

**Artificial Intelligence to Accelerate COVID-19
Identification from Chest X-rays**

**A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY**

Dyah Adila

**IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE**

Ju Sun

May, 2021

© Dyah Adila 2021
ALL RIGHTS RESERVED

Acknowledgements

I would like to express the deepest appreciation to my supervisor and mentor, Professor Ju Sun: he continually conveyed a genuine passion for research and a joyous spirit in advising his students. Without his guidance and persistent help, this thesis would not have been possible.

My sincere thanks also go to my committee members: Professor Daniel Boley and Professor Erich Kummerfeld, and my faculty mentors: Professor Vipin Kumar and Professor Shashi Shekhar, for their assistance and guidance throughout this project and my master's study.

Thanks and appreciation are also extended to Dr. Christopher Tignanelli and every collaborator from M Health Fairview for the resources and medical insights throughout the project.

Special thanks to my research lab colleagues and friends: Le Peng and Taihui Li. It was a great pleasure working with them, and I appreciate their rigor in working, ideas, help, and good humor.

These acknowledgments would not be complete without mentioning my parents and my amazing friends: Deka Auliya, Divya Nairy, and Kshitij Tayal, whose support has been an invaluable source of strength during my master's study.

Abstract

Importance: Clinical signs and symptoms for COVID-19 remain the mainstay of early diagnosis and initial management in the emergency department (ED) and inpatient setting at many hospitals due to delays in obtaining results of PCR testing and limitations in access to rapid antigen testing. The majority of many patients with COVID-19 will present with respiratory symptoms necessitating a chest x-ray (CXR) as a routine part of screening. An AI-based model to predict COVID-19 likelihood from CXR findings can serve as an important and immediate adjunct to accelerate clinical decision making.

Objective: To develop a robust AI-based diagnostic model to identify CXRs with COVID-19 compared with all non-COVID-19 CXRs.

Setting: Labeled frontal CXR images (samples of COVID-19 and non-COVID-19) from the M Health Fairview (Minnesota, USA), Valencian Region Medical ImageBank (Spain), MIMIC-CXR, Open-I 2013 Chest X-ray Collection, GitHub COVID-19 Image Data Collection (International).

Main Outcome and Measure: Model performance assessed via Area under the Receiver Operating Curve (AUROC) and Area Under the Precision and Recall Curve (AUPRC).

Results: Patients with COVID-19 had significantly higher COVID-19 Diagnostic Scores than patients without COVID-19 on both real-time electronic health record and external (non-publicly available) validation. The model performed well across all four methods for model validation with AUROCs ranging between 0.7 – 0.96 and high PPV and specificity. The model performed had improved discrimination for patients with “severe” as compared to “moderate” COVID-19 disease. The model had unrealistic performance using publicly available databases, reflecting the inherent limitations in many previously developed models relying on publicly available data for training and validation.

Conclusions and Relevance: AI-based diagnostic tools may serve as an adjunct, but not replacement, to support COVID-19 diagnosis which largely hinges on exposure history, signs, and symptoms. Future research should focus on optimizing discrimination of “mild” COVID-19 from non-COVID-19 image findings.

Contents

Acknowledgements	i
Abstract	ii
List of Tables	vi
List of Figures	vii
1 Introduction	1
1.1 Motivations	1
1.2 Objectives and Context	3
2 Background	4
2.1 Deep Neural Networks	4
2.1.1 Convolutional Neural Networks (CNN)	5
2.1.2 CNN Applications in Medical Image Analysis	9
2.2 Generative Adversarial Network	11
2.2.1 What is GAN	11
2.2.2 How GAN Works	12
2.2.3 Conditional GAN	14
3 Methodology	15
3.1 Data Acquisition	15
3.1.1 M Health Fairview Datasets	15
3.1.2 Publicly Available Datasets	16

3.2	Image Pre-processing	17
3.3	Model Development (Training)	17
3.3.1	Lung Segmentation	18
3.3.2	Outlier Detection	19
3.3.3	Feature Extraction and Classification	19
4	Experimental Results	21
4.1	Validation Method	21
4.2	Results	22
4.2.1	Prospective Validation using M Health Fairview CXR	22
4.2.2	Validation using publicly available COVID-19 CXRs	24
5	Discussion and Conclusion	27
5.1	Summary of Key Finding	27
5.2	Results Interpretation	27
5.2.1	M Health Fairview Evaluation Results	27
5.2.2	Public Dataset Evaluation Results	28
5.3	Results Implications	28
5.4	Study Limitations	29
5.5	Conclusion	29
	References	30

List of Tables

4.1	Prospective Validation for July 2020 using M Health Fairview CXRs . . .	22
4.2	Prospective Validation for July 2020 using M Health Fairview CXRs for 0.04 – 0.4 thresholds	23
4.3	Prospective Validation for July 2020 using M Health Fairview CXRs for patients with “severe” COVID-19 disease	24
4.4	Prospective Validation for July 2020 using M Health Fairview CXRs for patients with “moderate” COVID-19 disease	24
4.5	External Validation using publicly available COVID-19 CXRs	25
4.6	Publicly-Available Validation using Github COVID-19 database	26
4.7	Publicly Available (BIMCV) COVID-19 CXR validation	26

List of Figures

2.1	Supervised Learning Framework [1]	6
2.2	Fully Connected Layer	7
2.3	CNN Network Example	8
2.4	Semantic segmentation illustration	11
2.5	Generative model vs Discriminative model [2]	12
2.6	GAN Structure [3]	13
2.7	Conditional GAN architecture [4]	14
3.1	Overview of COVID-19 Diagnostic Model Pipeline	18
4.1	Distribution of COVID-19 Diagnostic Scores (X-axis) for patients with PCR confirmed positive COVID-19 (purple bars) and non-COVID-19 patients (green bars) during the month of July 2020.	23
4.2	Distribution of COVID-19 Diagnostic Scores (X-axis) for patients with COVID-19 (purple bars) and non-COVID-19 patients (green bars) from publically available datasets, prevalence	25

Chapter 1

Introduction

1.1 Motivations

The coronavirus disease 2019 (COVID-19) outbreak started was first reported on December 31, 2019, when Chinese authorities described a cluster of cases of pneumonia linked to the Huanan Seafood Wholesale Market in the Chinese city of Wuhan, Hubei Province [5]. Shortly thereafter, the virus spread rapidly within China and across the globe, with large sustained outbreaks on all six continents. The World Health Organization designated COVID-19 a global pandemic on March 11, 2020, and as of November 23rd 2020, there have been 59,515,380 confirmed cases and 1,402,032 deaths worldwide [6]. The rapid and sustained transmission of the virus has overwhelmed health systems worldwide and resulted in a shortage of critical equipment and supplies [7]. In the absence of an effective treatment, rapid identification and isolation of infected individuals has emerged as a key tool in curtailing the pandemic [8].

The mainstay of COVID-19 diagnosis is nucleic acid testing of upper or lower respiratory tract swab specimens using reverse transcription polymerase chain reaction (RT-PCR)[9]. However, due to limited RT-PCR capacity and shortages of testing supplies, RT-PCR today remains a bottleneck and delay for COVID-19 diagnosis. While many tertiary centers have acquired the capacity for in-house testing, most still need to send their samples to external laboratories. Even when readily available, turnaround times may exceed to 24- 48 hours, especially and possibly more in low resource settings [10]. While recent rapid PCR kits providing results in as little as 5 minutes show

promise, they are still not widely available to the healthcare system at large. Diagnostic Variability in the type of kit used, quality of sample collected, and transport limitations can significantly influence the reliability of the test, with recent studies reporting sensitivity as low as 60-71% [11]. RT-PCR samples may be obtained very early in the disease course, before the viral load is high enough to be reliably and accurately detected, and thus can be highly dependent on when the sample is taken. Furthermore, a patient's PCR result can be negative (or revert to negative) after the patient has mounted a successful recovery from the illness, which could cause a missed diagnosis of COVID-19.

