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

Computers in Biology and Medicine 139 (2021) 105002

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

COVID-19 infection localization and severity grading from chest X-ray images

Anas M. Tahir [a,*], Muhammad E.H. Chowdhury [a,**], Amith Khandakar [a], Tawsifur Rahman [a], Yazan Qiblawey [a], Uzair Khurshid [a], Serkan Kiranyaz [a], Nabil Ibtehaz [b], M. Sohel Rahman [b], Somaya Al-Maadeed [c], Sakib Mahmud [a], Maymouna Ezeddin [a], Khaled Hameed [d], Tahir Hamid [e]

[a] Department of Electrical Engineering, Qatar University, Doha, 2713, Qatar
[b] Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, Dhaka, 1205, Bangladesh
[c] Computer Science and Engineering Department, Qatar University, Doha, 2713, Qatar
[d] Radiology Department, Reem Medical Center, Doha, Qatar
[e] Hamad General Hospital and Weill Cornell Medicine - Qatar, Doha, Qatar

ARTICLE INFO

Keywords:
COVID-19
Chest X-ray
Lung Segmentation
Infection Segmentation
Convolutional Neural Networks
Deep Learning

ABSTRACT

The immense spread of coronavirus disease 2019 (COVID-19) has left healthcare systems incapable to diagnose and test patients at the required rate. Given the effects of COVID-19 on pulmonary tissues, chest radiographic imaging has become a necessity for screening and monitoring the disease. Numerous studies have proposed Deep Learning approaches for the automatic diagnosis of COVID-19. Although these methods achieved outstanding performance in detection, they have used limited chest X-ray (CXR) repositories for evaluation, usually with a few hundred COVID-19 CXR images only. Thus, such data scarcity prevents reliable evaluation of Deep Learning models with the potential of overfitting. In addition, most studies showed no or limited capability in infection localization and severity grading of COVID-19 pneumonia. In this study, we address this urgent need by proposing a systematic and unified approach for lung segmentation and COVID-19 localization with infection quantification from CXR images. To accomplish this, we have constructed the largest benchmark dataset with 33,920 CXR images, including 11,956 COVID-19 samples, where the annotation of ground-truth lung segmentation masks is performed on CXRs by an elegant human-machine collaborative approach. An extensive set of experiments was performed using the *state-of-the-art* segmentation networks, U-Net, U-Net++, and Feature Pyramid Networks (FPN). The developed network, after an iterative process, reached a superior performance for lung region segmentation with Intersection over Union (IoU) of 96.11% and Dice Similarity Coefficient (DSC) of 97.99%. Furthermore, COVID-19 infections of various shapes and types were reliably localized with 83.05% IoU and 88.21% DSC. Finally, the proposed approach has achieved an outstanding COVID-19 detection performance with both sensitivity and specificity values above 99%.

1. Introduction

The novel coronavirus 2019 (COVID-19) is an acute respiratory syndrome that has already caused over 4.9 million causalities and infected more than 243 million people, as of October 27, 2021 [1]. The business, economic, and social dynamics of the whole world have been affected due to this pandemic. Governments have imposed flight restrictions, social distancing, and taken measures to increase awareness of hygiene. Several studies have been done to forecast the future conditions of the virus and to recede its impact [2,3]. However, COVID-19 is still spreading at a very rapid rate. The common symptoms of coronavirus include fever, cough, shortness of breath, and pneumonia [4].

* Corresponding author.
** Corresponding author.
E-mail addresses: a.tahir@qu.edu.qa (A.M. Tahir), mchowdhury@qu.edu.qa (M.E.H. Chowdhury), amitk@qu.edu.qa (A. Khandakar), tawsifurahman.1426@gmail.com (T. Rahman), yq1302932@student.qu.edu.qa (Y. Qiblawey), uk1506741@qu.edu.qa (U. Khurshid), mkiranyaz@qu.edu.qa (S. Kiranyaz), 1017052037@grad.cse.buet.ac.bd (N. Ibtehaz), msrahman@cse.ac.buet.bd (M.S. Rahman), s_alali@qu.edu.qa (S. Al-Maadeed), sm1512633@qu.edu.qa (S. Mahmud), maymouna@qu.edu.qa (M. Ezeddin), dr.khalid@reemmedicalcenter.com (K. Hameed).

https://doi.org/10.1016/j.compbiomed.2021.105002
Received 3 August 2021; Received in revised form 27 October 2021; Accepted 27 October 2021
Available online 30 October 2021

Severe cases of coronavirus disease result in acute respiratory distress syndrome (ARDS) or complete respiratory failure, which requires support from mechanical ventilation and an intensive-care unit (ICU). People with a compromised immune system or elderly people are more likely to develop serious illnesses, including heart and kidney failures and septic shock [4].

Reliable detection of COVID-19 is crucial. However, the diagnosis procedures thereof, particularly through clinical diagnosis, are not straightforward as the common symptoms of COVID-19 are generally indistinguishable from other viral infections [5,6]. Currently, the primary diagnostic tool to detect COVID-19 is reverse-transcription polymerase chain reaction (RT-PCR) arrays, where the presence of Severe Acute Respiratory Syndrome Related Coronavirus 2 (SARS-CoV-2) Ribonucleic acid (RNA) is tested on collected respiratory specimens from the suspected cases [7,8]. However, RT-PCR arrays have a high false alarm rate caused by sample contamination, and damage through the virus mutations in the COVID-19 genome [9,10]. Therefore, several studies have suggested using chest computed tomography (CT) imaging as a primary diagnostic tool since it has shown higher sensitivity compared to RT-PCR [11,12]. Besides, several studies [11–13] have suggested performing CT scans as a secondary test if the suspected patients show shortness of breath or other respiratory symptoms but the RT-PCR result comes negative. Despite the superior performance, CT scans do pose some difficulties and certain limitations. For example, their sensitivity is limited to early COVID-19 cases with no or minimum pneumonia symptoms, the corresponding image acquisition process is slow, and the whole process is costly. On the other hand, X-ray imaging is a cheaper, faster, and readily available method, where the body gets exposed to a much smaller amount of harmful radiation compared to CT [14]. Chest X-ray (CXR) imaging is widely used as an assistive diagnostic tool in COVID-19 screening, and it is reported to have high potential prognostic capabilities [15].

The majority of early COVID-19 cases have exhibited similar features on radiographic images, including bilateral, multi-focal, ground-glass opacities with posterior or peripheral distribution, mainly in the lower lung lobes, while it develops to pulmonary consolidation in the late stage [16,17]. Even though chest radiographs can help in the early screening of the suspected case, the images of several other types of viral pneumonia are similar. They show a high similarity with other inflammatory lung diseases as well. Therefore, it is difficult for medical doctors to distinguish COVID-19 infections from other viral pneumonia using only a chest X-ray. Hence, this symptom similarity can lead to a wrong diagnosis under the current situation, which may cause mistreatment leading to human causalities.

The tremendous development in Deep Learning techniques in recent years has led to many *state-of-the-art* performances in several Computer Vision tasks, such as image classification, object detection, and image segmentation. This breakthrough led to increased utilization of AI-based solutions in various life sciences fields, including the domain of biomedical health problems and complications. Specifically, Convolutional Neural Network (CNN) has been proven extremely beneficial in several biomedical imaging applications, such as skin lesion classification [18], brain tumor detection [19], breast cancer detection [20], and lung pathology screening [21,22]. Deep Learning techniques on chest X-ray images are gaining popularity with the availability of deep CNNs, showing promising results in various applications. Rajpurkar et al. [23] proposed the CheXNet network, one of the top-performing architectures for CXR, by training Densenet121 on the ChestX-ray14 dataset [24], one of the largest public CXR datasets with over 100 thousand X-ray images for 14 different pathologies. Rahman et al. [25] investigated several pre-trained CNNs to classify the CXR images as either healthy or having manifestations of pulmonary tuberculosis (TB). The proposed model was trained over a dataset of 3500 infected and 3500 Normal CXR images. The best performing model, DenseNet201, performed very well achieving 98.57% sensitivity and 98.56% specificity.

1.1. Related works

Recently, many studies have proposed Deep Learning approaches to automate COVID-19 detection from chest X-ray images [26–35] with high performance. Ozturk et al. [26] presented a modified version of DarkNet for binary classification (COVID-19 vs Normal) and multi-class classification (COVID-19 vs Non-COVID pneumonia vs Normal). They reported 90.65% sensitivity for the binary scheme and 85.35% sensitivity for the multi-class scheme on a dataset that includes 114 COVID-19 CXRs. Apostolopoulos et al. [27] evaluated MobileNetV2 on a dataset with 224 COVID-19 cases achieving high discrimination performance with 98.7% sensitivity. Wang et al. [28] introduced COVID-Net, a CNN architecture tailored for COVID-19 recognition. The network exhibited 91% sensitivity over a dataset with 358 COVID-19 CXRs. Waheed et al. [29] proposed a synthetic data augmentation technique to alleviate the scarcity of public COVID-19 CXR data using Auxiliary Classifier Generative Adversarial Network (ACGAN). Chowdhury et al. [30] investigated several deep CNNs (SqueezeNet, ResNet18, ResNet101, MobileNetV2, DenseNet201, and CheXNet) for both binary and multi-class schemes on a dataset that contains 423 COVID-19 CXR images. DenseNet201 showed the best classification performance with 99.7% and 97.9% sensitivity values for binary and multi-class schemes, respectively. Yamac et al. [31] utilized CheXNet as a feature extractor while a proposed classifier, Convolution Support Estimation Network (CSEN), discriminates the target CXR as COVID-19, Bacterial pneumonia, Viral Pneumonia, or Normal. The network produced satisfactory results with 98% sensitivity over the benchmark QaTa-COV19 dataset that includes 462 COVID-19 CXR images. Fan et al. [32] investigated the role of attention mechanism on the COVID-19 recognition scheme by introducing Multi-Kernel-Size Spatial-Channel Attention Network. The proposed network achieved 98.1% sensitivity and 98.3% specificity on a dataset that comprises 500 COVID-19 and 500 Non-COVID CXR images.

