RESEARCH

COVID-19 infection map generation and detection from chest X-ray images

Aysen Degerli[1*] , Mete Ahishali[1] , Mehmet Yamac[1] , Serkan Kiranyaz[2] , Muhammad E. H. Chowdhury[2] ,
Khalid Hameed[3] , Tahir Hamid[4] , Rashid Mazhar[4] and Moncef Gabbouj[1]

Abstract

Computer-aided diagnosis has become a necessity for accurate and immediate coronavirus disease 2019 (COVID-19) detection to aid treatment and prevent the spread of the virus. Numerous studies have proposed to use Deep Learning techniques for COVID-19 diagnosis. However, they have used very limited chest X-ray (CXR) image repositories for evaluation with a small number, a few hundreds, of COVID-19 samples. Moreover, these methods can neither localize nor grade the severity of COVID-19 infection. For this purpose, recent studies proposed to explore the activation maps of deep networks. However, they remain inaccurate for localizing the actual infestation making them unreliable for clinical use. This study proposes a novel method for the joint localization, severity grading, and detection of COVID-19 from CXR images by generating the so-called *infection maps*. To accomplish this, we have compiled the largest dataset with 119,316 CXR images including 2951 COVID-19 samples, where the annotation of the ground-truth segmentation masks is performed on CXRs by a novel collaborative human–machine approach. Furthermore, we publicly release the first CXR dataset with the ground-truth segmentation masks of the COVID-19 infected regions. A detailed set of experiments show that state-of-the-art segmentation networks can learn to localize COVID-19 infection with an F1-score of 83.20%, which is significantly superior to the activation maps created by the previous methods. Finally, the proposed approach achieved a COVID-19 detection performance with 94.96% sensitivity and 99.88% specificity.

Keywords: SARS-CoV-2, COVID-19 detection, COVID-19 infection segmentation, Deep learning

Introduction

Coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome Coronavirus-2 (SARS-CoV-2) was first reported in December 2019 in Wuhan, China. The highly infectious disease rapidly spread around the World with millions of positive cases. As a result, COVID-19 was declared as a pandemic by the World Health Organization in March 2020. The disease may lead to hospitalization, intubation, intensive care, and even death, especially for the elderly [1, 2]. Naturally, reliable detection of the disease has the utmost importance. However, the diagnosis of COVID-19 is not straight-forward since its symptoms, such as cough, fever, breathlessness, and diarrhea are generally indistinguishable from other viral infections [3, 4].

The diagnostic tools to detect COVID-19 are currently reverse transcription of polymerase chain reaction (RT-PCR) assays and chest imaging techniques, such as Computed Tomography (CT) and X-ray imaging. Primarily, RT-PCR has become the gold standard in the diagnosis of COVID-19 [5, 6]. However, RT-PCR arrays have a high false alarm rate which may be caused by the virus mutations in the SARS-CoV-2 genome, sample contamination, or damage to the sample acquired from the patient [7, 8]. In fact, it is shown in hospitalized patients that RT-PCR sensitivity is low and the test results are highly unstable [6, 9–11]. Therefore, it is recommended to perform chest CT imaging initially on the suspected COVID-19 cases [12], since it is a more reliable clinical tool in the diagnosis with higher sensitivity compared to RT-PCR.

*Correspondence: aysen.degerli@tuni.fi
[1] Faculty of Information Technology and Communication Sciences, Tampere University, Tampere, Finland
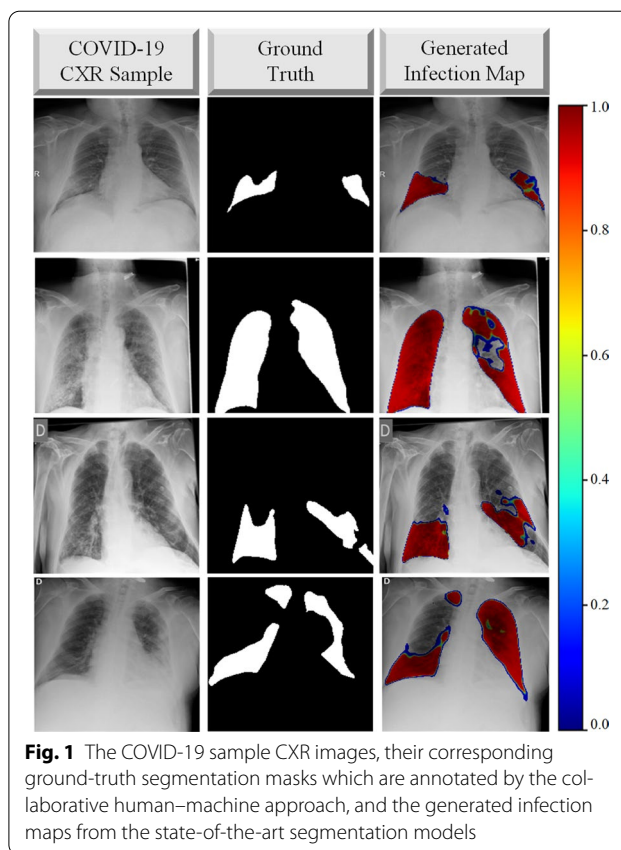Full list of author information is available at the end of the article

Degerli *et al. Health Inf Sci Syst (2021) 9:15*

Page 2 of 16

Hence, several studies [12–14] suggest performing CT on the negative RT-PCR findings of the suspected cases. However, there are several limitations of CT scans. Their sensitivity is limited in the early COVID-19 phase groups [15], and they are limited to recognize only specific viruses [16], slow in image acquisition, and costly. On the other hand, X-ray imaging is faster, cheaper, and less harmful to the body in terms of radiation exposure compared to CT [17, 18]. Moreover, unlike CT devices, X-ray devices are easily accessible; hence, reducing the risk of COVID-19 contamination during the imaging process [19]. Currently, chest X-ray (CXR) imaging is widely used as an assistive tool in COVID-19 prognosis, and it is reported to have a potential diagnosis capability in recent studies [20].
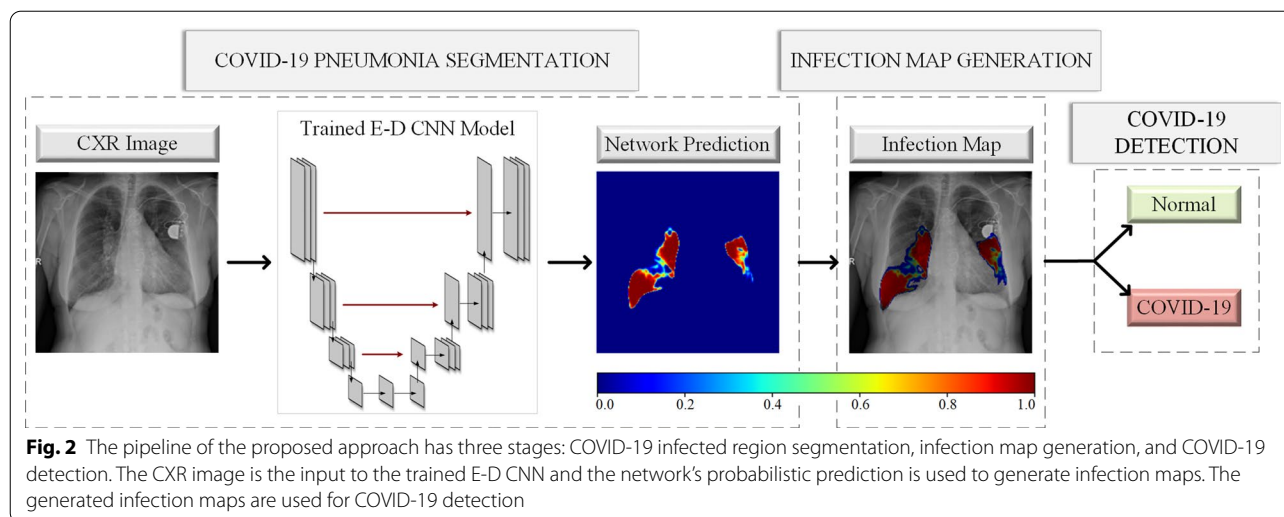
In order to automate COVID-19 detection/recognition from CXR images, several studies [21–23] have extracted features from the CXRs to utilize Support Vector Machines classifier. On the other hand, many studies [17, 24–32] have proposed to use deep Convolutional Neural Networks (CNNs). However, the main limitation of these studies is that the data is scarce for the target COVID-19 class. Such a limited amount of data degrades the learning performance of the deep networks. Two recent studies [33] and [34] have addressed this drawback with a compact network structure and achieved the state-of-the-art detection performance over the benchmark QaTa-COV19 and Early-QaTa-COV19 datasets that consist of 462 and 175 COVID-19 CXR images, respectively. Although these datasets were the largest available at that time, such a limited number of COVID-19 samples raises robustness and reliability issues for the proposed methods in general.

Moreover, all these previous machine learning solutions with X-ray imaging remain limited to only COVID-19 detection. However, as stated by Shi [35], COVID-19 pneumonia screening is important for evaluating the status of the patient and treatment. Therefore, along with the detection, COVID-19 related infection localization is another crucial problem. Hence, several studies [36–38] produced activation maps that are generated from different Deep Learning (DL) models trained for COVID-19 detection (classification) task to localize COVID-19 infection in the lungs. Infection localization has two vital objectives: an accurate assessment of the infection location and the severity of the disease. However, the results of previous studies show that the activation maps generated inherently from the underlying DL network may fail to accomplish both objectives, that is, irrelevant locations with biased severity grading appeared in many cases. To overcome these problems, two studies [39, 40] proposed to perform lung segmentation as the first

Fig. 1 The COVID-19 sample CXR images, their corresponding ground-truth segmentation masks which are annotated by the collaborative human–machine approach, and the generated infection maps from the state-of-the-art segmentation models

step in their approaches. This way, they have narrowed the region of interest down to the regions of lungs to increase the reliability of their methods. Overall, until this study, screening COVID-19 infection from such activation maps produced by classification networks was the only option for the localization due to the absence of ground-truth of the datasets available in the literature. Many studies [35, 39, 41–43] have COVID-19 infection ground-truths for CT images; however, ground-truth segmentation masks for CXR images are non-existent.

