Multi-Label Chest X-Ray Pathology Detection of COVID-19, with Transfer Learning

Colm Clancy (01365631)

School of Computer Science

National University of Ireland, Galway

Supervisors

Professor Michael G. Madden

In partial fulfillment of the requirements for the degree of

MSc in Computer Science (Data Analytics)

September 2020

DECLARATION I, Colm Clancy, do hereby declare that this thesis entitled Multi-Label Chest X-Ray Pathology Detection of COVID-19, with Transfer Learning is a bonafide record of research work done by me for the award of MSc in Computer Science (Data Analytics) from National University of Ireland, Galway. It has not been previously submitted, in part or whole, to any university or institution for any degree, diploma, or other qualification.

Signature: _____

Acknowledgements

I would like to thank my thesis supervisor, Professor Michael Madden for his time, knowledge and patience throughout the process. My fellow student and lecturers who throughout the year have made it an enjoyable, rewarding and engaging one. Finally, my friends and family for their support over the course of the year.

Abstract

Chest x-rays are recommended under certain circumstances for use as part of the diagnostic workup for suspected COVID-19 patients. For this work I combine two chest x-ray datasets, PadChest and BIMCV COVID-19+. PadChest a multi-label chest x-ray dataset (January 2019) that contains no COVID-19 positive patients, BIMCV+ COVID-19 (June 2020) is a multi-label dataset of patients who have at least one positive Polymerase Chain Reaction (PCR) test for COVID-19. I train two deep neural networks with transfer learning on the dataset of around 30k images for the detection of all 197 labels, with a test set of around 10k. As a new disease reporting is focused on detection of COVID-19 and its associated radiographic findings. In my best performing model I reports an AUC of 0.96 for the detection of the COVID-19 diagnosis. This compares favourably with existing research utilizing the same datasets to report a best model AUC of 0.94[37].

Keywords: COVID-19, Deep Learning, Transfer Learning

Contents

1 Introduction **1**

1.1 Background 1

1.2 World Health Organization Guidelines 2

1.3 Research Aims 2

1.4 Methodology 3

1.5 Structure of Thesis 3

2 Terminology **4**

2.1 Chest X-Ray Positions 4

2.2 Medical Imaging Standards 6

2.2.1 MIDS 6

2.2.2 DICOM 7

2.2.3 Monochrome 7

2.2.4 UMLS 7

2.3 Pathology Types 8

3 Literature Review **9**

3.1 CT Scans & X-Rays for Diagnosis 10

3.2 CNN Progress to Date 11

3.3 Current COVID-19 Analysis 12

3.4 Data Collection 13

4 Deep Learning **15**

4.1 Artificial Neural Networks 15

4.2 Convolutional Neural Networks 16

4.2.1 Activation Functions 18

4.2.2 Batch Normalization 18

4.2.3 Zero Padding 18

4.2.4 Pooling 19

4.2.5 Dropout 19

4.2.6 Dense/Fully Connected 19

4.3 DenseNet-121 19

4.4 Transfer Learning 22

5 Data Selection and Preparation **23**

5.1 X-Ray Datasets 23

5.2 Dataset Labels 24

5.3 Labels Inferred 25

5.4 Label Quality 26

5.5 Dataset Balance 26

5.6 Data Cleaning & Merging 30

5.6.1 Image Transformation 30

6 Methodology **32**

6.1 Train Test Split 32

6.2 Implementation 34

6.3 Class Imbalance 35

6.4 Model Summary 39

6.5 Training 40

7 Results **42**

8 Conclusion **45**

A Hierarchical Label Results **54**

List of Figures

2.1 X-Ray Positions(reproduced from[32]) 5
2.2 Medical Image Data Structure(reproduced from[31]) 6

4.1 Multi-Layer Neural Network(reproduced from[61]) 16
4.2 Convolution Layer(reproduced from[50]) 17
4.3 LeNet(reproduced from [5]) 17
4.4 5 Layer Dense Block(reproduced from[25]) 20
4.5 DenseNet121 Layers 21

5.1 Label Differences 24
5.2 Label Structure 26
5.3 Datasets Pathology balance 28
5.4 PadChest Demographic 29
5.5 BIMCV COVID-19+ Demographic 29
5.6 Monochrome 1 to Monochrome 2 31
5.7 Inverting Monochrome 2 31

6.1 Train Demographic 33
6.2 Validation Demographic 33
6.3 Test Demographic 33
6.4 CPU load 34

6.5 GPU load 34
6.6 Class Frequency 36
6.7 Class Imbalanced 37
6.8 Class Balanced 38
6.9 Training Loss 41
6.10 Training AUC 41

7.1 ROC 44

A.1 Results Part 1 55
A.2 Results Part 2 56
A.3 Results Part 3 57
A.4 Results Part 4 58
A.5 Results Part 5 59

List of Tables

4.1 DenseNet121 Layers Breakdown 22

5.1 X-Ray Datasets 23

6.1 Train Test Split 32
6.2 Model Layout 39
6.3 Model Parameters 39

7.1 Model Results 43

Chapter 1

Introduction

This project sets out the task of establishing a platform for COVID-19 and associated pathology detection in chest x-rays using deep neural networks, merged datasets and transfer learning.

1.1 Background

31/12/2019: The WHO(World Health Organisation) issues a statement about a pneumonia of unknown cause detected in the city of Wuhan in Hubei province, China[59].
13/01/2020: This new disease is detected in Thailand, the first case outside of China[59].
11/02/2020: The infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is given the name COVID-19[59].
11/03/2020: The WHO declared the COVID-19 outbreak to be a pandemic[59].

1.2 World Health Organization Guidelines

It is important to establish the positioning and appropriate use of x-rays in the wider context of COVID-19 detection. In accordance with WHO guidance:
"The diagnosis of COVID-19 is currently confirmed by laboratory testing through identification of viral RNA in reverse transcriptase polymerase chain reaction (RT-PCR). Chest imaging has been considered as part of the diagnostic workup of patients with suspected or probable COVID-19 disease where RT-PCR is not available, or results are delayed or are initially negative in the presence of symptoms suggestive of COVID-19. Imaging has been also considered to complement clinical evaluation and laboratory parameters in the management of patients already diagnosed with COVID-19[42]." For symptomatic patients with suspected COVID-19, WHO suggests using chest imaging for the diagnostic workup of COVID-19 when:

1. RT-PCR testing is not available;
2. RT-PCR testing is available, but results are delayed; and
3. initial RT-PCR testing is negative, but with high clinical of suspicion of COVID-19.

1.3 Research Aims

The research aims of this project are:

- Can the limited COVID-19 chest x-ray datasets available prove sufficient for transfer learning?
- Can COVID-19 diagnosis and its associated radiographic findings be identified by a deep learning model?

1.4 Methodology

The methodology is as follows:

- Create a dataset containing 50k chest X-rays by combining a pre-COVID-19 era chest x-ray dataset with with a COVID-19 era dataset.
- Implement a hierarchical labelling structure.
- Train/validate/test a CNN(Convolutional Neural Network) to detect multi-label findings associated with chest x-rays.

1.5 Structure of Thesis

Chapter 1 gives the background and sets out the problem, research aims and methodology for the thesis. Chapter 2 established the terminology needed to understand the forth coming chapters. Chapter 3 established the existing body of work in the associated with the project. Chapter 4 gives an overview of deep learning, CNNs and transfer learning. Chapter 5 cover data selection and preperation. Chapter 6 is a walk-though of the methodology of the model training process and parameter choices. Chapter 7 gives a factual representation on the model results. Chapter 8 reflects on the success of the project and its place in a wider context.

Chapter 2

Terminology

2.1 Chest X-Ray Positions

Chest x-ray datasets typically contain images from a variety of different positions. Figure 2.1 from Bustos et al. demonstrates the full complement of standard x-ray positions . P-A is the most common position as reflected in the composition of the datasets as described in chapter 5 and the front or back positions are most commonly the views used in training deep learning models. As we can see from Figure 2.1 P-A and A-P can appear quiet similar but heart size is the major tell with it being magnified in A-P. In the lordotic position the ribs will appear closer together. In x-ray datasets the position metadata may not always be available or complete as the data originates from the x-ray machines which involve human input, it is beneficial therefore to have some knowledge of the x-ray orientations when reviewing datasets.

- P-A (Posterior-Anterior)
- Lateral (side)
- Lordotic (curved position)

- A-P Surpine
- A-P (Anterior-Posterior)

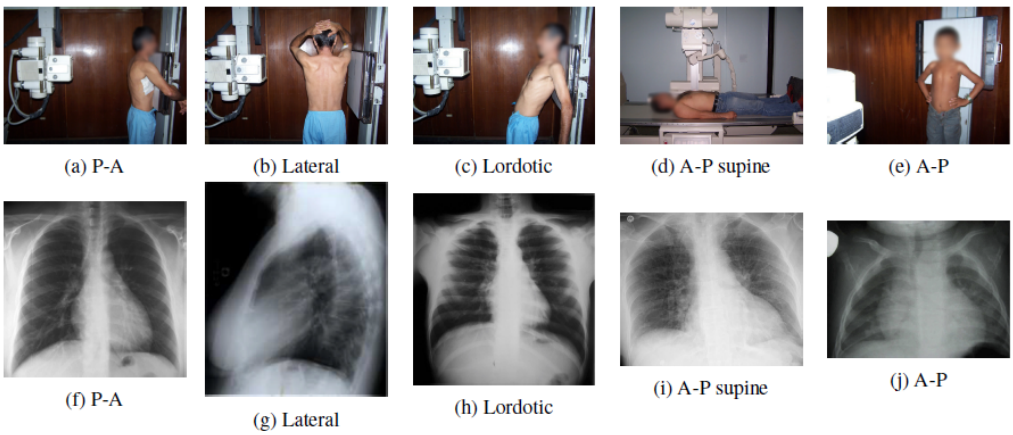


Figure 2.1: X-Ray Positions(reproduced from[32])

2.2 Medical Imaging Standards

2.2.1 MIDS

MIDS(Medical Imaging Data Structure) is a standard developed by the BIMCV team, its structure is based on the BIDS (Brain Imagining Data Structure) and is a proposal to set a standard for mixed image and metadata standard across medical imaging[31]. The BIMCV COVID-19+ dataset comes in this standard with a folder structure based on patient and imaging session with medical information structured in various files as illustrated in Figure 2.2 [31].