Oh et al. [33] proposed a patch-based deep CNN architecture for COVID-19 recognition. First, lung areas were extracted using a fully connected (FC)-DenseNet103 followed by patch-based classification using ResNet50, where a majority voting was utilized to make the final decision. The proposed pipeline achieved 95.5% Intersection over Union (IoU) for the lung segmentation task while it exhibited 96.9% sensitivity for the COVID-19 recognition task. In recent work [34], we investigated the ability of deep networks to distinguish between different Coronavirus family members (COVID-19, MERS-CoV, and SARS-CoV) using CXR images which is an extremely challenging task for medical doctors without the aid of clinical data. A cascaded system was proposed where first lung regions are segmented using U-Net model and then classified using a deep CNN classifier (SqueezeNet, ResNet18, InceptionV3, or DenseNet201). Our proposed pipeline achieved 93.1% IoU and 96.4% Dice Similarity Coefficient (DSC) for the segmentation task while it achieved 96.9% sensitivity for the recognition task. Motamed et al. [35] utilized a semi-supervised learning approach that only requires partial labels for the training data without the need for a single label from the positive class (COVID-19). The lung regions were first segmented using the U-Net model and then feed to the proposed randomized generative adversarial network (RANDGAN) for classification. Poor classification performance was achieved with 57% sensitivity and 80% specificity. Therefore, the introduced pipeline can have significant value in the very early stages of the emergence of a certain disease/pandemic where

annotated data are scarce. However, supervised approaches are still a preferable choice as soon as enough annotated data are created to train the deep CNN models. Despite the high classification performance achieved in most of the recent studies, they also have highlighted certain issues and drawbacks thereof as follows. First of all, all of these studies suffer from the issue of a small dataset, while the largest one has only a few hundred CXR samples. This makes their performance evaluation questionable and it is difficult to generalize their results in practice. Secondly, they only aimed for COVID-19 detection and/or classification among other types without further assessment and localization. These issues limit their usability, particularly in a real clinical setting.

On the other hand, few studies [36,37] considered lung segmentation as the first stage in their detection system. This ensures reliable decision-making in the classification phase and guards the network against irrelevant features from non-lung areas, such as heart, bones, background, or text. However, the previous segmentation approaches were trained on a mixture of medium and high-quality CXR images comprising a total of 704 X-ray images for Normal and TB cases, mainly collected from Montgomery [38] and Shenzhen [39] CXR lung mask datasets. Therefore, the segmentation performance degrades in unseen scenarios such as severe COVID-19 cases or low-quality images with poor signal-to-noise (SNR) levels. The lung areas can be partially or incompletely segmented for severe COVID-19 infections, such as, bilateral consolidation or fluid accumulation at lower-lung lobes, which degrades the classification performance. Therefore, creating a large benchmark CXR dataset with ground-truth lung segmentation masks is extremely important, and will help the research community to provide a more reliable detection system for COVID-19 and other lung pathologies.

Along with COVID-19 detection, infection localization is another crucial task that helps in evaluating the status of the patient and deciding on the treatment plan [40]. Therefore, several studies utilized class activation maps which are generated from Deep Learning models trained for COVID-19 classification tasks to localize infected lung regions. Those localized regions are potential signatures for COVID-19. However, more precise and reliable localization can be provided by ground-truth infection masks from expert radiologists. Therefore, Degerli et al. [41] proposed a novel approach for COVID-19 infection map generation by compiling a COVID-19 dataset consisting of 2951 CXR images with annotated ground-truth infection segmentation masks. Several encode-decoder (E-D) CNNs were trained and evaluated on the generated dataset, where the best performing network achieved an F1-score of 85.81% for infection localization. However, their proposed approach is limited only to COVID-19 infection localization. Therefore, there is certainly room for improvement particularly in the context of both localizing and quantifying infection regions by computing the overall percentage of infected area in the lungs. This can help medical doctors to quantify the severity and track the progression of COVID-19 pneumonia.

With the above backdrop, in this work, we attempt to overcome the aforementioned limitations and challenges. This paper makes the following key contributions:

- We present the largest COVID-19 benchmark dataset, namely, COVID-QU-Ex [65], having 11,956 COVID-19, 11,263 Non-COVID (but diseased), and 10,701 Normal (healthy) CXR images. It is expected that COVID-QU-Ex will be regarded as the most reliable benchmark hitherto available for reliable evaluation for COVID-19 detection, localization, and quantification models, particularly the ones involving *state-of-the-art* deep network architectures.
- We have prepared the ground-truth lung segmentation masks for the entire COVID-QU-Ex dataset applying an elegant human-machine collaborative approach that significantly reduces human labour to annotate the images. This is the first-ever attempt to provide ground-truth lung segmentation masks at such a large scale. Both the dataset and the ground-truth masks will be released along with this study as a public benchmark dataset. We believe that COVID-QU-Ex will be extremely beneficial for researchers, doctors, and engineers around the world to come up with innovative solutions for the early detection of COVID-19 with the help of the large benchmark COVID-19 CXR images with their ground-truth lung masks.
- Furthermore, we have experimented with three *state-of-the-art* image segmentation architectures, namely, U-Net [42], U-Net++ [43], and Feature Pyramid Networks (FPN) [44] with different backbone encoder structures for both lung and infection segmentation tasks thereby identifying which model is better suited for which task. As the backbone encoder, we started with shallow structures and went on to deeper ones thereby covering ResNet18, ResNet50 [45], DenseNet121, DenseNet161 [46], and InceptionV4 [47].
- Finally, we have proposed a novel and robust system for lung segmentation and COVID-19 localization with infection quantification from CXR images. This is a crucial accomplishment for a reliable diagnosis and assessment of the disease with the highest accuracy ever reached.

2. The benchmark COVID-QU-Ex dataset

In this section, we will first show the data compilation process; then, we will present the proposed approach for ground-truth lung mask generation.

2.1. Data compilation

Due to the emerging nature of the pandemic, initially, only limited efforts were being made by the highly infected countries on sharing clinical and radiography data publicly. Therefore, a group of researchers from Qatar University (QU) and Tampere University (TU), created two datasets, COVID-QU [48] and QaTa-Cov19 datasets [41]. The COVID-QU dataset consists of 3616 COVID-19, 8851 Non-COVID cases, and 6012 Normal cases, whereas the QaTa-Cov19 dataset comprises 2951 COVID-19 CXR along with their ground-truth infection masks. Gradually, more X-rays have become publicly available. Hence, we extended those datasets creating COVID-QU-Ex [65], which include over 33,000 CXR images, from three different classes:

1) 11,956 COVID-19 cases
2) 11,263 Non-COVID infections (viral or bacterial pneumonia) cases
3) 10,701 Normal (healthy) cases

In this study, only posterior-to-anterior (PA) or anterior-to-posterior (AP) chest X-rays were considered as this view of radiography is preferred and widely used by the radiologist, whereas a lateral image is usually taken to complement the frontal view. Besides, a very small portion of the compiled dataset were lateral X-rays. Thus, they were excluded from this study [49]. This dataset was created by utilizing numerous publicly available datasets and repositories, all of which are scattered, and with varying formats. The quality of the dataset was ensured through a rigorous quality control process where duplicates, extremely low-quality, and over-exposed images were identified and removed. The resulting dataset thus comprises images of high interclass dissimilarity with few varying resolutions, quality, and SNR levels (See

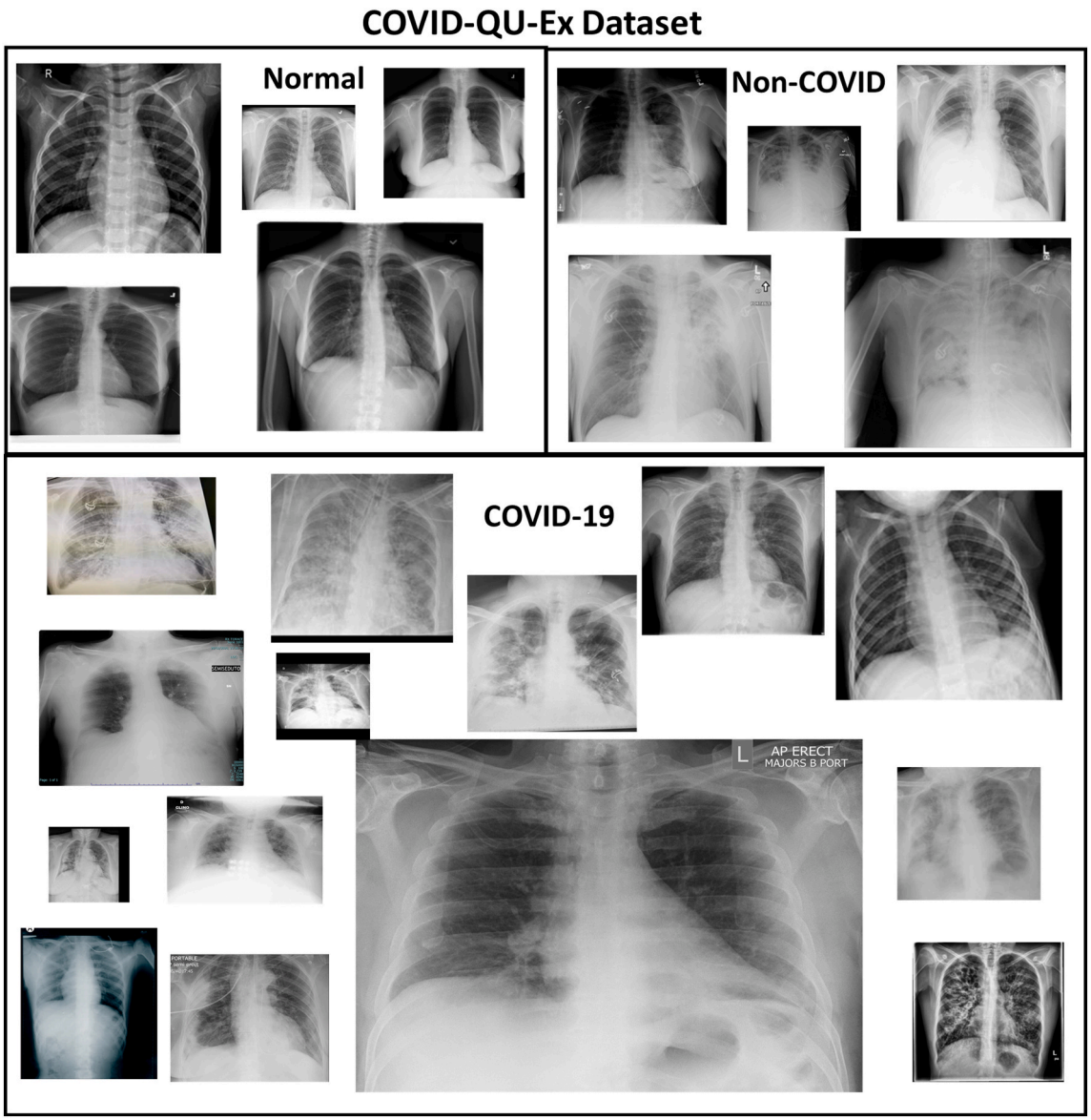

Fig. 1. Sample chest X-ray images from the COVID-QU-Ex dataset for Normal, Non-COVID, and COVID-19 classes. All images are rescaled with the same factor to illustrate the diversity of the dataset.

Fig. 1 for some representative samples).

Details of different data sources are given below:

COVID-19 CXR dataset: This dataset contains 11,956 positive COVID-19 CXR images among which 10,814 images are collected from the BIMCV-COVID19+ dataset [50], 183 CXR images from a German medical school [51], 559 CXR images from SIRM, Github, Kaggle, and Tweeter [52–55], and 400 CXR images from another COVID-19 CXR repository [56].

