un-aided by AI, where filled shapes represent expert with AI assistance.

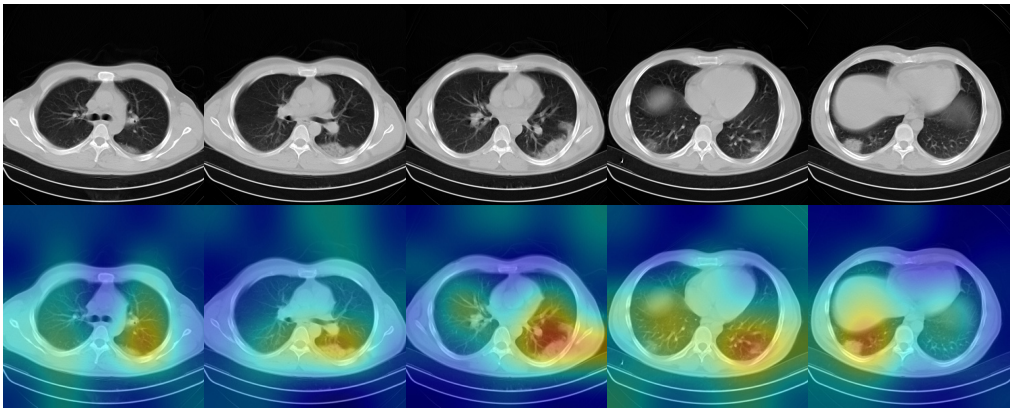


Fig 6. Class activation maps for *ai-corona* interpretation. This highlight which pixels in the images are important for the model's classification decision.

Conclusion and Discussion

We introduce *ai-corona*, a radiologist-assistant deep learning framework capable of accurate COVID-19 diagnosis in chest CT scans. Our deep learning framework was developed (training and tuning) on 5322 scans, 3985 subjects, gathered from cohorts from two countries, China and Iran, and was tested against three sets; the CC-CCII test set from China (395 scans, 252 subjects), MDH test set from Iran (357 scans, 346 subjects), and the MosMedData cohort from Russia (1110 scans, 1110 subjects). Our framework was able to learn to diagnose patients infected with COVID-19, as well being able to distinguish between COVID-19, other types of common pneumonia (CP) such as viral and bacterial, and other non COVID-19 abnormalities (NCA). Moreover, a set of 2672 subjects was used to calculate the sensitivity of RT-PCR.

The use of multiple datasets, each with scans differing in the number of slices, and a lack of slice-specific labeling, presented a challenge for this work. To address this, we dynamically select the middle 50% of slices in each scan and feed them to a single EfficientNetB3-based feature extractor, which after an average pooling operator, will result in a single feature vector that will be classified. This method, alongside the use of only one 2D CNN, will not only make our framework more robust, but it will also make its predictions faster and capable of running on slower hardware.

Our framework was compared to two other AI models, proposed by Zhang *et al.* [24] and Jin *et al.* [26] respectively. Its diagnosis performance is also compared to that of experts and other means of diagnosis in order to achieve a comprehensive and sensible image of the framework's abilities. In the end,*ai-corona* managed to outperform the two other AI models in COVID-19 diagnosis. Our framework achieves high sensitivity, while also having a high specificity.

Our framework achieved an AUC score of 0.997 on the CC-CCII test set and performed better than models proposed by Zhang *et al.* [24] and Jin *et al.* [26] on all four metrics. On the MDH test set,*ai-corona* gained an AUC score of 0.989 and performed mostly better in all of the metrics compared to the experts. It is worth mentioning that for our framework, diagnosing between the COVID-19 and CP classes was easier than between COVID-19 and NCA. Yet for the experts, it was the opposite. RT-PCR, as another method of diagnosis, had a sensitivity of 0.722 at best, worse than all the experts and the AI. At last, our framework gained a 0.954 AUC score on the MosMedData cohort, which outperforms Jin *et al.* [26]. A complete report of these evaluations can be found in Figure 3 through Figure 5 and Table 3.