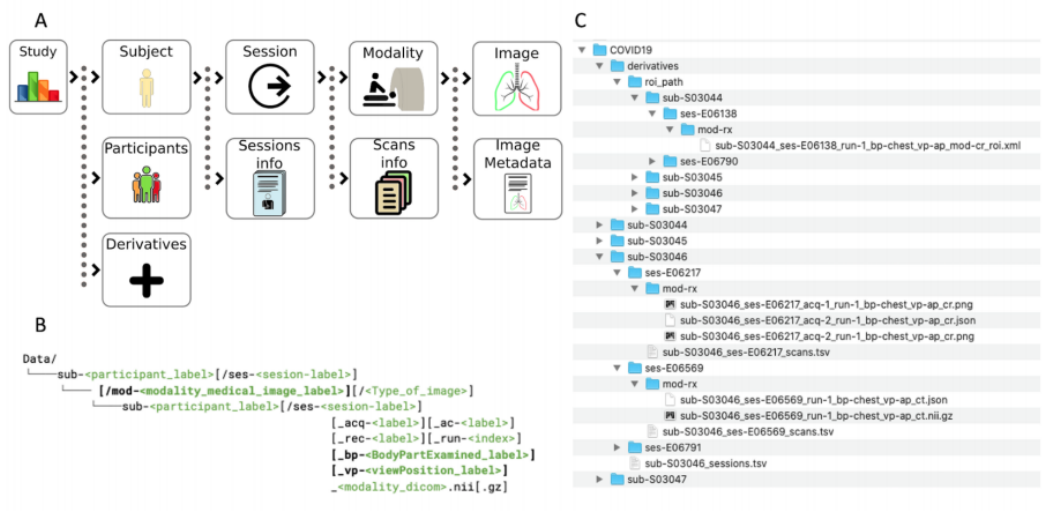


Figure 2.2: Medical Image Data Structure(reproduced from[31])

2.2.2 DICOM

Digital Imaging and Communications in Medicine (DICOM)[52] is the standard for the communication and management of medical imaging information and related data. Its purpose it to standardise medical imaging in hardware. Each image in the BIMCV COVID-19+ dataset contains an accompanying json file containing anonymised patient DICOM metadata that needs to be extracted for the project.

2.2.3 Monochrome

Each BIMCV COVID-19+ dataset image comes in one of two monochrome image file formats[41]:

- Monochrome 1: Reversed Monochrome image,Higher pixel values are displayed as blacker.
- Monochrome 2:Normal Monochrome image, Higher pixel values are displayed as whiter.

2.2.4 UMLS

The radiological labels in both PadChest and BIMCV COVID-19+ datasets are mapped to standard Unified Medical Language System (UMLS) terminology [6]. It serves two main purposes, each labels comes with an associated biomedical vocabulary unique identifier(CUI) code which makes it language agnostic as the original reports are in spanish and also when mapped each label then exists in the UMLS hierarchy, i.e. COVID-19 and pneumonia can be both two different labels but COVID-19 is also a subset of pneumonia.

2.3 Pathology Types

There are two main categories of labels in the PadChest and BIMCV COVID-19+ datasets, differential diagnosis and radiographic finding. Differential diagnosis are diseases that require more patient information or further lab tests for confirmation of diagnosis. Radiographic findings as determined upon inspection by a radiologists are findings that are completely observable in the images. COVID-19 is classified as a differential diagnosis, associated radiographic findings for COVID-19 patients include consolidation, ground-glass opacities and alveolar pattern which are radiographic findings[63][38].

Chapter 3

Literature Review

With the COVID-19 coronavirus pandemic sweeping the world real-time polymerase chain reaction (RT-PCR) testing of a nasal swab is established as the main method of diagnosis with computed tomography(CT) scans and chest x-rays analysed by radiologists offering additional information/confirmation on COVID-19 patient status[34].

The purpose of this literature review is to define the problem space and assess state-of-the-art deep learning methods as a diagnosis aid as applied to CT scans and chest x-rays images in relation to COVID-19.

Section 3.1 looks at the evidence around CT & x-rays as used in diagnosis, section 3.2 briefly looks at the recent success stories involving convolutional neural networks (CNN), section 3.3 is an assessment of recent COVID-19 deep learning attempts and similar methods that may be applied, section 3.4 points to ongoing data collection efforts.

3.1 CT Scans & X-Rays for Diagnosis

The standard method of checking for COVID-19 is the real-time reverse transcription polymerase chain reaction (RT-PCR) test[34].

Much of the early image testing around COVID-19 and thus some of the early papers mentioned below are centered around CT scans, but due to the rise in the number of cases, the complexity of testing a potentially COVID-19 positive patient in a CT scan room chest and the availability of such machines, chest x-rays have become more prevalent. Some analysis exists for both in terms of comparison to RT-PCR tests. It should be noted in the analysis that follows that sensitivity refer to correctly identifying those that have the disease (true positive rate) and specificity to correctly identifying those without the disease (true negative rate). A true positive is when the the positive class is correctly predicted, a true negative is when the negative class is correctly predicted.

A few small scale studies have emerged which allow us to compare RT-PCR testing with chest CT scans, a study by [48] found chest CT to be an important complement to RT-PCR testing, with a sensitivity of 97% and specificity of 25% with RT-PCR positive tests as a reference. In a study designed to compare COVID-19 with viral pneumonia on chest CT[49] using 7 different radiologists found that they had sensitivities ranging from 67% - 97% and specificities ranging from 7% - 100% in differentiating COVID-19 from viral pneumonia on chest CT with accuracies of 60% - 88%.

In [36] study of 51 patients, CT scans detected suspected COVID-19 earlier than some RT-PCR swab test which required 1-4 further tests before confirmation.

Yuen Frank Wong et al. had 64 patient chest x-rays reviewed by 2 radiologists, the patients had a 91% sensitivity RT-PRC test with 69% for the initial chest x-ray analysis. Consolidation followed by ground-glass opacities are found to be

the most common findings.

It should be noted that these early technical reports are being produced at the outset of the COVID-19 epidemic and that there exists potential for bias in the datasets based on the weight of COVID-19 hospital admissions[35].

3.2 CNN Progress to Date

ImageNet[56] is a large visual database designed for use in visual object recognition software research, its annual competition the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) can be reference to track the performance grains of convolutional neural network (CNN) throughout the 2010's. With the availability of increased CPU/GPU power and large annotated image datasets the popularity of the CNNs for image classification and object detection has grown a lot over the last 10 years, from [11] with AlexNet winning the 2012 ImageNet (ILSVRC) with an error rate to 15.3% to 26.2% for the next best effort to the popular [18] VGG model in 2014.

As the experimentation with CNN techniques have continued their potential in Computer-aided detection (CADe)/computer-aided diagnosis(CADx) has grown. Recent efforts by [40] in Google where they demonstrate an AI system "capable of surpassing human experts in breast cancer prediction have caught the public's attention reported that the area under the receiver operating characteristic curve (AUC-ROC)for the AI system was greater than the AUC-ROC for the average radiologist by an absolute margin of 11.5%", but it's just one of a number of impressive results in recent years with [30] applying CNNs to detect diabetic retinopathy (sensitivity of 87.4%, specificity of 89.5%) and [24] which claims dermatologist-level classification of skin cancer diagnosis.

3.3 Current COVID-19 Analysis

Given the work to date in deep learning on medical imaging not surprisingly several papers have appeared recently with regard to COVID-19 analysis, a summary of some of the methods, findings and issues are below.

Shin et al. [23] lay out various option for using CNNs in medical imaging, in the absence of large amounts of annotated medical images they hypothesise that transfer learning can be used with CNNs trained on the ImageNet dataset to build effective models for medical images, they test this by fine-tuning the AlexNet and GoogLeNet models with some success. Zhang et al. [46] use a dataset of 100 COVID-19 positive x-rays images combined with 1431 confirmed other pneumonia x-rays. Using [19] as pretrained on ImageNet as a feature extractor combined with a custom classification head and an anomaly detection head they obtain a sensitivity of 96% and specificity 70.65% on the dataset.

El-Din Hemdan, Shouman, and Karar [53] use a dataset of 50 images, with 25 COVID-19 positive a use 7 common pretrained models VGG19, DenseNet201, InceptionV3, ResNetV2, InceptionResNetV2, Xception, and MobileNetV2. They find VGG19 and DenseNet201 to be most effective.

Sethy et al. [43] combine a dataset of 127 COVID-19 , 127 pneumonia and 127 healthy images. The approach here is to use 11 popular pretrained models for feature detection cut off at specific CNN layers in combination with a support vector machine (SVM) as the multi-class classifier. They found ResNet50 with SVM calssifier to be optimal.

González et al. [37] report multi-label radiological findings on 23,159 test images obtaining an AUC of 0.94 for COVID-19 diagnosis.

Tartaglione et al. [44] use a dataset of 386 with 297 COVID-positive patients and develop a custom CNN named COVID-Net. The paper discusses various issues around bias in datasets and recommended image pre-processing steps,

namely lung segmentation before pre-training the CNN on a larger related dataset (pneumonia) to create a feature extractor and fine tuning it on the COVID-19 dataset. Pre-training and non-pre-training on different datasets is experimented with and they find that the sensitivity remains relatively constant but specificity is greatly improved with pre-training.

Maguolo and Nanni [39] take the bias concerns a setup further and investigate bias in the datasets, surmising that the neural networks could be learning patterns in the datasets not specific to COVID-19 but specific to the datasets. To achieve this, they remove the lungs entirely replacing them with black rectangles before training and testing their models. Training AlexNet on their dataset with no lungs they found a strong ability for the model to be able to detect a difference between the COVID-19 positive, pneumonia positive, and healthy images. Their paper highlights the issues around building a fair dataset and raise the question of exploring effective pre-processing that can remove the dataset specific features. The points they raise bring into question the validity of some the results in some of the early classification work carried out in relation to COVID-19 in some of the mentions papers.

3.4 Data Collection

From the early stages of the outbreak open source data collection efforts for COVID-19 positive CT and chest x-ray images have been underway with many of the deep learning models above using the data collected by Cohen, Morrison, and Dao [33] then sometimes in combination with many of the existing popular non-COVID-19 chest X-Ray datasets that exist for various model training or for classification testing purposes - e.g. the National Institute of Health of America (NIH) [55] , PadChest[32]. In June 2020 the BIMCV COVID-19+ [45] dataset

was released containing images from 1,311 COVID-19 patients, this dataset along with PadChest is used in the experiment and therefore covered in detail in chapter 5.

Chapter 4

Deep Learning

Machine Learning is the study of computer algorithms that improve automatically through experience[4]. Artificial neural networks(ANN) are a type of machine learning algorithm that when give data can learning structures and use that to make prediction on unseen data[7]. Deep learning in effect refers to multi-layer ANNs, "to learn representations of data with multiple levels of abstraction"[17].

4.1 Artificial Neural Networks

At its most basic a neural network consist of 3 sections, the input layer, hidden layer(s) and the output layer.Figure 4.1 shows a neural network with two hidden layers. Each layer can have multiple nodes. The input takes in the input features, the hidden layer consists of neurons or layers of neurons that connect the input to the output, and the output layer outputs at least one prediction. The connections between nodes have a weight value, and this weight value along with the input value feeds into an activation function to produce each nodes output value. Activation function are referenced in 4.2.1. The computation of all layers and nodes through to the last node is referred to as a forward pass. The difference

between the input and the output node is called the error, this error is apportioned backwards through the nodes in the network to update the weights, the weights values will change as an optimization algorithm selected for training converges, commonly using stochastic gradient descent[22]. This happens iteratively until some selected threshold is reached or training is stopped. This process is know as back-propagation[2]. This number of iterations allows us for refine our model weights. One forward pass and one back-propagation through the network for all training examples is known as an epoch.