RSNA CXR dataset (Non-COVID infections and Normal CXR): RSNA pneumonia detection challenge dataset [57] consists of 26,684 CXR images, where 8,851 images are Normal, 11,821 are abnormal, and 6,012 are lung opacity images. All images are in DICOM format. We have included 8,851 Normal and 6,012 lung opacity CXR images from this dataset in our COVID-QU-Ex dataset, where the latter is considered as Non-COVID images.

Chest-Xray-Pneumonia dataset: This is a Kaggle dataset [58] that comprises 1,300 viral pneumonia, 1,700 bacterial pneumonia, and 1,000 Normal CXR images. The viral and bacterial pneumonia images of this dataset are added as Non-COVID (diseased) images in our COVID-QU-Ex dataset.

PadChest dataset: PadChest [59] dataset comprises more than 160,000 CXR images from 67,000 patients that were collected and reported by radiologists at Hospital San Juan (Spain) from 2009 to 2017. We included 4,000 Normal, and 4,000 pneumonia/infiltrate (Non-COVID) cases from this dataset in our COVID-QU-Ex dataset.

Montgomery and Shenzhen CXR lung masks dataset: This dataset consists of 704 CXR images with their corresponding lung segmentation masks. In the first stage of the proposed human-machine collaborative approach, the lung masks from this dataset were used as the initial ground truth masks to train the lung segmentation models. The dataset was acquired by Shenzhen Hospital in China [39], and the tuberculosis control program of the Department of Health and Human Services of

Montgomery County, MD, USA [38]. Montgomery dataset consists of 80 Normal and 58 tuberculosis CXR with lung segmentation masks. On the other hand, the Shenzhen dataset comprises 326 Normal and 336 tuberculosis CXR, where 566 out of 662 CXR are provided with their corresponding masks.

QaTa-Cov19 CXR infection mask dataset [60]**:** This dataset was created by a research group from Qatar University and Tampere University. It consists of nearly 120,000 CXR images, including 2913 COVID-19 images with their corresponding ground-truth infection masks, but no ground-truth lung masks are provided. Thus, these ground-truth infection masks were used to train and evaluate the infection segmentation models.

2.2. Collaborative human-machine segmentation approach for lung ground-truth mask generation

Recent advancements in Deep Learning techniques have brought about remarkable success. However, supervised Deep Learning approaches require large and annotated data for training. Lack of adequate and quality data (including ground truth masks) often degrades the performance of the models, resulting in poor generalization capabilities. On the other hand, the process of producing ground truth segmentation masks is an exhaustive task, where human experts need to delineate pixel-wise masks. This process is bound to suffer from the varying subjectivity and hand-crafting levels of the human annotators. To overcome these issues, here, a *collaborative* human-machine segmentation approach is proposed to accurately produce the ground-truth lung segmentation masks for CXR images. The majority of the manual annotation process was assigned to biomedical engineering researchers from Qatar University (QU) team to reduce the load on medical collaborators from Hamad Medical Corporation (HMC). All researchers attended several training sessions conducted by MDs to grasp a general understating of Chest X-ray imaging and get exposed to a variety of cases with mild, moderate, or severe infections. This human-machine collaborative approach is performed in four main stages as follows.

Stage I (Initial Training):

In the first stage, three variants of the U-Net [42] segmentation model, are trained on 704 CXR images and ground-truth lung masks publicly available from Montgomery and Shenzhen dataset mentioned previously. The ground-truth CXR lung masks are referred to as the CXR-lung-mask-repository in Fig. 2, and it is enlarged throughout the mask creation process. Next, the best performing network in terms of Dice Similarity Coefficient (DSC) is selected as the main network for Stage II, which is referred to as the CXR-Segmentation network in Fig. 2.

Stage II (Collaborative Evaluation):

In the second stage, an iterative training is utilized to create lung masks for a subset of 3000 CXR samples (~10% of the full dataset) that well represent the diversity of the COVID-QU-Ex dataset. Firstly, a subset of 500 samples is selected and inferred using the CXR-Segmentation network. The predicted lung masks are then evaluated by researchers as "accept", "reject", "unsure", or "exclude". Accepted masks that accurately cover the lung areas are added to the CXR-lung-mask-repository. Rejected masks either miss certain parts of the lung or include irrelevant parts. These rejected masks are then manually examined by the researchers, and the corrected masks are finally added to the CXR-lung-mask-repository. The "unsure" masks are the severe cases with highly infected areas. These are usually consolidations or fluid accumulation at the lower lung lobes with a whitish color, which makes them indistinguishable from neighboring organs. The unsure masks are first assessed by MDs; then, researchers adjust the masks based on their recommendations. Finally, the "excluded" masks are the ones where the quality is extremely bad for proper lung segmentation. Eventually, the CXR-Segmentation network is re-trained on the extended mask dataset (extended through the above-mentioned protocol). Then the second subset of 500 samples is selected, and the steps of Stage II are repeated. This process is repeated until generating ground-truth masks for 3000 CXR samples is completed.

Stage III (Collaborative Selection):

In the third stage, six deep segmentation networks from the models of U-Net [42], U-Net++ [43], and FPN [44] are trained using the 3000 ground-truth masks generated in Stage II by the proposed approach. The trained networks are used to predict segmentation masks for the rest of the COVID-QU-Ex dataset, which is 30,920 unannotated samples (~90% of the full dataset). Among the six predictions, researchers selected the best one as the ground truth or discarded the sample for now if none of the masks segments the lung properly. The latter is a minority case that included less than 5% of the unannotated data. The network that registered the highest number of selection (as above) is considered as the best-performing network and used for a new training with the CXR-lung-masks-repository.

The discarded cases are then inferred by the best-performing segmentation network and evaluated manually following the steps in Stage II. As a result, the ground-truth masks for 33,920 CXR images are gathered to construct the benchmark COVID-QU-Ex lung masks dataset.

The proposed systematic collaboration ensured a good compromise between human intervention and machine training throughout the entire process. In Stage II, a smaller subset (~10%) of the dataset was annotated where manual modification was performed by RAs. On the other hand, a larger subset (~90%) of the dataset was annotated in Stage III, where the performance of the segmentation models has been enhanced. Thus, the load was reduced on the RAs, and they had to select among different network predictions rather than manually modifying the predicted masks. This approach saved valuable human labor time. Also, it enhanced the quality and reliability of the generated masks and reduced subjectivity.

Stage IV (Final Verification):

In the final stage, a final verification is performed by two radiologists on randomly selected 6788 CXR samples (20% of the full dataset). To ensure that the diversity of the COVID-QU-Ex dataset is well-captured during this verification, the samples are selected from COVID, Non-COVID, and Normal classes, with different resolution, quality, and SNR levels. Both radiologists accepted >97% of the annotated subset, while the rejected masks were modified by the radiologists then added to the dataset. Considering the noisy nature of the radiographic imaging and the subjectivity in the annotation process it is acceptable to have such a small rejection rate (~3%). Thus, the constructed COVID-QU-Ex dataset can be used as a reliable ground-truth lung segmentation masks dataset. In this study, the verified subset (20%) was considered as a test set for all the experimental evaluations, while the remaining data (80%) were used for training and validation.

3. Methods

In this section, we describe the proposed unified approach for lung segmentation and COVID-19 localization with infection quantification from the CXR images. The schematic representation of the pipeline of

the proposed COVID-19 recognition system is shown in Fig. 3. A binary lung mask is first generated from the input CXR image using the 1st encoder-decoder (E-D) CNN. In parallel, the input CXR is fed to the 2nd E-D CNN to generate COVID-19 infection masks. Then, the generated lung and infection masks are superimposed with the CXR image to localize and quantify COVID-19 infected lung regions. Finally, the generated infection mask is used to detect COVID-19 positive cases from COVID-19 negative cases. In what follows, we will describe these steps in detail.

The pseudo-code for training and evaluating the proposed COVID-19 recognition system is shown in Algorithm 1 and Algorithm 2, respectively.

3.1. Network models for lung and COVID-19 infection segmentation

Lung and COVID-19 pneumonia (infection) segmentation were performed on CXR images using three *state-of-the-art* deep E-D CNNs: U-Net [42], U-Net++ [43], and FPN [44], with different backbone (encoder) models using the variants of ResNet [45], DenseNet [46], and InceptionV4 [47] networks. Five variants of the backbone models were considered starting from shallow to deep structures: ResNet18, ResNet50, DenseNet121, DenseNet161, and InceptionV4.

The deployed encoder-decoder blocks provide a firm segmentation model that captures the context in the contracting path and empowers precise localization by the expanding path. The U-Net architecture has a classical decoder part that is symmetric to the encoder part, where max-pooling operations are replaced with up-sampling operations. Besides,

Algorithm Pseudo-Code

Algorithm 1: Train model

$X_{CXR,train}$ <- CXR images
$Y_{lung,train}$ <- Binary lung masks
$Y_{infc,train}$ <- binary infection masks
$Model_{EDCNN1}$ <- Lung segmentation model
$Model_{EDCNN2}$ <- Infection segmentation model
For all training batch <bx,by> in < $X_{CXR,train}$, $Y_{lung,train}$> do
    $Model_{EDCNN1}$.train_model(<bx,by>, Adam, Binary Cross Entropy)
end
For all training batch <bx,by> in < $X_{CXR,train}$, $Y_{linfc,train}$> do
    $Model_{EDCNN2}$.train_model(<bx,by>, Adam, Binary Cross Entropy)
end
Return $Model_{EDCNN1}$, $Model_{EDCNN2}$

Algorithm 2: Evaluate model

$X_{CXR,test}$ <- CXR image
$Model_{EDCNN1}$ <- Lung segmentation model
$Model_{EDCNN2}$ <- Infection segmentation model
$Y_{lung,test}$ = $Model_{EDCNN1}$.predict($X_{CXR,test}$)
$Y_{infc,test}$ = $Model_{EDCNN2}$.predict($X_{CXR,test}$)
For all pixel px in $Y_{lung,test}$ do
  If $Y_{lungtest}$[px] >0.5
    $Y_{lung,test}$[px] = 1
  Else
    $Y_{lung\text{-}test}$ [px] = 0
  End
End
For all pixel px in $Y_{infc,test}$ do
  If $Y_{infc,test}$ [px] >0.5
    $Y_{infc,test}$ [px] = 1
  Else
    $Y_{infc,test}$ [px] = 0
  End
End
p <-percentage of infected lungs per CXR image.
d <- detection of COVID-19 positive
p = sum($Y_{infc,test}$) /sum($Y_{lung,test}$) x 100
If p > 0 do
  d = 1
Else
  d = 0
End
Return p,d