To overcome the current limitations of testing, clinicians are increasingly relying on using more immediate clinical information to better manage patients suspected of harboring the infection, including first-line use of Chest computed tomography (CT) Scans for diagnosing COVID-19 in China [12] and chest radiographs (CXR) in the United Kingdom [10]. Previous studies have reported a high sensitivity (97-98%) for such an approach to identify COVID-19 [13], [14]. Despite its promise of high clinical utility, widespread use of CT scans as a means of rapidly identifying COVID-19 disease has limitations. This approach is resource-intensive, requires additional personal protective equipment (PPE), requires transportation of potentially infectious patients to the scanner, and is sensitive to human factors (delays in image reading and potential for misdiagnosis)[15]. As such, the American College of Radiology and the CDC recommend that CT should not be used for screening or as a first-line test to diagnose COVID-19 at this time [16].

The need persists for a readily available test to aid the emergency department, urgent care, and inpatient providers in identifying a patient at high-risk for COVID-19. Given widespread availability, low cost, and convenience, CXRs have emerged as the frontline diagnostic imaging test when combined with clinical history and key blood markers [10]. CXR is a safe and readily available resource in all hospitals and can be performed at the bedside. Automated methods to screen CXRs for COVID-19 in real-time, prioritize potential positive images for radiologist review, and deliver AI-enabled decision support to early hospital responders may improve provider accuracy in detecting and isolating potential COVID-19 patients, limit hospital spread of the virus, and save critical healthcare resources. In this study, we hypothesize that an AI-model using CXRs can accurately identify COVID-19 positive patients as well as differentiate

them from other pulmonary pathologies.

1.2 Objectives and Context

This study aims to develop a robust Artificial Intelligence (AI) model to help radiologists distinguish COVID-19 positive vs. negative cases given chest x-ray images. To ensure the model's generalizability, data for model development and evaluation comes from diverse sources, including private hospital data and open-source data.

Whilst medical considerations are also a vital component of this study, the purpose of this thesis is to satisfy the Computer Science graduate degree requirement. Thus, this thesis is focused solely on the technical background and aspects concerning the AI model.

Chapter 2

Background

Summary

This chapter will discuss the background knowledge to understand the methodology outlined in Chapter 3. First, we will discuss the basic high-level concept of deep neural network models, followed by the current state of the art of their usage in medical image analysis. Second, we will discuss the concept of the generative adversarial network (GAN).

2.1 Deep Neural Networks

Over the past few years, we have witnessed tremendous advances in machine learning using deep neural networks. Driven by the rapid increase in computational resources and available data, these neural network methods become state of the art in fundamental tasks in computer vision, such as image classification, recognition, and segmentation [17], [18], [19]. As these neural network models become increasingly powerful, their usage is becoming more popular in safety-critical domains, especially in medical image analysis [20].

There have been extensive reviews covering the use of deep learning methods in the medical domains. [21] covers deep neural network applications for fundamental tasks in the medical image analysis, such as object detection, image segmentation, and registration. [22] identifies the achievements made and the key challenges faced in this

line of research, along with the promising future direction. [23] focused on Convolutional Neural Network (CNN), the most commonly used neural network architecture for visual perception tasks, used in medical image analysis. [24] provided a comprehensive review of the state-of-the-art CNN techniques in this domain.

The methodology discussed in chapter 3 uses CNN architecture, on which this sub-chapter will mainly focus on. This sub-chapter is structured as follows: the first part serves as an introduction to the Convolutional Neural Networks architecture. The second part outlines the common practice of using CNN in medical image analysis, namely transfer learning. The third part presents the kinds of tasks CNNs are useful for, especially in the medical image analysis domain (e.g., image classification and segmentation).

2.1.1 Convolutional Neural Networks (CNN)

Supervised Learning

Before going in-depth to the CNN architecture, it is helpful to briefly discuss the supervised learning framework, the most standard method to train CNN models. In this framework, we are given pairs of data instances and an associated label for each instance, which can be illustrated as (data instance, label) pairs. For example, in this study, the data instances are chest x-ray images, and the labels are whether it shows symptoms of COVID-19 (e.g., binary yes/no). *Training* the CNN model is finding the parameter values as such when it receives the data instance (e.g., the chest x-ray), it correctly assigns the corresponding label (e.g., COVID positive or negative). Finding these parameter values requires performing iterative optimization using these labels as a form of *supervision*.

Neural network models mainly operate in one of the two modes: (1) Training and (2) Evaluation. After finding the optimum parameter values that assign correct labels to the training data samples during the *training* phase, in *evaluation*, we stop performing updates to these parameters. Now, given data samples never seen during training, the network performs operations between the data sample and the *learned parameters* to output its predicted label. Figure 2.1 illustrates the supervised learning framework.

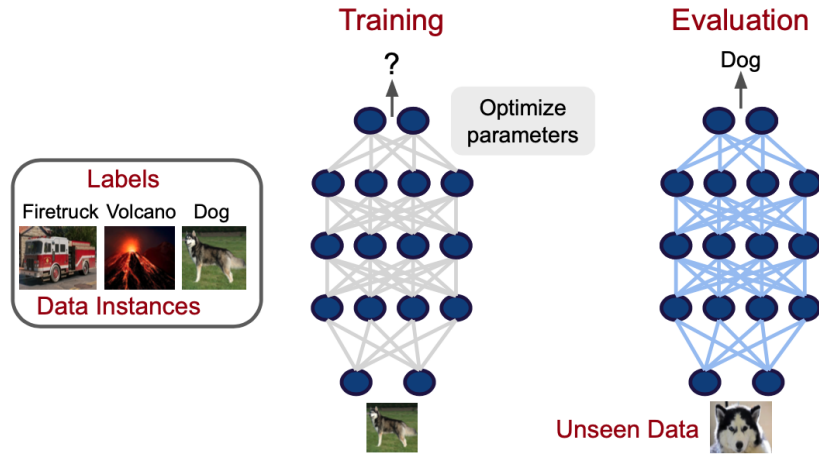


Figure 2.1: Supervised Learning Framework [1]

CNN Architecture

CNNs are arguably the most well-known family of neural networks when it comes to analyzing image data. They owe the popularity to the **convolutional layers**, which allow the neural network to share parameters across spatially close pixels within an image. This capability comes especially handy when it comes to extracting image features to feed into the classification layers. The extracted features automatically capture the encoded spatial information and relationship between image pixels, thus replacing the need for hand-crafted features. CNN architecture typically consists of convolution layers, pooling layers, and fully connected layers.

- **Convolutional layers.** Convolutional layer parameters consist of a set of learnable image filters, represented as 2-dimensional arrays. After receiving an input image, we slide each filter across the width and height of the image and compute the dot products between filter parameters and image pixel values at any position (i.e., the convolution operation). The output of this process is called the *activation maps*. Equation 2.1 shows the convolution operation for an image I of size $(m \times n)$ and filter K .

$$S(i, j) = (I * K)(i, j) = \sum_m \sum_n I(m, n) K(i - m, j - n) \quad (2.1)$$

The activation maps are then passed through an element-wise activation function to introduce non-linearity in the CNN weights. The most widely used activation function is the rectified linear unit (ReLU) [25]. The learned filter parameters at earlier layers will extract some type of visual features such as the object's edge or colors, and the filters at the deeper level of the network will learn to extract more high-level patterns such as the shape of the objects.