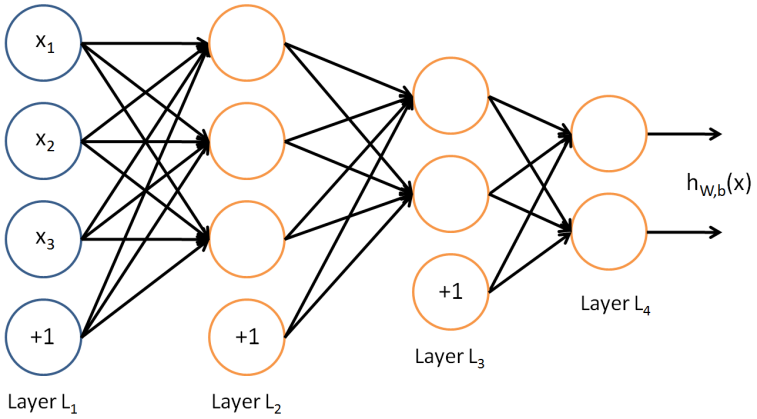


Figure 4.1: Multi-Layer Neural Network(reproduced from[61])

4.2 Convolutional Neural Networks

Convolutional Neural Networks(CNNs) are neural networks that through the use of hidden convolutional layers are shown to have the ability to extract useful features from image data[5].

Figure 4.2 is an illustration of a convolution layer in action, our 5*5 image in green, is traversed by a 3*3 filter in yellow with the filter values shown in red, to create the new feature value the area being traversed is multiplied by the corresponding filter value and the new convolved feature is displayed. The filter

is the element that allows us to detect patterns in an image.

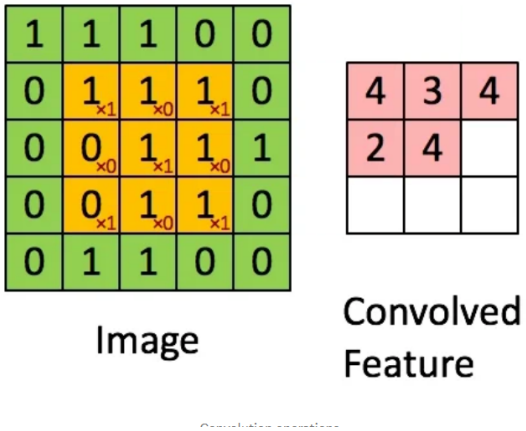


Figure 4.2: Convolution Layer(reproduced from[50])

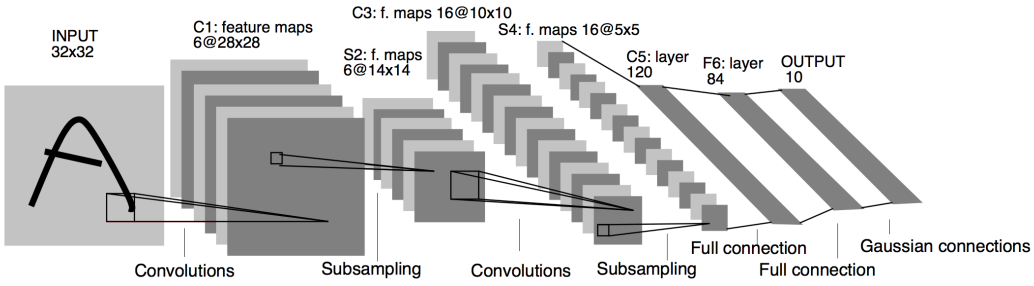


Figure 4.3: LeNet(reproduced from [5])

Figure 4.3 from Lecun et al. shows the LeNet CNN and is a good example of a simple CNN model. The LeNet-5 architecture consists of two sets of convolutional and average pooling layers, followed by a flattening convolutional layer, then two fully-connected layers and finally a softmax classifier. The first convolution layer has 6 feature maps and a 5*5 filter, the second layer is an average pooling layer, further reducing the image size. the third layer is another convolution layer this time with 16 feature maps, the forth layer is another average pooling later again reducing the image size. The effect of these filters with the convolution and pooling layers is to extract import features in an image, which

enables classification[5].

4.2.1 Activation Functions

The sigmoid activation function restricts the output between 0 and 1, popularised for nonlinearity and the computational simplicity of its derivative[3].

Sigmoid:

$$f(z) = \frac{1}{1 + \exp(-z)}$$

The rectified linear activation function (ReLu) is a simple calculation that is linear in its output for all positive inputs, and zero for all negative values, it allows for easier training of deep networks.[8].

Relu:

$$f(z) = \max(0, x)$$

4.2.2 Batch Normalization

Batch Normalization addressed the problem of internal covariate shift, parameters of layers changing as previous layers change during training. The normalization by batches between layers allows us to use larger learning rates and therefore speeds up training[16].

4.2.3 Zero Padding

Padding allows us to maintain the original input size by adding extra pixels at the edges an an image, when traversed by a filter the output image will remain the sane dimension as the input.

4.2.4 Pooling

Pooling is filter method similar to a convolutional layer but with a more simple computation, max or average pooling(selecting the max or average value in a filter area) will have the effect of extracting a higher level reduced spacial size representation of an image. Global Average Pooling computes a single average number for an entire feature map[14].

4.2.5 Dropout

Dropout is a regularization method, some probabilistic based number of layer outputs are randomly dropped. It can help reduce overfitting[13].

4.2.6 Dense/Fully Connected

A dense or gully connected layer is a layer that connects every neuron in one layer to every neuron in another layer[5].

4.3 DenseNet-121

As CNNs become deeper they become harder to train, DenseNet architecture addresses this by creating short paths from earlier layers to later layers[25].

Table 4.1 provides an outline of the 121 layers implicit in its name DenseNet-121. Initial convolution and pooling layers feed into dense blocks of size 6,12,24,16 each followed by transition layers until the classification layer. Detailed breakdown of a dense block layer can be seen in 4.5 highlighted in green consisting of batch normalisation, relu, 1*1 convolution on 4*k features, batch normalization, relu, 3*3 convolution layer. The transition layer 4.5 highlighted in purple consists of batch normalisation, relu, 1*1 convolution , 2*2 average pooling.

Focusing on the first dense block of size 6, each block output is carried forward and concatenated with the output of each of the blocks in front of it, as can be seen in Figure 4.4. The number of feature in each block is controlled by a growth rate k. This growth rate is a key feature in DenseNet, the $1 * 1$ convolution layer feature size in each dense block layer is of size $4 * k$. Controlling the feature size with a relatively small growth rate is shown to obtain state-of-the-art results on the ImageNet ILSVRC challenge[25]. Finally, the model ends with an average pooling layer before the classification layer.

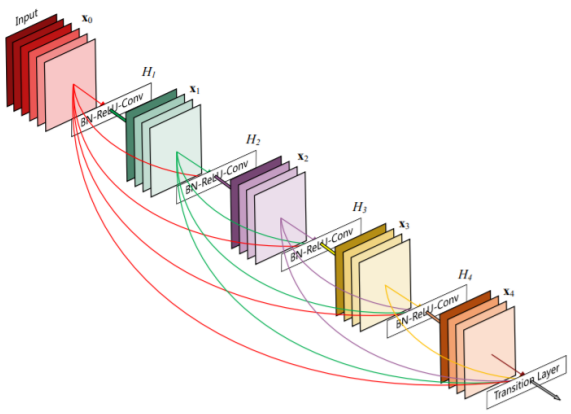


Figure 4.4: 5 Layer Dense Block(reproduced from[25])

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	[(None, 224, 224, 3)	0
zero_padding2d (ZeroPadding2D)	(None, 230, 230, 3)	0
conv1/conv (Conv2D)	(None, 112, 112, 64)	9408
conv1/bn (BatchNormalization)	(None, 112, 112, 64)	256
conv1/relu (Activation)	(None, 112, 112, 64)	0
zero_padding2d_1 (ZeroPadding2D	(None, 114, 114, 64)	0
pool1 (MaxPooling2D)	(None, 56, 56, 64)	0
conv2_block1_0_bn (BatchNormali	(None, 56, 56, 64)	256
conv2_block1_0_relu (Activation	(None, 56, 56, 64)	0
conv2_block1_1_conv (Conv2D)	(None, 56, 56, 128)	8192
conv2_block1_1_bn (BatchNormali	(None, 56, 56, 128)	512
conv2_block1_1_relu (Activation	(None, 56, 56, 128)	0
conv2_block1_2_conv (Conv2D)	(None, 56, 56, 32)	36864
conv2_block1_concat (Concatenat	(None, 56, 56, 96)	0
.....	.....	.....
pool2_bn (BatchNormalization)	(None, 56, 56, 256)	1024
pool2_relu (Activation)	(None, 56, 56, 256)	0
pool2_conv (Conv2D)	(None, 56, 56, 128)	32768
pool2_pool (AveragePooling2D)	(None, 28, 28, 128)	0
.....	.....	.....
bn (BatchNormalization)	(None, 7, 7, 1024)	4096
relu (Activation)	(None, 7, 7, 1024)	0
avg_pool (GlobalAveragePooling2	(None, 1024)	0
predictions (Dense)	(None, 197)	201925

Figure 4.5: DenseNet121 Layers

Layers	Output Size	DenseNet-121
Convolution	112*112	
Pooling	56*56	
Dense Block 1	56*56	[1*1->3*3]*6
Transitions layer 1	56*56->28*28	
Dense Block 2	28*28	[1*1->3*3]*12
Transitions layer 2	28*28->14*14	
Dense Block 3	14*14	[1*1->3*3]*24
Transitions layer 3	14*14->7*7	
Dense Block 4	7*7	[1*1->3*3]*16
Classification Layer	1*1	7*7, GAP, 197

Table 4.1: DenseNet121 Layers Breakdown

4.4 Transfer Learning

Many real world problems lack sufficient data to train a CNN from scratch, transfer learning is a learning framework that facilitates the reuse of features learned from training a model on one task to be applied to another task[9]. The theory being that low level features have an amount of generality to them and can be applied across domains allowing one to save time and to focus computational resources on training on the higher, additional or classification levels[15]. Pretrained ImageNet weights (1.4 million labeled images trained to identify 1,000 different classes) for a selection of popular CNN models are freely available from many deep learning frameworks including Keras, Tensorflow and PyTorch. [57][62][60].

Transfer learning has been applied successfully to various medial image tasks such as kidney detection problem in ultrasound images, interstitial lung disease classification and the detection of diabetic retinopathy. [21][23][30].