In this study, in order to overcome the aforementioned limitations and drawbacks, first, the benchmark dataset QaTa-COV19 proposed by the researchers of Qatar University and Tampere University in [33] and [34] is extended to include 2951 COVID-19 samples. This new dataset is 3–20 times larger than those used in earlier studies. The extended benchmark dataset, QaTa-COV19 with around 120K CXR images, is not only the largest ever composed dataset, but it is the first dataset that has the ground-truth segmentation masks for COVID-19 infection regions, as some samples are shown in Fig. 1. A crucial property of QaTa-COV19 dataset is that it contains CXRs with other (non-COVID-19) infections and anomalies such as pneumonia and pulmonary edema,

Degerli *et al. Health Inf Sci Syst (2021) 9:15*

Page 3 of 16

Fig. 2 The pipeline of the proposed approach has three stages: COVID-19 infected region segmentation, infection map generation, and COVID-19 detection. The CXR image is the input to the trained E-D CNN and the network's probabilistic prediction is used to generate infection maps. The generated infection maps are used for COVID-19 detection

both of which exhibit high visual similarity to COVID-19 infection in the lungs. Therefore, this is significantly more challenging task than distinguishing COVID-19 from the normal (healthy) cases as almost all studies in the literature did.

To obtain the ground-truth segmentation masks for the COVID-19 infected regions, a *human–machine collaborative* approach is introduced. The objective is to significantly reduce the human labor and thus to speed up and also to improve the segmentation masks because when they are drawn solely by medical doctors (MDs), human error due to limited perception, hand-crafting, and subjectivity will deteriorate the overall quality. This is an iterative process, where MDs initiate the segmentation by "manually-drawn" segmentation masks for a subset of CXR images. Then, the trained segmentation networks over this subset generate their own "competing" masks and the MDs are asked to compare them pair-wise (initial manual segmentation vs. machine-segmented masks) for each patient. Such a verification improves the quality of the generated masks as well as the (following) training runs. Over the best masks selected by experts, the networks are trained again this time over a larger set (or even perhaps over the entire dataset), and among the masks generated by the networks, the best masks are selected by the MDs. This human–machine collaboration process continues until the MDs are fully satisfied, i.e., a satisfactory mask can be found among the masks generated by the networks for all CXR images in the dataset. In this study, we show that even with two stages (iterations), highly superior infection maps can be obtained using

which an elegant COVID-19 detection performance can be achieved.

The rest of the paper is organized as follows. In "The benchmark QaTa-COV19 dataset", we introduce the benchmark QaTa-COV19[1] dataset. Our novel human–machine collaborative approach for the ground-truth annotation is explained in "Collaborative human–machine ground-truth annotation". Next, the details of COVID-19 infected region segmentation, and the infection map generation and COVID-19 detection are presented in "COVID-19 infected region segmentation" and "Infection map generation and COVID-19 detection", respectively[2]. The experimental setup and results with the benchmark dataset are reported in "Experimental setup" and "Experimental results", respectively. Finally, we conclude the paper in "Conclusions".

Materials and methodology

The proposed approach in this study is composed of three main phases: (1) training the state-of-the-art deep models for COVID-19 infected region segmentation using the ground-truth segmentation masks, (2) infection map generation from the trained segmentation networks, and (3) COVID-19 detection as it can be depicted in Fig. 2. In this section, we first detail the creation of the benchmark QaTa-COV19 dataset. Then, the proposed approach for collaborative human–machine ground-truth generation is introduced.

---

[1] The benchmark QaTa-COV19 is publicly shared at the repository https://www.kaggle.com/aysendegerli/qatacov19-dataset.

[2] The live demo of the proposed approach is implemented on http://qatacov.live/.

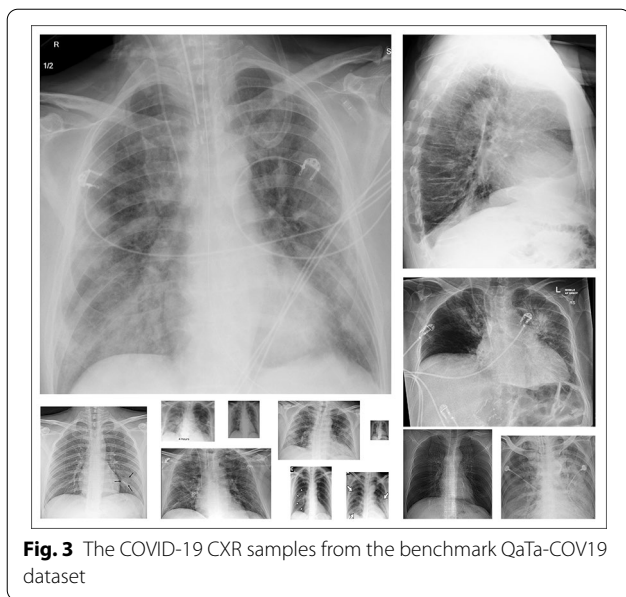Degerli *et al. Health Inf Sci Syst (2021) 9:15*

Page 4 of 16

Fig. 3 The COVID-19 CXR samples from the benchmark QaTa-COV19 dataset

The benchmark QaTa-COV19 dataset

The researchers of Qatar University and Tampere University have compiled the largest COVID-19 dataset up to date with nearly 120K CXR images: *QaTa-COV19* including **2951** COVID-19 CXRs. To create QaTa-COV19, we have utilized several publicly available, scattered, and different format datasets and repositories. Therefore, the collected images from the datasets had some duplicate, over-exposed and low-quality images that were identified and removed in the pre-processing stage. Consequently, the COVID-19 CXRs are from different publicly available sources resulting in high intra-class dissimilarity as depicted in Fig. 3. The image sources of COVID-19 and control group CXRs are detailed as follows:

COVID-19 CXRs

BIMCV-COVID19+ [44] is the largest publicly available dataset with 2473 COVID-19 positive CXR images. The CXR images of BIMCV-COVID19+ dataset were recorded with computed radiography (CR) and digital X-ray (DX) machines. Hannover Medical School and Institute for Diagnostic and Interventional Radiology [45] released 183 CXR images of COVID-19 patients. A total of 959 CXR images are from public repositories: Italian Society of Medical and Interventional Radiology (SIRM), GitHub, and Kaggle [40, 46–49]. As mentioned earlier, any duplication and low-quality images are removed since COVID-19 CXR images are collected from different public datasets and repositories. In this study, a total of 2951 COVID-19 CXRs are

gathered from the aforementioned datasets. Therefore, COVID-19 CXRs are of different age, group, gender, and ethnicity.

Control group CXRs

In this study, we have considered two control groups in the experimental evaluation. Group-I consists of only normal (healthy) CXRs with a smaller number of images compared to the second group. RSNA pneumonia detection challenge dataset [50] is comprised of about 29.7K CXR images, where 8851 images are normal. All CXRs in the dataset are in DICOM format, a popularly used format for medical imaging. Padchest dataset [51] consists of 160, 868 CXR images from 67, 625 patients, where 37, 871 images are from normal class. The images are evaluated and reported by radiologists at Hospital Sun Juan in Spain during 2009–2017. The dataset includes six different position views of CXR and additional information regarding image acquisition and patient demography. Paul Mooney [52] has released an X-ray dataset of 5863 CXR images from a total of 5856 patients, where 1583 images are from normal class. The data is collected from pediatric patients aging one to five years old at Guangzhou Women and Children's Medical Center, Guangzhou. The dataset in [53] consists of 7470 CXR images and the corresponding radiologist reports from the Indiana Network for Patient Care, where a total of 1343 frontal CXR samples are labeled as normal. In [54], there are 80 normal CXRs from the tuberculosis control program of the Department of Health and Human Services of Montgomery County and 326 normal CXRs from Shenzhen Hospital. In this study, a total of 12, 544 normal CXRs are included in control Group-I from the aforementioned datasets. On the other hand, Group-II consists of 116, 365 CXRs from 15 different classes. ChestX-ray14 [55] consists of 112, 120 CXRs with normal and 14 different thoracic disease images, which are atelectasis, cardiomegaly, effusion, infiltration, mass, nodule, pneumonia, pneumothorax, consolidation, edema, emphysema, fibrosis, PT, hernia, and normal (no findings). Additionally, from the pediatric patients [52], 2760 bacterial and 1485 viral pneumonia CXRs are included in Group-II.

Collaborative human–machine ground-truth annotation

Recent developments in the machine and deep learning techniques led to state-of-the-art performance in many computer vision (CV) tasks, such as image classification, object detection, and image segmentation. However, supervised DL methods require a huge amount of annotated data. Otherwise, the limited amount of data degrades the performance of the deep network
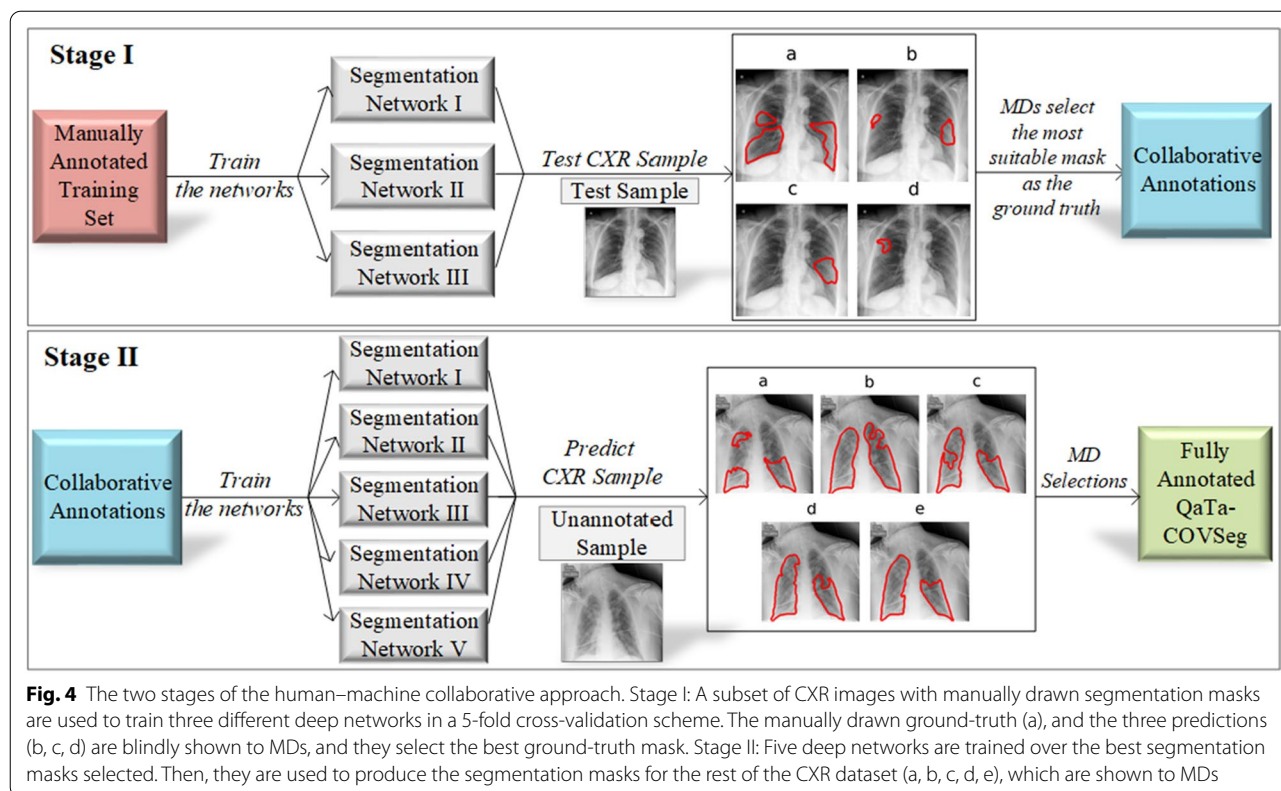
Degerli *et al. Health Inf Sci Syst (2021) 9:15*