- **Pooling layers.** It is a common practice to insert Pooling layers in between successive Convolution layers in a CNN. This layer reduces the spatial size of image representation as we go deeper in the network, reducing the number of parameters and computation load. It takes the output of the convolution layer and performs pooling through a similar filter sliding operation. However, here, the pooling layer takes the maximum of the pixels within the area covered by the filters. For example, using filters of size 2×2 would be taking the largest of 4 numbers in the 2×2 region. There are other less-common types of pooling, such as average pooling and global pooling.
- **Fully Connected layers.** After getting the 2-dimensional image representation through a series of successive convolution and pooling operations, this representation is then flattened into a 1-dimensional vector to be processed in the fully connected layer and get the predicted label. The structure of this layer

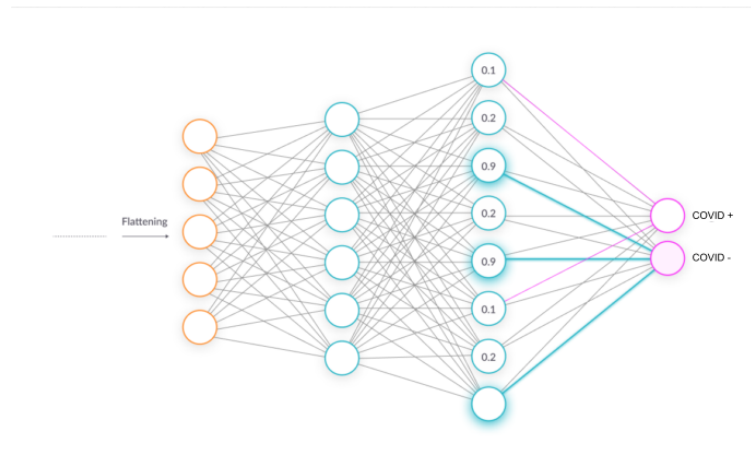


Figure 2.2: Fully Connected Layer

is a network of interconnected neurons [26] stacked in a layered fashion to form Artificial Neural network [27] as shown in Figure 2.2. The circles in Figure 2.2 illustrates neurons, and the interconnected lines represents the weights (parameters). The output of each neuron with a set of weight w and bias b , given an input vector x is illustrated in equation 2.2.

$$y = w^T x + b \quad (2.2)$$

The latest layer of neurons represents the classification labels. In the example, data samples that are assigned to "Covid +" label will have higher values in that neuron, and lower value in "Covid -" neuron.

Figure 2.3 shows an example of the full CNN operations pipeline for digit classification task.

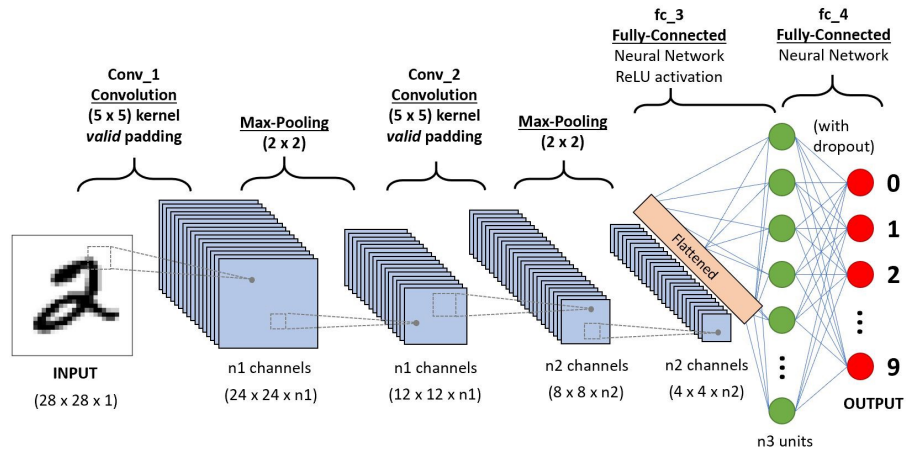


Figure 2.3: CNN Network Example

Transfer Learning

One of the biggest bottlenecks in training a deep neural network model is the need for an enormous amount of data to find the optimum parameters. One popular practice to tackle this is **transfer learning** - a two-step process comprised of a *pretraining* step and a *finetuning* step.

More specifically, we start from a neural network with randomly initialized parameters and train it on a generic task (e.g., classifying natural images)- the *pretraining* step. In computer vision applications, the most common pretraining step is using Imagenet [28], a large dataset of natural images and their labels. The pretraining step allows the network to have parameter values suitable for the generic tasks, allowing it to learn useful high-level features (e.g., recognizing object edges or colors). The *finetuning* step trains the neural network further on the true target task (e.g., classifying whether a chest x-ray is associated with a particular disease).

Previously in section 2.1.1, we have made references to the distinction between the types of features learned on the first and last layers of CNNs - first layer filters learn to extract low-level features such as object edge and colors, while deeper layer learns higher-level features like objects shape. The common practice of transfer learning follows this understanding. Usually, the *finetuning* step only modifies the parameters in the last few layers, freezing the initial layers' parameter. Intuitively, we reuse the layers that learn low level features from the generic task and further train the last layers to learn different high level features.

Previous works on building CNN models for disease recognition have commonly used publicly available pretrained models (e.g., AlexNet [29], DenseNet [30], ResNet [31], VGGNet [32], GoogleNet [33]) and finetune using open source medical image datasets (e.g., CheXpert [34] and MIMIC [35], the two popular chest-xray datasets with disease labels), or internal hospital data. [36] found improved performance on Computer-Aided Detection task with pretraining using ImageNet despite the disparity with the target medical dataset. [37] experimented on disease detection tasks using different medical imaging modalities and found that the transfer learning approach consistently performs better than training the network from scratch. [38] evaluated the performance of different pretrained CNNs on chest X-ray disease classification task.

2.1.2 CNN Applications in Medical Image Analysis

This part discusses two of the most popular applications of CNN in medical image analysis: *Image Classification* and *Semantic Segmentation*, which are the main components of the methodology in Chapter 3.

Image Classification

Classification is arguably the simplest and most widely used CNN application. Given images, classification models output a class label - one of the mutually exclusive pre-defined set of labels seen during model training. For example, in this study, we train the CNN using a dataset of (chest x-ray, COVID-19 label) pairs. In evaluation, given an unseen chest x-ray image, the CNN model outputs the predicted binary COVID-19 disease label.

The architecture characteristic for this task lies in the last fully connected (FC) layer, where each neuron represents one class, and their outputs are the likelihood of the data point belonging to that particular class. We can use a single neuron for binary classification, where the labels predictions will be based on whether the likelihood value is closer to zero or one.

A substantial body of work relevant to this study has looked at the uses of classification models in medical domains. Specifically, using CNN models to predict disease labels. [39], [40], and [41] are especially relevant as they build CNN models to classify diseases by looking into radiology images (i.e. 2-dimensional x-rays and 3-dimensional CT scans). A comprehensive survey about this line of work across various medical imaging modalities can be found in [42].

As these models become increasingly prevalent in actual clinical practice, [43] poses an important point that the models should serve as a second reader in clinical decision making, helping human readers produce pathology reports, but cannot be a permanent replacement.

Semantic Segmentation

Segmentation dives into the lowest possible of detail. Instead of classifying the image, we classify each *pixel* into higher-level groups they belong to. Because it categorizes every pixel, the output of the segmentation model is not a class label but a full image. For instance, in figure 2.4, the output of segmentation model is an image of three pixel values, each belongs to one of the "Person", "Bicycle" or "Background" class. Some popular open-source segmentation models are U-net [44] and SegNet [45].