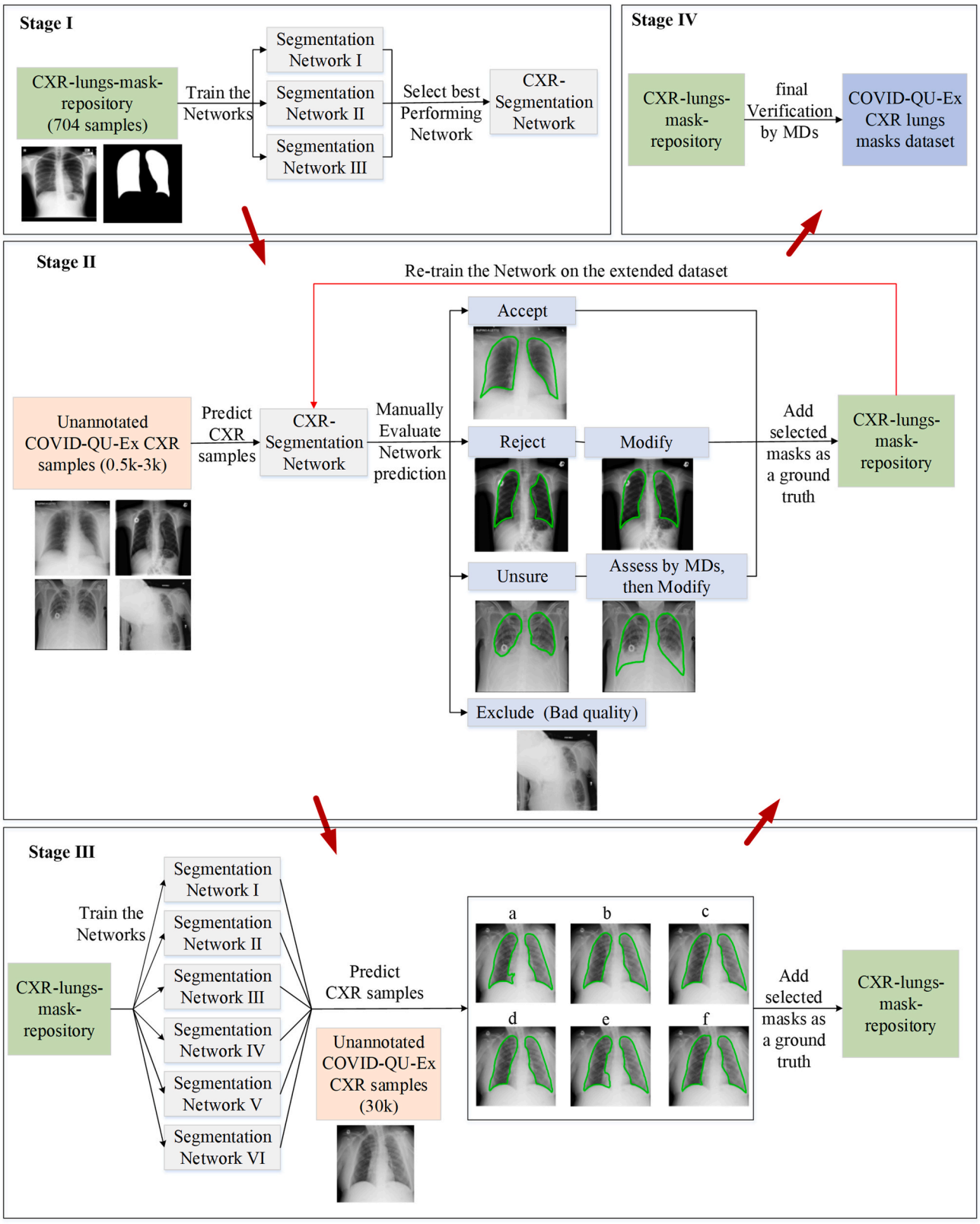

Fig. 2. Collaborative human-machine approach to create ground-truth lung segmentation masks for COVID-QU-Ex CXR dataset. Stage I: Three segmentation networks are trained on a repository of 704 CXR lung segmentation masks, and the best network in terms of DSC is selected for the subsequent stages. Stage II: An iterative training is utilized to create lung masks for a subset of 3000 CXR samples from the COVID-QU-Ex dataset. Firstly, A subset of 500 samples is inferred by the CXR segmentation model and the outputs are evaluated manually as accept, reject, modify, or exclude. Next, the modified masks are added to the lung repository and the network is re-trained on the extended dataset. These steps are repeated until generating ground-truth masks for the 3000 CXR samples is completed. Stage III: six deep segmentation networks are trained using the 3000 ground-truth masks generated in the previous stage. The trained networks are used to predict segmentation masks for the rest of the COVID-QU-Ex dataset (30,920 images). Stage IV: a final verification is performed by MDs on randomly selected 6788 CXR samples (20% of the full dataset) that well presents the diversity of the COVID-QU-Ex dataset.

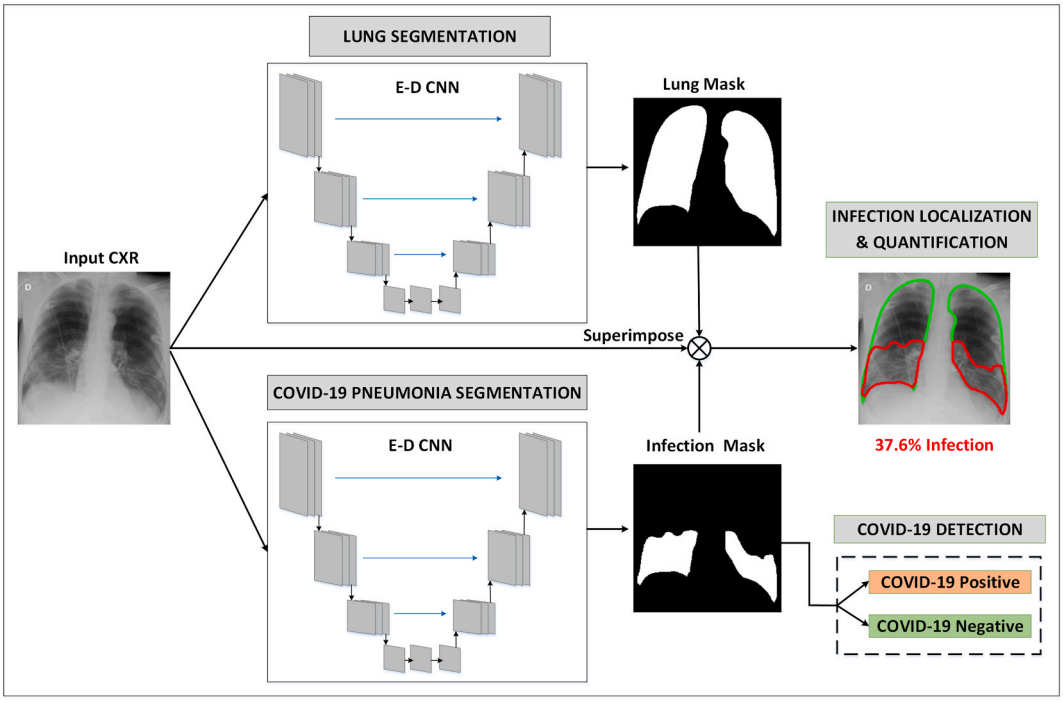

Fig. 3. Schematic representation of the pipeline of the proposed system. The input CXR image is fed to two ED-CNNs in parallel, to generate two binary masks: lung, and COVID-19 infection masks. Next, the generated masks are superimposed with the CXR image to localize and quantify COVID-19 infected lung regions. Finally, the generated infection mask is used to detect COVID-19 positive cases from COVID-19 negative cases.

high-resolution features from the encoder path are merged with the up-sampled output from the corresponding decoder path through skip connection. On the other hand, the U-Net++ is a recent implementation that has further developed the decoder block. The encoder and decoder blocks are connected through a series of nested dense convolutional blocks. This ensures a firm bridge between the encoder and decoder parts of the network, where information can be transferred to the final layers more intensively compared to the conventional U-Net. Both U-Net and U-Net++ architectures utilize 1×1 convolution to map the output from the last decoding block to two-channel feature maps, where a pixel-wise SoftMax activation function is applied to map each pixel into a binary class of background or lung for Lung segmentation task, and background or lesion for infection segmentation task. In contrast, the FPN employs the encoder-decoder as a pyramidal hierarchy by generating prediction masks at each spatial level of the decoder path. All predicted feature maps are up-sampled to the same size, concatenated, convolved with a 3×3 convolutional filter, and then SoftMax activation is applied to generate the final prediction mask.

To ensure efficient training and faster convergence, transfer learning was leveraged on the encoder side of the segmentation networks by initializing the convolutional layers with ImageNet [61] weights.

3.1.1. Segmentation loss function

The cross-entropy (CE) loss is used as the cost function for the segmentation networks:

$$CE = -\frac{1}{K}\sum_{k}\sum_{c} y_k \log(p(x_k)) \tag{1}$$

Here, x_k denotes the kth pixel in the predicted segmentation mask, $p(x_k)$ denotes its SoftMax probability, y_k is a binary random variable getting 1 if $y_k = c$, otherwise 0, and c denotes the class category, i.e., $c \in \{background, lung\}$ for the lung segmentation task, and $c \in \{background, lesion\}$ for the infection segmentation.

3.2. Post-processing

The predicted segmentation masks, $\widehat{Y}$, by the segmentation models are defined as $\widehat{Y}_{h,w} \in [0,1]$, where h and w represent the size of the image. In the post-processing step, binary segmentation masks are first generated by thresholding with a fixed value of 0.5. The predicted pixels are classified as lung if $\widehat{y} > 0.5$ for the lung segmentation task, while classified as COVID-19 infection if $\widehat{y} > 0.5$ for the infection segmentation task. The binary lung masks are further processed by hole filling and removal of small regions, <5% of the total positive predicted pixels. As a result, we increase the true-positives while minimizing the false-positives, i.e., non-lung regions that are falsely predicted as a lung. In contrast, infection masks are masked with post-processed lung masks to ensure that the infection region falls within the lung area and remove the false positives outside the lung region.

3.3. COVID-19 detection and quantification

The detection of COVID-19 is performed based on the prediction maps generated by the infection segmentation network. Accordingly, a CXR image is classified as COVID-19 positive if at least one pixel of lung areas is predicted as COVID-19 infection, i.e., $p(x\_k) > 0.5$. Otherwise, the image is considered as COVID-19 negative, i.e., it could be an image of a healthy person or a patient with Non-COVID pneumonia. Furthermore, COVID-19 infection is quantified by computing the overall percentage of infected lungs by dividing the sum of predicted infection pixels over the sum of predicted lung pixels. In addition, the infection percentage of each lung is computed in a similar manner, enabling doctors to assess the progression of COVID-19 for each lung individually.