In COVID-19 diagnosis, *ai-corona*'s impact on assisting experts' diagnosis was evaluated, which in COVID-19 diagnosis, mostly indicates a positive improvement on at least their sensitivity or specificity. This improvement is most noticeable for the junior radiologist and the radiology resident. Additionally, incorporation of the class activation maps in the experts' diagnosis can help them examine the involved regions better.

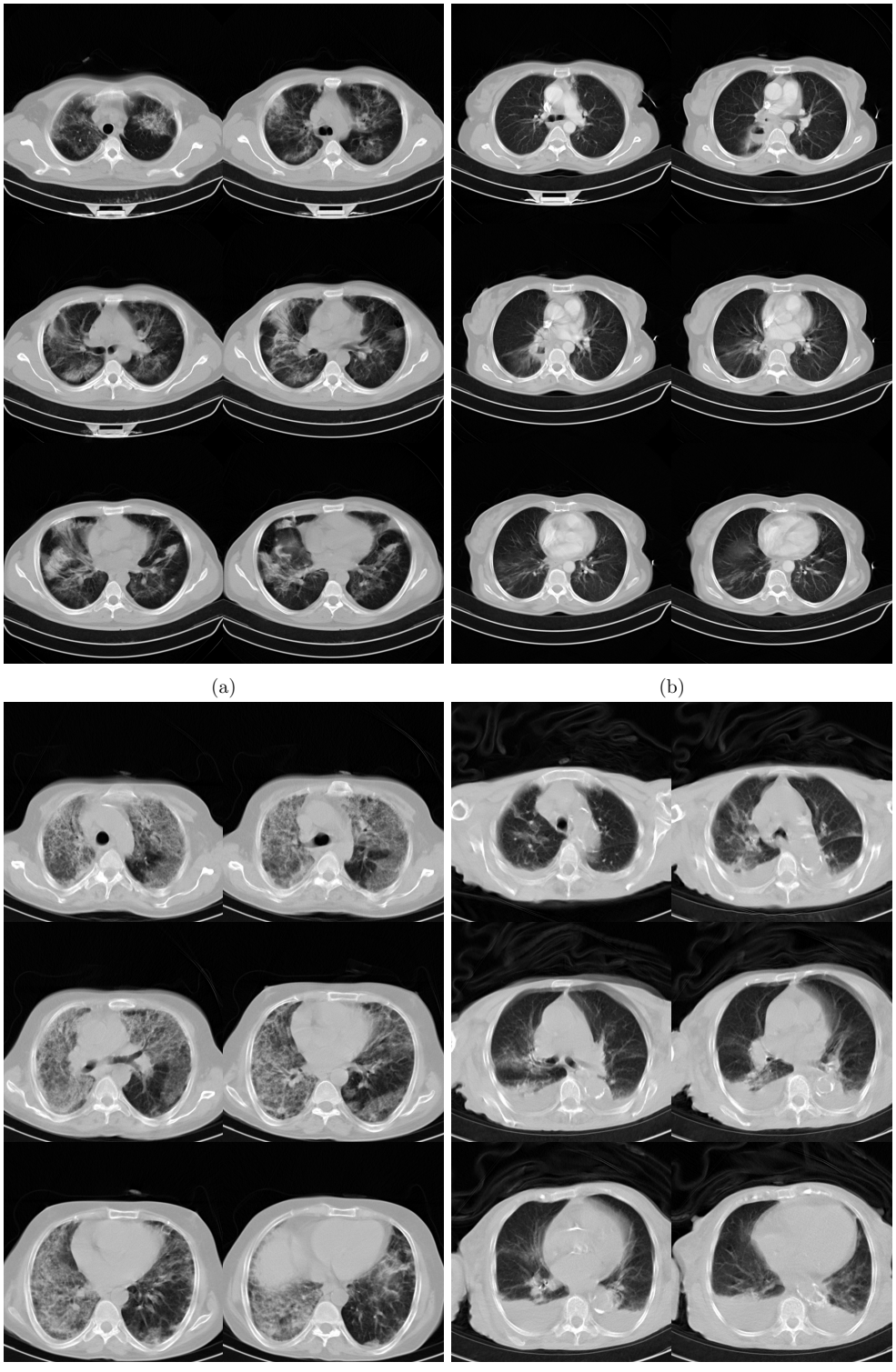
On having a positive impact on experts' diagnosis, two cases are discussed here to showcase how *ai-corona* made experts change their minds for good in suspicious cases. At least one expert misdiagnosed Figure 7a's case as NCA at first, but upon seeing the AI's diagnosis, correctly diagnosed as COVID-19. This expert cited seeing Peribrochovascular distribution, which is not common in COVID-19 (no subpleural distribution), as the reason for their misdiagnosis. In addition, Figure 7b's case was initially misdiagnosed as COVID-19 by at least one expert, but was changed correctly to NCA when seeing the AI's correct diagnosis. They cited that cavity, centrilobular nodule, mass, and mass like consolidations are not commonly seen in COVID-19 pneumonia and might implicate other diagnostics. Figure 3 On the other hand, the existence of error in CT-based diagnosis, both for *ai-corona* and experts, encourages us to study the cause for such errors, which might lead to better and more accurate predictions, or point out any if existing fundamental flaws in CT-based diagnosis. Figure 7a's case was misdiagnosed