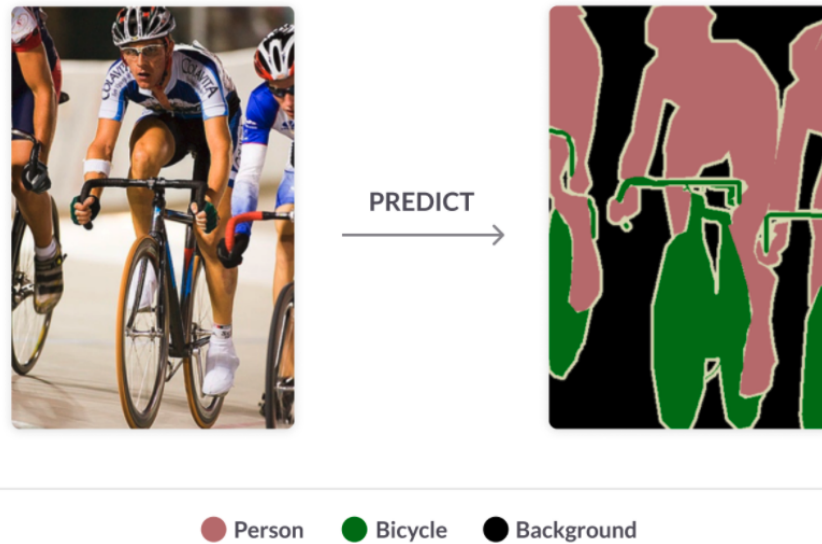


Figure 2.4: Semantic segmentation illustration

In medical image analysis, segmentation is used to extract more meaningful information in an image by dividing the original image into either background or meaningful objects. For instance, suppose we are given a chest x-ray, we would like to determine if a pixel is part of the lung, background, or device implants. Segmentation is often used as a preprocessing step before classification. Removing background area or device implants from the image before feeding it to the classification model can reduce the possible search area. Thus aiding the model in learning which feature is important for the classification task.

2.2 Generative Adversarial Network

This section discusses Generative Adversarial Network (GAN): (1) What is GAN, (2) How it works, (3) The type of GAN used in this study (i.e., Conditional GAN).

2.2.1 What is GAN

Generative Adversarial Network was first introduced in 2014 [46]. It is a member of the *generative model* family - given a training set of samples drawn from a data distribution

p_{data} , the goal of generative models is to represent an estimate distribution p_{model} of the original distribution p_{data} . More concretely, the expected result of training a generative model is to generate new data instances that resemble the training set (i.e., p_{model} as close as possible to p_{data}).

The classification models discussed in the preceding section are a family of *discriminative models*. Unlike discriminative models that tell us how likely a class label is to be assigned to a data instance (i.e., $P(Y|X)$), generative models capture the underlying data distribution (i.e., $P(X)$ or $P(X, Y)$).

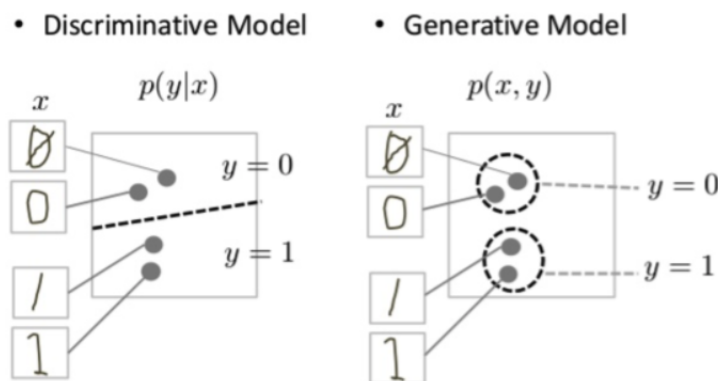


Figure 2.5: Generative model vs Discriminative model [2]

Figure 2.5 illustrates the distinction between the two model families. *Discriminative models* find a plane in the data space that separates the different classes. Finding this plane alone is sufficient to distinguish the classes of data samples without having to estimate where these samples lie in the data space. In contrast, *Generative models* estimates the location of these samples in the data space in order to produce new data samples that resemble the training data.

2.2.2 How GAN Works

Figure 2.6 illustrates the structure of GANs. It consists of two neural networks (1) *Generator* (G), and (2) *Discriminator* (D). The two exist within one GAN structure with competing aims. The *Generator*'s objective is to produce fake data instances such as they are not distinguishable from the real training instances. On the other hand,

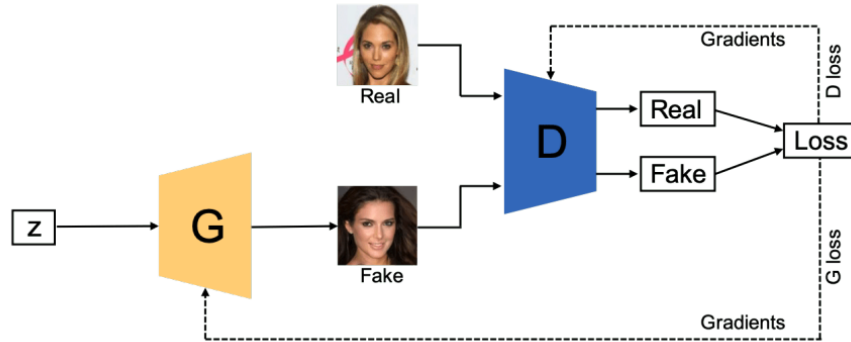


Figure 2.6: GAN Structure [3]

given a data instance, the discriminator aims to detect whether it comes from the real distribution (i.e., training data), or the 'fake' distribution modeled by the Generator.

Discriminator Network

Discriminator network (D) in GAN is simply a classification model. Its aim is to distinguish between the real training data samples from the fake samples produced by generator G. As shown in Figure 2.6, D network's inputs are the real training samples, and the fake samples produced by G. D's output is its prediction of whether the input image is "real" or "fake".

Generator Network

Given random noise z , Generator network G learns to generate fake data samples such that the discriminator D cannot distinguish them from the real training samples. During training, the D network's output is also incorporated into the generator (illustrated by the bottom dashed line in Figure 2.6). This serves as some sort of feedback - at the current training stage, how well can G's outputs fool D? In other words, how close are the generated "fake" outputs to the real training samples. Hence, as the training progresses, G will learn how to fool D better.

2.2.3 Conditional GAN

GAN enables us to generate data samples that resemble those in the training set. However, we cannot control data from which class the GAN network produces. Conditional GANs (CGANs) [47] lets us specify the label for each generated instance.

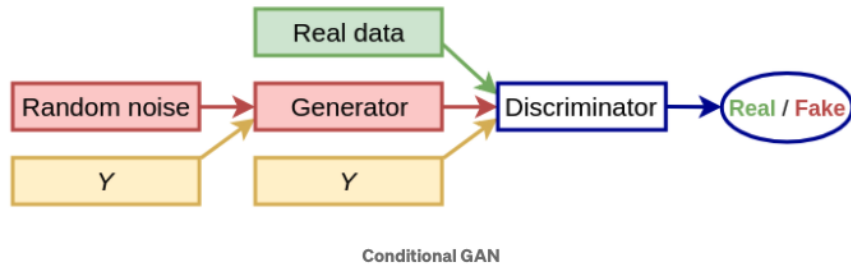


Figure 2.7: Conditional GAN architecture [4]

Figure 2.7 illustrates CGAN architecture. The *generator*'s input now consists of a random vector z and label information Y . Similarly, The *discriminator* has an additional Y information as input, together with the generator's output and the real training data.

Chapter 3

Methodology

This chapter details the experimental setups and methodology. First, the data acquisition sub-part will describe all internal and external data sources. Second, the image processing pipeline will be discussed. Third, the model development methodology will be outlined.