Chapter 5

Data Selection and Preparation

5.1 X-Ray Datasets

Dataset	#Images	#Labels	Diagnostic
ChestX-Ray 14(NIH)[55]	112k	14	Other
CheXpert[28]	224k	14	Other
MIMIC-CXR [29]	337k	14	Other
Padchest[32]	160k	297	Other
BIMCV COVID-19+[45]	5.3k	336	COVID-19
C19 Image Data Collection[33]	373	1	COVID-19

Table 5.1: X-Ray Datasets

A number of x-ray datasets were investigated as above in Table 5.1 before selecting the PadChest and BIMCV COVID-19+ datasets for use in the project. Sizeable datasets of chest x-ray and CT(computed tomography) scans have been around for a few of years. ChestX-Ray 14, CheXpert and MIMIC-CXR datasets with 1,2 and 300k+ images have similarities in their labelling methods, covering 14 major disease diagnosis. The Padchest dataset takes a more granular approach totalling 297 distinct labels. In terms of early COVID-19 datasets a number of early datasets were compiled, Cohen extracted a total of 373 at the time of his

paper from a range of open source avenues, including scraping from journals and published studies. The first substantial COVID-19 dataset became available in June 2020, BIMCV COVID-19+ with a total of 5318 chest x-ray and CT images of 1311 patients who tested positive for COVID-19 at least once. The differences in label methodology are explored in the next section.

5.2 Dataset Labels

NIH Chest X-ray Dataset of 14 Common Thorax Disease Categories:	
	PadChest label
Atelectasis	radiological finding
Cardiomegaly	radiological finding
Effusion	radiological finding
Infiltration	radiological finding
Mass	radiological finding
Nodule	radiological finding
Pneumonia	differential diagnosis
Pneumothorax	radiological finding
Consolidation	radiological finding
Edema	differential diagnosis
Emphysema	differential diagnosis
Fibrosis	differential diagnosis
Pleural_Thickening	radiological finding
Hernia	radiological finding

Figure 5.1: Label Differences

The PadChest dataset has 174 radiographic findings, 19 differential diagnosis, 104 anatomic locations, the BIMCV COVID-19+ dataset is build in the same fashion but contains two extra labels COVID-19 and COVID-19 uncertain, both labels have a unique CUI number and are recognised medical terms in the UMLS dictionary. I discard the anatomic locations in my combined dataset and focus on the radiographic and differential diagnosis labels.

The differentiation of radiographic findings and differential diagnosis as defined in section 2.3 are the standout differences from my data selection and other

main x-ray collections. In Figure 5.1 I contrast the label definition for the NIH chest x-ray dataset(labels are noted as 14 Common Thorax Disease Categories) with that of PadChest. Four of the NIH labels come under the heading of differential diagnosis as classified by UMLS labelling used by PadChest, which means a full clinical evaluation would be needed for diagnosis, not just x-ray evaluation. It should be noted that COVID-19 is classified as a differential diagnosis in the BIMCV dataset.

5.3 Labels Inferred

Of the labels that come with the PadChest dataset 27% are manually annotated by physicians, with the remaining 73% extracted using a supervised recurrent neural network, the BIMCV COVID-19+ dataset uses the same model to extract labels but is retrained to include COVID-19[32][45]. The labels are then mapped to match medical terms in UMLS dictionary. In the UMLS dictionary medical labels have a hierarchical structure [6]. In Figure 5.2, the category column shows the hierarchical label structure for COVID-19, it is a subset of viral pneumonia, which is a subset of pneumonia, which is a differential diagnosis. The report label column reflects the labelling that comes with the BIMCV COVID-19+ dataset for a COVID-19 positive patient. The inferred label column represents my application of the inferred label structure present in the UMLS dictionay,i.e. I take an x-ray labelled as COVID-19 positive and label its higher level label as also poitive. I propagate this methodology throughout the labelling of my dataset for all labels. This enables my model to be trained on labels at all levels of the UMLS hierarchy.

Category	Report Label	Inferred Label
differential diagnosis	0	1
└ pneumonia	0	1
└ atypical pneumonia	0	0
└ viral pneumonia	0	1
└ COVID 19	1	1

Figure 5.2: Label Structure

5.4 Label Quality

In assessing radiological reports the intended audience and the patient context should be kept in mind, the reports are written for other medical professionals to interpret. Some instances in my dataset come with the label no change(excluded from dataset), which only has meaning with knowledge of patient case history. Radiologists can often disagree on interpretation of x-rays so one approach to add confidence to labelling is carried out in the CheXNet pneumonia detection paper in which they add credence to their experiment results by usings a test set manually relabelled by radiologistsRajpurkar et al. The text mining process for extracting the labels used to train their model is said to be 90% accurate.

5.5 Dataset Balance

Figure 5.3 takes a look at the balance of the labels in PadChest and the BIMCV COVID-19+ datasets. The figure represents each pathology as a weight of its own dataset relative to the other dataset. Its an interesting way to learn more about the respective datasets, although the much smaller size of the BIMCV COVID-19+ dataset should be noted. Figure 5.3 for example shows that the label of central venous catheter is much more prevalent in the BIMCV COVID-19+ dataset, as its a catheter placed into a large vein its likely to be indicitive

of a patient in a serious condition. Known COVID-19 associated radiographic findings as referenced in section 3.3 consolidation, alveolar pattern and ground glass pattern are shown to be more heavily weighted in the BIMCV COVID-19+ dataset and thus add credence to the evidence that they are common COVID-19 associated radiographic findings.

The demographic of the datasets reflect what is commonly known about COVID-19, it tends to have more severe effects on the the old population and since our datasets are of people deemed to necessitate an x-ray the slightly older COVID-19 population as can be seen in Figures 5.4 and 5.5.

Figure 5.3: Datasets Pathology balance

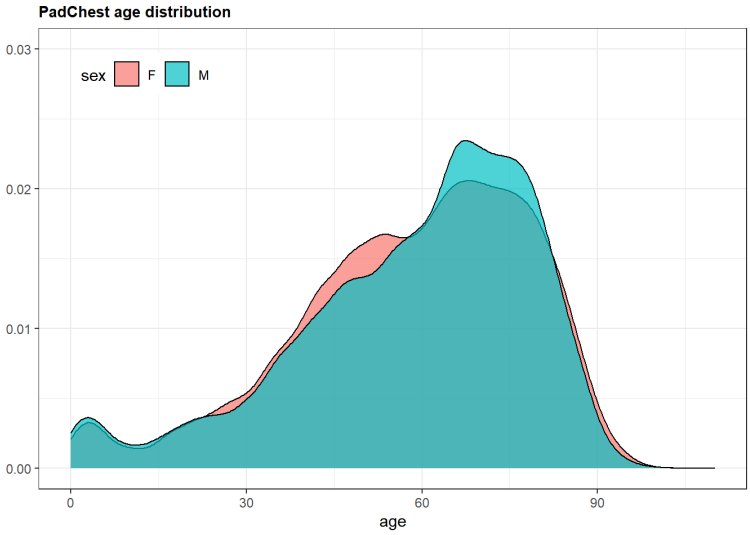


Figure 5.4: PadChest Demographic

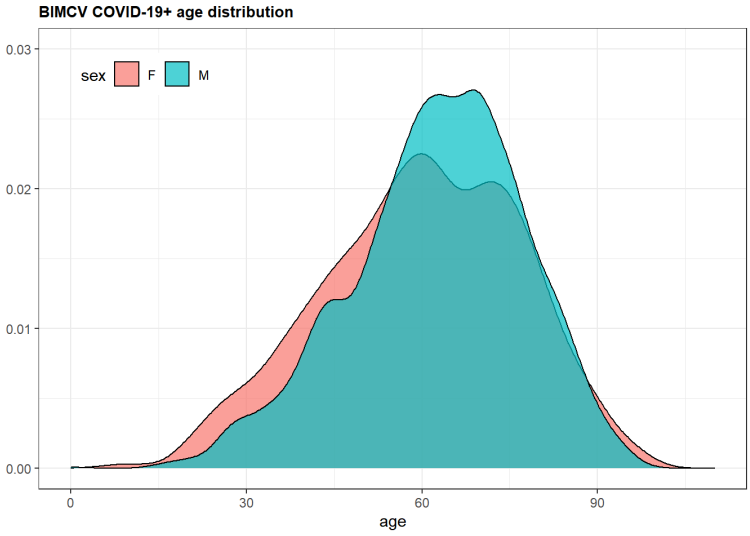


Figure 5.5: BIMCV COVID-19+ Demographic

5.6 Data Cleaning & Merging

The PadChest dataset is 4TB in total size and contains over 160,000 images, 50% of the dataset is used in this project. The images come in .png format and I convert them to .jpeg files for practical reasons around file transfer and training. The PadChest metadata age, sex, position, labels comes conveniently in a single .csv file.

BIMCV COVID 19+ follows the MIDS structure and each has a corresponding .json file containing metadata from the X-Ray machine in DICOM format. Data is merged from a number of different files to obtain the demographic information, the position and the labels. I also convert the images from .png to .jpeg. I then apply one hot encoding to the inferred labels. One hot encoding is the process of transforming categorical variables into a form more attuned to machine learning algorithm interpretation. It transforms d categorical variables into d distinct binary features[26]. Images are filtered out if their label is no change, only P-A and A-P positions are used for training the model. Images are filtered out if any metadata is missing.

5.6.1 Image Transformation

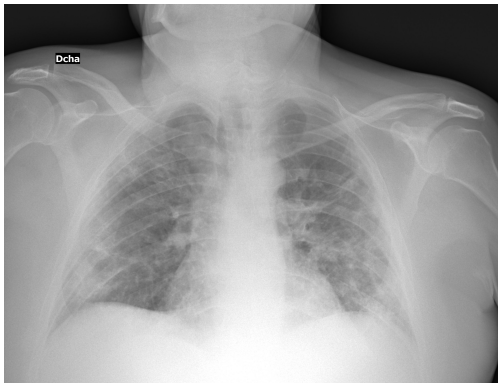
BIMCV COVID 19+ dataset images all need some image processing to bring them inline with the PadChest dataset. Approximately 1500 monochrome 1 images are converted and 750 monochrome 2 images are inverted using the Fiji image processing package as can be seen in Figure 5.6 and 5.7 [54] . In Figure 5.6(a) the contrast level is automatically normalized upon importation into Fiji and the saving to .jpeg completes the conversion. Note the image in5.6(a) is not black but of particularly low contrast and requires careful visual inspection.

(a) Monochrome 1

(b) Converted

Figure 5.6: Monochrome 1 to Monochrome 2

(a) Monochrome 2 inverted



(b) Monochrome 2 corrected

Figure 5.7: Inverting Monochrome 2

Chapter 6

Methodology

6.1 Train Test Split

My combined PadChest & BIMCV COVID-19+ dataset is split 60% train , 20% validate , 20% holdout or test set with exact figures in Table 6.1. Patients with multiple images are kept on the same side of the split to prevent data leakage and a balanced demographic of the splits can be seen in Figures 6.1, 6.2 and 6.3.