as COVID-19 by all the experts. Our framework, while correctly diagnosing for NCA, was not able to change the experts' minds. In a consensual final report, the experts cite that Mediastinal and bilateral hilar adenopathies were seen, as well as Anterior mediastinal soft-tissue density. In addition, Diffuse bilateral interstitial infiltrations were detected with crazy paving pattern, ground glass, and traction bronchiectasis, mainly in the right lung and partial volume loss of the right lung. Also, the position of central venous catheter tip was seen in the left brachiocephalic vein.

The success of AI in medical imaging-based diagnosis has been proven by this work and many others before it. *ai-corona* can positively influence an expert's opinion and improve the speed at which the subject screening process occurs, such that it helps critical cases get the care they urgently need faster.

But our work has its own drawbacks and shortcomings. Since the gathering of a dataset with a better labeling (one that alongside its accurate annotations also accompanies localization and slice labels) is time and resource consuming, we decided to opt for an approach that favours robustness and is capable of learning on simpler dataset. Developing our framework on a better dataset would certainly improve its performance. In addition, the CP class contains all kinds of conditions and diseases that cause pneumonia. As each of these conditions and diseases have their own distinct imaging features, having separate classes for them, especially Influenza-A, would improve the framework's performance. Lastly, our framework's learning would certainly benefit from more cases that are positive for COVID-19, yet have a negative RT-PCR result. As these cases are mostly experiencing the early stages of the infection, diagnosing them is more difficult. Moreover, classifying cases with a negative RT-PCR as non COVID-19 is illogical and their labeling protocol should be something else.

In future, approaches that do a better job incorporating clinical reports with the imaging data should be explored. In conclusion, with the individual drawbacks of diagnosing based on clinical representation, RT-PCR, and CT-based diagnosis, a method comprised of all three would definitely yield the most accurate diagnosis of COVID-19.



(a) (b)

(c) (d)

Fig 7. Panels (a), (b), and (c), are the chest CT scans of patients who were initially misdiagnosed by at least one radiologist but were then diagnosed correctly upon incorporating *ai-corona*'s correct prediction. Panel (d) shows the chest CT scans of patient that was misdiagnosed by *ai-corona* and radiologists.