Page 5 of 16

Fig. 4 The two stages of the human–machine collaborative approach. Stage I: A subset of CXR images with manually drawn segmentation masks are used to train three different deep networks in a 5-fold cross-validation scheme. The manually drawn ground-truth (a), and the three predictions (b, c, d) are blindly shown to MDs, and they select the best ground-truth mask. Stage II: Five deep networks are trained over the best segmentation masks selected. Then, they are used to produce the segmentation masks for the rest of the CXR dataset (a, b, c, d, e), which are shown to MDs

structures since their generalization capability depends on the availability of large datasets. Nevertheless, to produce ground-truth segmentation masks, pixel-accurate image segmentation by human experts can be a cumbersome and highly subjective task even for moderate size datasets.

In order to overcome this challenge, in this study, we propose a novel collaborative human–machine approach to accurately produce the ground-truth segmentation masks for infected regions directly from the CXR images. The proposed approach is performed in two main stages. First, a group of expert MDs manually segment the infected regions of a subset of (500 in our case) CXR images. Then, several segmentation networks that are inspired by the U-Net [56] structure with a 5-fold cross-validation scheme, are trained over the initial ground-truth masks. For each fold, the segmentation masks of the test samples are predicted by the networks. The network predicted masks along with the initial (MD drawn) ground-truth masks, and original CXR image are assessed by the MDs, and the best segmentation mask among them is selected. Steps of Stage-I are illustrated in Fig. 4 (top). At the end of the first stage, collaboratively annotated ground-truth masks for the subset of CXR images are formed, and they are obviously superior to the initial manually drawn masks since they are selected

by the MDs. An interesting observation in this stage was that MDs preferred the machine-generated masks over the manually drawn masks in the first stage in three out of five cases.

In the second stage five deep networks, inspired by U-Net [56], UNet++ [57], and DLA [58] architectures are trained over the collaborative masks, which were formed in Stage-I. The trained segmentation networks are used to predict the segmentation masks of the rest of the data, which is around 2400 unannotated COVID-19 images. Among the five predictions, the expert MDs select the best one as the ground-truth or deny all if none was found successful. For the latter case, MDs were asked to draw the ground-truth masks manually. However, we notice that this was indeed a minority case that included less than 5% of unannotated data. The steps of Stage-II are shown in Fig. 4 (bottom). As a result, the ground-truth masks for 2951 COVID-19 CXR images are gathered to construct the benchmark QaTa-COV19 dataset. The proposed approach does not only save valuable human labor time, but it also improves the quality and reliability of the masks by reducing the subjectivity with Stage-II verification step.

Degerli *et al. Health Inf Sci Syst (2021) 9:15*

Page 6 of 16

COVID-19 infected region segmentation

Segmentation of COVID-19 infection is the first step of our proposed approach as depicted in Fig. 2. Once the ground-truth annotation for QaTa-COV19 benchmark dataset is formed as explained in the previous section, we perform infected region segmentation extensively with 24 different network configurations. We have used three different segmentation models: U-Net, UNet++ and DLA, with four different encoder structures: CheXNet, DenseNet-121, Inception-v3 and ResNet-50, and frozen & not frozen encoder weight configurations.

Segmentation models

We have tried distinct segmentation model structures starting from shallow to deep structures with varied configurations as follows:

– **U-Net** [56] is an outperforming network for medical image segmentation applications with a u-shaped architecture as the encoder part is symmetric with respect to its decoder part. Therefore, this unique decoder structure with many feature channels allows the network to carry the information through its latest layers.
– **UNet++** [57] has further developed the decoder structure of U-Net by connecting the encoder to the decoder with the nested dense convolutional blocks. This way, the bridge between the encoder and decoder parts are more firmly knit; thus, the information can be transferred to its final layers more intensively compared to the classic U-Net.
– **DLA** [58] investigates the connecting bridges between the encoder and decoder, and proposes a way to fuse the semantic and spatial information with dense layers, which are progressively aggregated by iterative merging to deeper and larger scales.

Encoder selections for segmentation models

In this study, we use several deep CNNs to form the encoder part of the above-mentioned segmentation models as follows:

– **DenseNet-121** [59] is a deep network with 121 layers, each with additional input nodes connecting all the layers directly with each other. Therefore, the maximum information flow through the network is satisfied.
– **CheXNet** [60] is based on the architecture of DenseNet-121, which is trained over the ChestX-ray14 dataset [55] to detect pneumonia cases from CXR images. In [60], DenseNet-121 is initialized with the ImageNet weights and fine-tuned over 100K CXR

images resulting from the state-of-the-art results on the ChestX-ray14 dataset with a better performance compared to the conclusions of radiologists.

– **Inception-v3** [61] achieves state-of-the-art results with much less computational complexity compared to its deep competitors by factorizing the convolutions and pruning the dimensions inside the network. Despite the less complexity, it preserves a higher performance.
– **ResNet-50** [62] introduces a deep residual learning framework that forces the desired mapping of the input to a residual mapping. It is possible to achieve this goal by the shortcut connections on the stacked layers. These connections enable to merge the input and output of the stacked layers by addition operations; therefore, the problem of gradient vanishing is prevented.

We perform transfer learning on the encoder side of the segmentation models by initializing the layers with the ImageNet weights, except for CheXNet which is pretrained on the ChestX-ray14 dataset. We tried two configurations, in the first we freeze the encoder layers while in the second, they are allowed to vary.

Hybrid loss function

In this study, we have performed training the segmentation networks with a hybrid loss function by combining *focal loss* [63] with *dice loss* [64] to achieve a better segmentation performance. We use focal loss since COVID-19 infected region segmentation is an imbalanced problem: the number of background pixels is superior to the foreground's. Let the ground-truth segmentation mask be **Y**, where each pixel class label is defined as y, and the network prediction as $\hat{y}$. We define the pixel class probabilities as for the positive class $P(y = 1) = p$, and for the negative class $P(y = 0) = 1 - p$. On the other hand, the network prediction probabilities are modeled by the logistic function using the sigmoid curve as,

$$P(\hat{y} = 1) = \frac{1}{1 + e^{-z}} = q \qquad (1)$$

$$P(\hat{y} = 0) = 1 - \frac{1}{1 + e^{-z}} = 1 - q \qquad (2)$$

where z is some function of the input CXR image **X**. Then, we define the cross-entropy (CE) loss as follows:

$$CE(p, q) = -p \log q - (1 - p) \log(1 - q). \qquad (3)$$

A common solution to address the class imbalance problem is to add a weighting factor $\alpha \in [0, 1]$ for the positive class, and $1 - \alpha$ for the negative class, which defines the balanced cross-entropy (BCE) loss as,

$$BCE(p,q) = -\alpha p \log q - (1-\alpha)(1-p)\log(1-q). \tag{4}$$

In this way, the importance of positive and negative samples are balanced. However, adding the α factor does not solve the issue for the large class imbalance scenario. This is because the network cannot distinguish outliers (hard samples) and inliers (easy samples) with the BCE loss. To overcome this drawback, focal loss [63] proposes to set focusing parameter $\gamma \geq 0$ in order to down-weight the loss of easy samples that occur with small errors; so that the model can be forced to learn hard negative samples. The focal (F) loss is defined as,

$$F(p,q) = -\alpha(1-q)^{\gamma} p \log q - (1-\alpha)q^{\gamma}(1-p)\log(1-q). \tag{5}$$

where F loss is equivalent to BCE loss when $\gamma = 0$. In our experimental setup, we use the default setting as $\alpha = 0.25$, and $\gamma = 2$ for all the networks. To achieve a good segmentation performance, we combined focal loss with dice loss, which is based on the dice coefficient (DC) defined as follows:

$$DC = \frac{2|\mathbf{Y} \cap \hat{\mathbf{Y}}|}{|\mathbf{Y}| \cup |\hat{\mathbf{Y}}|} \tag{6}$$

where $\hat{\mathbf{Y}}$ is the predicted segmentation mask of the network. Hence, the DC can be interpreted as a dice (D) loss as follows:

$$D(p,q) = 1 - \frac{2\sum p_{h,w}q_{h,w}}{\sum p_{h,w} + \sum q_{h,w}} \tag{7}$$

where h and w are the height and width of the ground-truth and prediction masks $\mathbf{Y}$ and $\hat{\mathbf{Y}}$, respectively. Finally, we combined D and F losses by summation to achieve the so-called hybrid loss function for the segmentation networks.