3.1 Data Acquisition

The model was trained and validated using two types of data sources: (1) M Health Fairview dataset, and (2) collection of publicly available datasets.

3.1.1 M Health Fairview Datasets

All CXRs obtained at M Health Fairview, Minnesota, USA (12 hospitals, 60 clinics) were considered eligible for inclusion. Images from patients with age less than 18 years old were excluded in training or validation.

M Health Fairview Model Development (Training) Dataset

We obtained 2,220 CXRs from patients with PCR confirmed COVID-19 (taken either 2 weeks prior COVID-19 diagnosis or during a COVID-19 associated hospitalization) and 36,288 non-COVID-19 CXRs from M Health Fairview for model training and optimization. All the COVID-19 positive case CXRs were obtained between March 2nd 2020, to June 30th, 2020, and the negative controls were obtained between October 25th, 2016

to March 3rd, 2020. All CXRs were taken in Minnesota, U.S.A at an M Health Fairview clinic or hospital.

The mean age in the positive controls was 59.8 (standard deviation, 16.2) years old and the mean age in the negative controls was 58.6 (standard deviation, 18.6) years old. 48.5% of CXRs in the positive controls were from males and 49.4% of CXRs in the negative controls were from females.

M Health Fairview Prospective Validation Dataset

Prospective validation included all CXRs within the M Health Fairview system obtained between July 1, 2020 – July 30, 2020. To account for prevalence, varying ratios of case imbalance were evaluated against using a ratio of 1:1 (50% positive:negative) to 1:20 (4.8%). During this prospective time period there were CXRs from 5,228 patients that were negative for COVID-19 and 1,777 patients that were confirmed PCR positive for COVID-19 for a prevalence rate of 25.4%). The mean age in the positive controls was 61.6 (standard deviation, 16.2) years old and the mean age in the negative controls was 57.5 (standard deviation, 18.5) years old. 68.6% of CXRs in the positive controls were from males and 31.4% of CXRs in the negative controls were from males.

3.1.2 Publicly Available Datasets

Publicly Available COVID-19 Datasets

COVID-19 positive cases were collected from two open source COVID-19 databases, namely BIMCV COVID-19+ [48] and COVID Chest X-ray Github [49]. BIMCV COVID+ contains 2261 CXRs (after taking out CT images) and were collected from 11 hospitals from the Valencian Region, Spain, and the positive cases were dated between February 26th and April 18th, 2020. We included all frontal X-Rays (Images with “view” column attribute values: ”PA” or ”AP” or ”AP Supine” or ”AP semi erect” in the Github metadata) with ”COVID-19” or ”COVID-19, ARDS” or ”SARS” labels from the COVID Chest X-Ray Github. In total, we have 504 images from this dataset.

Publicly Available non-COVID-19 Datasets

For COVID-19 negative cases, we collected cases and frontal images combined from: (1) 2011 – 2016 MIMIC-CXR [50] (random sample of 23,611 images); and (2) Open-I 2013

Indiana University (IU) Chest X-Ray Collection [51] (random sample of 3,814 images). Images in MIMIC-CXR and Open-I sets are dated prior to December 2019 resulting in 27,424 images of patients with no particular medical status except the absence of COVID-19.

Patient demographic information in publicly available datasets was not available as it was removed by the originating institutions to facilitate patient de-identification.

3.2 Image Pre-processing

All the original X-rays were in the DICOM format. We worked only with frontal X-rays (with DICOM position attribute “PA” or “AP”). All X-rays were converted into PNG images. All images were resized and zero-padded. First, the image ratio was locked so that the larger dimension of the height and the width was resized to 1024. Next, zero-padding was then performed to the other dimension, so that the resulting size becomes 1024×1024 . Each image was then min-max normalized by equation 3.1.

$$\frac{I - P_{min}}{P_{max} - P_{min}} \times 255 \tag{3.1}$$

Where I denotes image pixels and P_{max} and P_{min} denote the maximum and minimum values of the image. Image pre-processing was done using the Python scikit-image package [52].

3.3 Model Development (Training)

For model development, 38,508 (2,220 positives and 36,288 negatives) M Health Fairview CXR were used for training. Model training was supplemented to maximize model generalizability using publically available (9,592 total with a positive:negative ratio of 1:16) images of COVID-19 positive and negative patients. In the training set, 444 positives and 7,257 negatives were held out only for tuning the hyperparameters of the deep learning models and rest were used to train the models.

Our main model pipeline consisted of: (1) lung segmentation, (2) outlier detection, and (3) feature extraction/classification part. Figure 3.1 illustrates the complete model pipeline.

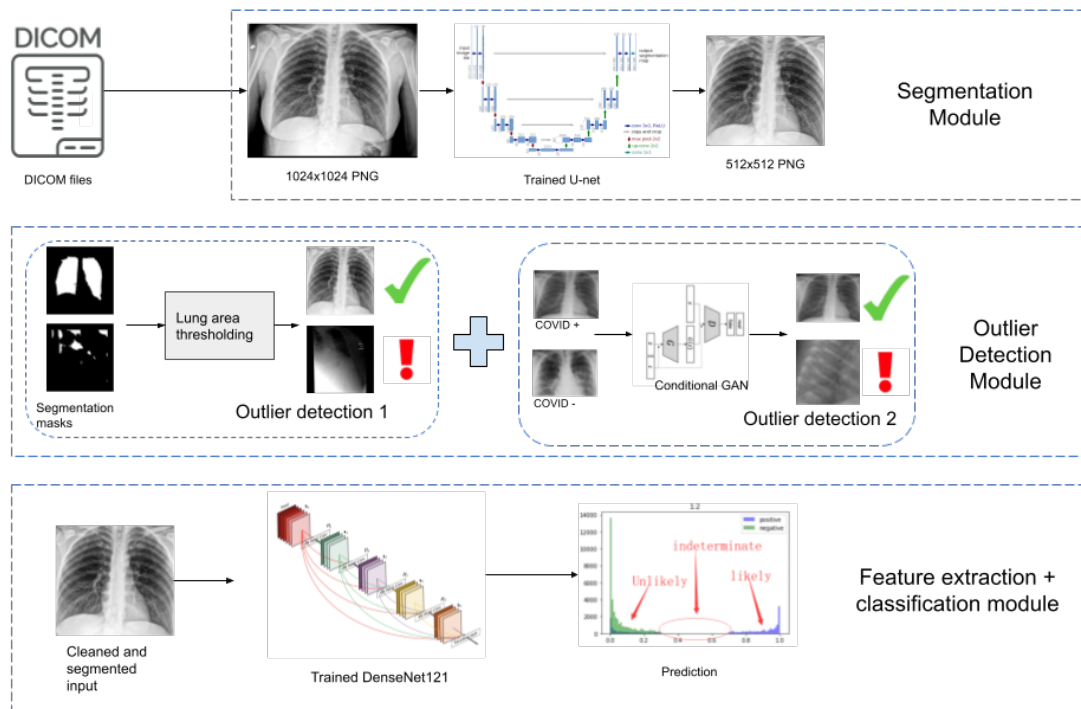


Figure 3.1: Overview of COVID-19 Diagnostic Model Pipeline

3.3.1 Lung Segmentation

We performed lung segmentation to focus the learning around the lung area, where the COVID-19 radiomic features are located. For this, the U-net model [44] that had been popular for biomedical image segmentation was adopted. The segmentation model was trained using three publicly available lung segmentation datasets: Montgomery [53], HIN [54], and JSRT [55]. The three datasets provided manual segmentation masks (i.e., segmentation labels).