Split	Total	PadChest	BIMCV
Train	30590	29500	1090
Validation	10255	9890	366
Test	10159	9781	377

Table 6.1: Train Test Split

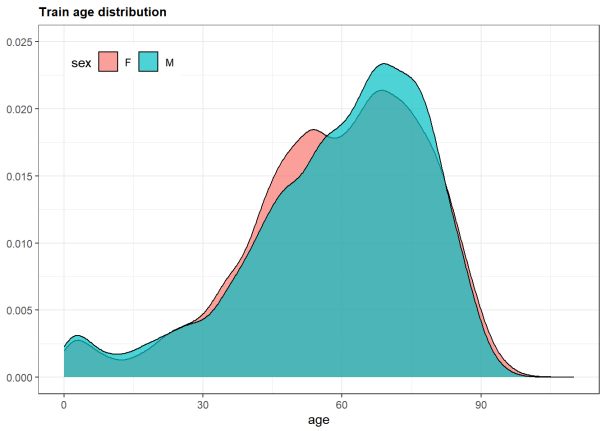


Figure 6.1: Train Demographic

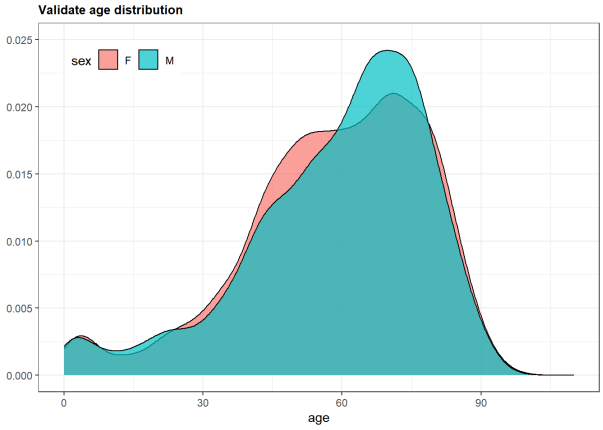


Figure 6.2: Validation Demographic

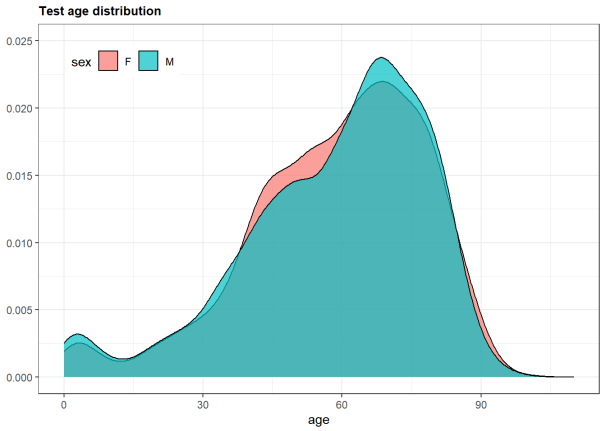


Figure 6.3: Test Demographic

6.2 Implementation

Keras a deep learning API that sits on top of TensorFlow(an end-to-end open source platform for machine learning) is used to implement my 2 deep learning models[57][62].

For computing resources Google AI Platform Notesbooks[47] is used. Data is first uploaded to Google Cloud Storage, then re-imported into an instance of JupyterLab running on AI Platform Notebooks. For my cloud based notebook training I found a Nvidia Telsa T4 GPU, Figure 6.5, with 16 CPUs, Figure6.4, to be optimal, the high number of CPUs enabling a good flow of data to keep the GPU busy. Average training time took 10 minutes per epoch, with 30 epochs per model training. Figure 6.4 shows a snapshot taken during training, displaying utilization of all 16 CPUs. Figure 6.5 is a snapshot during training of the GPU usage, 65% in this moment, both important metrics to monitor when training to ensure resources are being utilised efficiently.

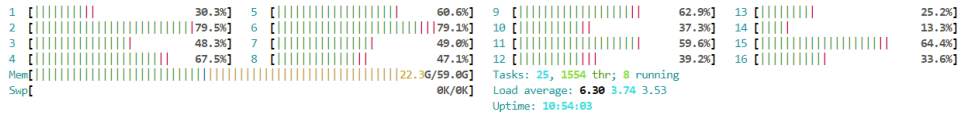


Figure 6.4: CPU load

```
+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID   Type   Process name                             Usage      |
|=============================================================================|
|    0      2052      C   /opt/conda/bin/python                      14719MiB |
+-----------------------------------------------------------------------------+
(base) jupyter@tensorflow-2-1-20200818-162126:~$ nvidia-smi
Fri Sep  4 20:07:26 2020
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 418.87.01    Driver Version: 418.87.01    CUDA Version: 10.1     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla T4            Off  | 00000000:00:04.0 Off |                    0 |
| N/A   68C    P0    52W /  70W |  14729MiB / 15079MiB |     65%      Default |
+-------------------------------+----------------------+----------------------+
```

Figure 6.5: GPU load

6.3 Class Imbalance

The frequency of positive to negative labels in out dataset is highly imbalanced as we can see in Figure 6.6. With COVID-19 for example only 5% of our data has a COVID-19 positive label. When it comes to training and monitoring the model on a loss function, if we have a label which is dominated by either a positive or negative label the model will train toward the dominant label as it will contribute more to the loss. To make sure less prevalent class will be detected the imbalance has to be addressed. I used the method employed by [51] to balance the class frequency.

Figure 6.6 shows the frequency of each class in the training set, Figure 6.7 highlights the imbalance, and Figure 6.8 the result of reweighing the class imbalance after implementation of the below formula.

Normal cross-entropy loss formula is:

$$\mathcal{L}_{cross-entropy}(x_i) = -(y_i \log(f(x_i)) + (1 - y_i) \log(1 - f(x_i))),$$

A weight factor is required such that;

$$w_{pos} \times freq_p = w_{neg} \times freq_n,$$

this is achieved by setting;

$$w_{pos} = freq_{neg}, w_{neg} = freq_{pos}$$

The resulting weight function introduced for each class such that the new cross-entropy loss functions is:

$$\mathcal{L}^w_{cross-entropy}(x) = -(w_p y \log(f(x)) + w_n(1 - y) \log(1 - f(x))).$$