Infection map generation and COVID-19 detection

Having the training set of COVID-19 CXR images via the collaborative human–machine approach explained in "The benchmark QaTa-COV19 dataset", we train the aforementioned segmentation networks to produce infection maps. After training the segmentation networks, we feed each test CXR sample $\mathbf{X}$ into the trained network. Then, we obtain the network prediction mask $\hat{\mathbf{Y}}$, which is used to generate an infection map that is a measure of infected region probabilities on the input $\mathbf{X}$. Each pixel in $\hat{\mathbf{Y}}$ is defined as $\hat{\mathbf{Y}}_{h,w} \in [0,1]$, where h and w represent the size of the image. We then apply an RGB-based color transform, i.e., the jet color scale to obtain the RGB version of the prediction mask, $\hat{\mathbf{Y}}_{R,G,B}$ as shown in Fig. 5 for a pseudo-colored probability measure visualization. The
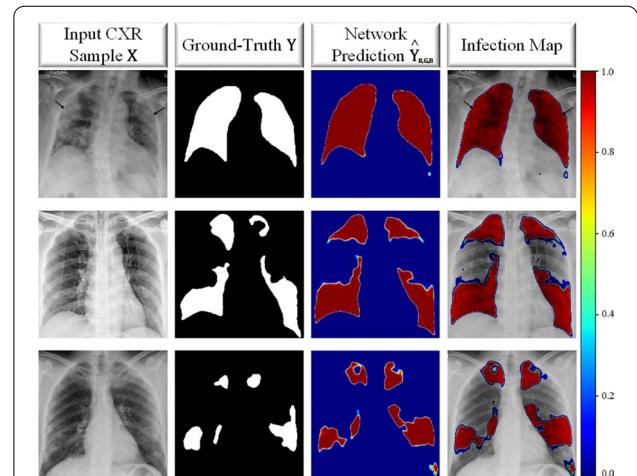

Fig. 5 The three COVID-19 CXR test samples, **X** with the corresponding ground-truth masks, **Y**. The color-coded network predictions, $\hat{\mathbf{Y}}_{R,G,B}$ are reflected translucent onto the **X** to generate an infection map on the lungs, where $\hat{\mathbf{Y}} > 0$

infection map is generated as a reflection of the network prediction $\hat{\mathbf{Y}}_{R,G,B}$ onto the CXR image $\mathbf{X}$. Hence, for visualization, we form the imposed image by concatenating the hue and saturation components of $\hat{\mathbf{Y}}_{H,S,V}$, and value component of $\mathbf{X}_{H,S,V}$. Finally, the imposed image is converted back to RGB domain. In the infection map, we do not show the pixels/regions with zero probabilities for a better visualization effect. This way, the infected regions, where $\hat{\mathbf{Y}} > 0$ are shown translucent as in Fig. 5.

Along with the infection map generation, which already provides localization and segmentation of COVID-19 infection, COVID-19 detection can easily be performed using the proposed approach. The detection of COVID-19 is performed based on the predictions of the trained segmentation networks. Accordingly, a test sample is classified as COVID-19 class if $\hat{\mathbf{Y}} \geq 0.5$ at any pixel location.

Experimental results

In this section, first, the experimental setup is presented. Then, both numerical and visual results are reported with an extensive set of comparative evaluations over the benchmark QaTa-COV19 dataset. Finally, visual comparative evaluations are presented between the infection maps and the activation maps extracted from state-of-the-art deep models.

Experimental setup

Quantitative evaluations for the proposed approach are performed for both COVID-19 infected region segmentation and COVID-19 detection. COVID-19 infected region segmentation is evaluated on a pixel-level, where

Degerli *et al. Health Inf Sci Syst (2021) 9:15*

Page 8 of 16

we consider the foreground (infected region) as the positive class, and background as the negative class. For COVID-19 detection, the performance is computed per CXR sample, and we consider COVID-19 as the positive class and the control group as the negative class. Overall, elements of the confusion matrix are formed as follows: true positive (TP): the number of correctly detected positive class members, true negative (TN): the number of correctly detected negative class samples, false positive (FP): the number of misclassified negative class members, and false negative (FN): the number of misclassified positive class samples. The standard performance evaluation metrics are defined as follows:

$$Sensitivity = \frac{TP}{TP + FN} \qquad (8)$$

where *sensitivity* (or Recall) is the rate of correctly detected positive samples in the positive class samples,

$$Specificity = \frac{TN}{TN + FP} \qquad (9)$$

where *specificity* is the ratio of accurately detected negative class samples to all negative class samples,

$$Precision = \frac{TP}{TP + FP} \qquad (10)$$

where *precision* is the rate of correctly classified positive class samples among all the members classified as positive samples,

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (11)$$

where *accuracy* is the ratio of correctly classified elements among all the data,

$$F(\beta) = (1 + \beta^2)\frac{(Precision \times Sensitivity)}{\beta^2 \times Precision + Sensitivity} \qquad (12)$$

where *F*-score is defined by the weighting parameter β. The *F1*-Score is calculated with $\beta = 1$, which is the harmonic average of *precision* and *sensitivity*. The *F2*-score is calculated with $\beta = 2$, which emphasizes FN minimization over FPs. The main objective of both COVID-19 segmentation and detection is to maximize *sensitivity* with a reasonable *specificity* in order to minimize FP COVID-19 cases or pixels. Equivalently, maximized *F2*-score is targeted with an acceptable *F1*-Score value. The performance with 95% confidence interval (CI) for COVID-19 detection is given in Table 3. The range of values can be calculated for each performance as follows:

$$r = \pm z \sqrt{metric(1 - metric)/N}, \qquad (13)$$

Table 1 Number of CXR samples in control Group-I per fold before and after data augmentation

Data	Number of samples	Training samples	Augmented training samples	Test samples
COVID-19	2951	2361	10035	590
Group-I	12544	10035	10035	2509
Total	15495	12396	**20070**	**3099**

The numbers of training and test samples are denoted in bold

where z is the level of significance, *metric* is any performance evaluation metric, and N is the number of samples. Accordingly, z is set to 1.96 for 95% CI.

We have implemented the deep networks with Tensorflow library [65] using Python on NVidia ®GeForce RTX 2080 Ti GPU card. For training, Adam optimizer [66] is used with the default momentum parameters, $\beta1 = 0.9$ and $\beta2 = 0.999$ using the aforementioned hybrid loss function. The segmentation networks are trained with 50-epochs with a learning rate of $\alpha = 10^{-4}$ and a batch size of 32.

For comparing the computed infection maps, the activation maps are computed as follows: the encoder structures of the segmentation networks are trained for the classification task with a modification at the output layer by adding 2-neurons for the number of total classes. The activation maps extracted from the classification models are then compared with the infection maps of the segmentation models. The classification networks, CheXNet, DenseNet-121, Inception-v3 and ResNet-50 are fine-tuned using categorical cross-entropy as loss function with 10 epochs and a learning rate of $\alpha = 10^{-5}$, which is a sufficient setting to prevent over-fitting, based on our previous study [34]. Other settings of the classifiers are kept the same with the segmentation models.

Experimental results

The experiments are carried out for both COVID-19 infected region segmentation and COVID-19 detection. We extensively tested the benchmark QaTa-COV19 dataset using three different state-of-the-art segmentation networks with four different encoder options for the initial dataset consisting of control Group-I. We also investigated the effect of frozen encoder weights on the performance. On the other hand, the leading model is selected and evaluated on the extended dataset, which includes more negative samples with the control Group-II.

Group-I experiments

We have evaluated the networks in a stratified 5-fold cross-validation scheme with a ratio of 80% training

Degerli *et al. Health Inf Sci Syst (2021) 9:15*

Page 9 of 16

Table 2 Average performance metrics (%) for COVID-19 infected region segmentation computed on the Group-I test (unseen) set from 5-folds with three state-of-the-art segmentation models, four encoder architectures, and weight initializations

Model	Encoder	Encoder layers (frozen *or* not frozen)	Sensitivity	Specificity	Precision	F1-Score	F2-Score	Accuracy	AUC
U-Net	CheXNet	✓	81.20	99.55	83.78	82.47	81.70	99.03	99.19
	CheXNet	✗	82.23	99.56	84.54	83.34	82.66	99.08	99.18
	DenseNet-121	✓	82.29	99.61	86.02	84.11	83.01	99.13	**99.35**
	DenseNet-121	✗	**84.00**	99.66	87.77	**85.81**	**84.71**	**99.22**	99.19
	Inception-v3	✓	80.42	99.59	84.94	82.62	81.28	99.05	99.20
	Inception-v3	✗	82.34	**99.70**	**88.87**	85.43	83.54	99.21	98.82
	ResNet-50	✓	81.43	99.62	86.07	83.67	82.31	99.11	99.30
	ResNet-50	✗	79.90	**99.70**	88.64	83.89	81.43	99.15	98.98
UNet++	CheXNet	✓	80.29	99.59	85.19	82.64	81.21	99.05	99.01
	CheXNet	✗	81.45	99.60	85.60	83.47	82.24	99.09	99.01
	DenseNet-121	✓	82.38	99.61	85.99	84.14	83.08	99.13	99.19
	DenseNet-121	✗	82.36	**99.68**	**88.07**	85.08	83.42	99.19	**99.30**
	Inception-v3	✓	82.87	99.57	84.83	83.81	83.24	99.10	99.21
	Inception-v3	✗	**83.49**	99.66	87.60	**85.45**	**84.22**	**99.20**	99.18
	ResNet-50	✓	82.07	99.59	85.41	83.71	82.72	99.10	99.15
	ResNet-50	✗	82.64	99.62	86.52	84.45	83.33	99.14	99.27
DLA	CheXNet	✓	79.99	99.61	85.57	82.66	81.04	99.06	99.12
	CheXNet	✗	82.84	99.56	84.63	83.71	83.19	99.09	99.17
	DenseNet-121	✓	82.48	99.62	86.40	84.36	83.21	99.14	99.16
	DenseNet-121	✗	82.84	99.56	84.63	83.71	83.19	99.09	99.17
	Inception-v3	✓	80.28	99.63	86.43	83.19	81.41	99.09	99.02
	Inception-v3	✗	**83.44**	**99.68**	**88.18**	**85.73**	**84.34**	**99.22**	**99.29**
	ResNet-50	✓	81.26	99.63	86.48	83.78	82.25	99.12	99.08
	ResNet-50	✗	82.07	99.65	86.99	84.45	83.00	99.15	99.31

The initialized encoder layers are set to *frozen* (✓) and *not frozen* (✗) states during the investigation

The leading performances of each metric are denoted in bold

to 20% test (unseen folds) over the benchmark QaTa-COV19 dataset. The input CXR images are resized to 224 × 224 pixels. Table 1 shows the number of CXRs per fold in the dataset. Since the two classes are imbalanced, we have applied data augmentation in order to balance the classes. Therefore, COVID-19 samples are augmented up to the same number of samples as the control Group-I in the training set for each fold. The data augmentation is performed using Image Data Generator in Keras: the CXR samples are augmented by randomly shifting them both vertically and horizontally by 10% and randomly rotating them in a range of 10 degrees. After shifting and rotating the images, blank sections are filled using the *nearest* mode.