The segmentation was not perfect. The resulting output mask often contains only part of the lung area and tend to be scattered over the whole lung area. To minimize the possibility of missing COVID-19 related radiomic features, we cropped out the

smallest square area that enclosed the predicted mask. All such square lung areas were subsequently resized into 512×512 , whether they were larger or smaller.

3.3.2 Outlier Detection

Practical X-rays have large variations and some of the extreme cases, e.g., caused by high/low exposure, skewed positions, or wrong position attributes. These variations can substantially contaminate the model training or the prediction process. To avoid overburdening the model, we chose to isolate these infrequent extreme cases for human screening.

We implemented two sequential procedures for this. First, before lung segmentation, we trained a Conditional Generative Adversarial Network (CGAN) [47] on the training CXRs to separate potential outliers. The class labels were fed into the conditional GAN as the “conditional” information. After training, any samples that were assigned scores lower than 0.1 by the discriminator with corresponding both positive and negative “conditional” information were declared as outliers. After the lung segmentation, we calculated the ratio of the area of the predicted lung mask and the area of the whole X-ray image for the remaining samples. Any CXR with a ratio below 0.1 or above 0.9 would be removed as outliers. The two procedures rejected about 10% of all input images, most of which were visually confirmed as challenging cases.

3.3.3 Feature Extraction and Classification

Previously in section 2, we have discussed two main bottlenecks in deep neural network training: (1) the need of huge amount of data, and (2) slow training time. In this study, reduce the training time by performing transfer learning (refer to chapter 2.1.1). We took a pre-trained DenseNet-121 [30], and further train the model using our CXR datasets to fine-tune it to recognize COVID-19. The difference between the prediction and the target (1 for positive and 0 for negative) was measured using the standard cross-entropy loss. The network was implemented using the deep learning package PyTorch 1.5.0 [56].

The positive cases and negative controls ratio in our dataset was imbalanced, reflecting the intrinsically biased distribution of COVID-19 cases in the population. To

counter the adverse effects of the imbalance on learning, we set our training objective as the maximum of averaged loss over the positive and the negative cases.

Chapter 4

Experimental Results

4.1 Validation Method

Two methods of validation were performed. First, model performance was evaluated prospectively using data retrospectively collected during the month of July, 2020 within the M Health Fairview System. Second, model performance was evaluated using publically available COVID-19 and non-COVID-19 imaging databanks. The COVID-19 Diagnostic Score from the model ranged from 0 to 1, indicating the likelihood of COVID-19.

To represent the performance of the model on different ratios, we calculated all the metrics on 1:1, 1:2, 1:5, 1:12, and 1:20 respectively for M Health Fairview and publically-available data validations. These ratios were obtained by subsampling the remaining validation data and all numbers reported were the average over 10 random repetitions. 10 random sub-samples were selected from this pool of images to construct mean estimates of AUROC (area under the receiver operating characteristic curve) and AUPRC (area under the precision recall curve). Specificity, sensitivity, positive predictive value (PPV), and negative predictive value (NPV) are provided for various threshold options.

4.2 Results

4.2.1 Prospective Validation using M Health Fairview CXR

This model was first evaluated via prospective validation using M Health Fairview images collected between July 1 – July 30, 2020. 7,005 CXRs were obtained in adults age 18 and older during this time period. 1777 images were from patients with PCR confirmed COVID-19 positive for a 25.4% prevalence. The mean AUROC and AUPRC are shown in the table 4.1.

Ratio	AUROC	AUPRC
1	0.800	0.842
2	0.801	0.759
5	0.803	0.631
12	0.801	0.497
20	0.797	0.412

Table 4.1: Prospective Validation for July 2020 using M Health Fairview CXRs

Distribution of COVID-19 Diagnostic Scores for both positive and negative cases during the month of July 2020 are provided in Figure 4.1.

Using the following thresholds for unlikely (score < 0.04), indeterminate (score $0.04 - 0.4$, 22.1% of images) and likely (score > 0.4) the following Specificity, sensitivity, PPV, and NPV are obtained (see table 4.2). Around 20% of the test images predictions were discarded from outlier detection module and indeterminate score combined.

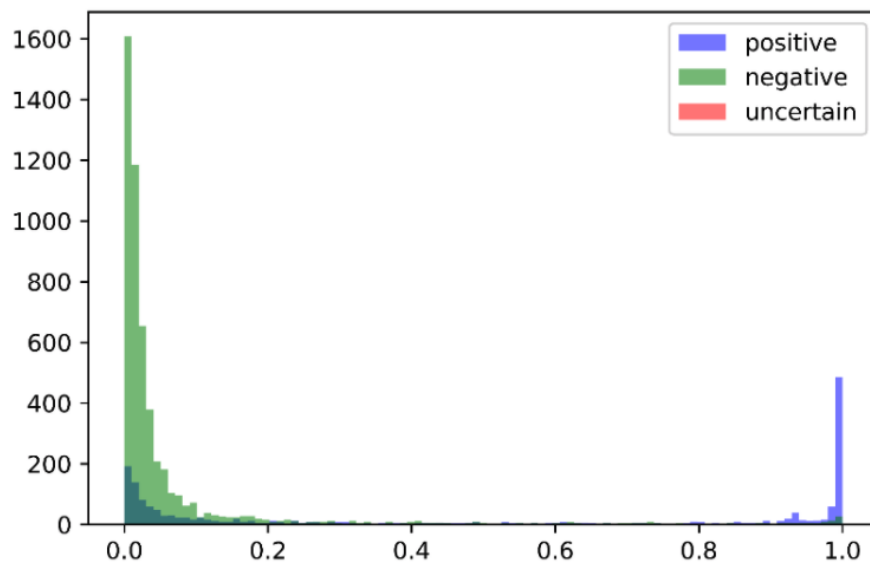


Figure 4.1: Distribution of COVID-19 Diagnostic Scores (X-axis) for patients with PCR confirmed positive COVID-19 (purple bars) and non-COVID-19 patients (green bars) during the month of July 2020.

Ratio	Specificity	Sensitivity	PPV	NPV
1	0.934	0.659	0.909	0.733
2	0.935	0.659	0.835	0.846
5	0.934	0.664	0.668	0.933
12	0.934	0.649	0.449	0.970
20	0.934	0.677	0.338	0.983

Table 4.2: Prospective Validation for July 2020 using M Health Fairview CXRs for 0.04 – 0.4 thresholds

To better understand these results we performed separated evaluation on images from patients with “severe” disease (patients that required ICU admission) and “moderate” disease (patients that required hospital admission).

Ratio	AUROC	AUPRC
1	0.834	0.871
2	0.834	0.799
5	0.834	0.688
12	0.837	0.562
20	0.829	0.483

Table 4.3: Prospective Validation for July 2020 using M Health Fairview CXRs for patients with “severe” COVID-19 disease

Ratio	AUROC	AUPRC
1	0.704	0.742
2	0.706	0.619
5	0.705	0.440
12	0.706	0.289
20	0.705	0.218

Table 4.4: Prospective Validation for July 2020 using M Health Fairview CXRs for patients with “moderate” COVID-19 disease

4.2.2 Validation using publicly available COVID-19 CXRs

Similarly, the model was validated using a sample of publicly available COVID-19 CXRs. Table 4.5 summarizes the quantitative performance, and figure 4.2 shows the distribution of COVID-19 Diagnostic Scores for both positive and negative cases.