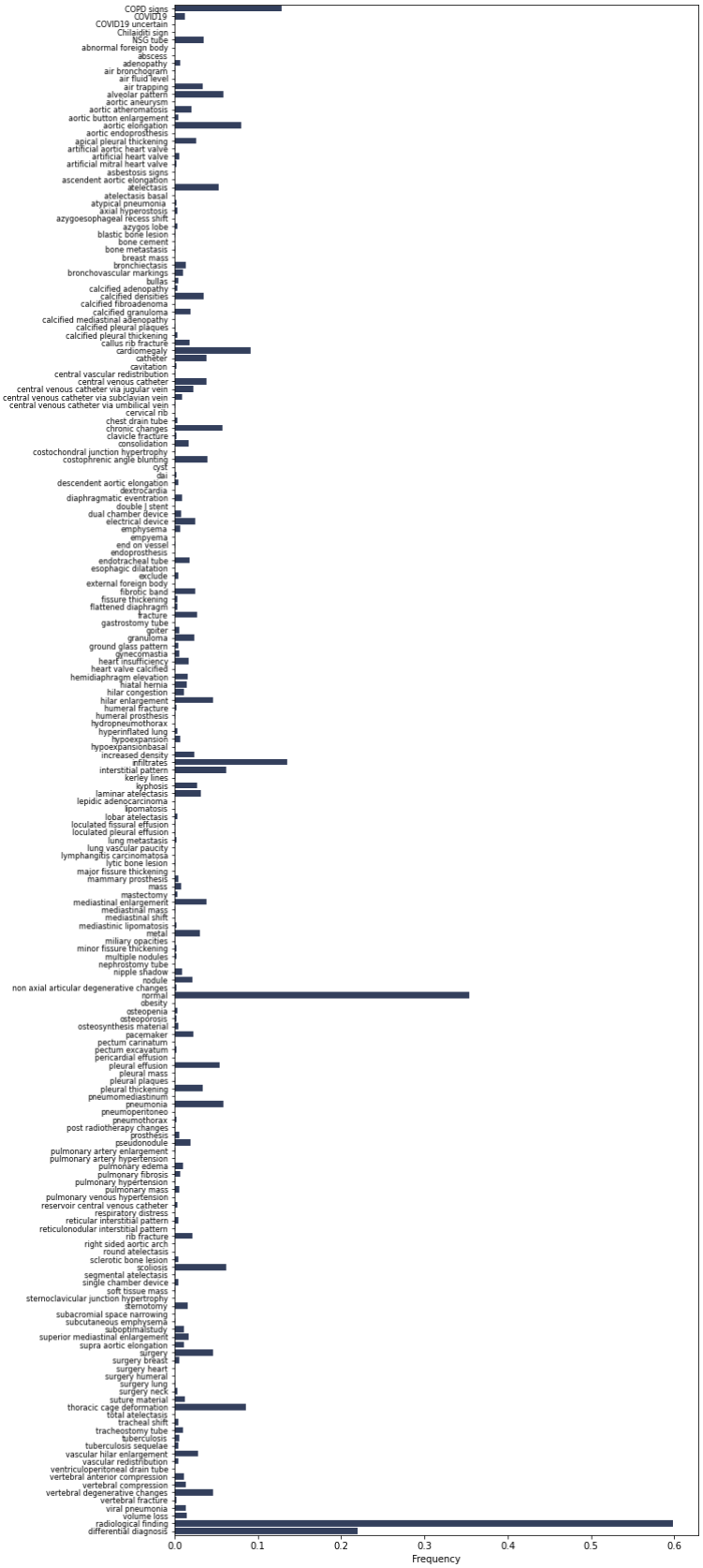


Figure 6.6: Class Frequency

Figure 6.7: Class Imbalanced

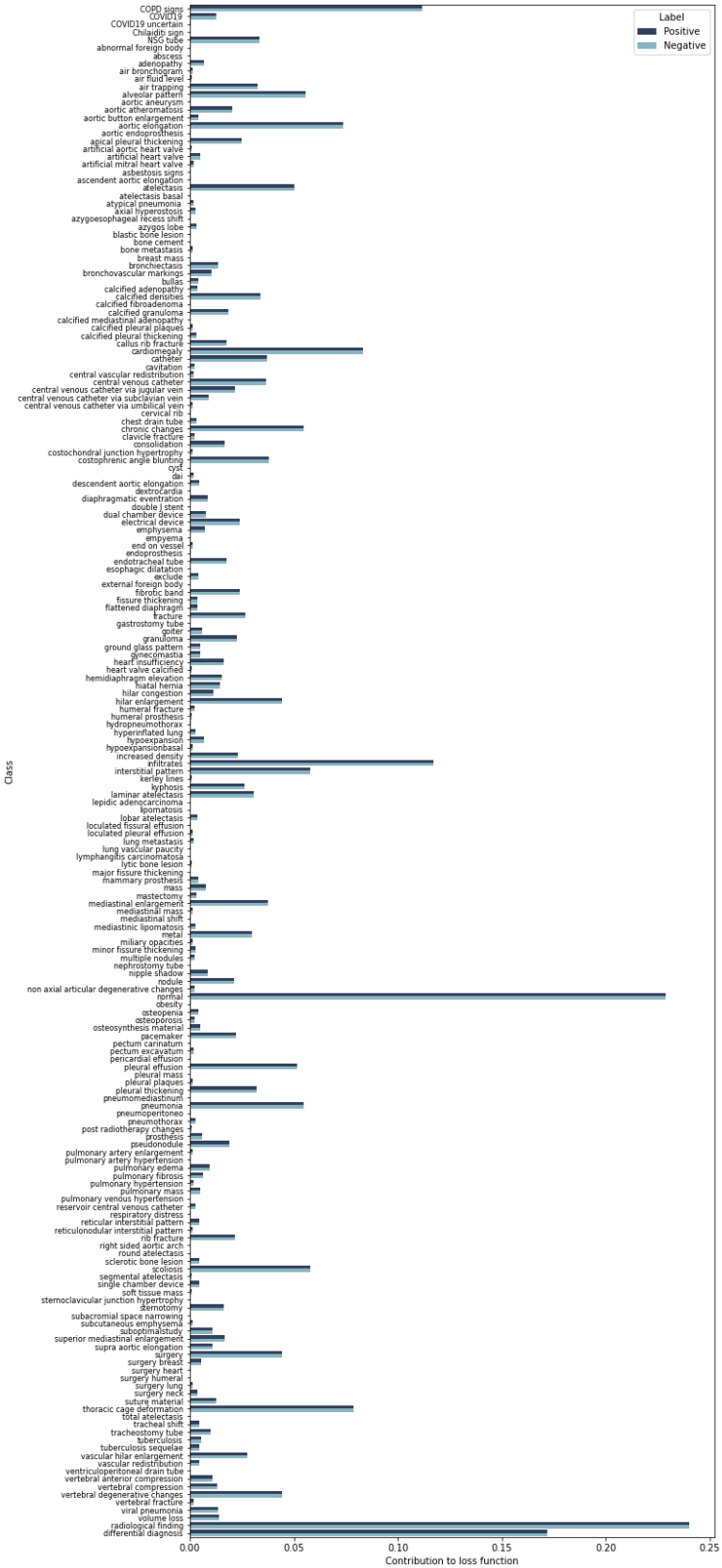


Figure 6.8: Class Balanced

6.4 Model Summary

Two models are trained on the dataset. A model consisting of a DenseNet-121 architecture with pre-trained ImageNet weights with a sigmoid classification layer to perform multi-layer classification on the 197 labels. Only the parameters on classification layer are trained, with the convolutional base frozen. A second model consisting of a DenseNet121 architecture with pre-trained ImageNet weights with additional dense and dropout layers and global average pooling layers as structured in table 6.2. The extra layers create extra parameters to be trained as in table 6.3 to be referred to as DenseNet121+ from here on. The convolutional base is frozen with the parameters in the additional layers to be trained.

The SGD(Stochastic gradient descent) optimizer with momentum is used with a learning rate of 0.01 with a decay rate of 0.01/30(the no. of epochs). The learning rate controls the size of the step taken in the direction taken by gradient descent[10]. Table 6.3 shows the total and trainable parameters for both models.

Densenet121	Densenet121+
Last Layer	Dense 512 relu
sigmoid 197	Dropout 0.2
—	Dense 512 relu
—	Dropout 0.2
—	GlobalAveragePooling2D
—	sigmoid 197

Table 6.2: Model Layout

	Densenet121	Densenet121+
Total Parameters:	7,239,429	7,926,021
Trainable Parameters:	201,925	888,517
Non-trainable Parameters:	7,037,504	7,037,504

Table 6.3: Model Parameters

6.5 Training

The training loss for my two models DenseNet121 and DenseNet121+ can be see in figure 6.9. DenseNet121+ performs best and has a lower loss on both the training and validation datasets in comparison to DenseNet121.

To to evaluate the models performance on label identification I use the AUROC(Area Under the Receiver Operating Characteristics) curve [1]. A higher AUC if representative of a models ability to distinguish between the two classes, disease or no disease in this case. The training AUC metric is an averaged figure across all labels. The training and validation AUC for my two models DenseNet121 and DenseNet121+ can be see in figure 6.10. On the training dataset DenseNet121+ performs slightly better, but on the validation dataset DenseNet121 performs slightly better, the model with fewer trainable parameters performing better.

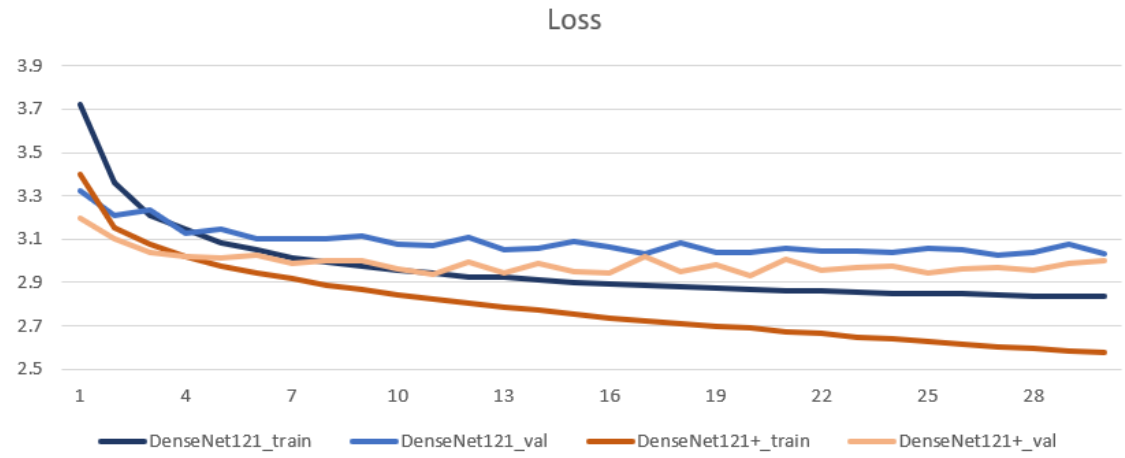


Figure 6.9: Training Loss

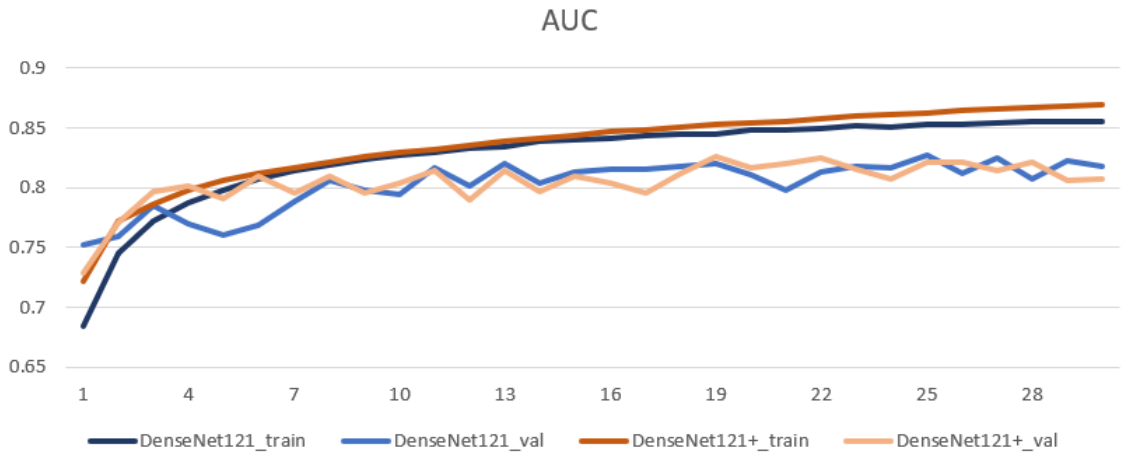


Figure 6.10: Training AUC

Chapter 7

Results

The DenseNet121 and the DenseNet121+ trained models are run on the test set of 10,158 chest x-rays for 197 labels and the results for COVID-19 and associated radiographic findings, consolidation, alveolar pattern and ground glass pattern are presented in Figure 7.1 and Table 7.1. Results for all 197 differential diagnosis and radiographic labels can be found in appendix A(labels with NA are presented to maintain the structure of the table, but are labels that do not appear in any of the reports in the dataset).

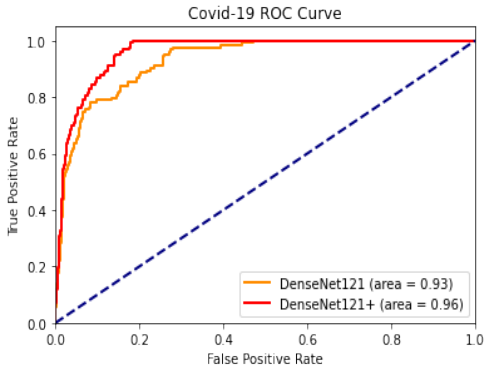
The performance of the models are broadly similar, with slightly higher AUC for DenseNet121+ for the highlighted diagnosis, but interestingly the average AUC is slightly higher for DenseNet121. A look at appendix A show better performance for DenseNet121 for some individual diagnosis. A longer training period or more data would be needed further establish a significant differences in performance, as it may refine the extra trainable parameters in DenseNet121+.

In terms of comparative results González et al. train 3 models on a combination of the PadChest and BIMCV COVID-19+ datasets, their best performing model EfficientNetB4 returns a COVID-19 AUC of 0.94, and 0.92, 0.84, 0.92 for consolidation, ground glass pattern and alveolar pattern respectively, performing

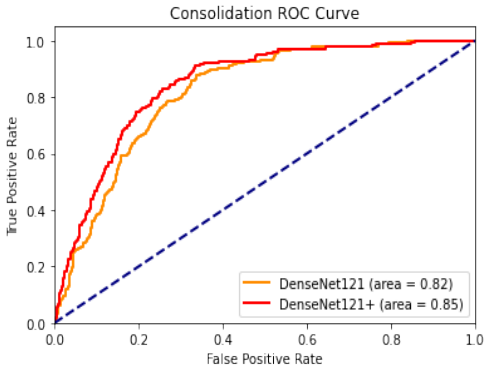
only better for ground glass pattern.