The performance of the segmentation models for COVID-19 infected region segmentation are presented in Table 2. Each model structure is evaluated with two configurations: *frozen* and *not frozen* encoder layers. We have used transfer learning on the encoder layers with ImageNet weights, except for the CheXNet model, which is pre-trained on the ChestX-ray14 dataset. The evaluation of the models with frozen encoder layers is also important since this process can lead to a better convergence and improved performance. However, as the results show, better performance is obtained when the network continues to learn on the encoder layers as well. For each model, we have observed that two encoders: DenseNet-121 and Inception-v3 are the top-performing ones for the infected region segmentation task. The U-Net model with DenseNet-121 encoder holds the leading performance by 84% sensitivity, 85.81% F1-Score, and 84.71% F2-Score. DenseNet-121 produces better results compared to other encoder types since it can preserve the information coming from earlier layers through the output by concatenating the feature maps from each dense layer. However, in the other segmentation models, Inception-v3 outperforms the other encoder types. The presented segmentation performances are obtained by

Degerli *et al. Health Inf Sci Syst (2021) 9:15*

Page 10 of 16

Table 3 Average COVID-19 detection performance results (%) computed from 5-folds over the Group-I test (unseen) set with three network models, four encoder architectures, and weight initializations. The initialized encoder layers are set to frozen (✓) and not frozen (×) states during the investigation

	Encoder	Encoder layers	Sensitivity	Specificity	Precision	F1-Score	F2-Score	Accuracy
U-Net	CheXNet	✓	97.56 ± 0.0056	91.10 ± 0.0050	72.07 ± 0.0071	82.90 ± 0.0059	91.11 ± 0.0045	92.33 ± 0.0042
	CheXNet	×	97.97 ± 0.0051	92.74 ± 0.0045	76.04 ± 0.0067	85.62 ± 0.0055	92.62 ± 0.0041	93.73 ± 0.0038
	DenseNet-121	✓	98.07 ± 0.0050	94.66 ± 0.0039	81.20 ± 0.0062	88.84 ± 0.0050	94.16 ± 0.0037	95.31 ± 0.0033
	DenseNet-121	×	**98.37** ± 0.0046	98.05 ± 0.0024	92.25 ± 0.0042	95.21 ± 0.0034	**97.08** ± 0.0027	98.12 ± 0.0021
	Inception-v3	✓	97.93 ± 0.0051	90.00 ± 0.0052	69.74 ± 0.0072	81.47 ± 0.0061	90.61 ± 0.0046	91.51 ± 0.0044
	Inception-v3	×	97.22 ± 0.0059	**98.37** ± 0.0022	**93.33** ± 0.0039	**95.24** ± 0.0034	96.42 ± 0.0029	**98.15** ± 0.0021
	ResNet-50	✓	98.24 ± 0.0047	93.88 ± 0.0042	79.06 ± 0.0064	87.61 ± 0.0052	93.69 ± 0.0038	94.71 ± 0.0035
	ResNet-50	×	96.37 ± 0.0067	97.82 ± 0.0026	91.21 ± 0.0045	93.72 ± 0.0038	95.30 ± 0.0033	97.54 ± 0.0024
UNet++	CheXNet	✓	97.80 ± 0.0053	91.70 ± 0.0048	73.49 ± 0.0069	83.92 ± 0.0058	91.73 ± 0.0043	92.86 ± 0.0041
	CheXNet	×	97.49 ± 0.0056	93.65 ± 0.0043	78.33 ± 0.0065	86.87 ± 0.0053	92.94 ± 0.0040	94.39 ± 0.0036
	DenseNet-121	✓	97.70 ± 0.0054	94.81 ± 0.0039	81.58 ± 0.0061	88.91 ± 0.0049	93.98 ± 0.0037	95.36 ± 0.0033
	DenseNet-121	×	96.51 ± 0.0066	**99.16** ± 0.0016	**96.44** ± 0.0029	**96.48** ± 0.0029	**96.50** ± 0.0029	**98.66** ± 0.0018
	Inception-v3	✓	**98.31** ± 0.0047	90.54 ± 0.0051	70.96 ± 0.0071	82.43 ± 0.0060	91.27 ± 0.0044	92.02 ± 0.0043
	Inception-v3	×	96.92 ± 0.0061	98.37 ± 0.0022	93.34 ± 0.0039	95.10 ± 0.0034	96.18 ± 0.0030	98.10 ± 0.0021
	ResNet-50	✓	97.80 ± 0.0053	93.39 ± 0.0043	77.69 ± 0.0066	86.59 ± 0.0054	92.98 ± 0.0040	94.23 ± 0.0037
	ResNet-50	×	96.78 ± 0.0064	97.43 ± 0.0028	89.87 ± 0.0048	93.20 ± 0.0040	95.31 ± 0.0033	97.31 ± 0.0025
DLA	CheXNet	✓	97.46 ± 0.0057	92.47 ± 0.0046	75.27 ± 0.0068	84.94 ± 0.0056	92.03 ± 0.0043	93.42 ± 0.0039
	CheXNet	×	97.32 ± 0.0058	94.93 ± 0.0038	81.87 ± 0.0061	88.93 ± 0.0049	93.78 ± 0.0038	95.39 ± 0.0033
	DenseNet-121	✓	97.36 ± 0.0058	95.66 ± 0.0036	84.08 ± 0.0058	90.23 ± 0.0047	94.38 ± 0.0036	95.99 ± 0.0031
	DenseNet-121	×	97.09 ± 0.0061	99.07 ± 0.0017	96.08 ± 0.0031	**96.58** ± 0.0029	**96.88** ± 0.0027	**98.69** ± 0.0018
	Inception-v3	✓	96.92 ± 0.0062	93.24 ± 0.0044	77.13 ± 0.0066	85.90 ± 0.0055	92.19 ± 0.0042	93.94 ± 0.0040
	Inception-v3	×	96.71 ± 0.0064	**99.13** ± 0.0016	**96.32** ± 0.0030	96.52 ± 0.0029	96.63 ± 0.0028	98.67 ± 0.0018
	ResNet-50	✓	**97.49** ± 0.0056	95.30 ± 0.0037	82.98 ± 0.0059	89.65 ± 0.0048	94.20 ± 0.0037	95.71 ± 0.0032
	ResNet-50	×	96.17 ± 0.0069	98.15 ± 0.0024	92.44 ± 0.0042	94.27 ± 0.0037	95.40 ± 0.0033	97.77 ± 0.0023

The leading performances of each metric are denoted in bold

setting the threshold value to 0.5 to compute the segmentation mask from the network probabilities. Precision-Recall curves of three leading deep models are plotted in Fig. 6a by varying this threshold value. Additionally, the Receiver Operating Characteristics (ROC) curve of these models, and their corresponding area under curve (AUC) scores are presented in Fig. 6b. Further investigation shows that AUC scores of the leading segmentation models are directly proportional to the COVID-19 detection performance as it can be depicted from the precision scores in Table 3.

The performances of the segmentation models for COVID-19 detection are presented in Table 3. All the models are evaluated by stratified a 5-fold cross-validation scheme, and the table shows the averaged results of these folds. The most crucial metric here is the sensitivity since missing any patient with COVID-19 is critical. In fact, the results indicate the robustness of the model as the proposed approach can achieve high sensitivity levels of 98.37% with a 97.08% F2-Score. Additionally, the proposed

approach achieves an elegant specificity of 99.16%, indicating a significantly low false alarm rate. It can be observed from Table 3 that DenseNet-121 encoder with the not frozen encoder layer setting gives the most promising results among the others. The confusion matrices, accumulated on each fold's test set, are presented in Table 4. The highest sensitivity in COVID-19 detection is achieved by the U-Net DenseNet-121 model (Table 4a). Accordingly, the U-Net DenseNet-121 model only misses 48 COVID-19 patients out of 2951. On the other hand, the highest specificity is achieved by UNet++ DenseNet-121 model (Table 4b). The UNet++ model only misses a minor part of the control class with 105 samples out of 12544.

Group-II experiments

We have selected the leading model from the Group-I experiments as U-Net with not frozen DenseNet-121 encoder setting. In Group-II experiments, we have gathered around 120K CXRs. The CXRs from the ChestX-ray14 dataset [55] are already divided into train and test
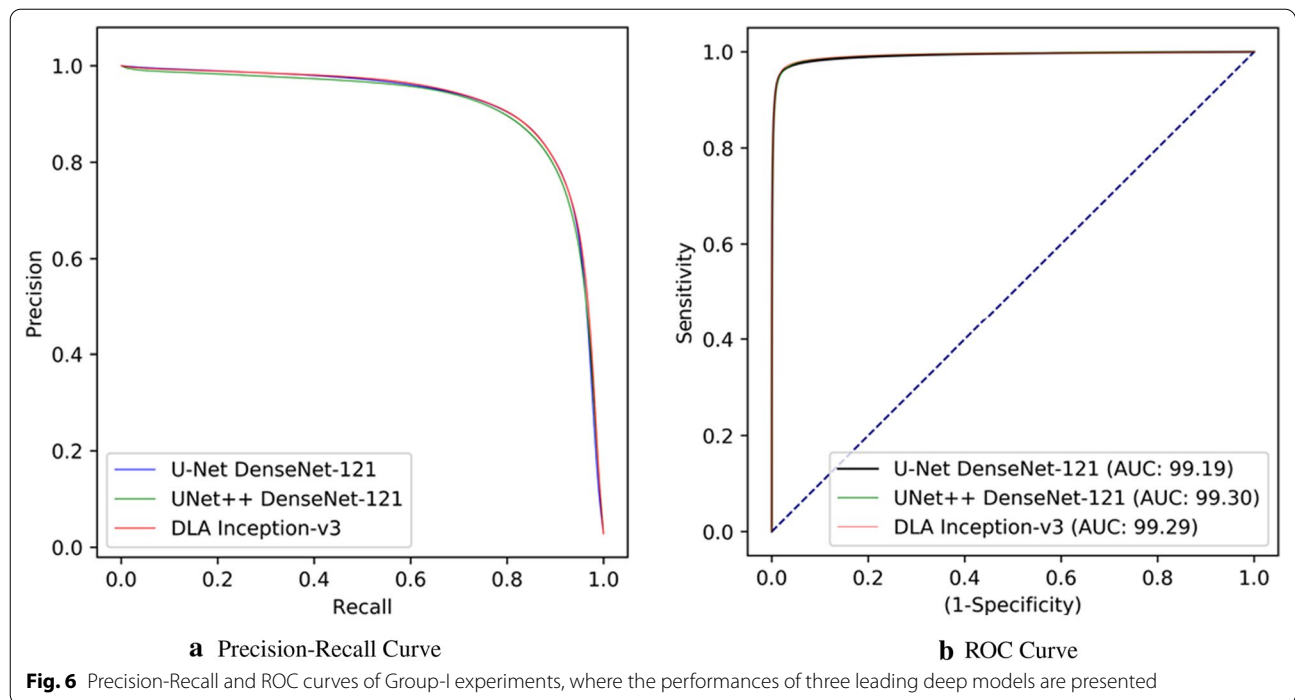
Fig. 6 Precision-Recall and ROC curves of Group-I experiments, where the performances of three leading deep models are presented

Table 4 Cumulative confusion matrices of COVID-19 detection by the best performing U-Net and UNet++ models with DenseNet-121 encoder

(a) U-Net DenseNet-121

U-Net		Predicted	
		Group-I	COVID-19
Ground Truth	Group-I	12,300	244
	COVID-19	48	2903

(b) UNet++ DenseNet-121

UNet++		Predicted	
		Group-I	COVID-19
Ground Truth	Group-I	12,439	105
	COVID-19	103	2848

sets. Accordingly, we have randomly separated the train and test sets of COVID-19, viral pneumonia, and bacterial pneumonia CXRs by keeping the same train/test ratio as in ChestX-ray14 [55]. Table 5 shows the number of training and test samples of the Group-II experiments. Additonally, we have applied augmentation to data except for ChestX-ray14 samples with the same set-up as in the Group-I experiments. In these experiments, we do not perform any cross-validation since ChestX-ray14 has predefined training and test sets.