Ratio	AUROC	AUPRC
1	0.960	0.975
2	0.959	0.963
5	0.961	0.949
12	0.956	0.927
20	0.965	0.929

Table 4.5: External Validation using publicly available COVID-19 CXRs

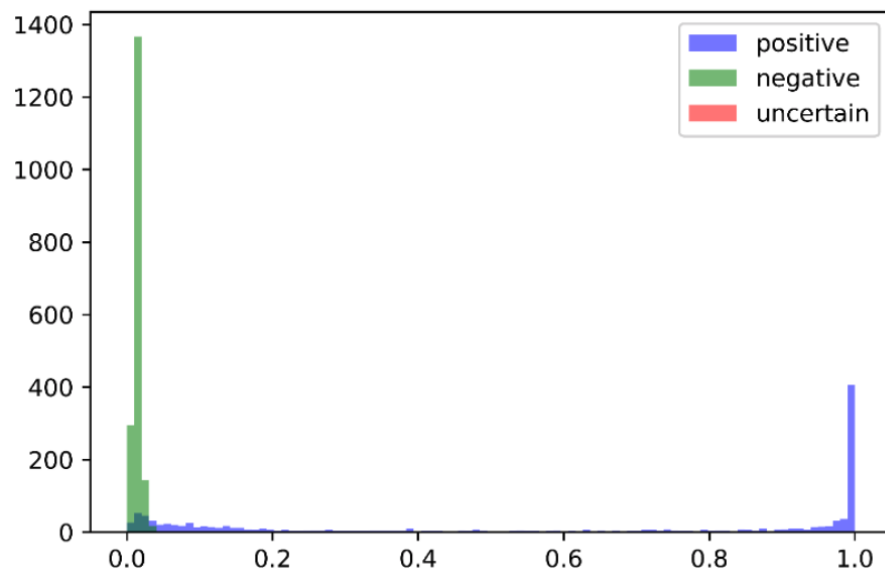


Figure 4.2: Distribution of COVID-19 Diagnostic Scores (X-axis) for patients with COVID-19 (purple bars) and non-COVID-19 patients (green bars) from publically available datasets, prevalence

To better understand these results we evaluated individual performance from the BIMCV COVID-19 and Github dataset.

Ratio	AUROC	AUPRC
1	0.966	0.980
2	0.965	0.970
5	0.965	0.956
12	0.964	0.942
20	0.963	0.935

Table 4.6: Publicly-Available Validation using Github COVID-19 database

Ratio	AUROC	AUPRC
1	0.958	0.975
2	0.9959	0.963
5	0.960	0.946
12	0.958	0.925
20	0.960	0.914

Table 4.7: Publicly Available (BIMCV) COVID-19 CXR validation

Chapter 5

Discussion and Conclusion

5.1 Summary of Key Finding

M Health Fairview validation results in AUROC of roughly 0.8 across all ratios and AUPRC range of 0.4 - 0.84. Evaluation on the publicly available datasets yields AUROC of roughly 0.96 across all ratios and 0.92 - 0.97 AUPRC.

Furthermore, more fine-grained evaluation on "severe" vs. "moderate" cases was performed on the M Health Fairview data, yielding non-trivial differences on the two cases' AUROC and AUPRC. "severe" cases evaluation yields 0.83 AUROC and 0.48 - 0.87 AUPRC, while "moderate" cases has AUROC of roughly 0.7 and 0.218 - 0.74 AUPRC. In contrast, individual evaluations on each public dataset source yield similar results across all sources (>0.9 AUROC and AUPRC).

5.2 Results Interpretation

5.2.1 M Health Fairview Evaluation Results

Model performance on the M Health Fairview dataset has > 0.93 specificity scores across all ratios and sensitivity scores between 0.65 - 0.67. This result suggests disparity in the model's capacity to identify negative cases vs. identifying the positives. This contrast is also apparent in figure 4.1. Negative prediction scores are concentrated at the leftmost regions, while some positive predictions are wrongly assigned low likelihood values (blue colored bars in leftmost regions).

We attempted to explain this disparity by looking at the more fine-grained experiments on the individual "severe" and "moderate" cases. The performance on "severe" cases outperforms the performance on "moderate" cases by 0.12 - 0.13 AUROC and AUPRC scores margins. This result implies that mixing the "severe" and "moderate" cases might contribute to the model's performance disparity. More severe COVID-19 cases show more visible characteristics on CXRs; hence, it is easier to distinguish from the negative cases.

5.2.2 Public Dataset Evaluation Results

The model consistently performs with > 0.9 AUROC and AUPRC on the combination of public datasets and the individual ones. This high performance is also implied by figure 4.2, showing only a small portion of the positive samples wrongly classified (blue bars in left regions) and almost all the negative predictions concentrated on the leftmost area.

While it is tempting to imply the model has a high generalization capacity, more rigorous analysis is needed to inspect this result. Machine learning models tend to resort to trivial information extracted from images [57], [58]. As the public evaluation data comes from multiple sources with different image characteristics (i.e., image contrast, organs size or device implants), the model might take advantage of this irrelevant information instead of making predictions based useful COVID-19 features on the public dataset.

5.3 Results Implications

The result shows promise that the performance of the model on the M Health Fairview data is adequate for the established role of CXR as a frontline tool for screening and triage. If properly incorporated as an AI-enabled clinical decision support into the workflow of current diagnostic modalities, the model has the potential to considerably improve the speed and reliability of current screening procedures. However, the generalization capacity of this model still need further inspection and improvement. Future work needs to assess the model's performance on different dataset, with less trivial distinguishing characteristics.

5.4 Study Limitations

This study is not without limitations. First, the number of our positive controls is relatively small. Second, our negative controls were not selected from a target population of suspected COVID-19 patients. Third, CXR findings for COVID-19 are nonspecific and overlap with a number of other infectious and non-infectious etiologies, which could complicate interpretation. Fourth, this model has not been tested in a clinical setting, which also introduces the additional uncertainty of not having a gold standard test(s) for true positive and true negative in cohort analysis. As such, it is difficult to say how well the model will perform in the real world and across different settings. Lastly, these models were trained and validated on fixed data. The models will evolve as new data arrive. It is possible to modify the models to make them gradually improve over time, leveraging advances in online machine learning. Finally, the integration of radiometric characteristics of COVID-19 positive patients will further improve models. Future directions will implement and prospectively evaluate these diagnostic models.

5.5 Conclusion

In conclusion, we provide an accurate and reliable model capable of differentiating between CXRs of patients with COVID-19 from non-COVID-19 CXRs. Delivery via an AI-enabled clinical decision support system for treating clinicians and radiologists may result in faster isolation, confirmatory PCR-based testing, and treatment. This could potentially decrease hospital contamination and provide better quality of care.

References

- [1] Maithra Raghu and Eric Schmidt. A survey of deep learning for scientific discovery. 2020, 2003.11755.
- [2] Background: What is a generative model?
- [3] Zhengwei Wang, Qi She, and Tomas E. Ward. Generative adversarial networks in computer vision: A survey and taxonomy, 2020, 1906.01529.
- [4] Manish Nayak. An introduction to conditional gans (cgans), May 2019.
- [5] Qun Li et al. Early transmission dynamics in wuhan, china, of novel coronavirus–infected pneumonia. *New England Journal of Medicine*, 382(13):1199–1207, 2020, <https://doi.org/10.1056/NEJMoa2001316>. PMID: 31995857.
- [6] World Health Organization. Coronavirus disease 2019 (covid-19)situation report –209, 2020.
- [7] Douglas B. White and Bernard Lo. A Framework for Rationing Ventilators and Critical Care Beds During the COVID-19 Pandemic. *JAMA*, 323(18):1773–1774, 05 2020, https://jamanetwork.com/journals/jama/articlepdf/2763953/jama-white_2020_vp_200068.pdf.
- [8] Joel Hellewell et al. Feasibility of controlling covid-19 outbreaks by isolation of cases and contacts. *The Lancet Global Health*, 8, 02 2020.
- [9] Centers for Disease Control and Prevention. Interim guidelines for collecting and handling of clinical specimens for covid-19 testing, 2020. Accessed 2020-04-26.
- [10] Lessons from the frontline of the covid-19 outbreak, Mar 2020.