Supporting information

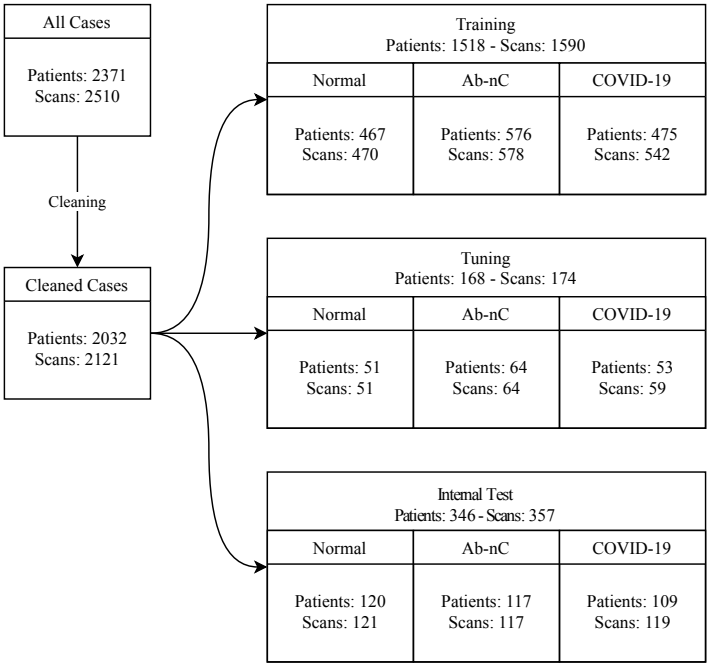
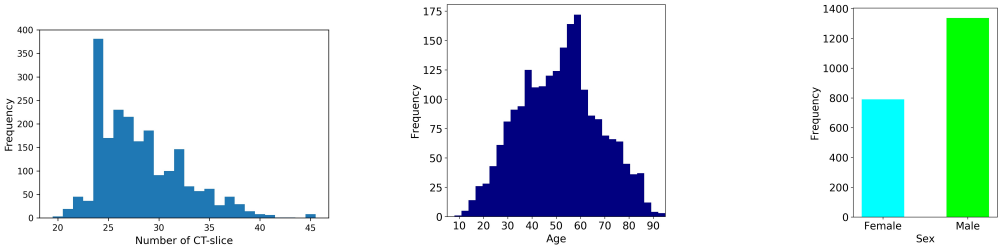


Fig 8. The cascade structure of the MDH. Number of subjects and scans in each split and set is indicated. The preliminary dataset was cleaned, by removing abdomen and high-resolution CT scans. The train and tuning sets were labeled by two expert radiologists. The NCA and normal classes of the test set was re-annotated by three expert radiologist (one new). The COVID-19 class are patients that meet our criteria and were hospitalized for more than three days.



(a) (b) (c)

Fig 9. The left panel corresponds to the distribution of image slices for cases in the MDH, middle panel shows the distribution of Age, while the right panel illustrates the sex distribution of cases in the MDH.

	Sensitivity (95% CI)	Specificity (95% CI)	F1-score (95% CI)	Kappa (95% CI)
ai-corona	0.983 (0.971, 0.995)	0.967 (0.951, 0.983)	0.975 (0.965, 0.985)	0.95 (0.929, 0.971)
Senior 1	0.958 (0.947, 0.969)	0.992 (0.987, 0.997)	0.974 (0.967, 0.981)	0.95 (0.938, 0.962)
Senior 1+AI	0.992 (0.987, 0.997)	0.983 (0.976, 0.99)	0.987 (0.983, 0.991)	0.975 (0.966, 0.984)
Senior 2	0.966 (0.957, 0.975)	0.942 (0.93, 0.954)	0.954 (0.944, 0.964)	0.908 (0.892, 0.924)
Senior 2+AI	0.975 (0.967, 0.983)	0.975 (0.966, 0.984)	0.975 (0.969, 0.981)	0.95 (0.938, 0.962)
Junior	0.983 (0.977, 0.989)	0.959 (0.949, 0.969)	0.971 (0.965, 0.977)	0.942 (0.93, 0.954)
Junior+AI	0.983 (0.976, 0.99)	0.95 (0.939, 0.961)	0.967 (0.96, 0.974)	0.933 (0.919, 0.947)
R. resident	0.966 (0.957, 0.975)	0.917 (0.904, 0.93)	0.943 (0.934, 0.952)	0.883 (0.867, 0.899)
R.res.+AI	0.966 (0.957, 0.975)	0.967 (0.956, 0.978)	0.966 (0.959, 0.973)	0.933 (0.919, 0.947)

Table 4. The quantitative evaluation of *ai-corona*, radiologists, and model-assisted radiologists performance results for distinguishing between the COVID-19 class and the normal class at a 95% confidence interval.