	DenseNet121	DenseNet121+
Average AUC	.82	.81
COVID-19	.93	.96
Consolidation	.82	.85
Alveolar Pattern	.89	.90
Ground Glass Pattern	.82	.85

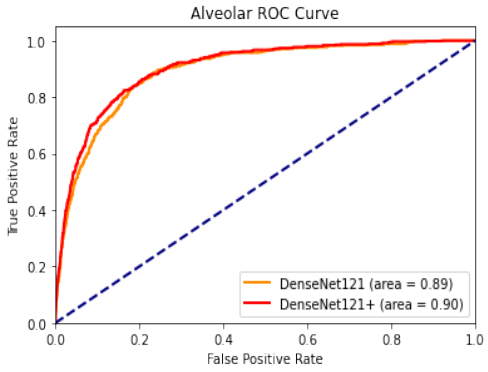
Table 7.1: Model Results



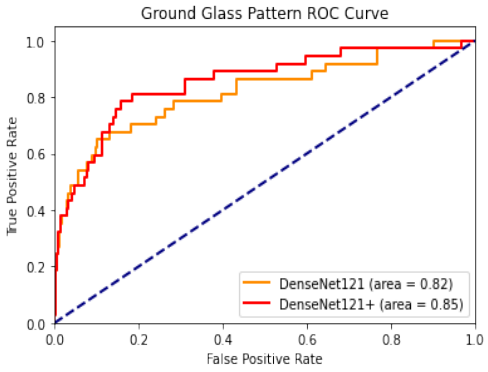
(a) COVID-19



(b) Consolidation



(c) Alveolar Pattern



(d) Ground Glass Pattern

Figure 7.1: ROC

Chapter 8

Conclusion

According to the WHO as at August 31st 2020, nearly 25 million cases and 800,000 death have been reported since the start of the COVID-19 outbreak[58]. Chest x-rays are the most commonly available medical imaging tool and according to Kesselman et al. there is a shortage of experts radiologists in developing and emerging countries leading to increase in mortality[12]. These circumstances provide the motivation to examine automated AI based diagnosis. I develop a model which detects COVID-19 with a AUC of 0.96, and in total for detection of 197 differential diagnosis and radiographic finding labels. As stipulated in the WHO guidelines there are circumstances where medical imaging can appropriately be used for COVID-19 detection. A test set independently labelled by expert radiologists would be the next step in the project to verify the results, but it is hoped that by carrying out research such as this another step can be taken in providing additional diagnosis tools in circumstances where the resources for lab testing is not available or the radiological expertise is limited.

Bibliography

[1] J.A. Hanley and Barbara Mcneil. "The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve". In: *Radiology* 143 (May 1982), pp. 29–36. DOI: 10.1148/radiology.143.1.7063747.

[2] D. Rumelhart, Geoffrey E. Hinton, and R. J. Williams. "Learning representations by back-propagating errors". In: *Nature* 323 (1986), pp. 533–536.

[3] Jun Han and Claudio Moraga. "The influence of the sigmoid function parameters on the speed of backpropagation learning". In: *From Natural to Artificial Neural Computation*. Ed. by José Mira and Francisco Sandoval. Berlin, Heidelberg: Springer Berlin Heidelberg, 1995, pp. 195–201.

[4] T.M. Mitchell. *Machine Learning*. McGraw-Hill International Editions. McGraw-Hill, 1997. ISBN: 9780071154673.

[5] Yann Lecun et al. "Gradient-Based Learning Applied to Document Recognition". In: *Proceedings of the IEEE* 86 (Dec. 1998), pp. 2278–2324. DOI: 10.1109/5.726791.

[6] Olivier Bodenreider. "The Unified Medical Language System (UMLS): Integrating Biomedical Terminology". In: *Nucleic acids research* 32 (Feb. 2004), pp. D267–70. DOI: 10.1093/nar/gkh061.

[7] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. 3rd. USA: Prentice Hall Press, 2009. ISBN: 0136042597.

[8] Vinod Nair and Geoffrey Hinton. "Rectified Linear Units Improve Restricted Boltzmann Machines Vinod Nair". In: vol. 27. June 2010, pp. 807–814.

[9] S. J. Pan and Q. Yang. "A Survey on Transfer Learning". In: *IEEE Transactions on Knowledge and Data Engineering* 22.10 (2010), pp. 1345–1359.

[10] Yoshua Bengio. "Practical Recommendations for Gradient-Based Training of Deep Architectures". In: *Neural Networks: Tricks of the Trade*. 2012.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems 25*. Ed. by F. Pereira et al. Curran Associates, Inc., 2012, pp. 1097–1105. URL: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

[12] Suhail Raoof et al. "Interpretation of Plain Chest Roentgenogram". In: *Chest* 141.2 (2012), pp. 545–558. ISSN: 0012-3692. DOI: https://doi.org/10.1378/chest.10-1302. URL: http://www.sciencedirect.com/science/article/pii/S0012369212600968.

[13] G. E. Dahl, T. N. Sainath, and G. E. Hinton. "Improving deep neural networks for LVCSR using rectified linear units and dropout". In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2013, pp. 8609–8613.

[14] M. Lin, Q. Chen, and S. Yan. "Network In Network". In: *CoRR* abs/1312.4400 (2014).

[15] A. S. Razavian et al. "CNN Features Off-the-Shelf: An Astounding Baseline for Recognition". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2014, pp. 512–519.

[16] Sergey Ioffe and Christian Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". In: *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*. Ed. by Francis R. Bach and David M. Blei. Vol. 37. JMLR Workshop and Conference Proceedings. JMLR.org, 2015, pp. 448–456. URL: http://proceedings.mlr.press/v37/ioffe15.html.

[17] Yann LeCun, Y. Bengio, and Geoffrey Hinton. *Deep Learning*. May 2015. DOI: 10.1038/nature14539.

[18] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *International Conference on Learning Representations*. 2015.

[19] K. He et al. "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778.

[20] Andrew Kesselman et al. "2015 RAD-AID Conference on International Radiology for Developing Countries: The Evolving Global Radiology Landscape". In: *Journal of the American College of Radiology* 13.9 (2016), pp. 1139–1144. ISSN: 1546-1440. DOI: https://doi.org/10.1016/j.jacr.2016.03.028. URL: http://www.sciencedirect.com/science/article/pii/S1546144016301727.

[21] Hariharan Ravishankar et al. "Understanding the Mechanisms of Deep Transfer Learning for Medical Images". In: *Deep Learning and Data Labeling for Medical Applications*. Ed. by Gustavo Carneiro et al. Cham: Springer International Publishing, 2016, pp. 188–196. ISBN: 978-3-319-46976-8.

[22] Sebastian Ruder. "An overview of gradient descent optimization algorithms". In: (Sept. 2016). arXiv: 1609.04747 [cs.LG].

[23] H. Shin et al. "Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning". In: *IEEE Transactions on Medical Imaging* 35.5 (2016), pp. 1285–1298.

[24] Andre Esteva et al. "Dermatologist-level classification of skin cancer with deep neural networks". In: *Nature* 542.7639 (Feb. 2017), pp. 115–118. ISSN: 14764687. DOI: 10.1038/nature21056.

[25] G. Huang et al. "Densely Connected Convolutional Networks". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 2017, pp. 2261–2269.

[26] Kedar Potdar, Taher S Pardawala, and Chinmay D Pai. "A comparative study of categorical variable encoding techniques for neural network classifiers". In: *International journal of computer applications* 175.4 (2017), pp. 7–9.

[27] Pranav Rajpurkar et al. "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning". In: *ArXiv* abs/1711.05225 (2017).

[28] J. Irvin et al. "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison". In: *AAAI*. 2019.

[29] Alistair Johnson et al. "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports". In: *Scientific Data* 6 (Dec. 2019), p. 317. DOI: 10.1038/s41597-019-0322-0.

[30] Rajiv Raman et al. *Fundus photograph-based deep learning algorithms in detecting diabetic retinopathy.* Jan. 2019. DOI: 10.1038/s41433-018-0269-y.

[31] BIMCV. "MIDS". In: (2020). URL: https://bimcv.cipf.es/bimcv-projects/mids/.

[32] Aurelia Bustos et al. "PadChest: A large chest x-ray image dataset with multi-label annotated reports". In: *Medical Image Analysis* 66 (2020), p. 101797. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2020.101797. URL: http://www.sciencedirect.com/science/article/pii/S1361841520301614.

[33] Joseph Paul Cohen, Paul Morrison, and Lan Dao. "COVID-19 Image Data Collection". In: *Preprint* (Mar. 2020). URL: http://arxiv.org/abs/2003.11597.

[34] Victor Corman et al. *Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR*. Jan. 2020. DOI: 10.2807/1560-7917.ES.2020.25.3.2000045.

[35] Beatriz Garcia Santa Cruz et al. *On the Composition and Limitations of Publicly Available COVID-19 X-Ray Imaging Datasets*. 2020. arXiv: 2008.11572 [eess.IV].

[36] Yicheng Fang et al. "Sensitivity of Chest CT for COVID-19: Comparison to RT-PCR". In: *Radiology* (May 2020), p. 200432. DOI: 10.1148/radiol.2020200432.

[37] Germán González et al. "UMLS-ChestNet: A deep convolutional neural network for radiological findings, differential diagnoses and localizations of COVID-19 in chest x-rays". In: *Preprint* (June 2020). URL: http://arxiv.org/abs/2006.05274.

[38] Aguiar.D Lobrinus.JA Schibler.M Fracasso.T Lardi.C. "Inside the lungs of COVID-19 disease". In: (2020). URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7248187/.

[39] Gianluca Maguolo and Loris Nanni. "A Critic Evaluation of Methods for COVID-19 Automatic Detection from X-Ray Images". In: *Preprint* (Apr. 2020). URL: http://arxiv.org/abs/2004.12823.

[40] Scott Mayer McKinney et al. "International evaluation of an AI system for breast cancer screening". In: *Nature* 577.7788 (Jan. 2020), pp. 89–94. ISSN: 14764687. DOI: 10.1038/s41586-019-1799-6.

[41] "Medical Connections". In: (2020). URL: https://www.medicalconnections.co.uk/kb/Photometric-Interpretations/.

[42] World Health Organization. "Use of chest imaging in COVID-19". In: (2020). URL: https://www.who.int/publications/i/item/use-of-chest-imaging-in-covid-19.

[43] Prabira Kumar Sethy et al. "Detection of coronavirus Disease (COVID-19) based on Deep Features and Support Vector Machine". In: *International Journal of Mathematical, Engineering and Management Sciences* 5.4 (Aug. 2020), pp. 643–651. DOI: 10.33889/ijmems.2020.5.4.052.

[44] Enzo Tartaglione et al. "Unveiling COVID-19 from Chest X-ray with deep learning: a hurdles race with small data". In: *Preprint* (Apr. 2020). URL: http://arxiv.org/abs/2004.05405.

[45] Maria de la Iglesia Vayá et al. "BIMCV COVID-19+: a large annotated dataset of RX and CT images from COVID-19 patients". In: *Preprint* (June 2020). URL: http://arxiv.org/abs/2006.01174.