The performance of the U-Net model for COVID-19 infected region segmentation and detection is presented in Table 6. The model achieved a segmentation performance by 81.72% sensitivity and 83.20% F1-Score. In comparison to initial experiments with the control Group-I data, the model can still achieve an elegant segmentation performance even with numerous samples in the test set. On the other hand, the COVID-19 detection performance with 27, 438 CXR images is very successful by 94.96% sensitivity, 99.88% specificity, and 96.40% precision. This indicates a very low false alarm rate of only 0.12%. Table 7 shows the confusion matrix on the test set. Accordingly, the model only misses 44 COVID-19 samples. In the control Group-II, only 31 CXR samples
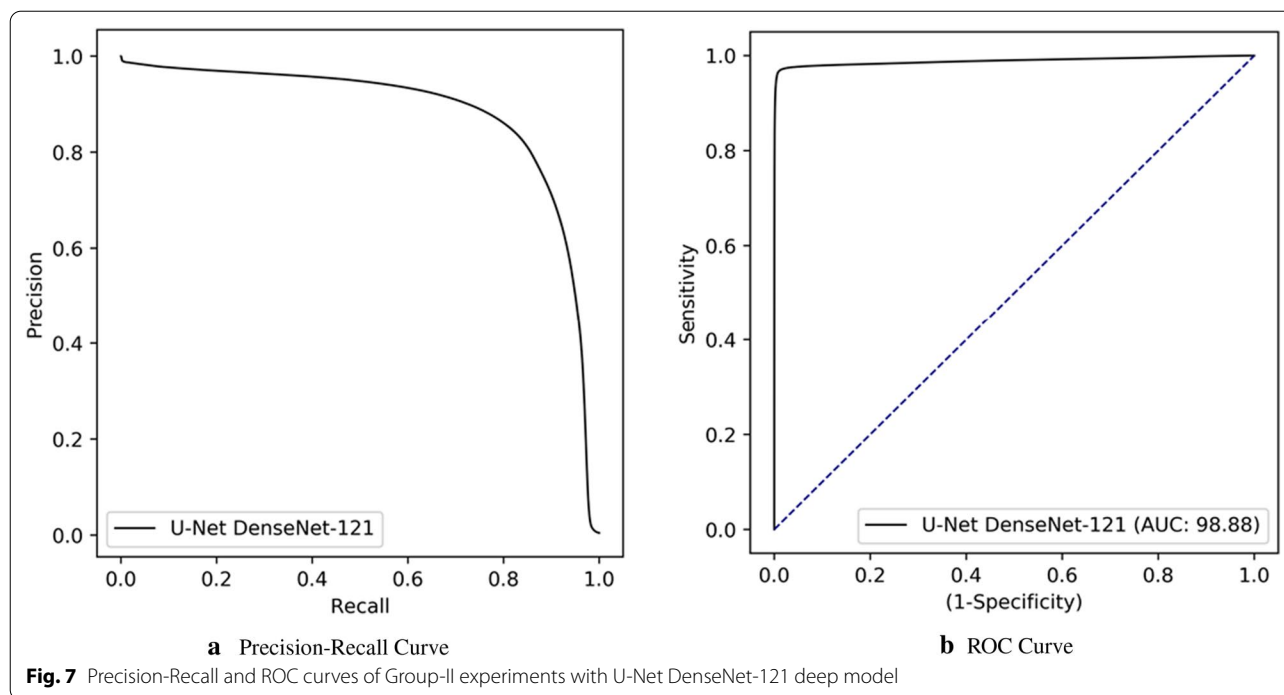
Degerli *et al. Health Inf Sci Syst (2021) 9:15*

Page 12 of 16

Fig. 7 Precision-Recall and ROC curves of Group-II experiments with U-Net DenseNet-121 deep model

Table 5 Number of CXR samples in control Group-II before and after data augmentation

Data	Training samples	Augmented	Augmented training samples	Test samples
COVID-19	2078	✓	10, 000	873
Bacterial Pneumonia	2130	✓	5000	630
ChestX-ray14	86, 524	✗	86, 524	25, 596
Viral Pneumonia	1146	✓	5000	339
Total	91,878		**106,524**	**27,438**

The numbers of training and test samples are denoted in bold

Table 6 COVID-19 infected region segmentation and detection results (%) computed on the Group-II test set from the U-Net model with DenseNet-121 encoder

	Performance metrics	Infected region segmentation	Detection
U-Net DenseNet-121	Sensitivity	81.72	94.96
	Specificity	99.93	99.88
	Precision	84.74	96.40
	F1-Score	83.20	95.67
	F2-Score	82.31	95.24
	Accuracy	99.85	99.73

Table 7 Cumulative confusion matrices of COVID-19 detection by the best performing U-Net model with DenseNet-121 encoder
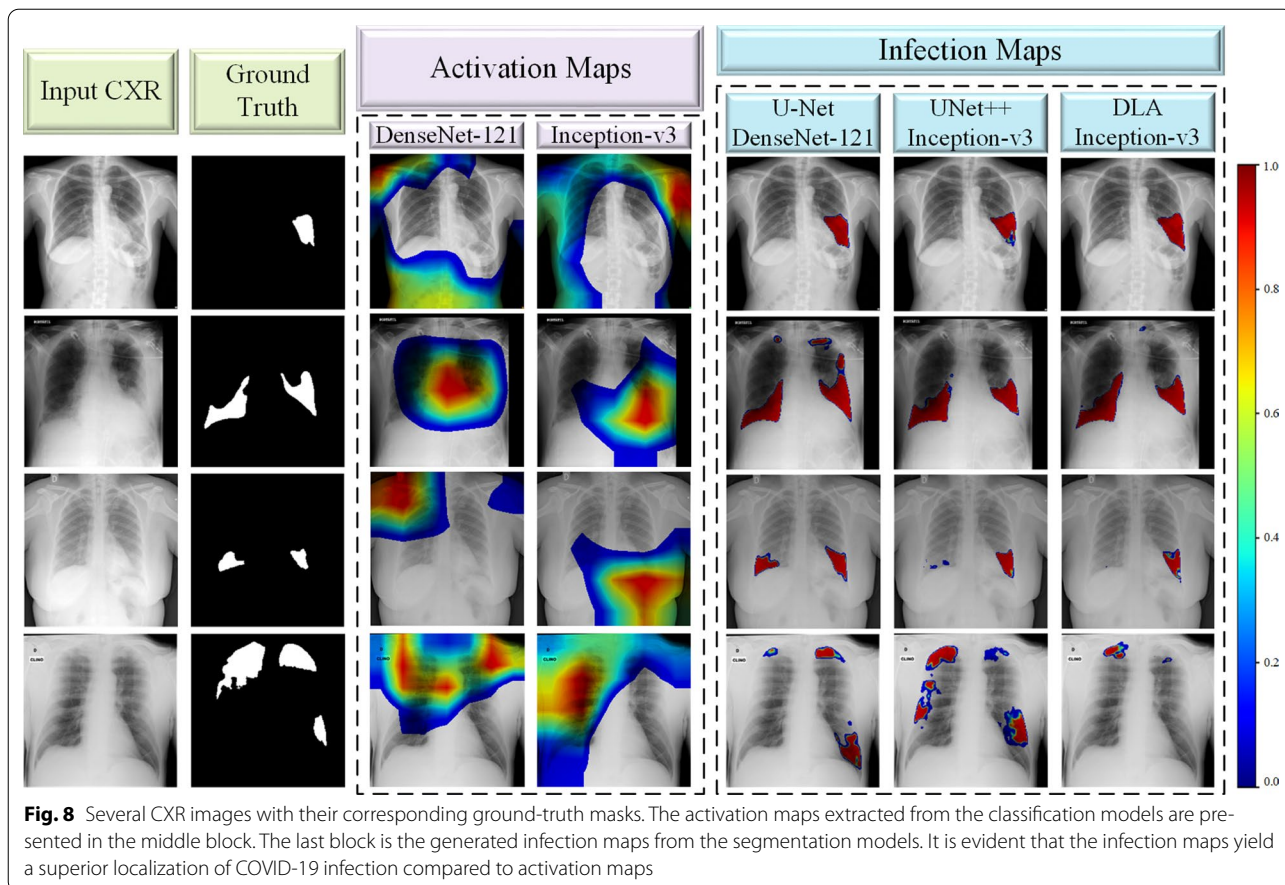
U-Net		Predicted	
		Group-II	COVID-19
Ground Truth	Group-II	26, 534	31
	COVID-19	44	829

samples. Lastly, Precision-Recall and ROC curves of Group-II experiments can be depicted from Fig. 7.

Infection vs activation maps

Several studies [36–38] propose to localize COVID-19 from CXRs by extracting activation maps from the deep classification models trained for COVID-19 detection.

are missed, which is a minor section in 26, 565 negative samples. The results show that the leading model is still robust on the extended data, where it consists of 15 different classes with 14 thoracic diseases and normal

Fig. 8 Several CXR images with their corresponding ground-truth masks. The activation maps extracted from the classification models are presented in the middle block. The last block is the generated infection maps from the segmentation models. It is evident that the infection maps yield a superior localization of COVID-19 infection compared to activation maps

Despite the simplicity of the idea, there are many limitations of this approach. First of all, without any infected region segmentation ground-truth masks, the network can only produce a rough localization, and the extracted activation maps may entirely fail to localize COVID-19 infection.

