QATAR UNIVERSITY

COLLEGE OF ENGINEERING

DEEP CONVNETS FOR COVID-19 RECOGNITION FROM CHEST X-RAYS

BY

ANAS M. TAHIR

A Thesis Submitted to

the College of Engineering

in Partial Fulfillment of the Requirements for the Degree of

Master of Science in Electrical Engineering

June 2021

COMMITTEE PAGE

The members of the Committee approve the Thesis of
Anas M. Tahir defended on 25/04/2021.

Prof. Mustafa Serkan Kiranyaz
Thesis/Dissertation Supervisor

Dr. Muhammad Enamul Hoque Chowdhury
Thesis/Dissertation Co-Supervisor

Prof. Selim Aksoy
Committee Member

Prof. Somaya Al-Madeed
Committee Member

Prof. Ridha Hamila
Committee Member

Approved:

Khalid Kamal Naji, Dean, College of Engineering

ABSTRACT

TAHIR, ANAS, M, Masters: June: 2021, Master of Science in Electrical Engineering

Title: Deep ConvNets for COVID-19 Recognition from Chest X-Rays

Supervisor of Thesis: Prof. Mustafa, Serkan, Kiranyaz.

Co-Supervisor of Thesis: Dr. Muhammad, Enamul Hoque, Chowdhury.

Coronavirus disease 2019 (COVID-19) is an extremely contagious and quickly spreading Coronavirus infestation. Severe Acute Respiratory Syndrome (SARS)-CoV and Middle East Respiratory Syndrome (MERS)-CoV, which outbreak in 2002 and 2011, and the current COVID-19 pandemic are all from the same family of coronavirus. The fatality rate due to SARS and MERS was higher than COVID-19. However, the spread of those was limited to few countries, while COVID-19 affected more than 200 countries worldwide, causing over 3 million causalities and infected more than 145 million people as of April 25, 2021. Given the effects of COVID-19 on pulmonary tissues, chest radiographic imaging has become a necessity for screening and monitoring the disease. Recently, numerous studies have proposed Deep Learning approaches based on Convolutional Neural Networks (CNNs, or ConvNets) for the automatic diagnosis of COVID-19 from chest X-rays (CXR). Although these methods achieved astonishing performance in early detection and diagnosis, they have used limited CXR repositories for evaluation, usually with a few hundred COVID-19 CXR images only. Thus, such data scarcity prevents reliable evaluation with the potential of overfitting. In addition, manual annotation of X-rays (delineation of the lung, or infection regions) is another challenge due to the extensive time and manual labor required from the physicians. Therefore, most of the proposed studies showed no or

limited performance in infection localization and severity grading of COVID-19 pneumonia.

In this thesis, in order to overcome the aforementioned limitations and challenges, we have conducted the following: (i) compiled the largest COVID-19 benchmark dataset, namely COVID-QU, which consists of 11,956 COVID-19, 11,263 non-COVID-19, 10,701 normal, 134 SARS, and 144 MERS CXR images, (ii) generated ground-truth lung segmentation masks for the entire COVID-QU dataset using an elegant human-machine collaborative approach, (iii) proposed a systematic approach to segment the lung, detect, localize, and quantify COVID-19 infections from CXR images, (iv) Trained and evaluated the proposed system for lung segmentation, infection segmentation, and two classification tasks: I) COVID-19 detection from the predecessor COVID family members, SARS, and MERS, II) COVID-19 detection from non-COVID-19 infections, and normal cases.

A detailed set of experiments using several state-of-the-art ConvNets showed top performance for the lung segmentation task with Intersection over Union (IoU) of 96.11% and Dice Similarity Coefficient (DSC) of 97.99%. Besides, COVID-19 infections of various shapes and types were reliably localized with 83.05% IoU and 88.21% DSC. Moreover, the proposed system was able to discriminate between different COVID family members, which is an extremely challenging task for medical doctors without the aid of clinical data. Sensitivities of 96.94%, 79.68%, and 90.26% were achieved for classifying COVID-19, MERS, and SARS classes, respectively. Furthermore, a good performance was obtained for the second classification scheme with sensitivities of 91.52%, 93.21%, and 91.12 for COVID-19, non-COVID, and normal classes, respectively.

ACKNOWLEDGMENTS

First, I thank Allah (GOD) Almighty for bestowing his many blessings upon me and granting me the will, and strength to fulfill this work.

Next, I would like to thank my supervisors Prof. Serkan Kiranyaz, and Dr. Muhammad Chowdhury, for their continuous support, guidance, and encouragement throughout my master period. I learned a lot from them, and I am still learning. They helped me to improve my technical skills and broaden my research knowledge.

I would like to extend my gratitude to everyone in our big research team, our collaborators from Hamad Medical Corporation, Dr. Khalid and Dr. Tahir, our collaborators from Tampere University, Prof. Moncef Gabbouj and his research team, and of course my teammates from QU-research group, Eng. Amith, Yazan, Uzair, Tawsifur, Sakib, Maymouna, and Nabil. Together we managed to create several benchmark datasets and provide different diagnostic tools that can help in fighting the current COVID-19 pandemic. Also, I would like to thank the Electrical Engineering Department for allowing me to work with them during my master's period.

Special thanks to my friends and brothers Abdalla Sewify, Ahmed Al-Obaidy, and Ayman Al-Kabaji for their continuous support.

Finally, words are not enough to express my endless gratitude to my parents, Dr. Mohammed Tahir and Dr. Nigar Abdul Ghafoor, and my siblings, Ahmed and Mariam. I would not have accomplished this without your patience, support, prayers, and advices.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

The World has experienced outbreaks of coronavirus infections during different points of time in the last two decades: (i) the Severe Acute Respiratory Syndrome (SARS)-CoV outbreak in 2002-2003 from Guangdong, China; (ii) the Middle East Respiratory Syndrome (MERS)-CoV outbreak in 2011 from Jeddah, Saudi Arabia; and (iii) Coronavirus Disease 2019 (COVID-19) or SARS-CoV-2 outbreak from Wuhan, China in December 2019. Even though all three diseases are from the same family of coronavirus [1], the genomic sequence of COVID-19 showed similar but distinct genome composition from its predecessors SARS and MERS [1, 2]. Despite a lower fatality rate of COVID-19, i.e., around 3% [3] when compared to SARS (10%) and MERS (35%), COVID-19 has resulted in many fold deaths (>3M already) than combined deaths of MERS and SARS (around 1700) [4]. The SARS-CoV epidemic has spread to 26 countries worldwide using person-to-person human contact [5]. In 2012, the infectious outbreak caused by MERS-CoV epidemic had spread to more than 1600 patients in 27 countries, resulting in over 600 deaths, 80% of which were reported in Saudi Arabia [6, 7]. The recent outbreak of COVID-19 was and still is an extremely infectious disease that has spread all over the world, forcing the World Health Organization (WHO) on 11th March 2020 to declare it as a pandemic [8].

1.1 Background

The business, economic, and social dynamics of the whole world were affected. Governments have imposed flight restrictions, social distancing, and increasing awareness of hygiene. However, COVID-19 is still spreading at a very rapid rate. The common symptoms of coronavirus include fever, cough, shortness of breath, and pneumonia. Severe cases of the CoV diseases include acute respiratory distress syndrome (ARDS) or complete respiratory failure, which requires support from

mechanical ventilation and intensive-care unit. People with a compromised immune system, elderly people, or people with other chronic diseases are more likely to develop serious illnesses, including organs failure, particularly kidneys or septic shocks [9].

Intuitively, reliable detection of COVID-19 disease has the utmost importance. However, the diagnosis procedures are not straightforward, as the common symptoms of COVID-19 are generally indistinguishable from the other viral infections [10, 11]. Currently, the primary diagnostic tool to detect COVI-19 is reverse-transcription polymerase chain reaction (RT-PCR) arrays, where the presence of SARS-CoV-2 RNA is tested on collected respiratory specimens from the suspected case [12, 13]. However, RT-PCR arrays have a high false alarm rate caused by sample contamination, damage to the sample, or virus mutations in the COVID-19 genome [14, 15]. Therefore, several studies suggested using chest computed tomography (CT) imaging as a primary diagnostic tool since it shows higher sensitivity values compared to RT-PCR [16, 17]. In addition, several studies [16-18] suggest performing CT as a secondary test if the suspected patients with shortness of breath or other respiratory symptoms showed negative RT-PCR findings. Despite the superior performance, CT scans are challenged by several limitations. Their sensitivity is limited for early COVID-19 cases, slow in imaging acquisition, and costly. On the other hand, X-ray imaging is a cheaper and faster method, where the body gets exposed to less radiation compared to CT [19]. Chest X-ray imaging is widely used as an assistive diagnostic tool in COVID-19 screening, and it is reported to have high potential prognostic capabilities [20].

Aiming to automize the COVID-19 recognition process from CXR images, recently, many studies [21-26] have proposed to use Deep Learning approaches based on Convolutional Neural Networks (CNNs, or ConvNets). These studies showed superior performance for early detection and diagnosis of COVID-19. However, the

data scarcity in these studies prevents a reliable evaluation with the potential of overfitting and limits the performance of deep networks. Moreover, several studies [27-29] proposed lung segmentation as a first-line in their COVID-19 recognition approach. Lung segmentation is an important pre-classification step, which narrows the region of interest from the entire CXR down to the region of lungs to increase network reliability. Thus, avoiding irrelevant areas in the decision-making process, such as heart, bones, background, or text. However, the reported segmentation performance is limited for COVID-19 cases. In general, the proposed lung segmentation networks miss highly COVID-19 infected lung regions, such as peripheral infection or fluid accumulation in lower lung lobes. Such poor performance takes place, as so far, there are no lung segmentation masks datasets available for COVID-19 X-ray images. Furthermore, most of these studies showed limited performance in infection localization and severity grading of COVID-19 pneumonia.

On the other hand, the majority of the proposed AI-based COVID-19 recognition approaches tries to distinguish COVID-19 from other viral/bacterial infections or normal X-rays. However, up to the author's knowledge, there is no work in the literature to recognize COVID-19 infection from the other two COVID-family members, MERS and SARS. Due to the overlapping patterns of lung infections, without the aid of clinical data, it is difficult for medical doctors (MDs) to distinguish between the images from different CoV family members using CXR only. Therefore, investigating the similarities of COVID family members in the eyes of AI can provide meaningful insights that can help in the medical diagnosis.

1.2 Thesis Objective

In order to overcome the aforementioned limitations and challenges, the key objectives of this thesis are:

1) Review the different proposed approaches in the literature for automatic COVID-19 recognition from CXR and investigate their gaps and limitations.
2) Compile the largest COVID-19 benchmark dataset, referred to as COVID-QU, which consists of 11,956 COVID-19, 11,263 non-COVID, 10,701 normal, 134 SARS, and 144 MERS chest X-rays images. This will help to investigate deep ConvNet on a comparatively larger dataset, which can provide a more reliable solution with better generalization capabilities.
3) Create ground-truth lung segmentation masks for the entire COVID-QU dataset using an elegant human-machine collaborative approach which can significantly reduce human labor and thus speed up the annotation process.
4) Propose a robust system to segment the lung, detect, localize, and quantify COVID-19 infections from chest X-ray images. This is a crucial task for accurate diagnosis and follow-up of COVID-19 patients.
5) Train and evaluate the proposed recognition system for lung segmentation, infection segmentation, and two classification schemes:
    - COVID-19 recognition from non-COVID-19 infections, and normal cases.
    - COVID-19 recognition from the predecessor COVID family members, SARS, and MERS.

1.3 Thesis Outline

The rest of the thesis is organized as follows: In Chapter 2, a comprehensive review is conducted on recent studies for AI-assisted diagnosis of COVID-19 from CXR. In Chapter 3, the benchmark COVID-QU CXR dataset is introduced with a novel collaborative human-machine approach for lung ground-truth saliency-map generation. Besides, the details of the proposed COVID-19 recognition system are explained in

Section 3.2. In Chapter 4, the experimental setup is defined, and the COVID-19 recognition system is evaluated on the benchmark dataset. Accordingly, the final results are discussed and analyzed. Finally, Chapter 5 draws the conclusion of the thesis and suggests some future directions.

CHAPTER 2: LITERATURE REVIEW

Readily available radiological imaging techniques such as chest CT and X-ray are crucial tools for COVID-19 detection. The majority of early COVID-19 cases show 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. Even though chest radiographs can help in the early screening of the suspected case, the images of several viral pneumonia are similar. They show a high overlap with other inflammatory lung diseases. Therefore, it is difficult for medical doctors to distinguish COVID-19 infection from other viral pneumonia. Hence, this symptom similarity can lead to wrong diagnosis in the current situation. Such an incorrect result can lead to non-COVID-19 viral pneumonia being falsely diagnosed as a highly suspicious COVID-19 case, thus delaying the treatment with consequent effort, costs, and risk of exposure to positive COVID-19 patients.

2.1 Related work

The tremendous development in Machine Learning and Deep Learning techniques in recent years led to state-of-the-art performance in several Computer Vision tasks, such as image classification, object detection, and image segmentation. This breakthrough in performance led to increased utilization of AI-based solutions in various fields, including biomedical health problems and complications. Specifically, ConvNet has been proven extremely beneficial in several biomedical imaging applications, such as skin lesion classification [30], brain tumor detection [31], and breast cancer detection [32], and lung pathology screening [33, 34]. Deep Learning techniques on chest X-ray images are gaining popularity with the availability of deep ConvNets, showing promising results in various applications. Rajpurkar et al. [35]

proposed CheXNet network, one of the top-performing architectures for CXR, by training DenseNet121 on the ChestX-ray14 dataset [36], the largest public CXR dataset with over 100 thousand X-ray images for 14 different pathologies. Vikash et al. [37] utilized the concept of transfer learning to recognize pneumonia infection from normal CXR by proposing an ensemble approach that combines the output of five pre-trained deep models achieving sensitivity values of >98% for pneumonia class. Lakhani et al. [38] reported an AUC of 0.99 on a dataset of 1,007 CXR by utilizing an assemble of GoogleNet and AlexNet to classify the CXR images as having manifestations of pulmonary tuberculosis TB or as healthy.

2.1.1 Deep Learning for COVID-19 Recognition from Chest X-rays

Recently, many studies have reported Machine Learning and Deep Learning approaches to automize COVID-19 detection from chest X-rays [21-26]. Ozturk et al. [21] presented a modified version of DarkNet, to provide a reliable diagnosis for binary classification (COVID-19 vs. Normal) and multi-class classification (COVID-19 vs. non-COVID-19 pneumonia vs. Normal). The introduced network was evaluated over a dataset that contains 114 COVID-19 CXR. However, low performance was reported with COVID-19 sensitivity values of 90.65% and 85.35% for binary and multi class schemes, respectively. Apostolopoulos et al. [22] utilized a dataset that consists of 224 COVID-19, 714 confirmed viral or bacterial pneumonia, and 504 normal X-rays. High discrimination accuracy of 96.7% and COVID-19 Sensitivity of 98.7% was obtained with MobileNetV2 model. Wang et al. [23] introduced a new ConvNet architecture (COVID-Net) tailored for COVID-19 recognition. COVID-Net was evaluated on a dataset with 358 COVID-19 CXR, where it achieved sensitivity values of 91%, 94%, and 95% for COVID-19, non-COVID-19 pneumonia, and normal classes, respectively. Waheed et al. [24] proposed a synthetic data augmentation technique to alleviate the scarcity of public

data available for COVI-19 X-rays. Auxiliary Classifier Generative Adversarial Network (ACGAN) model was introduced and implemented on 403 COVID-19, and 721 Normal CXR images. ACGAN model along with synthetic data augmentation yielded 95% accuracy and 90% COVID-19 sensitivity. Apostolopoulos et al. [25] trained MobileNetV2 on a 7-class dataset that includes 358 COVID-19, 1,342 Normal, and 1,199 x-ray images for five common thorax abnormalities. The followed approach achieved 87.66% 7-class Accuracy, 99.18% 2-class Accuracy (COVID-19 vs. remaining classes), and 97.36% COVID-19 sensitivity. Chowdhury et al. [26] compiled a dataset of 423 COVID-19, 1485 viral pneumonia and 1579 normal X-rays and have trained several deep ConvNets (SqueezeNet, ResNet18, ResNet101, MobileNetV2, DenseNet201 and CheXNet) for both 2-class (COVID-19 vs Normal) and 3-class schemes. DenseNet201 showed the best classification performance with 99.7% and 97.9% COVID-19 sensitivities for 2-class and 3-class schemes, respectively. However, most of the conducted studies used a rather small amount of data, e.g., the largest dataset includes only few hundred CXR samples. Therefore, it is difficult to generalize their results in practice.