[46] Jianpeng Zhang et al. "COVID-19 Screening on Chest X-ray Images Using Deep Learning based Anomaly Detection". In: *Preprint* (Mar. 2020). URL: http://arxiv.org/abs/2003.12338.

[47] *AI Platform Notebooks*. URL: https://cloud.google.com/ai-platform-notebooks.

[48] Tao Ai et al. *Correlation of Chest CT and RT-PCR Testing in Coronavirus Disease 2019 (COVID-19) in China: A Report of 1014 Cases*. Tech. rep.

[49] Harrison X Bai et al. *Performance of radiologists in differentiating COVID-19 from viral pneumonia on chest CT*. Tech. rep.

[50] *Convolution Image*. URL: https://towardsdatascience.com/build-your-own-convolution-neural-network-in-5-mins-4217c2cf964f.

[51] Coursera. *AI for Medical Diagnosis*. URL: https://www.coursera.org/learn/ai-for-medical-diagnosis.

[52] DICOM. *Digital Imaging and Communications in Medicine*. URL: https://www.dicomstandard.org/.

[53] Ezz El-Din Hemdan, Marwa A Shouman, and Mohamed Esmail Karar. *COVIDX-Net: A Framework of Deep Learning Classifiers to Diagnose COVID-19 in X-Ray Images*. Tech. rep.

[54] Fiji. *Fiji image processing*. URL: https://fiji.sc/.

[55] National Institutes of Health. *NIH Dataset*. URL: https://www.nih.gov/.

[56] ImageNet. *ImageNet*. URL: http://www.image-net.org/.

[57] *Keras*. URL: https://keras.io/.

[58] World Health Organization. *Situation Reports*. URL: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports.

[59] World Health Organization. *WHO Covid-19 Timeline*. URL: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/events-as-they-happen.

[60] *PyTorch*. URL: https://pytorch.org/.

[61] *Stanford NN Image*. URL: http://deeplearning.stanford.edu/tutorial/supervised/MultiLayerNeuralNetworks/.

[62] *Tensorflow*. URL: https://www.tensorflow.org/.

[63] Ho Yuen Frank Wong et al. *Frequency and Distribution of Chest Radiographic Findings in COVID-19 Positive Patients Authors*. Tech. rep.

Appendix A

Hierarchical Label Results

Label	DenseNet121	DenseNet121+
normal	0.81	0.82
exclude	0.55	0.5
suboptimal study	0.81	0.8
radiological finding	0.79	0.8
unchanged	NA	NA
obesity	0.92	0.88
chronic changes	0.78	0.79
calcified densities	0.67	0.68
calcified granuloma	0.68	0.69
calcified adenopathy	0.68	0.7
calcified mediastinal adenopathy	0.99	0.61
calcified pleural thickening	0.84	0.83
calcified pleural plaques	0.71	0.85
heart valve calcified	0.83	0.87
calcified fibroadenoma	0.76	0.67
calcified fibroadenoma	0.76	0.67
calcified granuloma	0.68	0.69
end on vessel	0.61	0.77
adenopathy	0.59	0.65
calcified adenopathy	0.68	0.7
nodule	0.66	0.66
multiple nodules	0.74	0.74
pseudonodule	0.66	0.67
nipple shadow	0.77	0.77
end on vessel	0.61	0.77
abscess	0.24	0.65
cyst	NA	NA
cavitation	0.81	0.81
fibrotic band	0.71	0.72
volume loss	0.76	0.79
hypoexpansion	0.86	0.88
bullas	0.76	0.81
pneumothorax	0.78	0.81
hydropneumothorax	NA	NA
pneumoperitoneo	0.71	0.69
pneumomediastinum	0.62	0.77
subcutaneous emphysema	0.89	0.91
hyperinflated lung	0.69	0.73
flattened diaphragm	0.77	0.76
lung vascular paucity	0.52	0.62
air trapping	0.77	0.79
bronchiectasis	0.77	0.79

Figure A.1: Results Part 1

infiltrates		0.82	0.83
interstitial pattern		0.78	0.8
ground glass pattern		0.82	0.85
reticular interstitial pattern		0.75	0.77
reticulonodular interstitial pattern		0.67	0.73
miliary opacities		0.49	0.65
alveolar pattern		0.89	0.9
consolidation		0.82	0.85
air bronchogram		0.82	0.8
air bronchogram		0.82	0.8
bronchovascular markings		0.83	0.86
air fluid level		0.6	0.73
increased density		0.77	0.8
atelectasis		0.74	0.75
total atelectasis		0.96	0.99
lobar atelectasis		0.78	0.83
segmental atelectasis		0.65	0.74
laminar atelectasis		0.69	0.69
round atelectasis		0.95	0.3
atelectasis basal	2	0.35	0.41
mediastinal shift		0.84	0.84
azygos lobe		0.64	0.66
fissure thickening		0.67	0.71
minor fissure thickening		0.69	0.72
major fissure thickening		0.66	0.68
loculated fissural effusion		0.75	0.87
pleural thickening		0.74	0.77
apical pleural thickening		0.75	0.77
calcified pleural thickening		0.84	0.83
pleural plaques		0.73	0.83
calcified pleural plaques		0.71	0.85
pleural effusion		0.92	0.93
loculated pleural effusion		0.83	0.89
loculated fissural effusion		0.75	0.87
hydropneumothorax		NA	NA
empyema		NA	NA
hemothorax		NA	NA
pleural mass		NA	NA
costophrenic angle blunting		0.77	0.79
vascular redistribution		0.81	0.83
central vascular redistribution		0.82	0.86
hilar enlargement		0.67	0.68
adenopathy		0.59	0.65
vascular hilar enlargement		0.68	0.71
pulmonary artery enlargement		0.52	0.67
hilar congestion		0.87	0.89
cardiomegaly		0.84	0.86
pericardial effusion		0.71	0.81

Figure A.2: Results Part 2

kerley lines	0.64	0.77
dextrocardia	0.71	0.69
right sided aortic arch	0.98	0.27
aortic atheromatosis	0.8	0.82
aortic elongation	0.84	0.86
descendent aortic elongation	0.86	0.87
ascendent aortic elongation	0.45	0.57
aortic button enlargement	0.71	0.74
supra aortic elongation	0.84	0.86
aortic aneurysm	0.71	0.58
mediastinal enlargement	0.79	0.79
superior mediastinal enlargement	0.79	0.8
superior mediastinal enlargement	0.79	0.8
superior mediastinal enlargement	0.79	0.8
descendent aortic elongation	0.86	0.87
ascendent aortic elongation	0.45	0.57
aortic aneurysm	0.71	0.58
mediastinal mass	0.64	0.78
hiatal hernia	0.83	0.84
tracheal shift	0.72	0.76
mass	0.77	0.78
mediastinal mass	0.64	0.78
breast mass	NA	NA
pleural mass	NA	NA
pulmonary mass	0.8	0.83
soft tissue mass	0.65	0.62
esophagic dilatation	NA	NA
azygoesophageal recess shift	0.7	0.71
pericardial effusion	0.71	0.81
mediastinic lipomatosis	0.77	0.83
thoracic cage deformation	0.75	0.77
scoliosis	0.75	0.78
kyphosis	0.83	0.85
pectum excavatum	0.81	0.79
pectum carinatum	0.81	0.69
cervical rib	0.63	0.8
vertebral degenerative changes	0.73	0.74
vertebral compression	0.75	0.77
vertebral anterior compression	0.76	0.77
lytic bone lesion	0.74	0.75
sclerotic bone lesion	0.59	0.62
blastic bone lesion	0.56	0.65
costochondral junction hypertrophy	0.8	0.78
sternoclavicular junction hypertrophy	NA	NA
axial hyperostosis	0.88	0.89
osteopenia	0.8	0.83
osteoporosis	0.88	0.92
non axial articular degenerative changes	0.75	0.77
subacromial space narrowing	0.78	0.73

Figure A.3: Results Part 3

fracture	0.69	0.71
clavicle fracture	0.43	0.49
humeral fracture	0.75	0.8
vertebral fracture	0.84	0.87
rib fracture	0.68	0.71
callus rib fracture	0.69	0.73
gynecomastia	0.83	0.83
hiatal hernia	0.83	0.84
Chilaiditi sign	0.65	0.71
hemidiaphragm elevation	0.76	0.76
diaphragmatic eventration	0.78	0.79
tracheostomy tube	0.96	0.96
endotracheal tube	0.97	0.97
NSG tube	0.97	0.97
chest drain tube	0.89	0.92
ventriculoperitoneal drain tube	0.46	0.7
gastrostomy tube	NA	NA
nephrostomy tube	0.33	0.92
double J stent	0.3	0.48
catheter	0.94	0.95
central venous catheter	0.94	0.95
central venous catheter via subclavian vein	0.93	0.95
central venous catheter via jugular vein	0.97	0.97
reservoir central venous catheter	0.74	0.8
central venous catheter via umbilical vein	1	1
electrical device	0.97	0.98
dual chamber device	0.97	0.99
single chamber device	0.95	0.97
pacemaker	0.98	0.99
dai	0.8	0.91
artificial heart valve	0.85	0.88
artificial mitral heart valve	0.85	0.89
artificial aortic heart valve	0.8	0.89
surgery	0.71	0.72
metal	0.74	0.75
osteosynthesis material	0.72	0.71
sternotomy	0.82	0.84
suture material	0.72	0.74
bone cement	0.65	0.56
prosthesis	0.76	0.72
humeral prosthesis	0.83	0.81
mammary prosthesis	0.82	0.82
endoprosthesis	0.91	0.89
aortic endoprosthesis	0.84	0.83
surgery breast	0.76	0.78
mastectomy	0.73	0.75
surgery neck	0.93	0.92
surgery lung	0.6	0.68
surgery heart	NA	NA
surgery humeral	0.65	0.27

Figure A.4: Results Part 4

abnormal foreign body	0.49	0.52
external foreign body	0.64	0.46
differential diagnosis	0.74	0.75
pneumonia	0.78	0.81
atypical pneumonia	0.73	0.81
viral pneumonia	0.94	0.96
COVID 19	0.93	0.96
COVID 19 uncertain	0.76	0.88
tuberculosis	0.72	0.76
tuberculosis sequelae	0.73	0.76
lung metastasis	0.64	0.71
lymphangitis carcinomatosa	0.82	0.97
lepidic adenocarcinoma	NA	NA
pulmonary fibrosis	0.82	0.86
post radiotherapy changes	0.77	0.79
asbestosis signs	NA	NA
emphysema	0.84	0.87
COPD signs	0.78	0.8
heart insufficiency	0.89	0.91
respiratory distress	0.97	0.98
pulmonary hypertension	0.91	0.93
pulmonary artery hypertension	0.38	0.53
pulmonary venous hypertension	0.9	0.89
pulmonary edema	0.92	0.92
bone metastasis	0.65	0.71

Figure A.5: Results Part 5