3.4. Experimental setup

The lung segmentation task was conducted over the COVID-QU-Ex dataset. In contrast, the infection segmentation and COVID-19 detection tasks were conducted over a subset of the COVID-QU-Ex dataset

comprising 2913 CXR samples with corresponding infection masks from the QaTa-Cov19 dataset [60]. The CXR images were resized to have a fixed dimension of 256 × 256 pixels to be used as the input for the deep networks. In all our experiments, we assumed an 80-20 split for train and test purposes respectively. Besides, 20% of training data was used as a validation set for model selection and to avoid overfitting. Table 1 summarizes the number of images per class used for training, validation, and testing.

Adam optimizer was used, with the initial learning rate, $\alpha = 10^{-4}$, momentum updates, $\beta_1 = 0.9$ and $\beta_2 = 0.999$, an adaptive learning rate that decreases the learning parameter by a factor of 5 if validation loss did not improve for 3 consecutive epochs, early stopping criterion of 8 epochs, where training stops if validation loss did not improve for 8 consecutive epochs, and mini-batch size of 4 images with 40 back-propagation epochs.

3.5. Evaluation metrics

We evaluate our approach as follows. The segmentation tasks are evaluated at the pixel level, where the foreground (lung or infected region) is considered as the positive class and the background as the negative class. For the COVID-19 detection task, the performance metric is computed per CXR sample, where X-rays with COVID-19 infection are considered as the positive class and X-rays of healthy people or patients with Non-COVID pneumonia are considered as the negative class.

The performance of deep CNNs is assessed using different evaluation metrics with a 95% confidence interval (CI). Notably, the CI (r) for each evaluation metric is computed as follows:

$$r = z\sqrt{metric(1 - metric)/N} \tag{2}$$

Here, N is the number of test samples, and z is the level of significance that is 1.96 for 95% CI.

3.5.1. Segmentation evaluation metrics

The performance of the lung and lesion segmentation networks is evaluated using three evaluation metrics, namely, Accuracy, Intersection over Union (IoU), and Dice Similarity Coefficient (DSC) as per the following equations.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

Here, *accuracy* is the ratio of the correctly classified pixels among the image pixels. *TP, TN, FP, FN* represent the true positive, true negative, false positive, and false negative, respectively.

$$Intersection\ over\ Union\ (IoU) = \frac{TP}{TP + FP + FN} \tag{4}$$

$$Dice\ Similarity\ Coefficient\ (DSC) = \frac{2TP}{2TP + FP + FN} \tag{5}$$

Here, both *IoU and DSC* are statistical measures of spatial overlap between the binary ground-truth and the predicted segmentation masks, where the main difference is that *the latter* considers double weight for *TP* pixels (true lung/lesion predictions) compared to *the former*.

3.5.2. COVID-19 detection evaluation metrics

The performance of the COVID-19 detection scheme is assessed using five evaluation metrics, namely, Accuracy, Precision, Sensitivity, F1-score, and Specificity as per the following equations.

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

Here, *precision* is the rate of correctly classified positive class CXR samples among all the samples classified as positive samples.

$$Sensitivity = \frac{TP}{TP + FN} \tag{7}$$

Here, *sensitivity* is the rate of correctly predicted positive samples from among the positive class samples.

$$F1 = 2\frac{Precision \times Sensitivity}{Precision + Sensitivity} \tag{8}$$

Here, $F1$ (i.e., F1-score) is the harmonic average of precision and sensitivity.

$$Specificity = \frac{TN}{TN + FP} \tag{9}$$

Here, *specificity* is the *sensitivity* of the negative class samples.

PyTorch [62] library with Python 3.7 was used to train and evaluate the deep CNN networks, running on a PC with Intel® Core™ i9-9900K CPU at 3.6 GHz, with 32 GB RAM, and with an 8-GB NVIDIA GeForce GTX 1080 GPU card.

4. Results

In this section, both quantitative and qualitative results are reported with an extensive set of comparative evaluations for lung segmentation, infection segmentation, and COVID-19 detection tasks.

4.1. Lung segmentation results

The performance of the lung segmentation models over the test (unseen) set is tabulated in Table 2. Recall that, each model was evaluated with five different encoder structures. For all models, it was observed that DenseNet encoders exhibit the top segmentation performance as they can share pieces of collective knowledge by densely connecting convolutional layers to their subsequent layers, thereby preserving the information coming from the earlier layer through the output layer. The FPN model with DenseNet121 encoder holds the leading position with 96.11% IoU, and 97.99% DSC.

The outputs of the three top-performing networks compared with the ground-truth are shown in Fig. 4. An interesting observation is that the three networks can reliably segment lung regions not only for COVID-19 cases, but for Non-COVID-19 pneumonia as well with different severity levels, i.e., mild, moderate, or severe. This elegant performance may be attributed to the large and diverse COVID-QU-Ex dataset (33,920

Table 1
Number of mages per class per train, validation, and test sets for each of the 5 folds used for lung segmentation, infection segmentation, and COVID-19 detection tasks.

Dataset Name	Task	Class	# of Samples	Training Samples	Validation Samples	Test Samples
COVID-QU-Ex dataset	Lung Segmentation	COVID-19	11,956	7658	1903	2395
		Non-COVID	11,263	7208	1802	2253
		Normal	10,701	6849	1712	2140
		Total	**33,920**	**21,715**	**5417**	**6788**
COVID-QU-Ex and QaTa-Cov19 [60] datasets	Infection Segmentation and COVID-19 Detection	COVID-19 positive	2913	1864	466	583
		COVID-19 negative	1457	932	233	292
		Non-COVID Normal	1456	932	233	291
		Total	**5826**	**3728**	**932**	**1166**

Table 2
Performance metrics (%) for lung region and COVID-19 infected region segmentation computed over test (unseen) set with three network models and five encoder architectures. x ± y means that the achieved metric value is x with standard deviation y.

Task	Model	Encoder	Accuracy	IoU	DSC
Lung Segmentation	U-Net	ResNet18	99.07 ± 0.23	95.91 ± 0.47	97.88 ± 0.34
		ResNet50	99.08 ± 0.23	95.93 ± 0.47	97.89 ± 0.34
		DenseNet121	**99.1 ± 0.22**	**96.06 ± 0.46**	**97.96 ± 0.34**
		DenseNet161	99.1 ± 0.22	96.02 ± 0.47	97.94 ± 0.34
		InceptionV4	99.07 ± 0.23	95.9 ± 0.47	97.88 ± 0.34
	U-Net ++	ResNet18	99.07 ± 0.23	95.9 ± 0.47	97.88 ± 0.34
		ResNet50	99.1 ± 0.22	96.04 ± 0.46	97.95 ± 0.34
		DenseNet121	**99.11 ± 0.22**	**96.1 ± 0.46**	**97.98 ± 0.33**
		DenseNet161	99.09 ± 0.23	95.98 ± 0.47	97.92 ± 0.34
		InceptionV4	99.08 ± 0.23	95.96 ± 0.47	97.91 ± 0.34
	FPN	ResNet18	99.06 ± 0.23	95.86 ± 0.47	97.86 ± 0.34
		ResNet50	99.07 ± 0.23	95.91 ± 0.47	97.88 ± 0.34
		DenseNet121	**99.12 ± 0.22**	**96.11 ± 0.46**	**97.99 ± 0.33**
		DenseNet161	99.09 ± 0.23	96.01 ± 0.47	97.94 ± 0.34
		InceptionV4	99.07 ± 0.23	95.92 ± 0.47	97.89 ± 0.34
Infection Segmentation	U-Net	**ResNet18**	**98.02 ± 0.8**	**82.92 ± 2.16**	**88.1 ± 1.86**
		ResNet50	97.84 ± 0.83	81.73 ± 2.22	87.02 ± 1.93
		DenseNet121	97.98 ± 0.81	82.53 ± 2.18	87.74 ± 1.88
		DenseNet161	97.86 ± 0.83	81.95 ± 2.21	87.19 ± 1.92
		InceptionV4	97.98 ± 0.81	82.03 ± 2.2	87.11 ± 1.92
	U-Net ++	ResNet18	97.9 ± 0.82	82.9 ± 2.16	88.06 ± 1.86
		ResNet50	97.93 ± 0.82	82.59 ± 2.18	87.78 ± 1.88
		DenseNet121	**97.97 ± 0.81**	**83.05 ± 2.15**	**88.21 ± 1.85**
		DenseNet161	97.95 ± 0.81	81.55 ± 2.23	86.66 ± 1.95
		InceptionV4	97.9 ± 0.82	81.13 ± 2.25	86.22 ± 1.98
	FPN	ResNet18	97.84 ± 0.83	81.9 ± 2.21	87.25 ± 1.91
		ResNet50	97.84 ± 0.83	80.83 ± 2.26	86.25 ± 1.98
		DenseNet121	97.99 ± 0.81	82.55 ± 2.18	87.71 ± 1.88
		DenseNet161	97.95 ± 0.81	81.89 ± 2.21	87.08 ± 1.93
		InceptionV4	**97.99 ± 0.81**	**83.08 ± 2.15**	**88.13 ± 1.86**

samples) comprising CXR samples with different quality, resolution, and SNR levels from COVID-19, Non-COVID-19, and Normal classes. Thus, our benchmark dataset is expected to help researchers to overcome the challenges and limitations faced, mainly in the lung segmentation phase for COVID-19 or other lung pathology problems. As most of the previous approaches were trained over Montgomery [38] and Shenzhen [39] CXR lung mask datasets that comprise medium and high-quality X-ray images from Normal and TB classes, the previous segmentation approaches were falling in unseen scenarios, such as, severe infection or low-quality images [37].

4.2. Infection segmentation results

The infection segmentation model has been first evaluated over two different configurations: cascaded and parallel segmentation. For the cascaded scheme, the lung region was first segmented using the lung segmentation model; then the segmented CXR was fed to the infection segmentation model whereas the plain CXR was fed to both models independently for the parallel scheme.

FPN model with DenseNet161 encoder was trained and evaluated on both schemes. The parallel scheme showed slightly better results with 87.08% DSC compared to 86.84% DSC for the cascaded scheme. Therefore, the parallel scheme was used as the main configuration for the remaining experiments. The performance of the infection segmentation models is presented in Table 2. U-Net++ model with DenseNet121 encoder showed the best performance with IoU and DSC values of 83.05% and 88.21%, respectively. Besides, the InceptionV4 encoder showed the best performance among FPN models with 83.08% IoU and 88.13% DSC. In contrast, the shallowest encoder, ResNet18 did better among U-Net models with IoU and DSC values of 82.92% and 88.1%, respectively.