Yamac et al. [39] introduced a compact architecture that utilizes the state-of-the-art pneumonia detection network, 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% COVID-19 sensitivity over the benchmark QaTa-COV19 dataset that contains 462 COVID-19 CXR images. In a recent approach [40] the same group of researchers, as in [39], proposed a reliable advance warning system to diagnose early-stage COVID-19 cases with limited or no infection signs from normal cases. Several deep learning and compact classifier approaches were evaluated over

Early-QaTa-COV19 datasets with 1,065 early-stage COVID-19, and 12,544 Normal CXR images. Satisfactory results were obtained with >97% and 95% COVID-19 sensitivity for the best deep learning (CheXNet) and compact (CSEN) approaches, respectively.

2.1.2 Lung Segmentation as a First Stage in the COVID-19 Recognition System

Moreover, several studies [27-29] considered lung segmentation as the first stage in their recognition system. This ensures reliable decision-making in the classification phase and guards the network against irrelevant features from non-lung areas. Rajaraman et al. [29] proposed a two-stage COVID-19 recognition model. In the first stage, U-Net segmentation network was utilized to segment the lung regions. Secondly, several pre-trained deep models (VGG16, VGG19, InceptionV3, etc.) were iteratively pruned to reduce network complexity while maintaining a satisfactory classification performance. The obtained results showed that the weighted average of top-3 pruned models improves the performance significantly, resulting in 99% COVID-19 sensitivity and 99.72% AUC. In a similar approach, Oh et al. [28] proposed a patch-based deep ConvNet 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 system achieved an overall classification accuracy and sensitivity of 88.9% and 85.9%, respectively. However, the previous segmentation approaches were trained on a mixture of medium and high-quality CXR, mainly from Montgomery [41] and Shenzhen [42] CXR lung mask dataset, which combinedly creates 704 X-ray images for normal and tuberculosis (TB) cases. Therefore, the segmentation performance degrades in unseen scenarios such as severe COVID-19 cases or low-quality images with poor SNR levels. Hence, lung areas can be partially or incompletely segmented for severe infections such as bilateral consolidation or fluid

accumulation at lower-lung lobes. Consequently, the classification performance degrades. Therefore, creating a large benchmark CXR dataset with ground-truth lung segmentation masks is of high importance, and will help the research community to provide more reliable detection system for COVID-19 and other thorax pathologies.

2.1.3 COVID-19 Infection Localization and Severity Grading

In addition, along with COVID-19 detection, infection localization is another crucial point that helps in evaluating the status of the patient and in the treatment process [43]. 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. Even though those localized regions are potential biomarkers for COVID-19, more precise and reliable localization can be provided by ground-truth infection mask from expert radiologists. Therefore, Degerli et al. [44] proposed a novel approach for COVID-19 infection map generation by compiling the largest COVID-19 dataset consisting of 2,951 CX images with annotated ground-truth infection segmentation masks. Several encode-decoder (E-D) ConvNets were trained and evaluated on the generated dataset, where the best performing network achieved an 85.81% f1-score for infection localization. However, their proposed approach is limited to infection localization. Therefore, there is room to revisit the problem with both lung and infection segmentation models to localize and quantify infection regions by computing the overall percentage of infected lungs. This can help medical doctors to better assess the severity and progress of COVID-19 pneumonia.

2.1.4 Classification of Coronavirus Family Members: COVID-19, MERS and SARS

On the other hand, worldwide researchers have presented numerous clinical and experimental information regarding the SARS and MERS, which could be useful in the fight against COVID-19 [54]. There have been studies in the literature

investigating the similarities between the genome structure of SARS, MERS, and COVID-19 [55]. However, up to the author's knowledge, COVID-19 recognition from the other two family members, MERS and SARs, using CXR has never been investigated. Owing to the overlapping pattern of lung infections, it is very challenging for MDs to diagnose the COVID type without the aid of clinical data, specifically RT-PCR. Although, SARS epidemic was contained in July 2003, and no case has been reported since May 2004 [45]. However, MERS still exists, where the most recent laboratory-confirmed cases were reported by Riyadh in March 2020 [46]. Therefore, investigating the similarities and uniqueness of COVID family members in the eyes of AI can bridge the knowledge gaps and provide MDs with meaningful insights that help in the diagnosis process.

In a nutshell, many studies have reported Deep Learning approaches to automate COVID-19 detection from CXR images. They have reported high detection performance for the disease; however, they also present certain issues and drawbacks. First of all, majority of them have used a limited amount of COVID-19 data, e.g., the largest dataset includes only a few hundred CXR samples. As mentioned earlier, such a data scarcity yields a lack of proper evaluation, and thus it is difficult to generalize their results in practice. Moreover, they only aimed for COVID-19 detection and/or classification among other types without further assessment and localization. Due to these issues, their usability and robustness for a clinical usage will be very limited. In this study, an end-to-end solution will be provided to segment the lung, detect, localize, and quantify COVID-19 infections from CXR images. Besides, the largest benchmark CXR dataset, named COVID-QU, will be compiled with over 33 thousand CXRs including 11,956 COVID-19 CXRs. Ground-truth lung segmentation masks will be created for the entire dataset using a novel collaborative human-machine approach,

which can save valuable human labor time and minimize subjectivity in the annotation process. This will help to investigate the state-of-the-art deep network models more reliably and accurately for COVID-19 and other lung pathology problems.

CHAPTER 3: MATERIALS AND METHODOLOGY

In this thesis, in order to overcome the aforementioned limitations and drawbacks, first, a large benchmark dataset so-called COVID-QU, is compiled with over 33,000 CXR images from five different classes: SARS, MERS, COVID-19, non-COVID-19 infections, and normal. Besides, in Section 3.1, a novel human-machine collaborative approach is proposed to generate lung saliency maps for the entire COVID-QU dataset. In section 3.2, a systematic approach is proposed to segment the lung, detect, localize, and quantify COVID-19 infections from chest X-ray images. Furthermore, two classification schemes are tackled: COIVD-19 recognition from non-COVID-19 and normal CXR, and COVID-19 discrimination from the other COVID family members MERS and SARs.

3.1 The Benchmark COVID-QU Dataset

Sharing COVID-19 data will help researchers, doctors, and engineers around the world to come up with innovative solutions for the early detection of COVID-19. In this section, we first show the data compilation process; then, we propose a novel approach for ground-truth lung saliency map generation.

3.1.1 Data Compilation

Due to the emerging nature of the pandemic, initially, little efforts have been made by highly infected countries on sharing clinical and radiography data publicly. Therefore, a group of researchers from Qatar University (QU) and Tampere University (TU), including the author of this thesis, have created two datasets, the so-called COVID-Family [47] and QaTa-Cov19 datasets [39]. The COVID-Family dataset consists of 462 COVID-19, 144 MERS, and 134 X-ray images. While QaTa-Cov19 dataset contains the same 462 COVID-19 samples included in the COVID-Family dataset, along with 2,760 bacterial pneumonia, 1,485 viral pneumonia, and 1,579

normal X-rays. The QaTa-Cov19 dataset was extended in a recent study [44] to include 2,951 COVID-19 CXR along with their ground-truth infection masks. Gradually, more X-rays become available publicly, and QU group managed to collect additional X-rays images. Hence, QU group compiled the largest COVID-19 CXR dataset with over 33 thousand samples, called COVID-QU. The dataset includes X-rays from five different classes:

1) 134 SARS X-rays
2) 144 MERS X-rays
3) 11,956 COVID-19 X-rays
4) 11,263 non-COVID-19 infections (viral or bacterial pneumonia) X-rays
5) 10,701 normal X-rays

In this study, only posterior-to-anterior (PA) or anterior-to-posterior (AP) chest X-rays were considered as this view of radiography is widely used by the radiologist. This dataset was created by utilizing numerous publicly available, scattered, and different format datasets and repositories. Authors ensured the quality of the provided information; duplicates, extremely low-quality, and over-exposed images were identified and removed in the preprocessing stage. Consequently, the dataset encapsulates images of high intraclass dissimilarity with varying resolution, quality, and SNR levels, as shown in Figure 1. Details of different data sources are listed below:

COVID-19 CXR dataset: The dataset contains 11,956 positive COVID-19 CXR images: 10,814 images are collected from BIMCV-COVID19+ dataset [48], 183 images from a Germany medical school [49], 559 X-ray image from SIRM, Github, Kaggle and Tweeter [50-53], and 400 X-ray images from another COVID-19 chest X-ray repository [54].

RSNA CXR dataset (Lung opacity and normal CXR): RSNA pneumonia

detection challenge dataset [55] consists of 26,684 chest X-ray images, where 8,851 images are normal, 11,821 are abnormal, and 6,012 are lung opacity images. All images are in DICOM format. In this study, we used 8,851 normal and 6,012 lung opacity X-ray, where lung opacity images are used as non-COVID-19 class.

Chest-Xray-Pneumonia dataset: This is a Kaggle dataset [56] that encapsulates 1,300 viral pneumonia, 1,700 bacterial pneumonia, and 1,000 normal X-rays. In this study, the viral and bacterial pneumonia are considered as non-COVID-19 class.

PadChest dataset: PadChest [57] dataset comprises more than 160,000 X-ray images from 67,000 patients that were collected and reported by radiologists at Hospital San Juan (Spain) from 2009 to 2017. In this study, we used 4,000 normal, and 4,000 pneumonia/infiltrate cases as non-COVID-19 class.

SARS and MERS CXR dataset: SARS and MERS X-ray images are even scarcer compared to COVID-19. Therefore, we collected and indexed X-ray images from different publicly available online resources and articles. SARS and MERS radiographic images were collected from 55 different articles (25-MARS, 30-SARS). A total of 260 images was collected from articles, and 18 images were from Joseph Paul Cohens' GitHub database [58]. Out of these, 70 MERS X-ray images were collected from [59], while 16 SARS X-ray images were collected from [60].

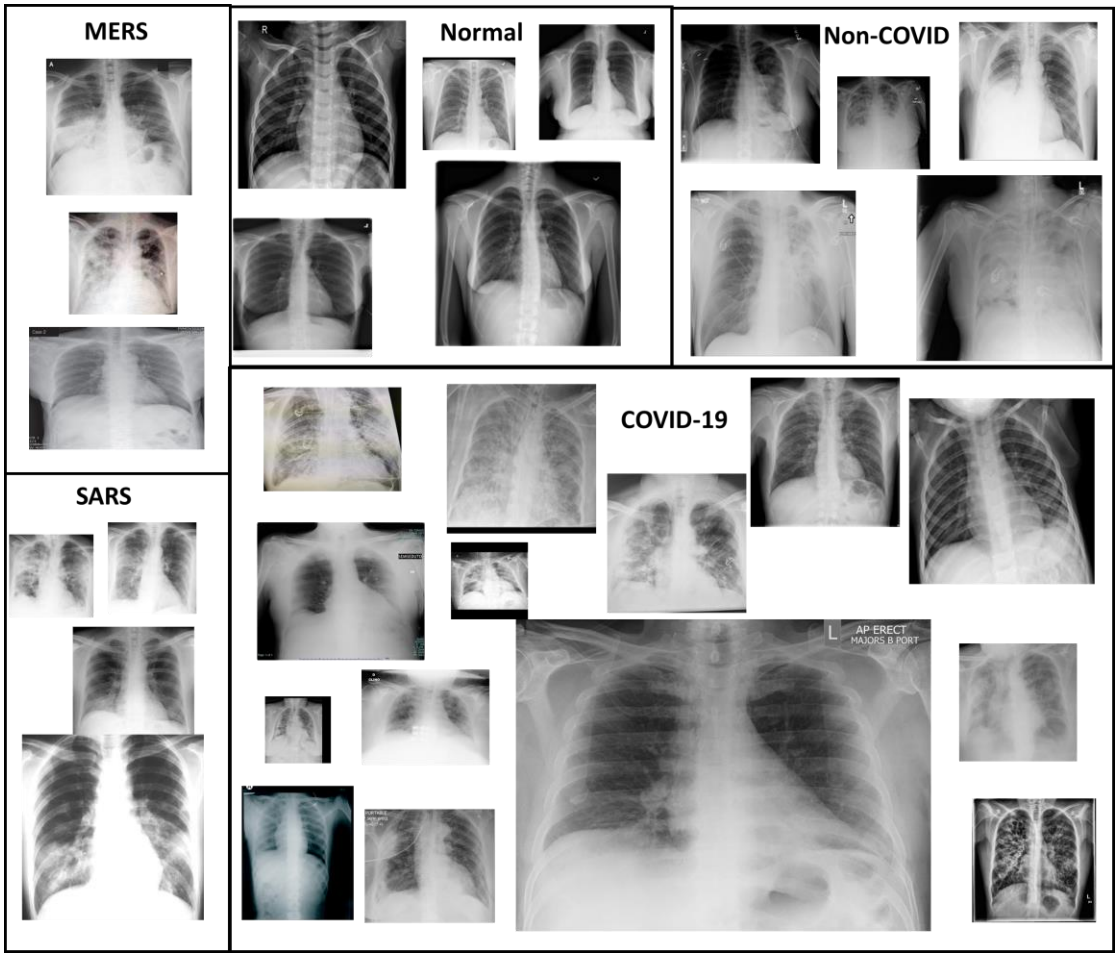

Figure 1. Sample X-ray images from the COVID-QU Dataset from five different classes: COVID-19, MERS, SARS, non-COVID-19 infections, and normal. The dataset encapsulates images from several countries around the world with different resolution, quality, and SNR levels.

Montgomery and Shenzhen CXR lung mask dataset: This dataset consists of 704 CXR images with their corresponding lung segmentation masks. However, it was not included in the COVID-QU dataset. Still, it was used as initial ground truth masks to train the segmentation model in the first stage of the proposed human-machine collaborative approach. The dataset was acquired by Shenzhen Hospital in China [42], and the tuberculosis control program of the Department of Health and Human Services of Montgomery County, MD, USA [41]. Montgomery dataset consists of 80 normal and 58 tuberculosis CXR with lung segmentation masks. While Shenzhen dataset

compromises 326 normal and 336 tuberculosis CXR, where 566 out of 662 CXR are provided with their corresponding masks.

QaTa-Cov19 CXR infection mask dataset [44]: This dataset was created by a research group from Qatar University and Tampere University. It consists of nearly 120K CXR images, including 2,913 COVID-19 images with their corresponding ground-truth infection masks. In this study, the ground-truth infection masks were used to train and evaluate the infection segmentation models.

3.1.2 Collaborative Human-Machine Ground-Truth Annotation

The process of producing ground truth segmentation masks is an exhaustive task, where human experts need to delineate pixel-wise masks with high accuracy levels. Besides, with the emergent of the current pandemic, it is even more challenging to assign such a task to medical experts, as they are busy fighting the disease.