- [11] Harrison X. Bai et al. Performance of radiologists in differentiating covid-19 from non-covid-19 viral pneumonia at chest ct. *Radiology*, 296(2):E46–E54, 2020, <https://doi.org/10.1148/radiol.2020200823>. PMID: 32155105.
- [12] Chinese Society of Radiology. Radiological diagnosis of new coronavirus infected pneumonitis: Expert recommendation from the chinese society of radiology (first edition). *Chin J Radiol*, 2020.
- [13] Tao Ai, Zhenlu Yang, Hongyan Hou, Chenao Zhan, Chong Chen, Wenzhi Lv, Qian Tao, Ziyong Sun, and Liming Xia. Correlation of chest ct and rt-pcr testing for coronavirus disease 2019 (covid-19) in china: A report of 1014 cases. *Radiology*, 296(2):E32–E40, 2020, <https://doi.org/10.1148/radiol.2020200642>. PMID: 32101510.
- [14] Yicheng Fang, Huangqi Zhang, Jicheng Xie, Minjie Lin, Lingjun Ying, Peipei Pang, and Wenbin Ji. Sensitivity of chest ct for covid-19: Comparison to rt-pcr. *Radiology*, 296(2):E115–E117, 2020, <https://doi.org/10.1148/radiol.2020200432>. PMID: 32073353.
- [15] Geoffrey D. Rubin et al. The role of chest imaging in patient management during the covid-19 pandemic: A multinational consensus statement from the fleischner society. *Radiology*, 296(1):172–180, 2020, <https://doi.org/10.1148/radiol.2020201365>. PMID: 32255413.
- [16] Scott Simpson et al. Radiological society of north america expert consensus document on reporting chest ct findings related to covid-19: Endorsed by the society of thoracic radiology, the american college of radiology, and rsna. *Radiology: Cardiothoracic Imaging*, 2(2):e200152, 2020, <https://doi.org/10.1148/ryct.2020200152>.
- [17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.
- [18] Xiaolong Wang, Ross B. Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. *CoRR*, abs/1711.07971, 2017, 1711.07971.

- [19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017, 1703.06870.
- [20] H. Greenspan, B. van Ginneken, and R. M. Summers. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Transactions on Medical Imaging*, 35(5):1153–1159, 2016.
- [21] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, 19(1):221–248, 2017, <https://doi.org/10.1146/annurev-bioeng-071516-044442>. PMID: 28301734.
- [22] Berkman Sahiner, Aria Pezeshk, Lubomir M. Hadjiiski, Xiaosong Wang, Karen Drukker, Kenny H. Cha, Ronald M. Summers, and Maryellen L. Giger. Deep learning in medical imaging and radiation therapy. *Medical Physics*, 46(1):e1–e36, 2019, <https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1002/mp.13264>.
- [23] Justin Ker, Lipo Wang, Jai Rao, and Tchoyoson Lim. Deep learning applications in medical image analysis. *IEEE Access*, PP:1–1, 12 2017.
- [24] Syed Anwar, Muhammad Majid, Adnan Qayyum, Muhammad Awais, Majdi Alnowami, and Khurram Khan. Medical image analysis using convolutional neural networks: A review. *Journal of Medical Systems*, 42:226, 10 2018.
- [25] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *CoRR*, abs/1505.00853, 2015, 1505.00853.
- [26] Warren Mcculloch and Walter Pitts. A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:127–147, 1943.
- [27] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, pages 65–386, 1958.
- [28] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.
- [29] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou,

- and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.
- [30] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *CoRR*, abs/1608.06993, 2016, 1608.06993.
- [31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015, 1512.03385.
- [32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv 1409.1556*, 09 2014.
- [33] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.
- [34] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie S. Shpankaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *CoRR*, abs/1901.07031, 2019, 1901.07031.
- [35] Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. 2019, 1901.07042.
- [36] Hoo-chang Shin, Holger Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Noguees, Jianhua Yao, Daniel Mollura, and Ronald Summers. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, 35, 02 2016.
- [37] Nima Tajbakhsh, Jae Y. Shin, Suryakanth R. Gurudu, R. Todd Hurst, Christopher B. Kendall, Michael B. Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *CoRR*, abs/1706.00712, 2017, 1706.00712.

- [38] Keno K. Bressen, Lisa C. Adams, Christoph Erxleben, Bernd Hamm, Stefan M. Niehues, and Janis L. Vahldiek. Comparing different deep learning architectures for classification of chest radiographs. *Scientific Reports*, 10(1), Aug 2020.
- [39] Koichiro Yasaka, Hiroyuki Akai, Osamu Abe, and Shigeru Kiryu. Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced ct: A preliminary study. *Radiology*, 286(3):887–896, 2018, <https://doi.org/10.1148/radiol.2017170706>. PMID: 29059036.
- [40] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Yi Ding, Aarti Bagul, Curtis Langlotz, Katie S. Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *CoRR*, abs/1711.05225, 2017, 1711.05225.
- [41] M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, and S. Mougiakakou. Lung pattern classification for interstitial lung diseases using a deep convolutional neural network. *IEEE Transactions on Medical Imaging*, 35(5):1207–1216, 2016.
- [42] Eric Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25, 01 2019.
- [43] Clara Mosquera-Lopez, Sos Aгаian, Alejandro Velez Hoyos, and Ian Thompson. Computer-aided prostate cancer diagnosis from digitized histopathology: A review on texture-based systems. *IEEE reviews in biomedical engineering*, 8, 07 2014.
- [44] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015, 1505.04597.
- [45] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017.
- [46] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets.

- In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.
- [47] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014, 1411.1784.
- [48] Maria de la Iglesia Vayá, Jose Manuel Saborit, Joaquim Angel Montell, Antonio Pertusa, Aurelia Bustos, Miguel Cazorla, Joaquin Galant, Xavier Barber, Domingo Orozco-Beltrán, Francisco García-García, Marisa Caparrós, Germán González, and Jose María Salinas. Bimcv covid-19+: a large annotated dataset of rx and ct images from covid-19 patients. 2020, 2006.01174.
- [49] Joseph Paul Cohen, Paul Morrison, Lan Dao, Karsten Roth, Tim Q Duong, and Marzyeh Ghassemi. Covid-19 image data collection: Prospective predictions are the future. *arXiv 2006.11988*, 2020.
- [50] Alistair Johnson, Tom Pollard, Seth Berkowitz, Nathaniel Greenbaum, Matthew Lungren, Chih-ying Deng, Roger Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6:317, 12 2019.
- [51] OpenI. Indiana university - chest x-rays (png images).
- [52] van der Walt et al. scikit-image: image processing in Python. *PeerJ*, 2:e453, 6 2014.
- [53] Sema Candemir et al. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE Transactions on Medical Imaging*, 33:577–590, 02 2014.
- [54] Sergii Stirenko et al. Chest x-ray analysis of tuberculosis by deep learning with segmentation and augmentation. *CoRR*, abs/1803.01199, 2018, 1803.01199.
- [55] Shiraishi J Katsuragawa et al. Development of a digital image database for chest radiographs with and without a lung nodule: Receiver operating characteristic

- analysis of radiologists' detection of pulmonary nodules. *American Journal of Roentgenology*, 174:71–74, 2000.
- [56] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.
- [57] Jason Jo and Yoshua Bengio. Measuring the tendency of cnns to learn surface statistical regularities. *CoRR*, abs/1711.11561, 2017, 1711.11561.
- [58] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.