Fig. 5(a) shows the robustness of three top-performing networks to reliably segment COVID-19 infections of various shapes (small, medium, or large) with different severity levels (mild, moderate, severe, or critical). In general, the FPN models produced smoother masks with better localization of infected regions compared to U-Net and U-Net++ models. This can be inspired by the hierarchy architecture of FPN where predictions are made on each spatial level of the decoder path, then merged to produce the final prediction mask, whereas only the final decoder block is used to generate the prediction mask in U-Net and U-Net ++ models. Fig. 5(b) shows infection localization and severity grading of COVID-19 pneumonia for a 42-year female patient on the 1st day (of hospital admission), 2nd day, and 3rd day using the proposed COVID-19 recognition system, where two parallel FPN with DenseNet121 encoders models were used for the lung and the infection segmentation tasks.

4.3. COVID-19 detection results

The performance of infection segmentation networks for COVID-19 detection from the CXR images is presented in Table 3. The sensitivity was considered as the primary metric for the detection task, as missing any COVID-19 positive case is critical. All the networks achieved high sensitivity values (>97%), where U-Net with DenseNet121 backbone and FPN with ResNet18 backbone achieved the best performance with a sensitivity of 99.66%. Similarly, all models showed high specificity values (>97%), where U-Net++ with ResNet18 backbone exhibited the best performance with 100% specificity, indicating the absence of any false alarm.

4.4. Computational complexity analysis

Table 4 compares the segmentation models in terms of inference time and the number of trainable parameters. The results present the inference time per CXR sample. It can be noticed that, due to their shallow and close structures, FPN and U-Net models are faster than U-Net ++ models. FPN with ResNet18 encoder is the fastest network taking up to 5.74 ms per image. In contrast, the U-Net++ model is the slowest with the highest number of trainable parameters. The most computationally demanding model is UNet++ with InceptionV4 encoder having a

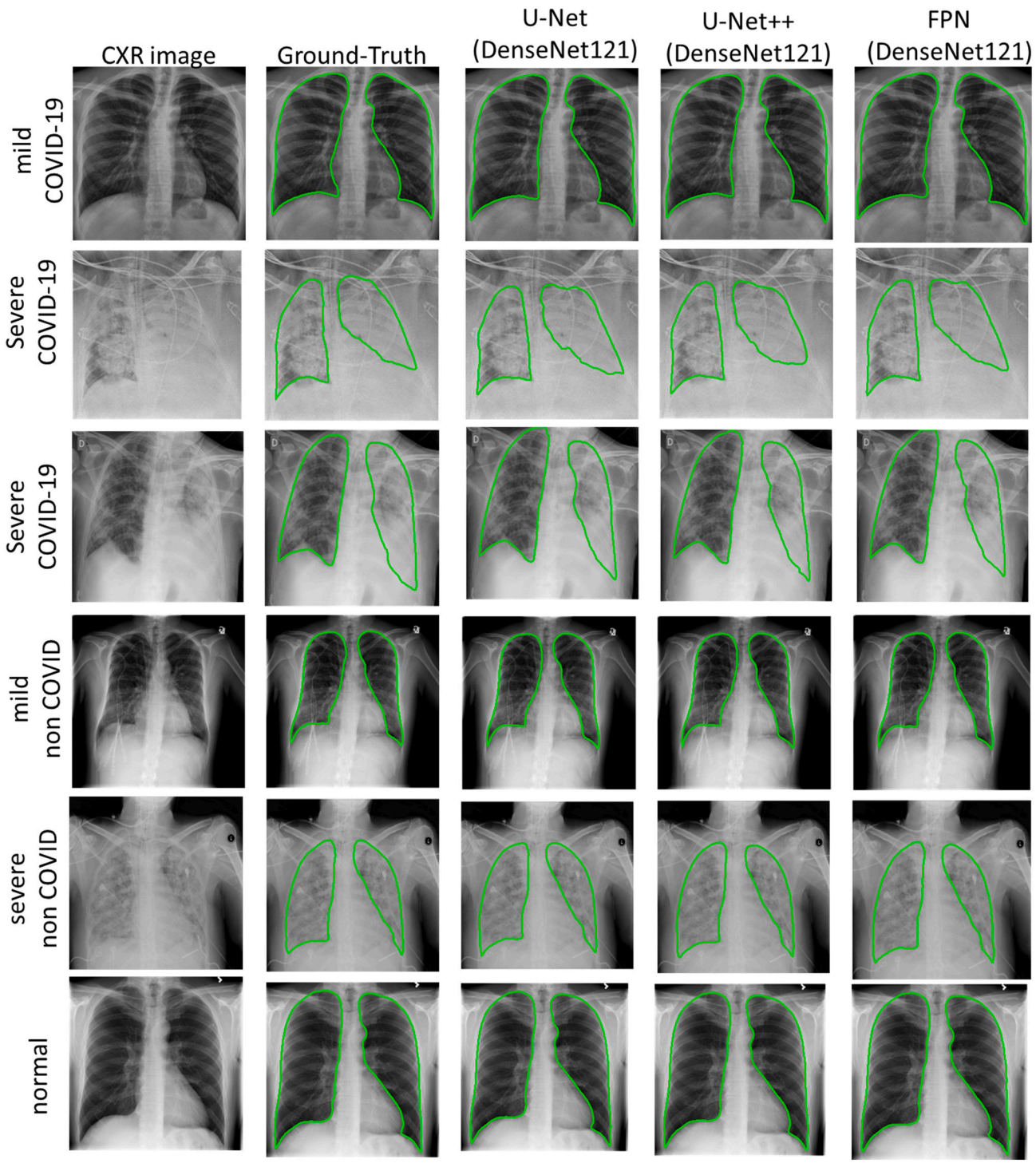

Fig. 4. Sample qualitative evaluation of generated lung masks by the three top-performing networks. Column 1 shows the CXR image, Column 2 shows ground truths, and the lung masks of the top three networks are shown in Columns 3–5, respectively.

staggering 59.35 M trainable parameters. However, UNet++ with DenseNet161 encoder is the slowest, with an inference time of 48.62 ms as it is the deepest model with 161 layers. Note that, for systems with limited computational capabilities, where both lung and infection segmentation cannot be used in parallel, the two models can be used consecutively. This will double (×2) the inference time. However, we can still say that the full system can be used for real-time clinical applications as the overall inference time is still less than 100 ms in the worst case, which means that multiple images can be processed within a second.

4.5. Comparison with related work

Table 5 compares the proposed work with recent literature about automatic COVID-19 pneumonia diagnosis from CXR images for three main tasks: classification, localization, and quantification. First, despite the superior classification performance achieved in most of the studies, small datasets have been used, with few hundred samples only, except [41] where they used 2951 COVID-19 CXRs. On the other hand, we evaluated our pipeline on four times larger cohort datasets with 11,956 COVID-19 CXRs, where >97% sensitivity and specificity values were

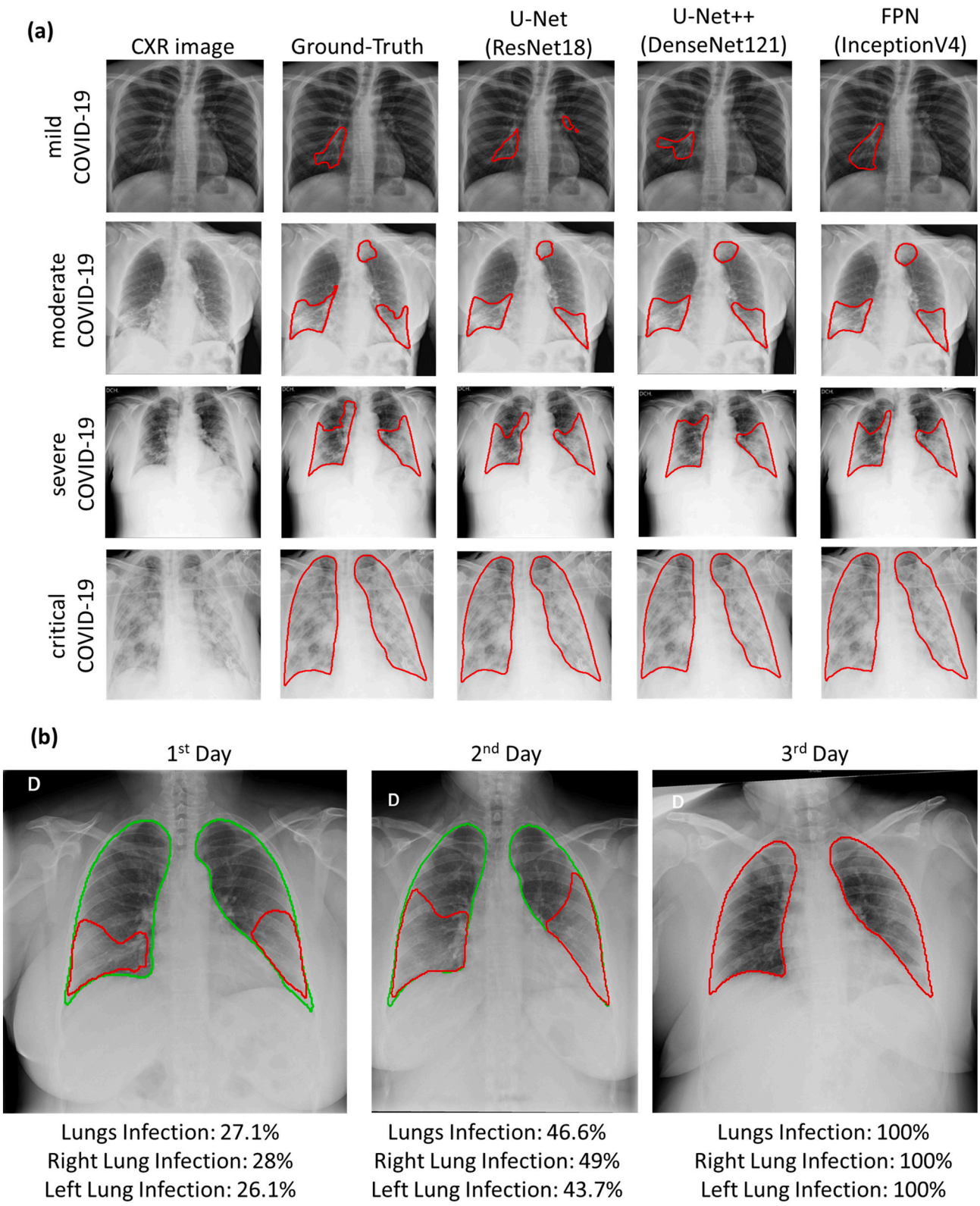

Fig. 5. **(a)** Sample qualitative evaluation of generated infection masks by the three top-performing networks. Column 1 shows the CXR image, Column 2 shows ground truths, and the lung masks of the top three networks are shown in Columns 3–5, respectively. **(b)** Infection localization and severity grading of COVID-19 pneumonia for a 42-year female patient on the 1st, 2nd, and 3rd days of admission using the proposed system.

achieved. This elegant performance is exhibited by the high diversity in the COVID-QU-Ex dataset which ensured good generalization capabilities by the deep CNN models. In addition, we provided a robust lung segmentation model which guards the detection and localization schemes against irrelevant features from non-lung areas. Therefore, empowered by the largest ever ground-truth lung segmentation mask dataset (33,920 samples), an outstanding performance was achieved with 97.9% DSC. Finally, only a single study [41] provided precise and

Table 3
COVID-19 detection performance results (%) computed over test (unseen) set with three network models, and five encoder architectures. x ± y means that the achieved metric value is x with standard deviation y.