In order to overcome this issue, a collaborative human-machine approach is proposed to produce ground-truth lung segmentation masks for CXR images accurately. The majority of the manual annotation process was assigned to biomedical engineering researcher assistants (RAs) from QU team to reduce the load on medical collaborators from Hamad Medical Corporation (HMC). Unlike infection segmentation, which needs precise delineation by medical experts, lung segmentation can be done by non-medical people with proper supervision by MDs. Therefore, before starting the annotation process, all RAs attended several training sessions conducted by MDs to grasps a general understating of Chest X-ray imaging and get exposed to a variety of cases with mild, moderate, or severe infections. Hight attention was given to different types of abnormalities, such as lung opacity consolidation and fluid accumulation, which can make border detection more difficult.

The intended approach was performed in four main stages.

3.1.2.1 Stage I (Initial Training):

In the first stage, three variants of U-Net [61] segmentation model, are trained on 704 ground-truth CXR lung masks from Montgomery and Shenzhen dataset mentioned previously. The ground-truth CXR lung masks are referred to as CXR lung mask repository in Figure 2, and it is enlarged throughout the mask creation process. Next, the best performing network in terms of DSC is selected as the main network for Stage II, which is referred to as CXR-Segmentation network in Figure 2.

3.1.2.2 Stage II (Collaborative Evaluation):

In the second stage, an iterative training is utilized to create lung masks for a subset of 3,000 CXR samples (10% of the full dataset) that well present the diversity of COVID-QU dataset. Firstly, A subset of 500 samples is selected and inferred using CXR-Segmentation model. The predicated lung masks are then evaluated by researchers from QU group: as accept, reject, unsure, or exclude. Accepted masks that cover the lung areas are directly added to CXR-lung-mask-repository. Rejected mask are incomplete ones which miss parts of the lung areas or include extra areas. Those rejected masks are first modified by RAs then added to CXR-lung-mask-repository. Unsure masks are severe cases with highly infected areas; those are usually consolidations or fluid accumulation at lower lung lobes with a whitish color, which makes them indistinguishable from neighboring organs. The doubtful areas are first assessed by MDs; then, RAs adjust the masks based on their recommendations. While the generated masks and corresponding X-rays are excluded only if the quality is extremely bad such as the case shown in Figure 2, where the right lung is blurred and corrupted by extra lighting. Finally, the segmentation network is re-trained on the extended mask dataset. Then a second subset of 500 samples is selected, and the steps of Stage II are repeated. This process is repeated until ground-truth masks for all 3,000

CXR samples are generated.

3.1.2.3 Stage III (Collaborative Selection):

In the third stage, six deep segmentation networks, inspired by U-Net [61], U-Net++ [62], and FPN [63] architectures, are trained on the 3,000 collaborative masks generated in Stage II. The trained networks are used to predict segmentation masks for the rest of COVID-QU dataset, which is around 30,900 unannotated samples (90% of the full dataset). Among the six predictions, RAs select the best one as a ground-truth or deny if none of the masks segments the lung properly. The latter case was a minority case that included less than 5% of unannotated data. The most selected network was considered as the main network and re-trained with the extended masks repository. The denied cases were then inferred by the main 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 lung masks dataset. The proposed collaborative approach saves valuable human labor time. Also, it enhances the quality and reliability of the generated masks and reduces subjectivity.

3.1.2.4 Stage IV (Final Verification):

In the final stage, a final verification is performed by MDs on 6,788 CXR samples (20% of the full dataset) that well presents the diversity of COVID-QU dataset. The samples are selected from COVID, non-COVID-19 and normal classes, with different resolution, quality, and SNR levels. Even though checking the entire dataset will result in higher quality masks. However, it is not a feasible solution with the large number of images that we have as it will add an extra burden for MDs. In this study, the verified subset (20%) was considered as a test set for all the experiments, while the remaining data (80%) was considered as train and validation sets.

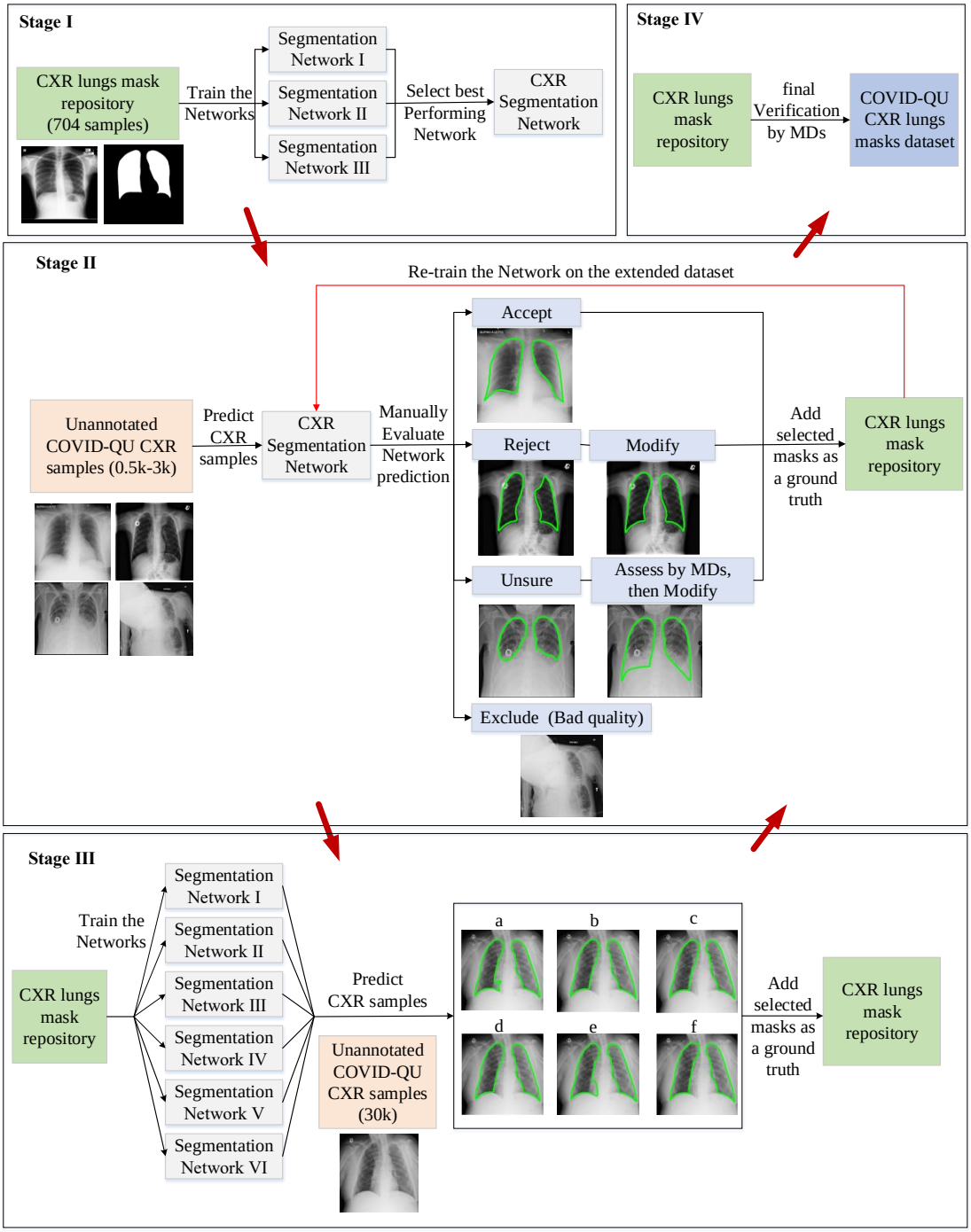

Figure 2. Collaborative human-machine approach to create ground-truth lung segmentation masks for COVID-QU CXR dataset

3.2 COVID-19 Recognition System

In this section, we describe the proposed system to segment the lung, detect, localize, and quantify COVID-19 infections from CXR images (Figure 3). First, a

binary lung mask is generated from the input CXR image using the 1st encoder-decoder (E-D) ConvNet. In parallel, the input CXR is fed to the 2nd E-D ConvNet to generate COVID-19 infection masks. Next, generated lung and infection masks are superimposed with the CXR image to localize and quantify COVID-19 infected lung regions. The generated infection mask is then used to detect COVID-19 positive cases from COVID-19 negative cases, where the CXR is classified positive if at least one pixel of lung area is predicted as COVID-19 infection. Furthermore, a 3rd ConvNet is trained and evaluated on two classification schemes:

- COVID-19 recognition from non-COVID-19 infections and normal cases
- COVID-19 recognition from the other coronaviruses SARS and MERS.

Additionally, Score-CAM visualization method is deployed to provide an interpretable result and investigate the reasoning behind the specific decisions of the deep ConvNet classifier.

The classification network will be removed in future studies, once ground-truth infections masks are created for non-COVID-19 cases using the proposed collaborative human-machine approach. Therefore, the infection segmentation model can be used to generate a 3-channel infection mask, where the channels represent: background, non-COVID-19 lesion, and COVID-19 lesions. Thus, the 3-channel infection masks can be used to detect COVID-19 from non-COVID-19, or normal cases.

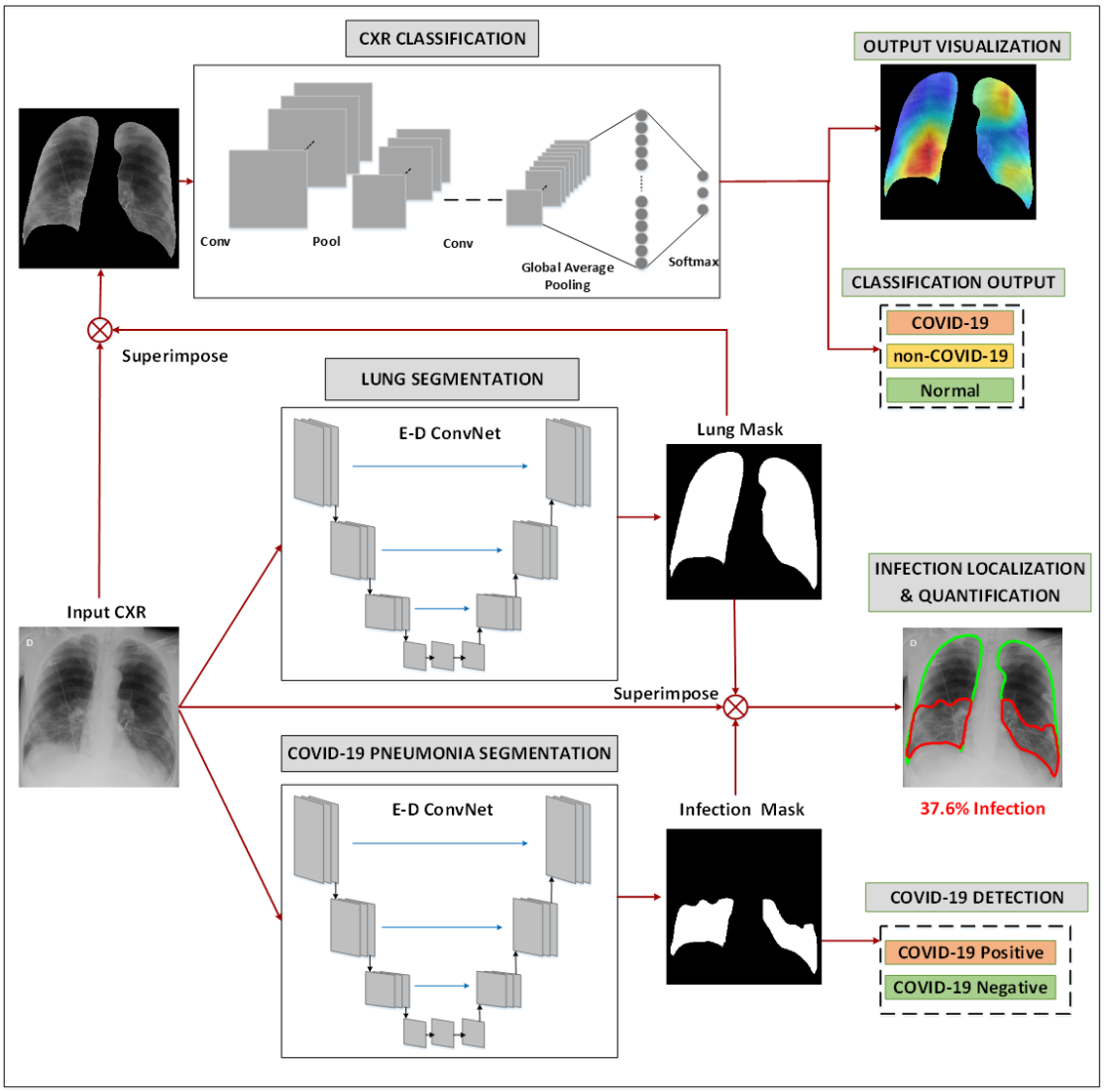

Figure 3. Schematic representation of the proposed COVID-19 recognition system

3.2.1 Image Pre-Processing

Medical images are sometimes poor in contrast and often corrupted by noise due to different sources of interference, such as the imaging process and data acquisition. As a result, it may become harder to evaluate them visually. Contrast enhancement methods can play an important role in improving the image quality to provide a better interpretable image to the medical doctors. Besides, it can boost the performance of deep recognition systems. In order to investigate potential enhancement on the classification performance, four pre-processing schemes were evaluated in this

study: original chest X-ray image, which did not undergo any form of pre-processing, contrast limited adaptive histogram equalization (CLAHE), image complementation, and finally, the combination of the three (original, CLAHE, complemented) schemes applied altogether to form a 3-channel approach.

3.2.1.1 CLAHE Technique

Histogram equalization (HE) is a technique mainly used with images that are predominantly dark to enhance the contrast by effectively spreading out the most frequent intensity values [64]. The HE transformation can be defined as follows:

$$y = T(x) = (L-1)\sum_{i=0}^{x} p_x(X=i) \quad (1)$$

where x denotes the random variable representing the original pixel intensities, $p_x(X=x)$ is the probability of having the pixel intensity x, $T(.)$ is the transformation function, y are the new intensities after transformation, and $L = 2^N$ is the intensity values for an N-bit image, i.e., for 8-bit gray-scale image, L-1=255 is the maximum intensity value. A closer look at equation (1) will reveal the fact that $T(x)$ is the approximation of the cumulative distribution functions [65]. An improved HE variant is called Adaptive Histogram Equalization (AHE). The adaptive equalization performs HE over small regions (patches) in the image. It improves local contrast and edges adaptively in each patch according to the local distribution of pixel intensities instead of the global information of the image. However, AHE could over amplify the noise component in the image [66]. To address this difficulty, Contrast-Limited Adaptive Histogram Equalization (CLAHE) limits the amount of contrast enhancement that can be produced within the selected region by a threshold parameter. Therefore, produced images are more natural in appearance than those produced by AHE [67]. Besides, the clarification of image details is improved.

When the HE technique was applied to the X-ray images, it was observed that

it might saturate certain regions. However, CLAHE technique can address this drawback in general. For instance, Figure 4 shows the application of CLAHE and HE techniques over a sample X-ray image. The histogram for the equalized images shows that the values are redistributed across all pixels compared with the histogram of the original image. The CLAHE image showed bell-shaped histogram as Rayleigh distribution was used for transformation, while the HE showed a flat histogram with an uniform distribution. However, the image was saturated in the center of the lungs when HE technique was applied. In addition, some regions of the HE image show a sharp brightness difference, whereas the CLAHE image exhibits a smooth transition of intensities for adjacent pixels. As a result, in this study, CLAHE was used for pre-processing the X-ray images instead of HE.