	Sensitivity (95% CI)	Specificity (95% CI)	F1-score (95% CI)	Kappa (95% CI)
ai-corona	0.924 (0.901, 0.947)	0.974 (0.959, 0.989)	0.949 (0.934, 0.964)	0.898 (0.872, 0.924)
Senior 1	0.857 (0.836, 0.878)	0.957 (0.944, 0.97)	0.903 (0.89, 0.916)	0.814 (0.794, 0.834)
Senior 1+AI	0.908 (0.892, 0.924)	0.974 (0.965, 0.983)	0.939 (0.929, 0.949)	0.881 (0.865, 0.897)
Senior 2	0.899 (0.884, 0.914)	0.957 (0.946, 0.968)	0.926 (0.916, 0.936)	0.856 (0.837, 0.875)
Senior 2+AI	0.899 (0.881, 0.917)	0.983 (0.976, 0.99)	0.939 (0.929, 0.949)	0.881 (0.863, 0.899)
Junior	0.765 (0.743, 0.787)	0.983 (0.974, 0.992)	0.858 (0.843, 0.873)	0.746 (0.723, 0.769)
Junior+AI	0.857 (0.839, 0.875)	1.0 (1.0, 1.0)	0.923 (0.912, 0.934)	0.856 (0.837, 0.875)
R. resident	0.882 (0.863, 0.901)	0.855 (0.838, 0.872)	0.871 (0.858, 0.884)	0.737 (0.711, 0.763)
R.res.+AI	0.899 (0.882, 0.916)	0.94 (0.927, 0.953)	0.918 (0.906, 0.93)	0.839 (0.815, 0.863)

Table 5. The quantitative evaluation of *ai-corona*, radiologists, and model-assisted radiologists performance results for distinguishing between the COVID-19 class and the NCA class at a 95% confidence interval.

	Sensitivity (95% CI)	Specificity (95% CI)	F1-score (95% CI)	Kappa (95% CI)
ai-corona	0.915 (0.883, 0.947)	0.929 (0.893, 0.965)	0.922 (0.894, 0.95)	0.831 (0.793, 0.869)
Senior 1	0.897 (0.876, 0.918)	0.946 (0.925, 0.967)	0.894 (0.877, 0.911)	0.841 (0.815, 0.867)
Senior 1+AI	0.949 (0.934, 0.964)	0.95 (0.93, 0.97)	0.925 (0.911, 0.939)	0.887 (0.87, 0.904)
Senior 2	0.949 (0.934, 0.964)	0.938 (0.916, 0.96)	0.914 (0.9, 0.928)	0.869 (0.848, 0.89)
Senior 2+AI	0.974 (0.963, 0.985)	0.95 (0.932, 0.968)	0.938 (0.926, 0.95)	0.906 (0.89, 0.922)
Junior	0.923 (0.901, 0.945)	0.871 (0.843, 0.899)	0.844 (0.824, 0.864)	0.757 (0.733, 0.781)
Junior+AI	0.983 (0.974, 0.992)	0.912 (0.89, 0.934)	0.909 (0.896, 0.922)	0.86 (0.838, 0.882)
R. resident	0.821 (0.793, 0.849)	0.925 (0.897, 0.953)	0.831 (0.81, 0.852)	0.75 (0.723, 0.777)
R.res.+AI	0.923 (0.904, 0.942)	0.954 (0.935, 0.973)	0.915 (0.901, 0.929)	0.873 (0.854, 0.892)

Table 6. The quantitative evaluation of *ai-corona*, radiologists, and model-assisted radiologists performance results for distinguishing between the NCA class and the other two classes at a 95% confidence interval.

	Sensitivity (95% CI)	Specificity (95% CI)	F1-score (95% CI)	Kappa (95% CI)
ai-corona	0.942 (0.917, 0.967)	0.919 (0.883, 0.955)	0.931 (0.907, 0.955)	0.841 (0.801, 0.881)
Senior 1	0.992 (0.985, 0.999)	0.949 (0.929, 0.969)	0.949 (0.937, 0.961)	0.92 (0.904, 0.936)
Senior 1+AI	0.983 (0.975, 0.991)	0.983 (0.971, 0.995)	0.975 (0.967, 0.983)	0.963 (0.952, 0.974)
Senior 2	0.942 (0.926, 0.958)	0.979 (0.964, 0.994)	0.95 (0.941, 0.959)	0.925 (0.909, 0.941)
Senior 2+AI	0.975 (0.964, 0.986)	0.983 (0.971, 0.995)	0.971 (0.962, 0.98)	0.956 (0.944, 0.968)
Junior	0.959 (0.946, 0.972)	0.962 (0.943, 0.981)	0.943 (0.931, 0.955)	0.913 (0.896, 0.93)
Junior+AI	0.95 (0.933, 0.967)	0.983 (0.97, 0.996)	0.958 (0.947, 0.969)	0.937 (0.921, 0.953)
R. resident	0.917 (0.899, 0.935)	0.966 (0.949, 0.983)	0.925 (0.911, 0.939)	0.887 (0.866, 0.908)
R.res.+AI	0.967 (0.954, 0.98)	0.975 (0.962, 0.988)	0.959 (0.95, 0.968)	0.938 (0.923, 0.953)