Model	Encoder	Accuracy	Precision	Sensitivity	F1-score	Specificity
U-Net	ResNet18	98.89 ± 0.6	99.14 ± 0.53	98.63 ± 0.67	98.88 ± 0.6	99.14 ± 0.53
	ResNet50	98.89 ± 0.6	98.47 ± 0.7	99.31 ± 0.48	98.89 ± 0.6	98.46 ± 0.71
	DenseNet121	98.8 ± 0.62	97.98 ± 0.81	**99.66 ± 0.33**	98.81 ± 0.62	97.94 ± 0.82
	DenseNet161	98.71 ± 0.65	97.97 ± 0.81	99.49 ± 0.41	98.72 ± 0.65	97.94 ± 0.82
	InceptionV4	98.03 ± 0.8	98.28 ± 0.75	97.77 ± 0.85	98.02 ± 0.8	98.28 ± 0.75
U-Net ++	ResNet18	99.23 ± 0.5	100 ± 0	98.46 ± 0.71	99.22 ± 0.5	**100 ± 0**
	ResNet50	99.14 ± 0.53	99.83 ± 0.24	98.46 ± 0.71	99.14 ± 0.53	99.83 ± 0.24
	DenseNet121	99.23 ± 0.5	99.14 ± 0.53	**99.31 ± 0.48**	99.22 ± 0.5	99.14 ± 0.53
	DenseNet161	98.2 ± 0.76	97.95 ± 0.81	98.46 ± 0.71	98.2 ± 0.76	97.94 ± 0.82
	InceptionV4	98.2 ± 0.76	98.45 ± 0.71	97.94 ± 0.82	98.19 ± 0.77	98.46 ± 0.71
FPN	ResNet18	98.54 ± 0.69	97.48 ± 0.9	**99.66 ± 0.33**	98.56 ± 0.68	97.43 ± 0.91
	ResNet50	98.46 ± 0.71	98.46 ± 0.71	98.46 ± 0.71	98.46 ± 0.71	98.46 ± 0.71
	DenseNet121	98.97 ± 0.58	99.65 ± 0.34	98.28 ± 0.75	98.96 ± 0.58	99.66 ± 0.33
	DenseNet161	98.11 ± 0.78	97.3 ± 0.93	98.97 ± 0.58	98.13 ± 0.78	97.26 ± 0.94
	InceptionV4	99.23 ± 0.5	99.31 ± 0.48	99.14 ± 0.53	99.22 ± 0.5	99.31 ± 0.48

Table 4
The number of trainable parameters of the models with their inference time (ms) per CXR sample.

Model	Encoder	Trainable parameters	Inference Time (ms)
U-Net	ResNet18	14.32 M	5.78
	ResNet50	32.5 M	10.44
	DenseNet121	13.60 M	22.86
	DenseNet161	38.73 M	29.74
	InceptionV4	48.79 M	26.53
U-Net ++	ResNet18	15.96 M	8.30
	ResNet50	48.97 M	19.90
	DenseNet121	30.06 M	25.13
	DenseNet161	79.04 M	48.62
	InceptionV4	59.35 M	32.53
FPN	ResNet18	13.04 M	5.74
	ResNet50	26.11 M	10.34
	DenseNet121	9.29 M	22.68
	DenseNet161	29.49 M	29.62
	InceptionV4	43.57 M	26.08

reliable localization of COVID-19 infected lung regions based on ground-truth annotation from medical experts, where the proposed model achieved 83.2% DSC for localizing infected regions. In contrast, our model showed higher localization performance with 88.1% DSC. Moreover, our deployment of lung and infection segmentation models enabled both localization and quantification of infected regions. Therefore, our system could facilitate early intervention and provide a unified solution that helps doctors to access the severity and track the progression of the disease.

5. Conclusion

Early identification and isolation of highly infectious COVID-19 cases play a vital role in treatment as well as preventing the spread of the virus. X-ray imaging is a low-cost, easily accessible, and fast method that can be an excellent alternative for conventional diagnostic methods such as RT-PCR and CT scans. Therefore, numerous studies proposed AI-based solutions for automatic and real-time detection of COVID-19. In general, these methods showed outstanding performance for early detection and diagnosis. However, they have used limited CXR repositories for evaluation with a small number, a few hundreds, of COVID-19 samples. Thus, the generalization of the achieved results on a large cohort dataset is not guaranteed. In addition, they showed limited performance in infection localization and severity grading of COVID-19 pneumonia. In this study, we proposed a robust and comprehensive system to segment the lung, detect, localize, and quantify COVID-19 infections from the CXR images. To accomplish this, we compiled the largest CXR dataset hitherto known, namely, COVID-QU-Ex [65], which consists of 11,956 COVID-19, 11,263 Non-COVID pneumonia, and 10,701 Normal CXR images. Moreover, we constructed ground-truth lung segmentation masks for the benchmark dataset using an elegant collaborative human-machine approach, which saved valuable human labour time and minimized subjectivity in the annotation process. The publicly shared dataset will help researchers to investigate deep CNN models on a comparatively larger dataset, which can provide more reliable solutions for COVID-19 and other lung pathology problems. Extensive experiments on COVID-QU-Ex showed superior lung segmentation performance with 96.11% IoU and 97.99% DSC. Moreover, the proposed system proved reliable in localizing COVID-19 infection of various severity, achieving IoU and DSC values of 83.05% and 88.21%, respectively. Furthermore, unprecedented COVID-19 detection performance was achieved with sensitivity and specificity values $> 99\%$. To the best of our knowledge, this is the first study that utilizes both lung and infection segmentation to detect, localize and quantify COVID-19 infection from X-ray images. Therefore, it can assist the medical doctors to better diagnose the severity of COVID-19 pneumonia and follow up the progression of the disease easily.

In the future, we plan to explore robust quantization and model compression techniques to further reduce the model complexity and accelerate the inference process, using the new generation of heterogeneous network models such as Self-Organized Operational Neural Networks [63,64].

Data availability

The COVID-QU-Ex chest X-ray datasets and corresponding lung mask created during the current study are available in the following Kaggle repository: www.kaggle.com/dataset/cf77495622971312010dd5934ee91f07ccbcfdea8e2f7778977ea8485c1914df.

Author contributions

Experiments were designed by AMT, MEHC, and SK. Experiments were performed by AMT, AK, TR, YQ, and UK. Data were compiled and created by AMT, AK, TR, YQ, UK, NI, SM, ME, KH, and TH. Results were analyzed by AMT, MEHC, SK, MSR, SAM, KH, and TH. The project is supervised by MEHC and SK. All the authors were involved in the interpretation of data and paper writing and revision of the article.

Funding

Qatar University COVID19 Emergency Response Grant (QUERG-CENG-2020-1) from Qatar University, and UREP28-144-3-046 grant from Qatar National Research Fund provided the support for the work and the claims made herein are solely the responsibility of the authors.

Table 5
Comparing the proposed work with recent literature about automatic COVID-19 diagnosis using CXR images, in terms of utilized Dataset, whether Lung and/or Infection Segmentation models are used, deployed Networks, and achieved Results.

Ref.	Dataset (# of subjects)	Lung Seg.	Infection Seg.	Network	Results
[27]	COVID19 (224) Non-COVID (714) Healthy (504)			Classification model MobileNetV2	Class. Sens. 98.7% Spe. 96.5%
[28]	COVID19 (358) Non-COVID (8,066) Healthy (5,538)			Classification model COVID-Net	Class. Sens. 91.0% Spe. 99.5%
[29]	COVID-19 (403) Healthy (721)			Augmentation model CovidGAN Classification model VGG16	Class. Sens. 90.0% Spe. 97.0%
[30]	COVID-19 (423) Non-COVID (1,485) Healthy (1,579)			Classification model DenseNet201	Class. Sens. 99.7% Spe. 99.6%
[31]	COVID-19 (462) Non-COVID (2,485) Healthy (1,579)			Feature extraction model CheXNet Classification model Convolutional Support Estimation Network	Class. Sens. 98.5% Spe. 94.7%
[32]	COVID-19 (500) Non-COVID (500)			Classification model Multi-Kernel-Size Spatial-Channel Attention Network	Class. Sens. 98.1% Spe. 98.3%
[33]	COVID-19 (180) Non-COVID (74) TB (57) Healthy (191)	✓		Segmentation model FC-DenseNet103 Patch-based classification model ResNet50	Lung Seg. IoU 95.5% Class. Sens. 85.9% Spe. 96.4%
[34]	COVID-19 (423) MERS-CoV (144) SARS-CoV (134)	✓		Segmentation model U-Net Classification model InceptionV3	Lung Seg. IoU 93.1% DSC 96.4% Class. Sens. 96.9% Spe. 91.7%
[35]	COVID-19 (573) Non-COVID (5,559) Healthy (8,066)	✓		Classification model RAND-GAN	Lung Seg. DSC 83.0% Class. Sens. 57.0% Spe. 80.0%
[41]	COVID-19 (2951) Non-COVID (116,365)		✓	Segmentation models U-Net, U-Net++, and DLA Backbone Encoders: CheXNet, DenseNet121, InceptionV3, and ResNet50	Infection Seg. DSC 83.2% Class. Sens. 94.9% Spe. 99.9%
This work	COVID-19 (11,956) Non-COVID (11,263) Healthy (10,701)	✓	✓	Segmentation models U-Net, U-Net++, and FPN Backbone Encoders ResNet18, ResNet50, DenseNet121, DenseNet161, InceptionV4	Lung Seg. IoU 96.1% DSC 97.9% Class. Sens. 99.6% Spe. 97.4% Infection Seg. IoU 83.1% DSC 88.1%

Open access publication is supported by Qatar National Library (QNL).

Conflicts of interest

The authors report no declarations of interest.

References

[1] World Health Organization, WHO coronavirus disease (COVID-19) Dashboard, Available, https://covid19.who.int/?gclid=Cj0KCQjwtZH7BRDzARIsAGjbK2ZXWRpJROEl97HGmSOx0_ydkVbc02Ka1FlcysGjEI7hnaIeR6xWhr4aAu57EALw_wcB, 2020.

[2] S. Shastri, K. Singh, S. Kumar, P. Kour, V. Mansotra, Time series forecasting of Covid-19 using deep learning models: India-USA comparative case study, Chaos, Solit. Fractals 140 (2020) 110227, 2020/11/01/.

[3] S. Shastri, K. Singh, M. Deswal, S. Kumar, V. Mansotra, CoBiD-net: a tailored deep learning ensemble model for time series forecasting of covid-19, in: Spatial Information Research, 2021/06/12, 2021.