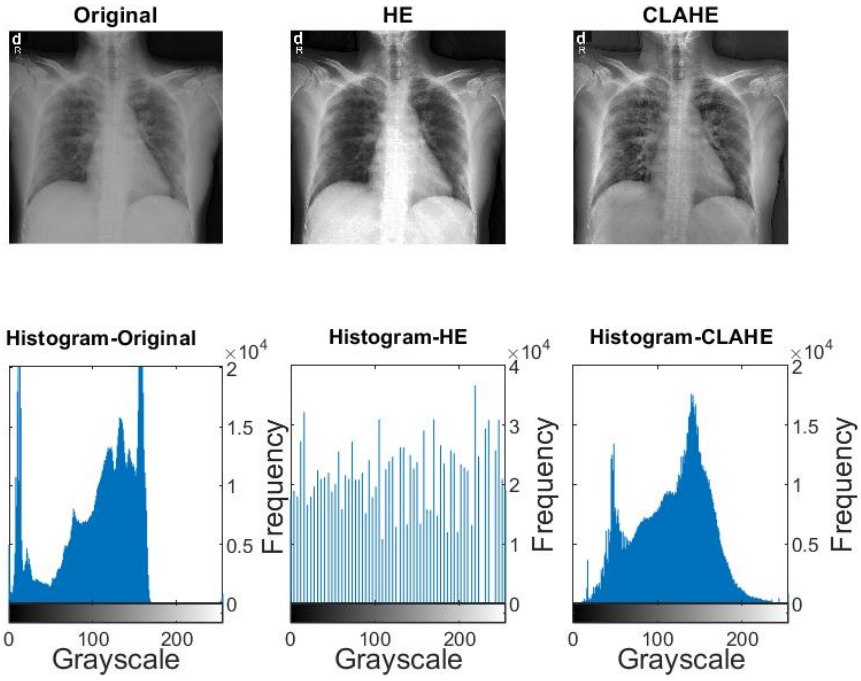

Figure 4. Comparison between original, HE, and CLAHE equalized X-ray images with corresponding histograms

3.2.1.2 Image Complementation

The image inversion or complement is a technique where the zeros become ones and ones become zeros so black and white are reversed in a binary image. For an 8-bit greyscale image, the original pixel is subtracted from the highest intensity value, 255, the difference is considered as pixel values for the new image. The mathematical expression is:

$$y = 225 - x \tag{2}$$

where x and y are the intensity values of the original and the transformed (new) images. This technique shows the lungs area (i.e., the region of interest) lighter and the bones are darker. As this is a standard procedure, which is used widely by radiologists, it may equally help deep networks for a better classification. It can be noted that the histogram for the complemented image is a flipped copy of the original image (Figure 5).

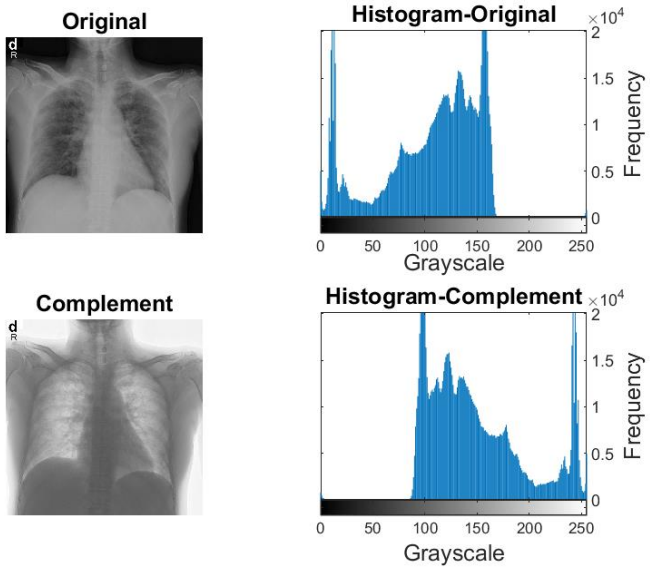

Figure 5. Comparison between an original X-ray and its image complement.

3.2.1.3 3-Channel Scheme

Finally, as shown in Figure 6, the 3-channel scheme was used as the input to the

deep networks, where original, CLAHE, and complement images were used altogether. The pixel values for each image are concatenated into a single matrix in order to create a new image. This 3-channel approach is expected to enhance the network performance compared to grayscale X-ray images as the utilized deep ConvNet classifiers were initially pre-trained on RGB images from the ImageNet dataset.

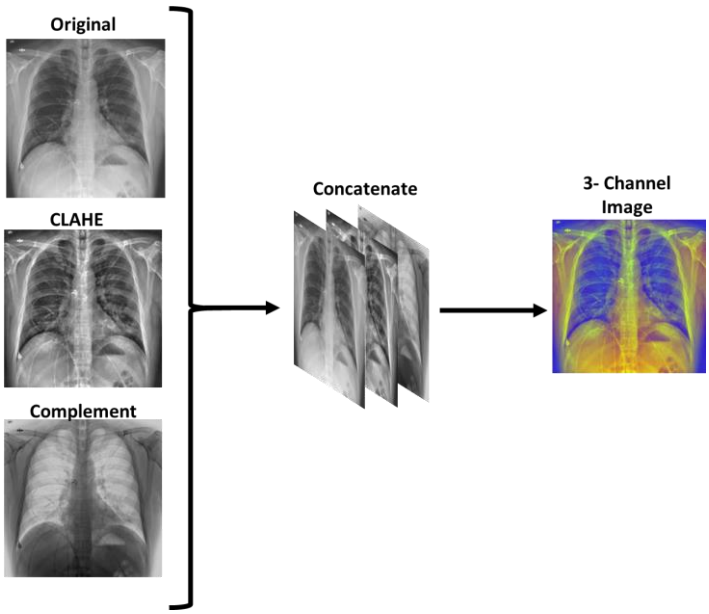

Figure 6. Illustration of 3-channel scheme

3.2.2 ConvNet Models for Lung and COVID-19 Infection Segmentation

Lung parenchyma and COVID-19 infections segmentation were performed on CXR images using three state-of-the-art deep E-D ConvNets: U-Net [61], U-Net++ [62], and FPN [63] with different backbone (encoder) models using the variants of ResNet, DenseNet, and InceptinV4 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 structures provide a firm segmentation model that captures the context in the contracting path and empowers precise localization by

the expanding path. 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. In addition, high-resolution features from the encoder path are merged with the up-sampled output from the corresponding decoder path through skip connection. Moreover, U-Net++ is a recent implementation that has further developed the decoder structure. 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 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, 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 upsampled to the same size, concatenated, convolved with 3x3 convolutional filter, and then SoftMax activation is applied to generate the final prediction mask.

3.2.2.1 Segmentation Loss Function

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

$$CE = -\frac{1}{K}\sum_k \sum_c y_k \log\big(p(x_k)\big) \tag{3}$$

where x_k denotes the k^{th} 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.

3.2.2.2 Post-processing for Segmentation Masks

The predicted segmentation masks, $\hat{Y}$, by the segmentation models are defined as $\hat{Y}_{h,w} \in [0,1]$, 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 $\hat{y} > 0.5$ for the lung segmentation task, while classified as COVID-19 infection if $\hat{y} > 0.5$ 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, non-lung regions that are falsely predicted as lung. In contrast, infection masks are and operated 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.

3.2.2.3 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, healthy people or patients with non-COVID-19 pneumonia. Furthermore, COVID-19 infection is quantified by computing the overall percentage of infected lungs. Equivalently, the sum of predicted infection pixels over the sum of predicted lung pixels. In addition, the infection percentage of each lung is computed, enabling doctors to assess the progress of COVID-19 for each lung individually.

3.2.3 ConvNet Models for Chest X-ray Classification

Choosing the best ConvNet for a specific problem is usually a tradeoff between the following two criteria: computational complexity and classification accuracy.

Therefore, we investigated several Deep Learning models, starting from shallow to deep models with sequential, residual, and dense connections. In this study, two classification schemes were considered: (i) COVID-19 recognition from SARS and MERS coronaviruses, (ii) COVID-19 recognition from non-COVID-19 infections and normal cases. For the first classification scheme, four pretrained models were investigated: SqueezeNet, ResNet18, DenseNet201, and InceptionV3. While for the second classification scheme, five pretrained models were investigated: ResNet18, ResNet50, DenseNet121, Densenet161, and InceptionV4. The output layer of each network was replaced by a SoftMax layer with three neurons to classify the X-ray images into one of the two 3-class schemes. Details of the employed models are given below:

SqueezeNet [68]: is the smallest network considered in this study, with 18 layers only. Introducing fire modules, where a squeeze convolutional layer with 1x1 kernels is fed to an expand layer that has a mix of 3x3 and 1x1 kernels. The network begins with a standalone convolutional layer, followed by eight fire block and end with a convolutional layer followed by a SoftMax layer. The number of kernels per fire module is increased gradually through the network. The network performs max-pooling after the first convolutional layer, 4th fire module, and 8th fire module. The compact architecture of SqueezeNet makes it favorable over other networks for such problems that it can achieve a comparable performance level.

ResNet [69]: Overfitting is a well-known paradigm for training deep ConvNets that can drastically degrade the network performance when trained with scarce data. The overfitting problem becomes worse when higher number of training epochs are performed and eventually, the network saturates due to the vanishing gradients problem. ResNet introduces the concept of residual blocks, where the input and output

of stacked layers are merged by simple add-up, providing an extra path for the backward gradient flow. This prevents the vanishing gradients problem and enhances the generalization capabilities of the network. In this study three variants of ResNet model were considered: ResNet18, ResNet50, and ResNet152.

DenseNet [70]: aggregates all feature maps instead of summing residuals. All layers in a dense block are densely connected to their subsequent layers, receiving extra supervision from previous layers. The dense structure creates a compact layer with little redundancy in the learned feature, where different network parts can share pieces of collective knowledge. In this study three variants of DenseNet model were considered: DeneNet121, DenseNet161, and DenseNet201.

InceptionV3 [71]: showed improved performance compared to its deeper competitors in classifying different types of problems. Typically, larger kernels are favored for global features that are distributed over a large area of images. In contrast, smaller kernels are preferred for an area-specific feature that is distributed over image frame. This inspired the idea of inception layers, where kernels of different sizes (1x1, 3x3, and 5x5) are concatenated within the same layer instead of going deeper into the network. The Inception architecture increases the network space, where the best features can be selected by training.

InceptionV4 [72]: further factorized the convolutions and pruned the dimensions of the InceptionV3 network. Despite the lower complexity, it preserves a higher performance.

3.2.3.1 Classification Loss Function

The cross-entropy loss is used as the cost function for the classification networks.

3.2.4 Transfer Learning

Transfer learning is a well-established Deep Learning approach, where gained knowledge from one problem is applied to a different but a related problem. To ensure an efficient training and faster convergence, transfer learning was utilized for the classification networks and on the encoder side of the segmentation networks by initializing the convolutional layers with ImageNet [73] weights.

3.2.5 ConvNet Output Visualization

Visualization techniques help in understanding the internal mechanisms of ConvNet and the reasoning behind the network making a specific decision. In addition, it interprets the results in a way that is easily understandable to human, thereby increasing the confidence of ConvNet outcomes. The main visualization technique employed in literature is Gradient-weighted class Activation Map (Grad-CAM) [74], where activation maps are generated by backward passing the gradients of the target class back to the final convolutional layer in the network to produce the localization map. The localization map $\text{Grad}-CAM\ L^c_{Grad-CAM} \in \mathbb{R}^{hxw}$ of height h and width w for class c is obtained by first computing the gradients of the score of target class with respect to the feature map A^k as $\frac{\partial y^c}{\partial A^k}$ where y^c is the network output before SoftMax. Next, the gradients are backward passed through global average pooling to compute the α weights, which highlights the importance of feature map k for the decision making of target class c:

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{4}$$

Finally, a weighted combination of activation maps A^k is followed by ReLU to obtain Grad-CAM map:

$$L^c_{Grad-CAM} = ReLU(\sum_k \alpha_k^c A^k) \tag{5}$$

Recently, Score-CAM [75] was proposed as a promising alternative to GRD-

CAM. Score-CAM gets rid of the dependencies on gradients by obtaining the weight of each activation map through forward passing scores of the target class. Given a ConvNet model $y^c = f(CXR)$ that takes an input CXR image and outputs a scalar y^c. The contribution of a specific feature map A^k toward output y^c is defined as follows:

$$\alpha_k^c = f\left(CXR \circ H_l^k\right) - f(CXR) \tag{6}$$

where

$$H_l^k = n(Up(A_l^k)) \tag{7}$$

$Up(.)$ denotes the up-sampling operation of A into the input (CXR) size, $n(.)$ is a normalization function that maps elements of input matrix into [0,1], and $\circ$ is the element-wise multiplication. Finally, the Score-CAM saliency map is computed using the same equation as Grad-CAM, Equation (5).

In this study, Score-CAM method is deployed to visualize the classification outputs of the proposed COVID-19 recognition system.

CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION

In this section, first, the experimental setup is presented. Then, both numerical and visual results are reported with an extensive set of comparative evaluations.

The findings of this thesis are published in two journal articles [47] and [76], which will be referred to as study I and study II in the upcoming section, respectively. In study I, we released a unique dataset, COVID-family, consisting of 701 CXR images with their corresponding ground-truth lung masks targeting a challenging classification task to distinguish among different COVID family members: SARS, MERS, COVID-19. A robust 2-stage system was employed, where first lung regions are segmented and then classified. In study II, we proposed a more practical diagnosis solution for the current pandemic, recognizing COVID-19 positive cases from non-COVID-19 infections or normal cases. We released the largest CXR lung mask dataset, COVID-QU, with over 33k CXR images. A reliable end-to-end solution was provided not for CXR classification only but to localize and quantify COVID-19 infections as well using a robust 3-stage system (Figure 3).

4.1 Experimental Setup

In Study I, a 2-stage image recognition system was proposed using the concatenation of lung segmentation and classification networks. The U-Net segmentation network was pre-trained and validated on the Montgomery [41] and Shenzhen [42] lung masks dataset. The pre-trained U-Net model was used to create lung masks for COVID-19, MERS, and SARS chest X-ray images. Next, the lung masks were fine-tuned by the MDs to develop ground truth masks for the COVID-family dataset. The deep classification networks were evaluated on the compiled COVID-family dataset and their corresponding lung masks. Two classification schemes were considered: plain CXR classification and segmented CXR classification. Both networks were trained using 5-

fold cross-validation (CV), with 80% train and 20% test (unseen folds), where 20% of training data is used as a validation set to avoid overfitting. The CXR images were resized to have a fixed dimension of 256x256 pixels to be used as the input for deep networks. The imbalance class distribution ratio of the dataset has a major impact on the performance of deep models. Therefore, the size of each class was balanced in the train set using data augmentation. We performed data augmentation by applying rotations of 5, 10, 20, and 25 degrees. In addition, horizontal and vertical image translations were used within the interval [-0.1, +0.1]. Table 1 summarizes the number of images per class used for training, validation, and testing at each fold.

Table 1. Number of Images per Class and per Fold Used for Study I

Task	Dataset	Class	# of Samples	Training Samples	Augmented Training Samples	Validation Samples	Test Samples
Lung Segmentation	Montgomery [41] and Shenzhen [42]	-	704	450	-	112	142
CXR Classification	COVID-family	COVID-19	423	270	1,890	68	85
		MERS	144	92	1,932	23	29
		SARS	134	89	1,806	21	26
		Total	701	451	5,628	112	140

In study II, both CXR classification and lung segmentation task were conducted over the constructed benchmark COVID-QU lung masks dataset. In contrast, the infection segmentation and COVID-19 detection tasks were conducted over a subset of 2,913 CXR samples from COVID-QU dataset with corresponding infection masks from QaTa-Cov19 dataset [44]. All tasks were performed with 20% test set, and 80% train set, where 20% of training data was used as a validation set. Table 2 summarizes the number of images per class used for training, validation, and testing.