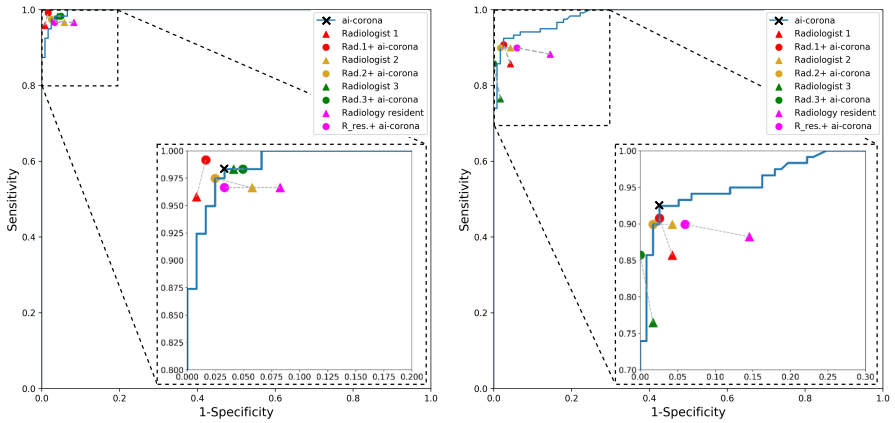
Table 7. The quantitative evaluation of *ai-corona*, radiologists, and model-assisted radiologists performance results for distinguishing between the normal class and the other two classes at a 95% confidence interval.

	Sensitivity (95% CI)	Specificity (95% CI)	F1-score (95% CI)	Kappa (95% CI)
ai-corona	0.906 (0.878, 0.934)	0.917 (0.891, 0.943)	0.912 (0.893, 0.931)	0.823 (0.789, 0.857)
Senior 1	0.94 (0.927, 0.953)	0.992 (0.986, 0.998)	0.965 (0.958, 0.972)	0.933 (0.921, 0.945)
Senior 1+AI	0.974 (0.966, 0.982)	0.983 (0.976, 0.99)	0.979 (0.973, 0.985)	0.958 (0.948, 0.968)
Senior 2	0.991 (0.987, 0.995)	0.942 (0.93, 0.954)	0.967 (0.96, 0.974)	0.933 (0.92, 0.946)
Senior 2+AI	0.991 (0.986, 0.996)	0.975 (0.966, 0.984)	0.983 (0.978, 0.988)	0.966 (0.957, 0.975)
Junior	0.94 (0.928, 0.952)	0.959 (0.949, 0.969)	0.948 (0.938, 0.958)	0.899 (0.879, 0.919)
Junior+AI	0.983 (0.976, 0.99)	0.95 (0.939, 0.961)	0.966 (0.96, 0.972)	0.933 (0.919, 0.947)
R. resident	0.966 (0.957, 0.975)	0.917 (0.903, 0.931)	0.942 (0.933, 0.951)	0.882 (0.866, 0.898)
R.res.+AI	0.983 (0.977, 0.989)	0.967 (0.957, 0.977)	0.975 (0.969, 0.981)	0.95 (0.94, 0.96)