[4] A. Pormohammad, et al., Comparison of confirmed COVID-19 with SARS and MERS cases - clinical characteristics, laboratory findings, radiographic signs and outcomes: a systematic review and meta-analysis, Rev. Med. Virol. 30 (4) (2020), https://doi.org/10.1002/rmv.2112 p. e2112, 2020/07/01.

[5] T. Singhal, A Review of Coronavirus Disease-2019 (COVID-19), 2020, pp. 1–6.

[6] C. Sohrabi, et al., World Health Organization Declares Global Emergency: A Review of the 2019 Novel Coronavirus (COVID-19), 2020.

[7] P. Kakodkar, N. Kaka, M.J.C. Baig, A Comprehensive Literature Review on the Clinical Presentation, and Management of the Pandemic Coronavirus Disease 2019 (COVID-19), vol. 12, 2020 no. 4.

[8] Y. Li, et al., Stability Issues of RT-PCR Testing of SARS-CoV-2 for Hospitalized Patients Clinically Diagnosed with COVID-19, 2020.

[9] A. Tahamtan, A. Ardebili, in: Real-time RT-PCR in COVID-19 Detection: Issues Affecting the Results, Taylor & Francis, 2020.

[10] J. Xia, J. Tong, M. Liu, Y. Shen, D.J. Guo, Evaluation of Coronavirus in Tears and Conjunctival Secretions of Patients with SARS-CoV-2 Infection, vol. 92, 2020, pp. 589–594, no. 6.

[11] T. Ai, et al., Correlation of Chest CT and RT-PCR Testing in Coronavirus Disease 2019 (COVID-19) in China: a Report of 1014 Cases, 2020, p. 200642.

[12] S. Salehi, A. Abedi, S. Balakrishnan, A.J. Gholamrezanezhad, Coronavirus Disease 2019 (COVID-19): a Systematic Review of Imaging Findings in 919 Patients, 2020, pp. 1–7.

[13] Y. Fang, et al., Sensitivity of Chest CT for COVID-19: Comparison to RT-PCR, 2020, p. 200432.

[14] D.J. Brenner, E.J. Hall, Computed Tomography—An Increasing Source of Radiation Exposure, vol. 357, 2007, pp. 2277–2284, no. 22.
[15] F. Shi, et al., Review of Artificial Intelligence Techniques in Imaging Data Acquisition, Segmentation and Diagnosis for Covid-19, 2020.
[16] C. Huang, et al., Clinical Features of Patients Infected with 2019 Novel Coronavirus in Wuhan, China, vol. 395, 2020, pp. 497–506, no. 10223.
[17] M. Hosseiny, S. Kooraki, A. Gholamrezanezhad, S. Reddy, L.J. Myers, Radiology Perspective of Coronavirus Disease 2019 (COVID-19): Lessons from Severe Acute Respiratory Syndrome and Middle East Respiratory Syndrome, vol. 214, 2020, pp. 1078–1082, no. 5.
[18] A. Esteva, et al., Dermatologist-level Classification of Skin Cancer with Deep Neural Networks, vol. 542, 2017, pp. 115–118, no. 7639.
[19] H. Dong, G. Yang, F. Liu, Y. Mo, Y. Guo, Automatic brain tumor detection and segmentation using U-Net based fully convolutional networks, in: Annual Conference on Medical Image Understanding and Analysis, Springer, 2017, pp. 506–517.
[20] L. Shen, L.R. Margolies, J.H. Rothstein, E. Fluder, R. McBride, W. J. S. r Sieh, Deep Learning to Improve Breast Cancer Detection on Screening Mammography, vol. 9, 2019, pp. 1–12, no. 1.
[21] D. Ardila, et al., End-to-end Lung Cancer Screening with Three-Dimensional Deep Learning on Low-Dose Chest Computed Tomography, vol. 25, 2019, pp. 954–961, no. 6.
[22] A. Tahir, et al., Coronavirus: Comparing COVID-19, SARS and MERS in the Eyes of AI, 2020 arXiv preprint arXiv:1706.05587.
[23] P. Rajpurkar, et al., Deep Learning for Chest Radiograph Diagnosis: A Retrospective Comparison of the CheXNeXt Algorithm to Practicing Radiologists, vol. 15, 2018 no. 11, p. e1002686.
[24] Stanford ML Group, CheXpert: A Large Dataset of Chest X-Rays and Competition for Automated Chest X-Ray Interpretation, Available, https://stanfordmlgroup.github.io/competitions/chexpert/.
[25] T. Rahman, et al., Reliable Tuberculosis Detection Using Chest X-Ray with Deep Learning, Segmentation and Visualization, vol. 8, 2020, pp. 191586–191601.
[26] T. Ozturk, M. Talo, E.A. Yildirim, U.B. Baloglu, O. Yildirim, U. Rajendra Acharya, Automated detection of COVID-19 cases using deep neural networks with X-ray images, Comput. Biol. Med. 121 (2020) 103792, 2020/06/01/.
[27] I.D. Apostolopoulos, T.A. Mpesiana, Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks, Phys. Eng. Sci. Med. 43 (2) (2020) 635–640, 2020/06/01.
[28] L. Wang, Z.Q. Lin, A. Wong, COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images, Sci. Rep. 10 (1) (2020) 19549, 2020/11/11.
[29] A. Waheed, M. Goyal, D. Gupta, A. Khanna, F. Al-Turjman, P.R. Pinheiro, CovidGAN: data augmentation using auxiliary classifier GAN for improved covid-19 detection, IEEE Access 8 (2020) 91916–91923.
[30] M.E.H. Chowdhury, et al., Can AI help in screening viral and COVID-19 pneumonia? IEEE Access 8 (2020) 132665–132676.
[31] M. Yamaç, M. Ahishali, A. Degerli, S. Kiranyaz, M.E.H. Chowdhury, M. Gabbouj, Convolutional sparse support estimator-based COVID-19 recognition from X-ray images, IEEE Trans. Neural Netw. Learn. Syst. 32 (5) (2021) 1810–1820.
[32] Y. Fan, J. Liu, R. Yao, X. Yuan, COVID-19 detection from X-ray images using multi-kernel-size Spatial-Channel attention network, Pattern Recogn. 119 (2021) 108055, 2021/11/01/.
[33] Y. Oh, S. Park, J.C. Ye, Deep learning COVID-19 features on CXR using limited training data sets, IEEE Trans. Med. Imag. 39 (8) (2020) 2688–2700.
[34] A. Tahir, et al., Deep learning for reliable classification of COVID-19, MERS, and SARS from chest X-ray images, Cogn. Comput. (2021).
[35] S. Motamed, P. Rogalla, F. Khalvati, RANDGAN: randomized generative adversarial network for detection of COVID-19 in chest X-ray, Sci. Rep. 11 (1) (2021) 8602, 2021/04/21.
[36] S. Rajaraman, J. Siegelman, P.O. Alderson, L.S. Folio, L.R. Folio, S.K. Antani, "Iteratively pruned deep learning ensembles for COVID-19 detection in chest X-rays, IEEE Access 8 (2020) 115041–115050.
[37] Y. Oh, S. Park, J.C. Ye, Deep Learning Covid-19 Features on Cxr Using Limited Training Data Sets, 2020.
[38] S. Jaeger, et al., Automatic Tuberculosis Screening Using Chest Radiographs, vol. 33, 2013, pp. 233–245, no. 2.
[39] S. Candemir, et al., Lung Segmentation in Chest Radiographs Using Anatomical Atlases with Nonrigid Registration, vol. 33, 2013, pp. 577–590, no. 2.
[40] F. Shi, et al., Large-scale Screening of Covid-19 from Community Acquired Pneumonia Using Infection Size-Aware Classification, 2020.
[41] A. Degerli, et al., COVID-19 infection map generation and detection from chest X-ray images, Health Inf. Sci. Syst. 9 (1) (2021) 15, 2021/04/01.
[42] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.
[43] Z. Zhou, M.M.R. Siddiquee, N. Tajbakhsh, J. Liang, Unet++: a nested u-net architecture for medical image segmentation, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer, 2018, pp. 3–11.
[44] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125.
[45] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[46] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.
[47] C. Szegedy, S. Ioffe, V. Vanhoucke, A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31, 2017 no. 1.
[48] T. Rahman, et al., Exploring the Effect of Image Enhancement Techniques on COVID-19 Detection Using Chest X-Rays Images, 2020.
[49] J. Corne, Chest X-Ray Made Easy E-Book, Elsevier Health Sciences, 2015.
[50] Medical Imaging Databank of the Valencia Region, BIMCV-COVID19+: a large annotated dataset of RX and CT images of COVID19 patients, Available, https://bimcv.cipf.es/bimcv-projects/bimcv-covid19/#1590858128006-9e640421-6711.
[51] GitHub, covid-19-image-repository, Available, https://github.com/ml-workgroup/covid-19-image-repository/tree/master/png, 2020.
[52] Eurorad, Available, https://www.eurorad.org/.
[53] GitHub, covid-chestxray-dataset, Available, https://github.com/ieee8023/covid-chestxray-dataset, 2020.
[54] SIRM, COVID-19 DATABASE, Available, https://www.sirm.org/category/senza-categoria/covid-19/, 2020.
[55] Kaggle, COVID-19 radiography Database, Available, https://www.kaggle.com/tawsifurrahman/covid19-radiography-database, 2020.
[56] GitHub, COVID-CXNet, Available, https://github.com/armiro/COVID-CXNet, 2020.
[57] Kaggle, RSNA pneumonia detection challenge, Available, https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data, 2018.
[58] Kaggle, Chest X-ray images (pneumonia), Available, https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia, 2018.
[59] Medical Imaging Databank of the Valencia Region, PadChest: A large chest x-ray image dataset with multi-label annotated reports, Available, https://bimcv.cipf.es/bimcv-projects/padchest/.
[60] A. Degerli, et al., COVID-19 Infection Map Generation and Detection from Chest X-Ray Images, 2020.
[61] Image-net.org, ImageNet, Available, http://www.image-net.org/.
[62] Pytorch.org, PyTorch, Available, https://pytorch.org/.
[63] J. Malik, S. Kiranyaz, M.J.N.N. Gabbouj, Self-organized Operational Neural Networks for Severe Image Restoration Problems, vol. 135, 2021, pp. 201–211.
[64] S. Kiranyaz, J. Malik, H.B. Abdallah, T. Ince, A. Iosifidis, M.J. Gabbouj, Self-Organized Operational Neural Networks with Generative Neurons, 2020.
[65] A.M. Tahir, M.E.H. Chowdhury, Y. Qiblawey, A. Khandakar, T. Rahman, S. Kiranyaz, COVID-QU-Ex, Kaggle, 2021, https://doi.org/10.34740/KAGGLE/DSV/2759090.