Table 2. Number of Images per Class and per Fold Used for study II

Dataset Name	Task	Class	# of Samples	Training Samples	Validation Samples	Test Samples
COVID-QU dataset	Lung Segmentation and CXR Classification	COVID-19	11,956	7,658	1,903	2,395
		non-COVID-19	11,263	7,208	1,802	2,253
		Normal	10,701	6,849	1,712	2,140
		Total	33,920	21,715	5,417	6,788
COVID-QU and QaTa-Cov19 [44] datasets	Infection Segmentation and COVID-19 Detection	COVID-19 positive	2,913	1,864	466	583
		COVID-19 negative: non-COVID-19	1,457	932	233	292
		COVID-19 negative: Normal	1,456	932	233	291
		Total	5,826	3,728	932	1,166

The performance of different ConvNet models was assessed using different evaluation metrics with 95% confidence intervals (CIs). Accordingly, CI for each evaluation metric was computed as follows:

$$r = z\sqrt{metric(1 - metric)/N} \tag{8}$$

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

4.1.1 Segmentation Evaluation Metrics

The performance of the segmentation models is evaluated on a pixel-level, where the foreground (lung or infected region) was considered as the positive class and background as the negative class. Three evaluation metrics were computed to evaluate the segmentation performance:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{9}$$

where $accuracy$ 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.

$$Intersection\ over\ Union\ (IoU) = \frac{TP}{TP+FP+FN} \tag{10}$$

$$Dice\ Similarity\ Coefficient\ (DSC) = \frac{2*TP}{2*TP+FP+FN} \quad (11)$$

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

4.1.2 Classification Evaluation Metrics

The performance of different classification networks is assessed using five evaluation metrics: Accuracy, Precision, Sensitivity, F1-score, and Specificity.

Per-class values were computed over the overall confusion matrix that accumulates all test fold results of the 5-fold cross-validation.

$$Accuracy_{class\_i} = \frac{TP_{class\_i}+TN_{class\_i}}{TP_{class\_i}+TN_{class\_i}+FP_{class\_i}+FN_{class\_i}} \quad (12)$$

where accuracy is the ratio of correctly classified CXR samples among all the data.

$$Precision_{class\_i} = \frac{TP_{class\_i}}{TP_{class\_i}+FP_{class\_i}} \quad (13)$$

where precision is the rate of correctly classified positive class CXR samples among all the samples classified as positive samples.

$$Sensitivity_{class_i} = \frac{TP_{class_i}}{TP_{class_i}+FN_{class_i}} \quad (14)$$

where sensitivity is the rate of correctly predicted positive samples in the positive class samples,

$$Specificity_{class\_i} = \frac{TN_{class\_i}}{TN_{class\_i}+FP_{class\_i}} \quad (15)$$

where specificity is the sensitivity of the negative class.

$$F1\_score_{class_i} = 2\frac{Precision_{class_i} \times Sensitivity_{class_i}}{Precision_{class_i}+Sensitivity_{class_i}} \quad (16)$$

where F1_score is the harmonic mean of precision and sensitivity.

Besides $class_i$ = COVID-19, MERS or SARS for the first classification scheme, or

COVID-19, non-COVID-19 or Normal for the second scheme.

The overall performance for each metric is computed using the weighted average values of each class.

$$metric_x = \frac{\sum_{class_i} n_{class_i}(metric_{x_{class_i}})}{N} \quad (17)$$

where $metric_x = \text{Accuracy}, \text{Precision}, \text{Sensitivity}, \text{F1}_{\text{score}}, \text{or specificity}.$

Finally, n_{class_i} is the total number of cases per class, and N is the total number of all cases.

PyTorch [77] library with Python 3.7 was used to train and evaluate the deep ConvNet networks, with an 8-GB NVIDIA GeForce GTX 1080 GPU card. 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 which 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.

4.2 Experimental Results

In this section, numerical and quantitative evaluation results for study I and study II are presented.

4.2.1 Experimental Results for Study I

The performance of the proposed 2-stage image recognition system is detailed in this section. The deep ConvNet based classification networks were evaluated on the benchmark COVID-family dataset. Two classification schemes (plain and segmented CXR classification) were evaluated, and the outcome was interpreted with the help of Score-CAM visualization technique.

4.2.1.1 Lung Segmentation Results

The U-Net segmentation model was trained and evaluated on 704 CXR samples with ground-truth lung masks of the Montgomery and Shenzhen dataset, as shown in Table 3. The model showed promising segmentation performance with IoU and DSC of 93.11% and 96.35%, respectively, on the two publicly available datasets.

Table 3. Performance Metrics for Lung Region Segmentation Using U-Net Model

Network	Accuracy (%)	IoU (%)	DSC (%)
U-Net	98.21 ± 0.98	93.11 ± 1.87	96.35 ± 1.39

The qualitative evaluation of the trained U-Net model on the compiled COVID-family dataset images is presented in Figure 7. The model can reliably segment the lung images if the lung areas are distinguishable; however, the segmentation network suffers from severely infected lungs due to the whitened infection area in the lungs. Predicted lung masks by the U-Net model were revised by medical doctors to ensure that the segmentation masks encapsulate the entire lung region.

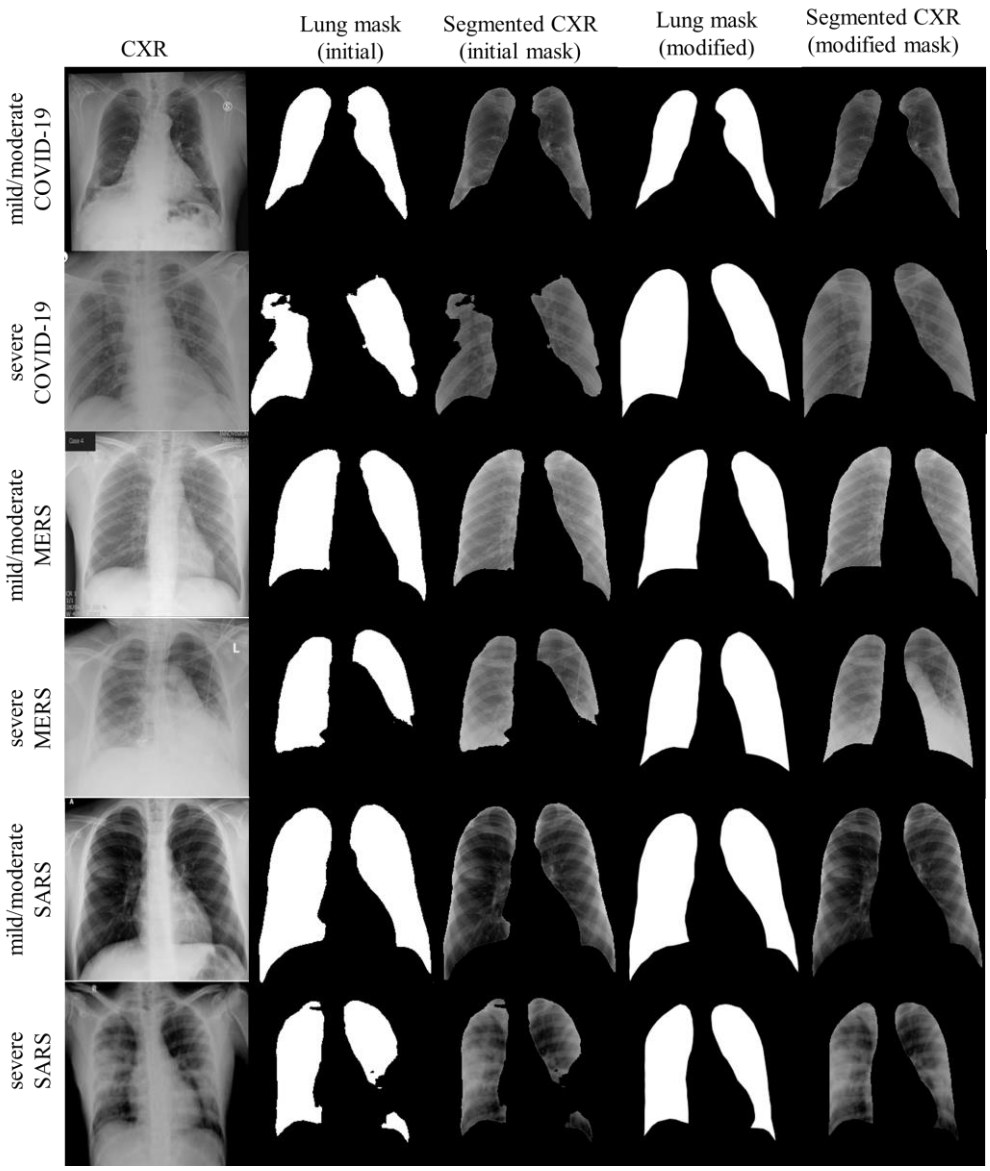

Figure 7. Qualitative evaluation of the U-Net model. Original X-ray images (left), lung mask generated by the trained U-Net model and corresponding segmented lung, fine-tuned mask by the radiologist, and their corresponding lung segment

4.2.1.2 Classification Results

Table 4 summarizes the classification performances of the deep ConvNet models in-terms of the per-class performance metrics for plain and segmented X-ray image classifications. For each network, four different pre-processing schemes (original, CLAHE, complemented, and 3-channel) were compared, and the best performing scheme is presented in Table 4. For plain X-ray images, it was observed

that SqueezeNet achieved the best classification performance on original images, while ResNet18 and Inceptionv3 outperformed on 3-channel images. For the segmented lung X-ray images, SqueezeNet and InceptionV3 showed the best performance with the original lung images without any pre-processing, and InceptionV3 outperformed all the networks. On the other hand, ResNet18 and DenseNet201 performed better on 3-Channel images. In general, the investigated ConvNet models showed high COVID-19 sensitivity values (>96%) for segmented data, while it showed varying results with plain X-rays. For instance, with plain X-ray, SqueezeNet showed 91.97% COVID-19 sensitivity, while InceptionV3 showed 99.53% COVID-19 sensitivity. For SARS and MERS, the InceptionV3 network achieved the highest sensitivities for a plain and segmented lung X-ray images. The sensitivity for MERS and SARS detection were 93.1%/79.68% and 97.04%/90.26% for plain/segmented lung CXRs, respectively. It is evident that the overall performance for MERS detection significantly degrades with segmentation. This is most likely due to a large number of lower-quality chest X-ray images in the MERS dataset. Even though the performance degrades with segmented lungs, as the network learns from the main region of interest (lung area), the results obtained from the segmented lungs are much more reliable.

Table 4. Performance Metrics (%) for Four Classification Networks: SqueezeNet, ResNet18, InceptionV3, and DenseNet201. The Best Preprocessing Technique is Reported for Each Network.

	Network	Class	Accuracy	Precision	Sensitivity	F1-score	Specificity
Plain	SqueezeNet	COVID-19	88.27 ± 3.07	89.31 ± 2.94	91.97 ± 2.59	90.48 ± 2.8	82.63 ± 3.61
X-rays	(Original)	MERS	91.56 ± 4.54	84.97 ± 5.84	72.09 ± 7.33	77.58 ± 6.81	96.58 ± 2.97
		SARS	91.86 ± 4.63	77.32 ± 7.09	81.25 ± 6.61	78.9 ± 6.91	94.36 ± 3.91
		Overall	89.77 ± 2.24	86.13 ± 2.56	85.84 ± 2.58	85.98 ± 2.57	88.02 ± 2.4
	ResNet18	COVID-19	94.04 ± 2.26	92.99 ± 2.43	97.88 ± 1.37	95.29 ± 2.02	88.21 ± 3.07
	(3-Channel)	MERS	96.03 ± 3.19	94.34 ± 3.77	85.49 ± 5.75	89.5 ± 5.01	98.75 ± 1.81
		SARS	97.16 ± 2.81	96.17 ± 3.25	88.89 ± 5.32	91.97 ± 4.6	99.12 ± 1.58
		Overall	95.02 ± 1.61	93.88 ± 1.77	93.61 ± 1.81	93.74 ± 1.79	92.41 ± 1.96
	Inceptionv3	COVID-19	97.87 ± 1.38	97.13 ± 1.59	99.53 ± 0.65	98.29 ± 1.24	95.36 ± 2
	(3-Channel)	MERS	98.3 ± 2.11	98.4 ± 2.05	93.1 ± 4.14	95.56 ± 3.36	99.64 ± 0.98
		SARS	99.29 ± 1.42	99.2 ± 1.51	97.04 ± 2.87	98.08 ± 2.32	99.82 ± 0.72
		Overall	98.22 ± 0.98	97.79 ± 1.09	97.73 ± 1.1	97.76 ± 1.1	97.07 ± 1.25
	DenseNet201	COVID-19	96.17 ± 1.83	96.55 ± 1.74	97.18 ± 1.58	96.85 ± 1.66	94.64 ± 2.15
	(complement)	MERS	97.02 ± 2.78	93.57 ± 4.01	91.72 ± 4.5	92.63 ± 4.27	98.39 ± 2.06
		SARS	98.86 ± 1.8	97.23 ± 2.78	97.04 ± 2.87	97.05 ± 2.86	99.3 ± 1.41
		Overall	96.84 ± 1.29	96.07 ± 1.44	96.03 ± 1.45	96.05 ± 1.44	96.28 ± 1.4
Segmented	SqueezeNet	COVID-19	92.12 ± 2.57	91.51 ± 2.66	96.22 ± 1.82	93.71 ± 2.31	85.83 ± 3.32
X-rays	(Original)	MERS	91.26 ± 4.61	83.01 ± 6.13	71.31 ± 7.39	75.92 ± 6.98	96.41 ± 3.04
		SARS	92.88 ± 4.35	82.58 ± 6.42	80.6 ± 6.7	81.28 ± 6.6	95.78 ± 3.4
		Overall	88.13 ± 2.39	88.05 ± 2.4	88.13 ± 2.39	88.09 ± 2.4	89.89 ± 2.23
	ResNet18	COVID-19	93.01 ± 2.43	91.74 ± 2.62	97.16 ± 1.58	94.37 ± 2.2	86.69 ± 3.24
	(3-Channel)	MERS	92.44 ± 4.32	85.27 ± 5.79	76.39 ± 6.94	80.59 ± 6.46	96.59 ± 2.96
		SARS	95.44 ± 3.53	91.13 ± 4.81	84.33 ± 6.16	87.6 ± 5.58	98.06 ± 2.34
		Overall	91.12 ± 2.11	91.2 ± 2.1	91 ± 2.12	91 ± 2.12	93.58 ± 1.81
	Inceptionv3	COVID-19	94.84 ± 2.11	94.85 ± 2.11	96.94 ± 1.64	95.82 ± 1.91	91.63 ± 2.64
	(Original)	MERS	93.41 ± 4.05	86.87 ± 5.52	79.68 ± 6.57	82.62 ± 6.19	96.95 ± 2.81
		SARS	96 ± 3.32	88.97 ± 5.3	90.26 ± 5.02	89.58 ± 5.17	97.35 ± 2.72
		Overall	92.12 ± 1.99	92.08 ± 2	92.12 ± 1.99	92.1 ± 2	93.81 ± 1.78
	DenseNet201	COVID-19	94.12 ± 2.24	93.2 ± 2.4	97.64 ± 1.45	95.3 ± 2.02	88.74 ± 3.01
	(3-Channel)	MERS	93.27 ± 4.09	89.75 ± 4.95	75.57 ± 7.02	81.93 ± 6.28	97.84 ± 2.37
		SARS	94.86 ± 3.74	86.38 ± 5.81	87.21 ± 5.65	86.42 ± 5.8	96.66 ± 3.04
		Overall	91.12 ± 2.11	91.18 ± 2.1	91.12 ± 2.11	91.15 ± 2.1	92.11 ± 2

Figure 8 shows the comparative ROC curves for different networks for different pre-processing schemes with plain and segmented X-rays. For plain X-rays, it is apparent from Figure 8(A) that Inceptionv3 outperforms other models over the original dataset while DenseNet201 and ResNet18 obtain a close performance, even though DenseNet201 is a very deep network compared to ResNet18. In contrast, the performance of SqueezeNet is comparable to the significantly deeper network, DenseNet201. Interestingly, the performances of InceptionV3, ResNet18, and DenseNet201 are comparable in the case of CLAHE images, and SqueezeNet shows a promising performance as well (Figure 8(B)). However, there is no notable performance improvement observed by this pre-processing scheme rather than making the classification less network independent. Figure 8(C) shows that significant performance improvement can be achieved using deeper networks with the complemented image. In contrast, the performance degrades for ResNet18 and especially for SqueezeNet. Figure 8(D) clearly depicts that the 3-channel scheme significantly improves the classification performance of InceptionV3 and ResNet18. However, this is not the case for DenseNet201 and SqueezeNet. On the other hand, with segmented X-rays, the four networks showed close performance for different pre-processing techniques (Figure 8(E-H)). Therefore, it can be concluded that proper segmentation can guide the network to learn from lung regions mainly. Thus, it makes the classification problem less dependent on the preprocessing technique. In addition, it eases the recognition task for the shallow networks, allowing them to achieve comparable results to their deeper competitors. Consequently, InceptionV3 using the original dataset without any preprocessing, showed the best classification performance for segmented X-rays.