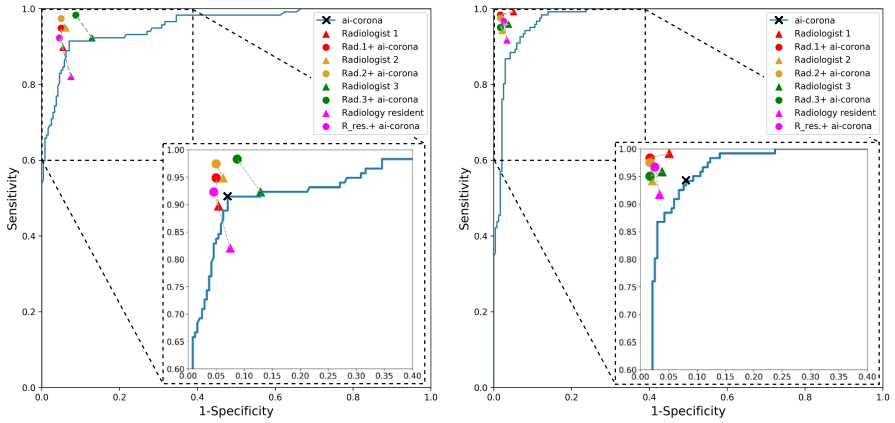
Table 8. The quantitative evaluation of *ai-corona*, radiologists, and model-assisted radiologists performance results for distinguishing between the NCA class and the normal class at a 95% confidence interval.

	AUC
abnormal vs normal + COVID	0.959 (0.944, 0.974)
normal vs abnormal + COVID	0.978 (0.968, 0.988)
COVID vs normal	0.997 (0.995, 0.999)
COVID vs abnormal	0.986 (0.981, 0.991)
normal vs abnormal	0.961 (0.951, 0.971)

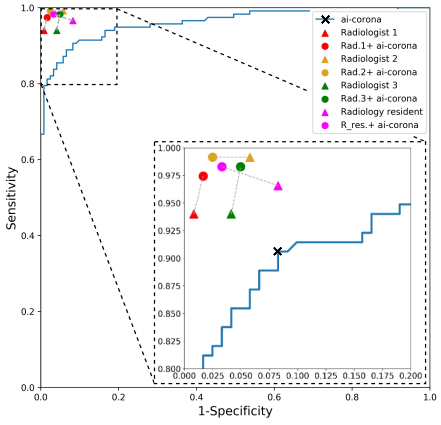
Table 9. Area Under Curve (AUC) summary



(a) COVID-19 vs. normal **(b)** COVID-19 vs. NCA



(c) NCA vs. others **(d)** normal vs. others



(e) normal vs. NCA

Fig 10. The ROC diagram representing the performance of various pipelines for the different combinations of comparison. The Solid blue line is for *ai-corona* by adapting different discrimination threshold value which is used to convert the continuous probability to binary "Yes" or "No" results. The filled triangle symbols are the (1-specificity, sensitivity) for the individual clinical expert, while the filled circle symbols are for the model-assisted radiologist. The inset plots magnify the highest part of sensitivity and specificity.

Acknowledgement

Our framework is available to expert professionals and the public health-care via the website at ai-corona.com for free and unlimited use, where they can upload a chest CT scan and have it diagnosed for COVID-19 infection. The authors would like to express their gratitude to Masih Daneshvari Hospital and Zahra Yousefi for all their hard work and assistance in this project. The computational part of this work was carried out on the High-Performance Computing Cluster of the Institute for Research in Fundamental Sciences (IPM). Our project has received the ethical license of IR.SBMU.NRITLD.REC.1399.024 from the Iranian National Committee for Ethics in Biomedical Research.

References

1. Huang, Chaolin, et al. "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China." The Lancet 395.10223 (2020): 497-506.

2. Chen, Nanshan, et al. "Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study." The Lancet 395.10223 (2020): 507-513.

3. Kujawski, Stephanie A., et al. "characteristics of the first 12 patients with coronavirus disease 2019 (COVID-19) in the United States." Nature Medicine (2020): 1-7.

4. Centers for Disease Control and Prevention. Interim Guidelines for Collecting, Handling, and Testing Clinical Specimens from Persons Under Investigation (PUIs) for Coronavirus Disease 2019 (COVID-19). 2020. www.cdc.gov/coronavirus/2019-ncov/lab/guidelines-clinical-specimens.html. Published February 14, 2020. Accessed April 14, 2020.