In this study, we check the reliability of our proposed COVID-19 detection approach by comparing it with DL models trained for the classification task. In order to achieve this objective, we compare the infection map and activation map of CXR images, which are generated from the segmentation and classification networks, respectively. Therefore, we have trained the encoder structures of the segmentation networks, which are CheXNet, DenseNet-121, Inception-v3, and ResNet-50 to perform COVID-19 classification task. We have extracted activation maps from these trained models by the Gradient-weighted Class Activation Mapping (Grad-CAM) approach proposed in [67]. The localization Grad-CAM $L^c_{\text{Grad-CAM}} \in \mathbb{R}^{h \times w}$ of height h and width w for class c is calculated by the gradient of m^c before the softmax with respect to the convolutional layer's feature maps A^k as $\frac{\partial m^c}{\partial A^k}$. The gradients are

passed through from the global average pooling during back-propagation;

$$\alpha^c_k = \frac{1}{Z} \sum_i \sum_j \frac{\partial m^c}{\partial A^k}, \tag{14}$$

where α is the weight that shows the important feature map k from A for a target class c. Then, the linear combination is performed following by ReLU to obtain the Grad-CAM;

$$L^c_{\text{Grad-CAM}} = ReLU(\sum_k \alpha^c_k A^k). \tag{15}$$

Despite their elegant performance, activation maps extracted from deep classification networks are not suitable for localizing COVID-19 infection as depicted in Fig. 8. In fact, infections found by the activation maps are highly irrelevant indicating false locations outside of the lung areas. On the other hand, infection maps can generate a highly accurate location with an elegant severity grading of COVID-19 infection. The proposed infection maps can conveniently be used by medical experts for an enhanced assessment of the disease. Real-time

Degerli *et al. Health Inf Sci Syst (2021) 9:15*

Page 14 of 16

Table 8 The number of trainable and non-trainable parameters of the models with their inference time (ms) per sample

	Encoder	Encoder Layers	Train-able	Non-Traina-ble	Time (ms)
U-Net	CheXNet	✓	5.19M	6.96M	2.56
	DenseNet-121	✓	5.19M	6.96M	2.58
	Inception-v3	✓	8.15M	21.79M	**2.53**
	ResNet-50	✓	9.06M	23.50M	2.54
	CheXNet	✓	12.06M	85.63K	2.62
	DenseNet-121	✗	12.06M	85.63K	2.58
	Inception-v3	✗	29.9M	36.42K	2.61
	ResNet-50	✗	32.51M	47.56K	2.64
UNet++	CheXNet	✓	7.53M	6.96M	5.17
	DenseNet-121	✓	7.53M	6.96M	5.10
	Inception-v3	✓	8.68M	21.79M	5.32
	ResNet-50	✓	10.88M	23.51M	5.58
	CheXNet	✗	14.40M	88.45K	5.24
	DenseNet-121	✗	14.40M	88.45K	5.25
	Inception-v3	✗	30.43M	39.23K	5.32
	ResNet-50	✗	34.34M	50.37K	5.46
DLA	CheXNet	✓	6.27M	6.96M	4.65
	DenseNet-121	✓	6.27M	6.96M	4.63
	Inception-v3	✓	7.20M	21.79M	4.70
	ResNet-50	✓	8.74M	23.51M	4.90
	CheXNet	✗	13.15M	88.45K	4.63
	DenseNet-121	✗	13.15M	88.45K	4.65
	Inception-v3	✗	28.96M	39.23K	4.72
	ResNet-50	✗	32.2M	50.37K	4.90

The initialized encoder layers are set to frozen (✓) *or* not frozen (✗)

The inference time of the fastest model is denoted in bold

implementation of the infection maps will obviously speed up the detection process, can also monitor the progression of COVID-19 infection in the lungs.

Computational complexity analysis

In this section, we present the computational times of the networks and their number of trainable & non-trainable parameters. Table 8 shows the elapsed time in milliseconds (ms) during the inference step for each network used in the experiments. The results in the table represent the running time per sample. It can be observed from the table that the U-Net model is the fastest among the others due to its shallow structure. The fastest network is U-Net Inception-v3 with frozen encoder layers taking up 2.53 ms. On the other hand, the slowest model is UNet++ structure since it has the largest number of trainable parameters. The most computationally demanding model is UNet++ ResNet-50 with frozen

encoder layers, which takes 5.58 ms. We, therefore, conclude that all models can be used as real-time clinical applications.

Conclusions

The immediate and accurate detection of highly infectious COVID-19 plays a vital role in preventing the spread of the virus. In this study, we used CXR images since X-ray imaging is cheaper, easily accessible, and faster than the conventional methods commonly used such as RT-PCR and CT. As a major contribution, the largest CXR dataset, QaTa-COV19, which consists of 2951 COVID-19, and 116, 365 control group images, has been compiled and is shared publicly as a benchmark dataset. Moreover, for the first time in the literature, we release the ground-truth segmentation masks of the infected regions along with the introduced benchmark QaTa-COV19. Furthermore, we proposed a human–machine collaborative approach, which can be used when a fast and accurate ground-truth annotation is desired but manual segmentation is slow, costly, and subjective. Finally, this study reveals the first approach ever proposed for infection map generation in CXR images. Our extensive experiments on QaTa-COV19 show that a reliable COVID-19 diagnosis can be achieved by generating infection maps, which can locate the infection on the lungs by 81.72% sensitivity, and 83.20% F1-Score. Moreover, the proposed joint approach can achieve an elegant COVID-19 detection performance with 94.96% sensitivity and 99.88% specificity. Many COVID-19 detectors proposed in the literature reported similar or even better detection performances. However, not only they are evaluated over small-size datasets, but also they can only discriminate between COVID-19 and normal (healthy) data, which is a straightforward task. The proposed joint approach is the only COVID-19 detector that can distinguish it from other thoracic diseases as being evaluated over the largest CXR dataset ever composed. Accordingly, the most important aspect of this study is that the generated infection maps can assist MDs for a better and objective COVID-19 assessment. For instance, it can show the time progress of the disease if the time series CXR data are generated by the proposed infection maps. It is clear that when compared with the activation maps extracted from deep models, the proposed infection maps are highly superior and reliable cues for COVID-19 infection.

Author details
[1] Faculty of Information Technology and Communication Sciences, Tampere University, Tampere, Finland. [2] Department of Electrical Engineering, Qatar

Degerli *et al. Health Inf Sci Syst (2021) 9:15*

Page 15 of 16

University, Doha, Qatar. [3] Reem Medical Center, Doha, Qatar. [4] Hamad Medical Corporation Hospital, Doha, Qatar.