In a nutshell, the performance gain from a specific pre-processing technique is

both problem and network dependent. Additionally, for future studies, it is worth investigating the effect of the ensemble technique on the X-ray classification scheme. The ensemble approach combines the output from several networks trained with different pre-processing techniques to generate the final classification output, rather than the 3-channel scheme used in this study, where the variants of the pre-processed input X-ray are combined and fed to a single network to make the final decision.

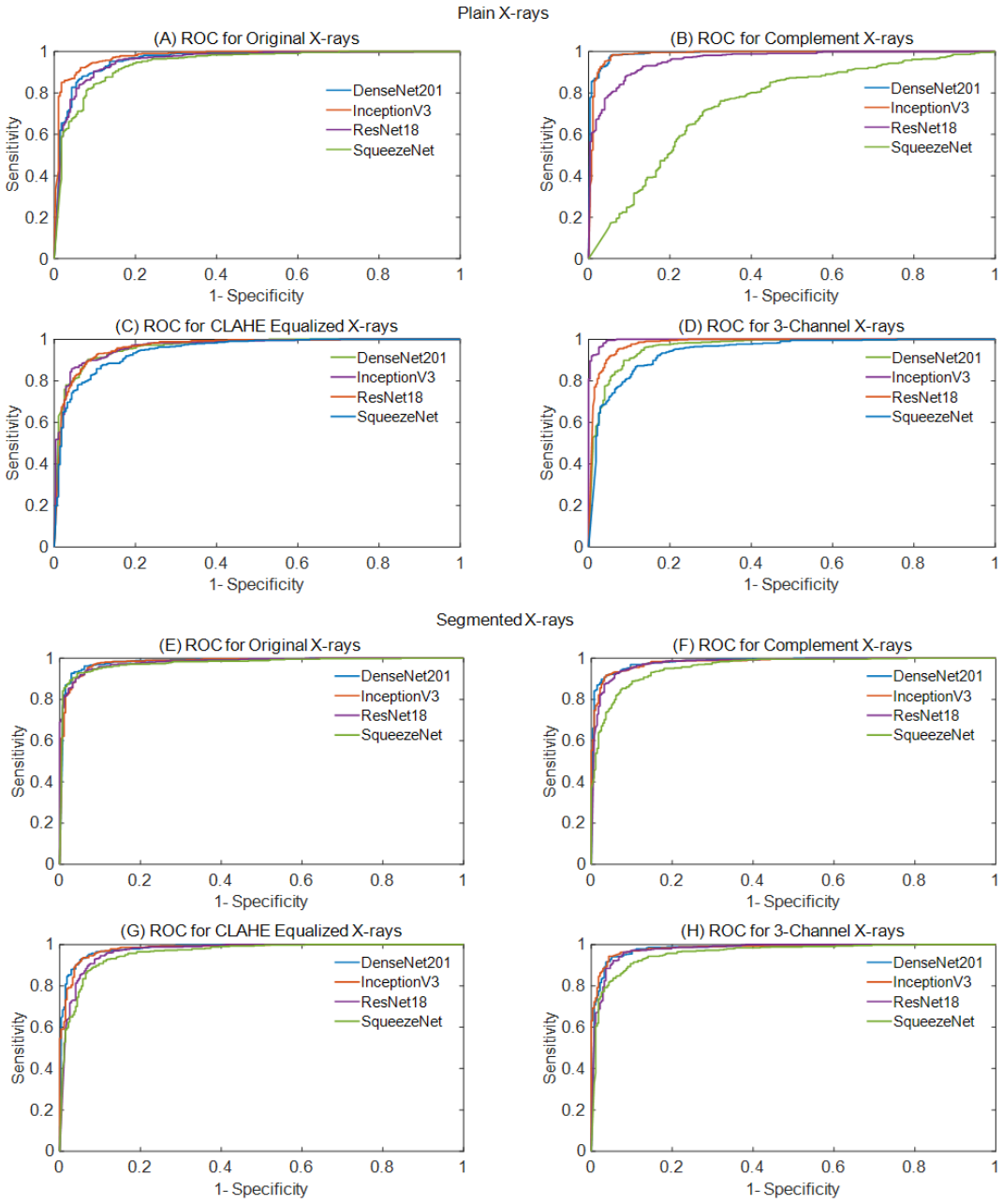

Figure 8. Comparison of the ROC for four networks using plain X-ray images (A-D) and segmented lung images (E-H): Original images (A/E), Complemented images (B/F), CLAHE images (C/G), and 3-channel images (D/H)

Score-CAM can help to localize the main regions of the input CXR that contributes to the ConvNet prediction. Figure 9 shows the Score-CAM saliency map for COVID-19, MERS, and SARS examples. It can be observed that with plain X-ray inputs, the ConvNets are learning irrelevant features from non-lung areas. In contrast,

with segmented X-rays, ConvNets are restricted to learn from the lung areas only. Therefore, the dominant contributing regions in the ConvNet decision-making are the lung areas.

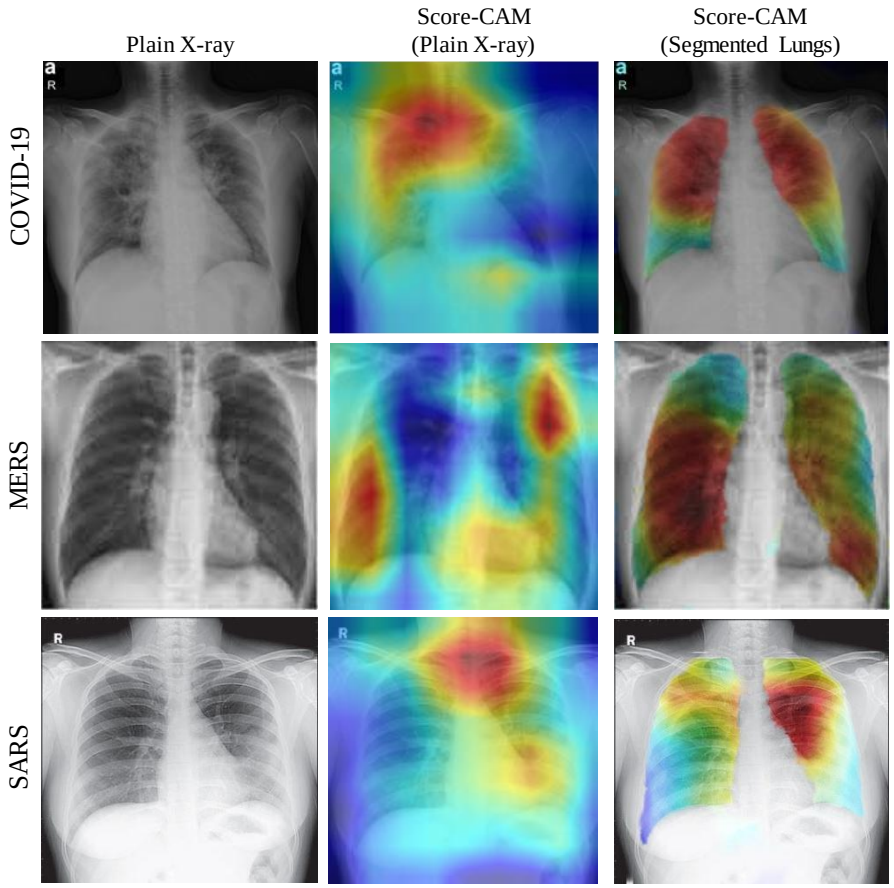

Figure 9. Examples of probabilistic saliency maps for COVID-19, MERS and SARS patients: (A) Plain CXR image, (B) Score-CAM for plain CXR inferred by InceptionV3 network, and (C) Score-CAM for segmented CXR inferred by InceptionV3 network

Figure 10 shows sample miss-classified X-ray images, corresponding lung image and Score-CAM visualization for COVID-19, MERS, and SARS images to identify the potential reasons of the network failure. It can be seen from Figure 10 that InceptionV3 failed to classify the lung images properly if the network did not learn from the lung areas exclusively whereas, for those images that are correctly classified

by the network, Score-CAM is showing that the ConvNet model is learning from the entire lung region. Therefore, it can be summarized that the reliable segmentation of lung from the X-ray images and the use of segmented X-ray images for the classification problem can significantly increase the reliability of AI-based computer-aided diagnosis applications.

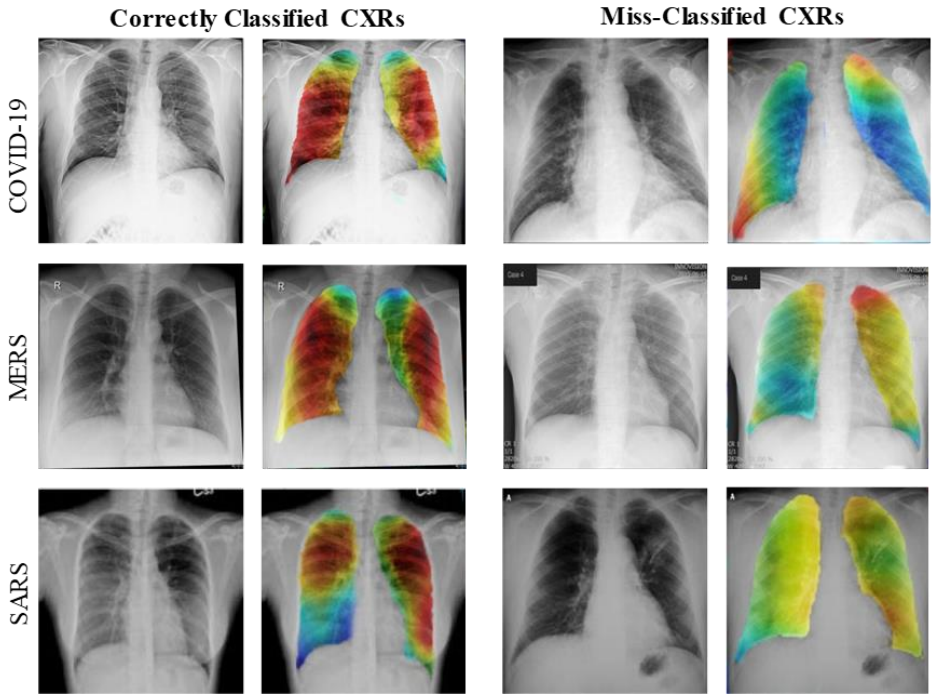

Figure 10. Comparison of the Score-CAM for correctly classified and miss-classified CXR images by InceptionV3

4.2.2 Experimental Results for Study II

In this section, qualitative and quantitative evaluation over the benchmark COVID-QU dataset is presented for the proposed end-to-end COVID-19 recognition system. An extensive set of experiments was performed for lung segmentation, lesion segmentation, COVID-19 detection, CXR classification tasks.

4.2.2.1 Lung Segmentation Results

The performance of the lung segmentation models over the test (unseen) set is tabulated in Table 5. Each model was evaluated with five different encoder structures. For all models, it was observed that DenseNet encoders give the top-segmentation performance as they can share pieces of collective knowledge by densely connecting convolutional layers to their subsequent layers, therefore, preserving the information coming from the earlier layer through the output layer. The FPN model with DenseNet121 encoder holds the leading performance with 96.11% IoU, and 97.99% DSC.

Table 5. Performance Metrics (%) for Lung Region Segmentation Computed over Test (Unseen) Set with Three Network Models and Five Encoder Architectures.

Model	Encoder	Accuracy	IoU	DSC
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

The outputs of the top three networks compared with the ground-truth are shown in Figure 11. 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: mild, moderate, or severe. This elegant performance is empowered by the large COVID-QU dataset (over 33k samples), which encapsulates CXR samples with different quality, resolution, and SNR levels from COVID-19, non-COVID-19 and normal classes. Therefore, the constructed benchmark dataset can 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 [41] and Shenzhen [42] CXR lung mask dataset which comprise of medium and high-quality X-ray images from normal and TB classes. Therefore, previous segmentation approaches were falling in unseen scenarios, such as severe infection or low-quality images [28].

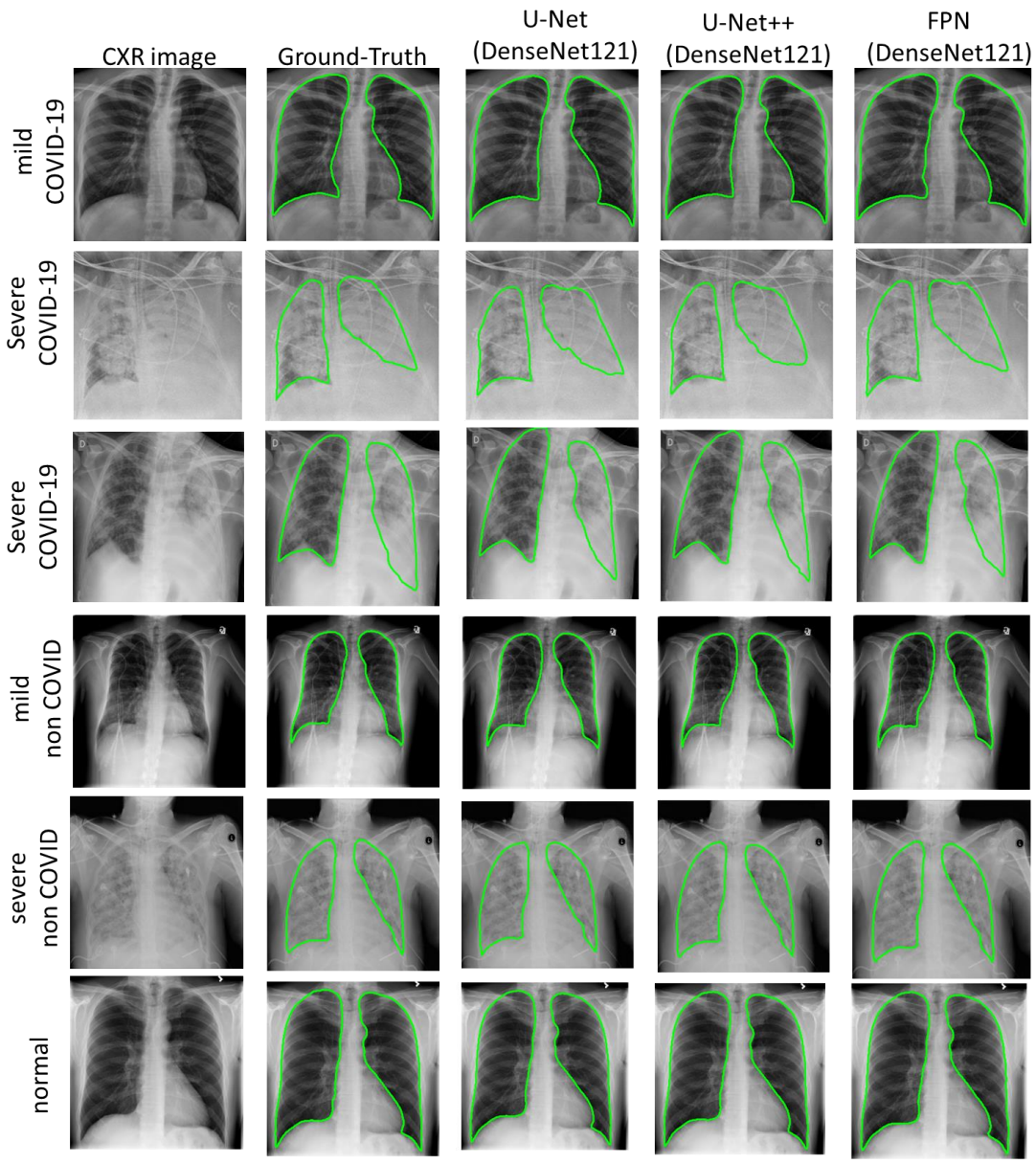

Figure 11. Qualitative evaluation of generated lung masks by top three networks. CXR image (1st column), ground truth (2nd column), and the lung masks of the top three networks (columns 3-5).