5. Wang, Wenling, et al. "Detection of SARS-CoV-2 in different types of clinical specimens." Jama (2020).

6. Mahdavi, Mahdi, et al. "Early detection of COVID-19 mortality risk using non-invasive clinical characteristics." (2020).

7. Ai, Tao, et al. "Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases." Radiology (2020): 200642.

8. Fang, Yicheng, et al. "Sensitivity of chest CT for COVID-19: comparison to RT-PCR." Radiology (2020): 200432.

9. Surkova, Elena, Vladyslav Nikolayevskyy, and Francis Drobniewski. "False-positive COVID-19 results: hidden problems and costs." The Lancet Respiratory Medicine 8.12 (2020): 1167-1168.

10. Shi, Heshui, et al. "Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: a descriptive study." The Lancet Infectious Diseases (2020).

11. Revel, Marie-Pierre, et al. "COVID-19 patients and the Radiology department–advice from the European Society of Radiology (ESR) and the European Society of Thoracic Imaging (ESTI)."

12. Bernheim, Adam, et al. "Chest CT findings in coronavirus disease-19 (COVID-19): relationship to duration of infection." Radiology (2020): 200463.

13. Liu, Xiaoxuan, et al. "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis." The lancet digital health 1.6 (2019): e271-e297.

14. Ardila, Diego, et al. "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography." Nature medicine 25.6 (2019): 954-961.

15. Coudray, Nicolas, et al. "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning." Nature medicine 24.10 (2018): 1559-1567.

16. Fourcade, A., and R. H. Khonsari. "Deep learning in medical image analysis: A third eye for doctors." Journal of stomatology, oral and maxillofacial surgery 120.4 (2019): 279-288.

17. Rajpurkar, Pranav, et al. "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning." arXiv preprint arXiv:1711.05225 (2017).

18. Wang, Xiaosong, et al. "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

19. Rajpurkar, Pranav, et al. "Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists." PLoS medicine 15.11 (2018): e1002686.

20. Irvin, Jeremy, et al. "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. 2019.

21. Bien, Nicholas, et al. "Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet." PLoS medicine 15.11 (2018): e1002699.

22. Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.

23. Li, Lin, et al. "Artificial Intelligence Distinguishes COVID-19 from Community Acquired Pneumonia on Chest CT." Radiology (2020): 200905.

24. Zhang, Kang, et al. "Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of covid-19 pneumonia using computed tomography." Cell (2020).

25. Maas, Benjamin, Erfan Zabeh, and Soroush Arabshahi. "QuickTumorNet: Fast Automatic Multi-Class Segmentation of Brain Tumors." arXiv preprint arXiv:2012.12410 (2020).

26. Jin, Cheng, et al. "Development and evaluation of an artificial intelligence system for COVID-19 diagnosis." Nature communications 11.1 (2020): 1-14.

27. Armato III, Samuel G., et al. "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.

28. Tianchi Competition. https://tianchi.aliyun.com/competition/entrance/231601/informatio (2017).

29. Morozov, S. P., et al. "MosMedData: Chest CT Scans With COVID-19 Related Findings Dataset." arXiv preprint arXiv:2005.06465 (2020).

30. Corman, Victor M., et al. "Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR." Eurosurveillance 25.3 (2020): 2000045.

31. Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009.

32. Huang, Gao, et al. "Densely connected convolutional networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

33. He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

34. Chollet, François. "Xception: Deep learning with depthwise separable convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

35. Tan, Mingxing, and Quoc V. Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." arXiv preprint arXiv:1905.11946. (2019).

36. Bianco, Simone, et al. "Benchmark analysis of representative deep neural network architectures." IEEE Access 6 (2018): 64270-64277.

37. Van Rossum, Guido, and Fred L. Drake. "PYTHON 3 Reference Manual." (2009).

38. Chollet, François. "keras." (2015).

39. Mason, Darcy. "SUET33: pydicom: an open source DICOM library." Medical Physics 38.6Part10 (2011): 3493-3493.

40. Zhou, Bolei, et al. "Learning deep features for discriminative localization." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

41. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

42. Mohammadi, Abbas, et al. "SARS-CoV-2 Detection in Different Respiratory Sites: A Systematic Review and Meta-Analysis." medRxiv (2020).