References

1. Severe Outcomes Among Patients with Coronavirus Disease 2019 (COVID-19) - United States, February 12-March 16, 2020. MMWR Morb Mortal Wkly Rep 2020;69:343-346. https://doi.org/10.15585/mmwr.mm6912e2
2. World Health Organization: Coronavirus disease 2019 (covid-19): situation report, 88 (2020)
3. Sohrabi C, Alsafi Z, O'Neill N, Khan M, Kerwan A, Al-Jabir A, Iosifidis C, Agha R. World health organization declares global emergency: a review of the 2019 novel coronavirus (covid-19). Int J Surg. 2020;73:71–6. https://doi.org/10.1016/j.ijsu.2020.02.034.
4. Singhal T. A review of coronavirus disease-2019 (covid-19). Indian J Pediatr. 2020;87(4):281–6. https://doi.org/10.1007/s12098-020-03263-6.
5. Kakodkar, P., Kaka, N., Baig, M.: A comprehensive literature review on the clinical presentation, and management of the pandemic coronavirus disease 2019 (covid-19). Cureus **12**(4) (2020). https://doi.org/10.7759/cureus.7560
6. Li Y, Yao L, Li J, Chen L, Song Y, Cai Z, Yang C. Stability issues of rt-pcr testing of sars-cov-2 for hospitalized patients clinically diagnosed with covid-19. J Med Virol. 2020;92(7):903–8. https://doi.org/10.1002/jmv.25786.
7. Tahamtan A, Ardebili A. Real-time rt-pcr in covid-19 detection: issues affecting the results. Expert Rev Mol Diagn. 2020;20(5):453–4. https://doi.org/10.1080/14737159.2020.1757437.
8. Xia J, Tong J, Liu M, Shen Y, Guo D. Evaluation of coronavirus in tears and conjunctival secretions of patients with sars-cov-2 infection. J Med Virol. 2020;92(6):589–94. https://doi.org/10.1002/jmv.25725.
9. Xiao AT, Tong YX, Zhang S. False-negative of rt-pcr and prolonged nucleic acid conversion in covid-19: rather than recurrence. J Med Virol. 2020;92(10):1755–6. https://doi.org/10.1002/jmv.25855.
10. Yang, Y., Yang, M., Shen, C., Wang, F., Yuan, J., Li, J., Zhang, M., Wang, Z., Xing, L., Wei, J., et al.: Laboratory diagnosis and monitoring the viral shedding of 2019-ncov infections. The Innovation **1**(100061) (2020). https://doi.org/10.1016/j.xinn.2020.100061
11. World Health Organization. Laboratory testing for coronavirus disease 2019 (covid-19) in suspected human cases: interim guidance, 2 March 2020. Tech. rep.: World Health Organization; 2020.
12. Salehi S, Abedi A, Balakrishnan S, Gholamrezanezhad A. Coronavirus disease 2019 (covid-19): a systematic review of imaging findings in 919 patients. AJR Am J Roentgenol. 2020;215(1):87–93. https://doi.org/10.2214/AJR.20.23034.
13. Fang Y, Zhang H, Xie J, Lin M, Ying L, Pang P, Ji W. Sensitivity of chest ct for covid-19: comparison to rt-pcr. Radiology. 2020;296(2):E115–7. https://doi.org/10.1148/radiol.2020200432.
14. Ai T, Yang Z, Hou H, Zhan C, Chen C, Lv W, Tao Q, Sun Z, Xia L. Correlation of chest ct and rt-pcr testing in coronavirus disease 2019 (covid-19) in china: a report of 1014 cases. Radiology. 2020;296(2):E32–40. https://doi.org/10.1148/radiol.2020200642.
15. Bernheim A, Mei X, Huang M, Yang Y, Fayad ZA, Zhang N, Diao K, Lin B, Zhu X, Li K, et al. Chest ct findings in coronavirus disease-19 (covid-19): relationship to duration of infection. Radiology. 2020;295(3):200463. https://doi.org/10.1148/radiol.2020200463.
16. Li Y, Xia L. Coronavirus disease 2019 (covid-19): role of chest ct in diagnosis and management. AJR Am J Roentgenol. 2020;214(6):1280–6. https://doi.org/10.2214/AJR.20.22954.
17. Narin, A., Kaya, C., Pamuk, Z.: Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks. arXiv preprint arXiv:2003.10849 (2020)
18. Brenner DJ, Hall EJ. Computed tomography-an increasing source of radiation exposure. N Engl J Med. 2007;357(22):2277–84. https://doi.org/10.1056/NEJMra072149.
19. Rubin GD, Ryerson CJ, Haramati LB, Sverzellati N, Kanne JP, Raoof S, Schluger NW, Volpi A, Yim JJ, Martin IB, et al. The role of chest imaging in patient management during the covid-19 pandemic: a multinational consensus statement from the fleischner society. Radiology. 2020;296(1):172–80. https://doi.org/10.1148/radiol.2020201365.
20. Shi, F., Wang, J., Shi, J., Wu, Z., Wang, Q., Tang, Z., He, K., Shi, Y., Shen, D.: Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for covid-19. IEEE Rev Biomed Eng pp. 1 (2020). https://doi.org/10.1109/RBME.2020.2987975
21. Ismael AM, Şengür A. Deep learning approaches for covid-19 detection based on chest x-ray images. Expert Syst Appl. 2021;164:114054. https://doi.org/10.1016/j.eswa.2020.114054.
22. Gomes, J.C., Barbosa, V.A.d.F., Santana, M.A., Bandeira, J., Valença, M.J.S., de Souza, R.E., Ismael, A.M., dos Santos, W.P.: Ikonos: An intelligent tool to support diagnosis of covid-19 by texture analysis of x-ray images. Res. Biomed. Eng. pp. 1–14 (2020). https://doi.org/10.1007/s42600-020-00091-7
23. Ismael AM, Şengür A. The investigation of multiresolution approaches for chest x-ray image based covid-19 detection. Health Inf Sci Syst. 2020;8(1):1–11. https://doi.org/10.1007/s13755-020-00116-6.
24. Chowdhury, N.K., Rahman, M.M., Kabir, M.A.: Pdcovidnet: a parallel-dilated convolutional neural network architecture for detecting covid-19 from chest x-ray images. Health Inf Sci Syst **8**(27) (2020). https://doi.org/10.1007/s13755-020-00119-3
25. Pham TD. Classification of covid-19 chest x-rays with deep learning: new models or fine tuning? Health Inf Sci Syst. 2021;9(1):1–11. https://doi.org/10.1007/s13755-020-00135-3.
26. Chowdhury ME, Rahman T, Khandakar A, Mazhar R, Kadir MA, Mahbub ZB, Islam KR, Khan MS, Iqbal A, Al-Emadi N, et al. Can ai help in screening viral and covid-19 pneumonia? IEEE Access. 2020;8:132665–76. https://doi.org/10.1109/ACCESS.2020.3010287.
27. Apostolopoulos ID, Mpesiana TA. Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks. Phys Eng Sci Med. 2020;43(2):635–40. https://doi.org/10.1007/s13246-020-00865-4.
28. Hall, L.O., Paul, R., Goldgof, D.B., Goldgof, G.M.: Finding covid-19 from chest x-rays using deep learning on a small dataset. arXiv preprint arXiv:2004.02060 (2020)
29. Wang, L., Lin, Z.Q., Wong, A.: Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. Sci Rep **10**(19549) (2020). https://doi.org/10.1038/s41598-020-76550-z
30. Sethy, P.K., Behera, S.K.: Detection of coronavirus disease (covid-19) based on deep features. Preprints **2020030300** (2020). https://doi.org/10.20944/preprints202003.0300.v1
31. Zhang, J., Xie, Y., Li, Y., Shen, C., Xia, Y.: Covid-19 screening on chest x-ray images using deep learning based anomaly detection. arXiv preprint arXiv:2003.12338 (2020)
32. Afshar, P., Heidarian, S., Naderkhani, F., Oikonomou, A., Plataniotis, K.N., Mohammadi, A.: Covid-caps: a capsule network-based framework for identification of covid-19 cases from x-ray images. arXiv preprint arXiv:2004.02696 (2020)
33. Yamac, M., Ahishali, M., Degerli, A., Kiranyaz, S., Chowdhury, M.E., Gabbouj, M.: Convolutional sparse support estimator based covid-19 recognition from x-ray images. arXiv preprint arXiv:2005.04014 (2020)
34. Ahishali, M., Degerli, A., Yamac, M., Kiranyaz, S., Chowdhury, M.E.H., Hameed, K., Hamid, T., Mazhar, R., Gabbouj, M.: Advance warning methodologies for covid-19 using chest x-ray images. IEEE Access 2021;9:41052–65. https://doi.org/10.1109/ACCESS.2021.3064927
35. Shi, F., Xia, L., Shan, F., Wu, D., Wei, Y., Yuan, H., Jiang, H., Gao, Y., Sui, H., Shen, D.: Large-scale screening of covid-19 from community acquired pneumonia using infection size-aware classification. arXiv preprint arXiv:2003.09860 (2020)
36. Yeh, C.F., Cheng, H.T., Wei, A., Liu, K.C., Ko, M.C., Kuo, P.C., Chen, R.J., Lee, P.C., Chuang, J.H., Chen, C.M., et al.: A cascaded learning strategy for robust covid-19 pneumonia chest x-ray screening. arXiv preprint arXiv:2004.12786 (2020)
37. Oh Y, Park S, Ye JC. Deep learning covid-19 features on cxr using limited training data sets. IEEE Trans Med Imaging. 2020;39(8):2688–700. https://doi.org/10.1109/TMI.2020.2993291.
38. Ozturk, T., Talo, M., Yildirim, E.A., Baloglu, U.B., Yildirim, O., Acharya, U.R.: Automated detection of covid-19 cases using deep neural networks with x-ray images. Comput Biol Med **121**(103792) (2020). https://doi.org/10.1016/j.compbiomed.2020.103792

Degerli *et al. Health Inf Sci Syst (2021) 9:15*

Page 16 of 16

39. Alom, M.Z., Rahman, M., Nasrin, M.S., Taha, T.M., Asari, V.K.: Covid_mtnet: Covid-19 detection with multi-task deep learning approaches. arXiv preprint arXiv:2004.03747 (2020)
40. Haghanifar, A., Majdabadi, M.M., Ko, S.: Covid-cxnet: Detecting covid-19 in frontal chest x-ray images using deep learning. arXiv preprint arXiv:2006.13807 (2020)
41. Shan, F., Gao, Y., Wang, J., Shi, W., Shi, N., Han, M., Xue, Z., Shi, Y.: Lung infection quantification of covid-19 in ct images with deep learning. arXiv preprint arXiv:2003.04655 (2020)
42. Zhang K, Liu X, Shen J, Li Z, Sang Y, Wu X, Zha Y, Liang W, Wang C, Wang K, et al. Clinically applicable ai system for accurate diagnosis, quantitative measurements, and prognosis of covid-19 pneumonia using computed tomography. Cell. 2020;181(6):1423–33. https://doi.org/10.1016/j.cell.2020.04.045.
43. Qiu, Y., Liu, Y., Xu, J.: Miniseg: An extremely minimum network for efficient covid-19 segmentation. arXiv preprint arXiv:2004.09750 (2020)
44. Vayá, M.d.l.I., Saborit, J.M., Montell, J.A., Pertusa, A., Bustos, A., Cazorla, M., Galant, J., Barber, X., Orozco-Beltrán, D., Garcia, F., et al.: Bimcv covid-19+: a large annotated dataset of rx and ct images from covid-19 patients. arXiv preprint arXiv:2006.01174 (2020)
45. COVID-19 Image Repository. https://github.com/ml-workgroup/covid-19-image-repository. Accessed 16 September 2020.
46. COVID-19 DATABASE. https://www.sirm.org/category/senza-categoria/covid-19/. Accessed 16 September 2020.
47. Cohen, J.P., Morrison, P., Dao, L.: Covid-19 image data collection. arXiv preprint arXiv:2003.11597 (2020)
48. COVID-19 Radiography Database. https://www.kaggle.com/tawsifurrahman/covid19-radiography-database. Accessed 16 September 2020.
49. Chest Imaging. https://www.eurorad.org/. Accessed 16 September 2020.
50. RSNA Pneumonia Detection Challenge. https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/overview. Accessed 22 September 2020.
51. Bustos A, Pertusa A, Salinas JM, de la Iglesia-Vayá M. Padchest: A large chest x-ray image dataset with multi-label annotated reports. Med Image Anal. 2020;66:101797. https://doi.org/10.1016/j.media.2020.101797.
52. Kermany DS, Goldbaum M, Cai W, Valentim CC, Liang H, Baxter SL, McKeown A, Yang G, Wu X, Yan F, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell. 2018;172(5):1122–31. https://doi.org/10.1016/j.cell.2018.02.010.
53. Demner-Fushman D, Kohli MD, Rosenman MB, Shooshan SE, Rodriguez L, Antani S, Thoma GR, McDonald CJ. Preparing a collection of radiology examinations for distribution and retrieval. J Am Med Inform Assoc. 2016;23(2):304–10. https://doi.org/10.1093/jamia/ocv080.
54. Jaeger S, Candemir S, Antani S, Wáng YXJ, Lu PX, Thoma G. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. Quant Imaging Med Surg. 2014;4(6):475–7. https://doi.org/10.3978/j.issn.2223-4292.2014.11.20.
55. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3462–3471 (2017). https://doi.org/10.1109/CVPR.2017.369

56. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Proceedings of the International Conference on Medical image computing and computer-assisted intervention, vol. 9351, pp. 234–241. Springer (2015)
57. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: Proceedings of the Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pp. 3–11. Springer (2018). https://doi.org/10.1007/978-3-030-00889-5_1
58. Yu, F., Wang, D., Shelhamer, E., Darrell, T.: Deep layer aggregation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2403–2412 (2018). https://doi.org/10.1109/CVPR.2018.00255
59. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2261–2269 (2017). https://doi.org/10.1109/CVPR.2017.243
60. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., et al.: Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225 (2017)
61. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818–2826 (2016). https://doi.org/10.1109/CVPR.2016.308
62. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
63. Lin, T., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2999–3007 (2017). https://doi.org/10.1109/ICCV.2017.324
64. Milletari, F., Navab, N., Ahmadi, S.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571 (2016). https://doi.org/10.1109/3DV.2016.79
65. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., et al.: Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467 (2016)
66. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
67. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), pp. 618–626 (2017). https://doi.org/10.1109/ICCV.2017.74