4.2.2.2 Infection Segmentation Results

The infection segmentation model was first evaluated over two different configurations: cascaded and parallel segmentation. For the cascaded scheme, first lung region was segmented using the lung segmentation model, then the segmented CXR

was fed to the infection segmentation model. While 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 (Table 6). 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.

Table 6. Performance Metrics (%) for COVID-19 Infected Region Segmentation Using Two Types of Inputs: Plain CXR, and Segmented CXR

Model	Encoder	Input	Accuracy	IoU	DSC
FPN	DenseNet161	Plain CXR	97.95 ± 0.81	81.89 ± 2.21	87.08 ± 1.93
		Segmented CXR	97.99 ± 0.81	81.86 ± 2.21	86.84 ± 1.94

The performance of the infection segmentation models is presented in Table 7. U-Net++ model with DenseNet121 encoder showed the best performance with IoU and DSC values of 83.05% and 88.21%, respectively. Besides, InceptionV4 encoder has achieved the highest performance among FPN models with 83.08% IoU and 88.13% DSC. In contrast, the shallowest encoder, ResNet18 presented the leading performance among U-Net models with IoU and DSC values of 82.92% and 88.1%, respectively.

Table 7. Performance Metrics (%) for COVID-19 Infected Region Segmentation Computed over Test (Unseen) Set with Three Network Models, and Five Encoder Architectures.

Model	Encoder	Accuracy	IoU	DSC
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

Figure 12(a) shows the robustness of top-three networks to reliably segment COVID-19 infections of various shapes (small, medium, or large infection) with different severity levels (mild, moderate, severe, or critical infection). 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. Figure 12(b) shows infection localization and severity grading of COVID-19 pneumonia for a 42-year female patient on the 1st day (admission to hospital), 2nd day, and 3rd day using the proposed COVID-19 recognition system.

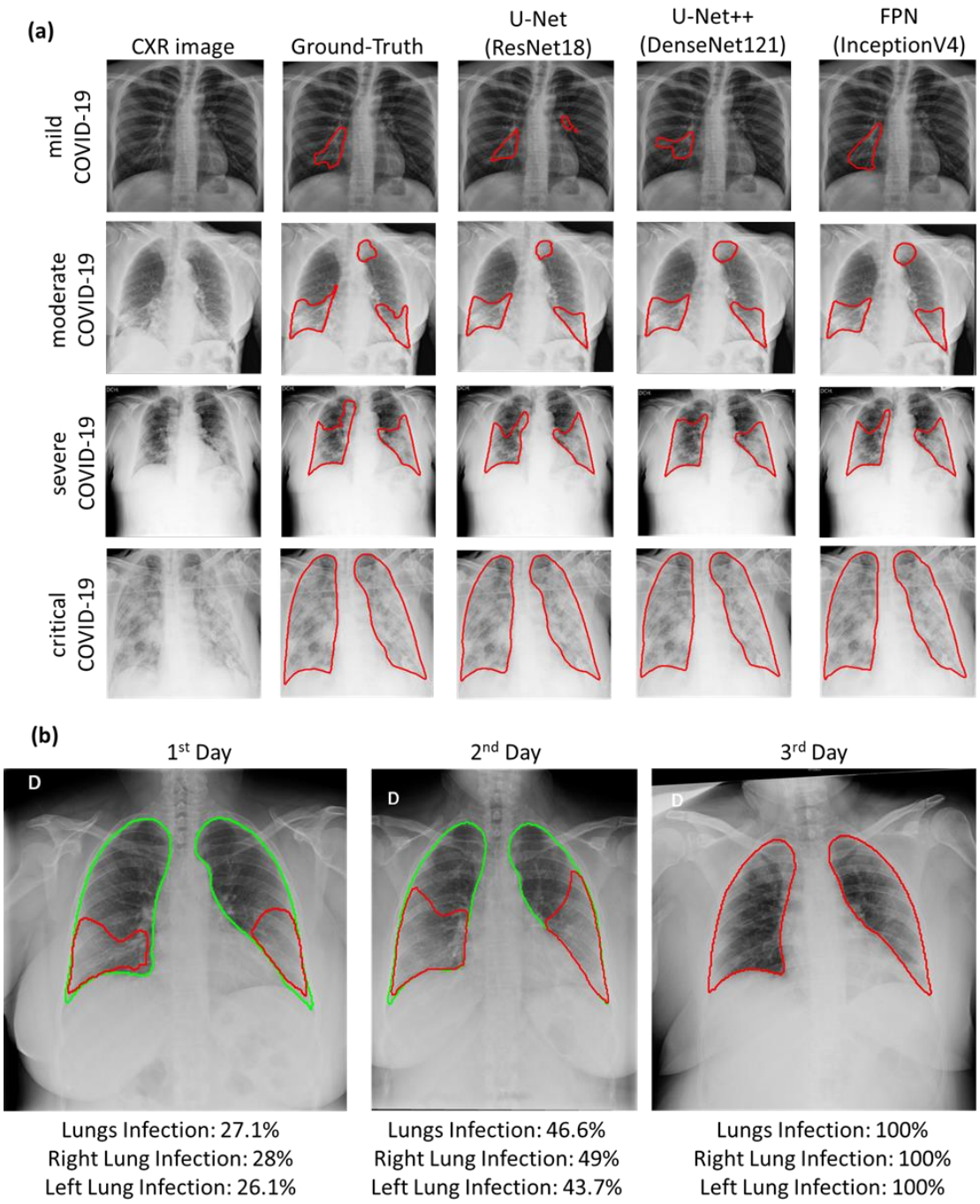

Figure 12. (a) Qualitative evaluation of generated infection masks by top three networks. CXR image (1st column), ground truth (2nd column), and the infection masks of the top three networks (columns 3-5). (b) Infection localization and severity grading of COVID-19 pneumonia for a 42-year female patient on the 1st, 2nd, and 3rd days using the proposed system.

4.2.2.3 COVID-19 Detection Results

The performance of infection segmentation networks for COVID-19 detection from the CXR images is presented in Table 8. 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 99.66% sensitivity. Similarly, all models showed elegant specificity values (>97%), where U-Net++ with ResNet18 backbone achieved the top performance with 100% specificity, indicating the absence of false alarm rate.

Table 8. COVID-19 Detection Performance Results (%) Computed Over Test (Unseen) Set with Three Network Models and Five Encoder Architectures.

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

4.2.2.4 Classification Results

The classification network was used for a 3-class recognition scheme to classify CXR images as COVID-19, non-COVID-19, or normal. However, this network will be

removed in future studies once ground-truth infection masks are created for non-COVID-19 cases. Therefore, the binary masks generated by infection segmentation models can be extended to a 3-channel infection mask, where the channels denote: background, non-COVID-19 lesion, and COVID-19 lesions. Thus, the multi-channel infection mask can be used to distinguish between COVID-19, non-COVID-19, or normal images.

The recognition was first evaluated on two schemes: plain CXR and segmented CXR classification. InceptionV4 was trained and evaluated on both schemes (Table 9). Similar findings to study I were observed, where the performance degrades with segmented CXR. An Overall sensitivity of 95.53% was achieved with plain CXR, while 91.95% sensitivity was achieved with segmented CXR. However, the results are more reliable with segmented CXR as the network learns from the main regions of interest, lung areas (Figure 13). Therefore, the cascaded configuration was used for the classification scheme with the concatenation of lung segmentation and classification models.

Table 9. Performance Metrics (%) for the 3-Class Recognition Scheme Using Two Types of Inputs: Plain CXR, and Segmented CXR

Model	Input	Accuracy	Precision	Sensitivity	F1-score	Specificity
InceptionV4	Plain CXR	97.07 ± 0.4	95.6 ± 0.49	95.53 ± 0.49	95.55 ± 0.49	97.84 ± 0.35
	Segmented CXR	94.65 ± 0.54	91.98 ± 0.65	91.95 ± 0.65	91.96 ± 0.65	96 ± 0.47

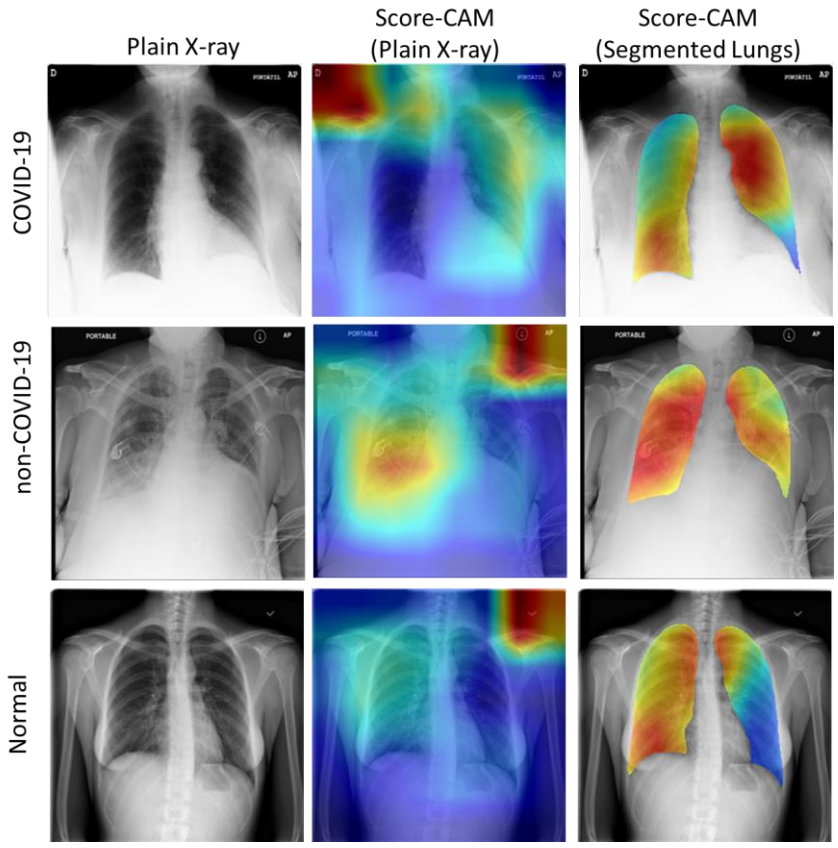

Figure 13. Examples of probabilistic saliency maps for COVID-19, non-COVID-19 and normal cases: (A) Plain CXR image, (B) Score-CAM for plain CXR inferred by InceptionV4 network, and (C) Score-CAM for segmented CXR inferred by InceptionV4 network.

The performance of different classification models is presented in Table 10. DenseNet models and InceptionV4 showed better performance compared to ResNet family models. InceptionV4 achieved the best overall sensitivity of 91.95%, with per class sensitivities of 91.52%, 93.21%, and 91.12 for COVID-19, non-COVID-19, and normal classes, respectively. DenseNet161 has achieved the highest sensitivity for the COVID-19 class, 92.82%. However, it showed lower sensitivity to non-COVID-19 cases, 87.71%.

Comparing the best achieved COVID-19 sensitivity from the classification scheme, 91.52%, with the achieved sensitivities using the COVID-19 detection scheme

based on infection masks, >97%, we can clearly see around 6% drop in performance. However, as mentioned previously, the infection segmentation models were trained and evaluated on a subset of COVID-QU dataset, 2913 CXR images, with mild, moderate, and severe cases. On the other hand, the full dataset was utilized for the classification task, including 11956 COVID-19 CXR images with different severity levels, including early cases with minimum or no signs of COVID-19 pneumonia. Therefore, the network can confuse those cases with normal or non-COVID-19 cases. Thus, the overall performance degrades.

Table 10. Performance Metrics (%) for the 3-Class Recognition Scheme Computed over Test (Unseen) Set with Four Network Models.

Model	Class	Accuracy	Precision	Sensitivity	F1-score	Specificity
ResNet18	Normal	92.1 ± 0.64	86.03 ± 0.82	89.49 ± 0.73	87.73 ± 0.78	93.31 ± 0.59
	COVID-19	92.37 ± 0.63	93.59 ± 0.58	84.13 ± 0.87	88.61 ± 0.76	96.86 ± 0.41
	non-COVID	92.84 ± 0.61	86.67 ± 0.81	92.68 ± 0.62	89.57 ± 0.73	92.92 ± 0.61
	Overall	92.44 ± 0.63	88.91 ± 0.75	88.66 ± 0.75	88.65 ± 0.75	94.43 ± 0.55
ResNet50	Normal	93.56 ± 0.58	88.34 ± 0.76	91.68 ± 0.66	89.98 ± 0.71	94.43 ± 0.55
	COVID-19	94.05 ± 0.56	91.5 ± 0.66	91.65 ± 0.66	91.57 ± 0.66	95.36 ± 0.5
	non-COVID	93.18 ± 0.6	91.28 ± 0.67	87.84 ± 0.78	89.53 ± 0.73	95.83 ± 0.48
	Overall	93.61 ± 0.58	90.43 ± 0.7	90.39 ± 0.7	90.39 ± 0.7	95.22 ± 0.51
DenseNet121	Normal	94.45 ± 0.54	88.54 ± 0.76	94.63 ± 0.54	91.48 ± 0.66	94.36 ± 0.55
	COVID-19	94.59 ± 0.54	95.97 ± 0.47	88.39 ± 0.76	92.02 ± 0.64	97.97 ± 0.34
	non-COVID	94.55 ± 0.54	91.02 ± 0.68	92.72 ± 0.62	91.86 ± 0.65	95.46 ± 0.5
	Overall	94.53 ± 0.54	91.98 ± 0.65	91.79 ± 0.65	91.8 ± 0.65	96 ± 0.47
DenseNet161	Normal	93.97 ± 0.57	88.28 ± 0.77	93.27 ± 0.6	90.71 ± 0.69	94.3 ± 0.55
	COVID-19	94.73 ± 0.53	92.28 ± 0.63	92.82 ± 0.61	92.55 ± 0.62	95.77 ± 0.48
	non-COVID	93.83 ± 0.57	93.3 ± 0.59	87.71 ± 0.78	90.42 ± 0.7	96.87 ± 0.41
	Overall	94.19 ± 0.56	91.36 ± 0.67	91.27 ± 0.67	91.26 ± 0.67	95.67 ± 0.48
InceptionV4	Normal	94.27 ± 0.55	90.74 ± 0.69	91.12 ± 0.68	90.93 ± 0.68	95.72 ± 0.48
	COVID-19	94.84 ± 0.53	93.72 ± 0.58	91.52 ± 0.66	92.61 ± 0.62	96.65 ± 0.43
	non-COVID	94.8 ± 0.53	91.3 ± 0.67	93.21 ± 0.6	92.25 ± 0.64	95.59 ± 0.49
	Overall	94.65 ± 0.54	91.98 ± 0.65	91.95 ± 0.65	91.96 ± 0.65	96 ± 0.47

4.2.2.5 Computational Complexity Analysis

Table 11 compares the segmentation models in terms of computational

inference time and the number of trainable parameters. The results present the running time per CXR sample. It can be seen that FPN and U-Net models are faster than U-Net ++ models, due to their shallow and close structures. FPN with ResNet18 encoder is the fastest network taking up to 5.74 ms per image. In contrast, U-Net++ model is the slowest with the largest number of trainable parameters. The most computationally demanding model is UNet++ with InceptionV4 encoder with 59.35M 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.

Table 11. The Number of Trainable Parameters of The Segmentation 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	32.50 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

The computational complexity and inference time of the classification models is presented in Table 12. ResNet18 has achieved the fastest speed performance with 3.83 ms, while DenseNet161 is the slowest with 27.40ms. Therefore, the overall inference time for the full system is <100ms, where lung and infection segmentation

models are used in parallel, and the classification system is used in a sequential manner, cascaded with the lung segmentation model. Moreover, for systems with limited computational capabilities, where one model can run at a time, the three models can be used in sequence. This will increase the inference time, <150ms. However, we can still say that the full system can be used for real-time clinical applications.

Table 12. The Number of Trainable Parameters of The Classification Models with Their Inference Time (ms) per CXR Sample.

Model	Trainable parameters	Inference Time (ms)
ResNet18	11.18 M	3.83
ResNet50	23.51 M	8.49
DenseNet121	6.96 M	20.42
DenseNet161	26.48 M	27.40
InceptionV4	41.15 M	23.98

To the best of our knowledge, this is the first work to utilize both lung and infection segmentation to detect, localize and quantify COVID-19 infection from X-ray images. Therefore, assisting medical doctors to better diagnose the severity of COVID-19 pneumonia and follow up the progression of the disease.

CHAPTER 5: CONCLUSION AND FUTURE WORK

Early identification and isolation of highly infectious COVID-19 cases play a vital role in 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. 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 large cohort dataset is not guaranteed. In addition, they showed limited performance in infection localization and severity grading of COVID-19 pneumonia. In this thesis work, we propose a robust system to segment the lung, detect, localize, and quantify COVID-19 infections from CXR images. To accomplish this, we compiled the largest CXR dataset, COVID-QU, which consists of 11,956 COVID-19, 11,263 non-COVID-19 pneumonia, and 10,701 normal, 134 SARS-CoV, and 144 MERS-CoV images. Moreover, we constructed ground-truth lung segmentation masks for the benchmark dataset using an elegant collaborative human-machine approach, which can save valuable human labor time and minimize subjectivity in the annotation process. The released dataset can help researchers to investigate deep ConvNet on a comparatively larger dataset, which can provide more reliable solutions for COVID-19 and other lung pathology problems. An extensive set of experiments using state-of-the-art ConvNets over COVID-QU dataset 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 sizes and shapes, achieving IoU and DSC values of 83.05% and 88.21%, respectively. Furthermore, two classification schemes were tackled:

(i) COVID-19 recognition from non-COVID-19 infections, and normal cases, (ii) COVID-19 recognition from other coronaviruses, SARS, and MERS. For the first classification scheme, the best network achieved sensitivities of 96.94%, 79.68%, and 90.26% for classifying COVID-19, MERS, and SARS images, respectively. For the second classification scheme we achieved sensitivities of 91.52%, 93.21%, and 91.12 for COVID-19, non-COVID, and normal classes, respectively.

In the future, we plan to modify our system to localize non-COVID-19 infections as well by creating ground-truth infection masks for non-COVID-19 CXR images using the collaborative human-machine approach. In addition, 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, Self-Organized Operational Neural Networks [78, 79].

REFERENCES

[1] E. Prompetchara, C. Ketloy, and T. J. A. P. J. A. I. Palaga, "Immune responses in COVID-19 and potential vaccines: Lessons learned from SARS and MERS epidemic," vol. 38, no. 1, pp. 1-9, 2020.

[2] S. Kumar, V. K. Maurya, A. K. Prasad, M. L. Bhatt, and S. K. J. V. Saxena, "Structural, glycosylation and antigenic variation between 2019 novel coronavirus (2019-nCoV) and SARS coronavirus (SARS-CoV)," pp. 1-9, 2020.

[3] World Health Organization. (2020). *WHO Coronavirus Disease (COVID-19) Dashboard*. Available: https://covid19.who.int/?gclid=Cj0KCQjwtZH7BRDzARIsAGjbK2ZXWRpJROEl97HGmSOx0_ydkVbc02Ka1FlcysGjEI7hnaIeR6xWhr4aAu57EALw_wcB

[4] E. Mahase, "Coronavirus: covid-19 has killed more people than SARS and MERS combined, despite lower case fatality rate," ed: British Medical Journal Publishing Group, 2020.

[5] W.H.O. (2020). *SARS (Severe Acute Respiratory Syndrome)*. Available: https://www.who.int/ith/diseases/sars/en/

[6] H. Gao, H. Yao, S. Yang, and L. J. F. o. m. Li, "From SARS to MERS: evidence and speculation," vol. 10, no. 4, pp. 377-382, 2016.

[7] W.H.O., "Middle East respiratory syndrome coronavirus (MERS-CoV)," 2019.

[8] W.H.O. (2020). *WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020*. Available: https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020

[9] 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," *Reviews in Medical Virology,* https://doi.org/10.1002/rmv.2112 vol. 30, no. 4, p. e2112, 2020/07/01 2020.

[10] T. J. T. I. J. o. P. Singhal, "A review of coronavirus disease-2019 (COVID-19)," pp. 1-6, 2020.

[11] C. Sohrabi *et al.*, "World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19)," 2020.

[12] P. Kakodkar, N. Kaka, and M. J. C. Baig, "A comprehensive literature review on the clinical presentation, and management of the pandemic coronavirus disease 2019 (COVID-19)," vol. 12, no. 4, 2020.

[13] Y. Li *et al.*, "Stability issues of RT-PCR testing of SARS-CoV-2 for hospitalized patients clinically diagnosed with COVID-19," 2020.

[14] A. Tahamtan and A. Ardebili, "Real-time RT-PCR in COVID-19 detection: issues affecting the results," ed: Taylor & Francis, 2020.

[15] J. Xia, J. Tong, M. Liu, Y. Shen, and D. J. J. o. m. v. Guo, "Evaluation of coronavirus in tears and conjunctival secretions of patients with SARS-CoV-2 infection," vol. 92, no. 6, pp. 589-594, 2020.

[16] 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," p. 200642, 2020.

[17] S. Salehi, A. Abedi, S. Balakrishnan, and A. J. A. J. o. R. Gholamrezanezhad, "Coronavirus disease 2019 (COVID-19): a systematic review of imaging findings in 919 patients," pp. 1-7, 2020.

[18] Y. Fang *et al.*, "Sensitivity of chest CT for COVID-19: comparison to RT-PCR," p. 200432, 2020.

[19] D. J. Brenner and E. J. J. N. E. J. o. M. Hall, "Computed tomography—an increasing source of radiation exposure," vol. 357, no. 22, pp. 2277-2284, 2007.

[20] F. Shi *et al.*, "Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for covid-19," 2020.

[21] T. Ozturk *et al.*, "Automated detection of COVID-19 cases using deep neural networks with X-ray images," vol. 121, p. 103792, 2020.

[22] I. D. Apostolopoulos, T. A. J. P. Mpesiana, and E. S. i. Medicine, "Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks," p. 1, 2020.

[23] L. Wang, Z. Q. Lin, and A. J. S. R. Wong, "Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images," vol. 10, no. 1, pp. 1-12, 2020.

[24] A. Waheed, M. Goyal, D. Gupta, A. Khanna, F. Al-Turjman, and P. R. J. I. A. Pinheiro, "Covidgan: Data augmentation using auxiliary classifier gan for improved covid-19 detection," vol. 8, pp. 91916-91923, 2020.

[25] I. D. Apostolopoulos, S. I. Aznaouridis, M. A. J. J. o. M. Tzani, and B. Engineering, "Extracting possibly representative COVID-19 Biomarkers from X-Ray images with Deep Learning approach and image data related to Pulmonary Diseases," p. 1, 2020.

[26] M. E. Chowdhury *et al.*, "Can AI help in screening viral and COVID-19 pneumonia?," 2020.

[27] A. Haghanifar, M. M. Majdabadi, and S. J. a. p. a. Ko, "COVID-CXNet: Detecting COVID-19 in Frontal Chest X-ray Images using Deep Learning," 2020.

[28] Y. Oh, S. Park, and J. C. J. I. T. o. M. I. Ye, "Deep learning covid-19 features

on cxr using limited training data sets," 2020.

[29] S. Rajaraman, J. Siegelman, P. O. Alderson, L. S. Folio, L. R. Folio, and S. K. J. a. p. a. Antani, "Iteratively Pruned Deep Learning Ensembles for COVID-19 Detection in Chest X-rays," 2020.

[30] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," vol. 542, no. 7639, pp. 115-118, 2017.

[31] H. Dong, G. Yang, F. Liu, Y. Mo, and Y. Guo, "Automatic brain tumor detection and segmentation using U-Net based fully convolutional networks," in *annual conference on medical image understanding and analysis*, 2017, pp. 506-517: Springer.

[32] L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, R. McBride, and W. J. S. r. Sieh, "Deep learning to improve breast cancer detection on screening mammography," vol. 9, no. 1, pp. 1-12, 2019.

[33] D. Ardila *et al.*, "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography," vol. 25, no. 6, pp. 954-961, 2019.

[34] A. Tahir *et al.*, "Coronavirus: Comparing COVID-19, SARS and MERS in the eyes of AI," *arXiv preprint arXiv:1706.05587*, 2020.

[35] P. Rajpurkar *et al.*, "Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists," vol. 15, no. 11, p. e1002686, 2018.

[36] 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/

[37] V. Chouhan *et al.*, "A novel transfer learning based approach for pneumonia

detection in chest X-ray images," vol. 10, no. 2, p. 559, 2020.

[38] P. Lakhani and B. J. R. Sundaram, "Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks," vol. 284, no. 2, pp. 574-582, 2017.

[39] M. Yamac, M. Ahishali, A. Degerli, S. Kiranyaz, M. E. Chowdhury, and M. J. a. p. a. Gabbouj, "Convolutional Sparse Support Estimator Based Covid-19 Recognition from X-ray Images," 2020.

[40] M. Ahishali *et al.*, "A Comparative Study on Early Detection of COVID-19 from Chest X-Ray Images," 2020.

[41] S. Jaeger *et al.*, "Automatic tuberculosis screening using chest radiographs," vol. 33, no. 2, pp. 233-245, 2013.

[42] S. Candemir *et al.*, "Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration," vol. 33, no. 2, pp. 577-590, 2013.

[43] F. Shi *et al.*, "Large-scale screening of covid-19 from community acquired pneumonia using infection size-aware classification," 2020.

[44] A. Degerli *et al.*, "COVID-19 Infection Map Generation and Detection from Chest X-Ray Images," 2020.

[45] World Health Organization. (2004). *China's latest SARS outbreak has been contained, but biosafety concerns remain – Update 7*. Available: https://www.who.int/csr/don/2004_05_18a/en/

[46] World Health Organization. (2020). *Middle East respiratory syndrome coronavirus (MERS-CoV) – Saudi Arabia*. Available: https://www.who.int/csr/don/05-may-2020-mers-saudi-arabia/en/

[47] A. Tahir *et al.*, "Deep Learning for Reliable Classification of COVID-19, MERS, and SARS from Chest X-Ray Images," *Cognitive Computation*, 2021.

[48] 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

[49] GitHub. (2020). *covid-19-image-repository*. Available: https://github.com/ml-workgroup/covid-19-image-repository/tree/master/png

[50] Eurorad. Available: https://www.eurorad.org/

[51] GitHub. (2020). *covid-chestxray-dataset*. Available: https://github.com/ieee8023/covid-chestxray-dataset

[52] SIRM. (2020). *COVID-19 DATABASE*. Available: https://www.sirm.org/category/senza-categoria/covid-19/

[53] Kaggle. (2020). *COVID-19 Radiography Database*. Available: https://www.kaggle.com/tawsifurrahman/covid19-radiography-database

[54] GitHub. (2020). *COVID-CXNet*. Available: https://github.com/armiro/COVID-CXNet

[55] Kaggle. (2018). *RSNA Pneumonia Detection Challenge*. Available: https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data

[56] Kaggle. (2018). *Chest X-Ray Images (Pneumonia)*. Available: https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia

[57] 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/

[58] J. P. Cohen, P. Morrison, and L. J. a. p. a. Dao, "COVID-19 image data collection," 2020.

[59] J.-Y. Rhee, G. Hong, and K. M. J. J. j. o. i. d. Ryu, "Clinical implications of five

cases of Middle East respiratory syndrome coronavirus infection in South Korea Outbreak," p. JJID. 2015.445, 2016.

[60] L. Grinblat, H. Shulman, A. Glickman, L. Matukas, and N. J. R. Paul, "Severe acute respiratory syndrome: radiographic review of 40 probable cases in Toronto, Canada," vol. 228, no. 3, pp. 802-809, 2003.

[61] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015, pp. 234-241: Springer.

[62] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and 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.

[63] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117-2125.

[64] R. Maini and H. Aggarwal, "A comprehensive review of image enhancement techniques," *arXiv preprint arXiv:1003.4053*, 2010.

[65] R. Maini and H. J. a. p. a. Aggarwal, "A comprehensive review of image enhancement techniques," 2010.

[66] S. M. Pizer *et al.*, "Adaptive histogram equalization and its variations," vol. 39, no. 3, pp. 355-368, 1987.

[67] J. B. Zimmerman, S. M. Pizer, E. V. Staab, J. R. Perry, W. McCartney, and B. C. J. I. T. o. M. I. Brenton, "An evaluation of the effectiveness of adaptive histogram equalization for contrast enhancement," vol. 7, no. 4, pp. 304-312,

1988.

[68] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. J. a. p. a. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size," 2016.

[69] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

[70] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700-4708.

[71] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818-2826.

[72] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017, vol. 31, no. 1.

[73] Image-net.org, "ImageNet."

[74] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618-626.

[75] H. Wang *et al.*, "Score-cam: Score-weighted visual explanations for convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 24-25.

[76] A. M. Tahir *et al.*, "COVID-19 Infection Localization and Severity Grading

from Chest X-ray Images," *Nature Scientific Reports* 2021.

[77] Pytorch.org. *PyTorch*. Available: https://pytorch.org/

[78] J. Malik, S. Kiranyaz, and M. J. N. N. Gabbouj, "Self-organized operational neural networks for severe image restoration problems," vol. 135, pp. 201-211, 2021.

[79] S. Kiranyaz, J. Malik, H. B. Abdallah, T. Ince, A. Iosifidis, and M. J. a. p. a. Gabbouj, "Self-Organized Operational Neural Networks with Generative Neurons," 2020.