COVID-19 Detection using Deep Learning

MSc Research Project
Data Analytics

Shovan Roy
Student ID: X18201156

School of Computing
National College of Ireland

Supervisor: Mr. Hicham Rifai

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Shovan Roy		
Student ID:	X18201156		
Programme:	Data Analytics	**Year:**	2019-2020
Module:	MSc Research Project		
Supervisor:	Mr. Hicham Rifai		
Submission Due Date:	17/08/2020		
Project Title:	COVID-19 Detection using Deep Learning		
Word Count:	**4091**	**Page Count: 16**	

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Shovan Roy

Date: 17/08/2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

COVID-19 Detection using Deep Learning

Shovan Roy
X18201156

Abstract

Since the emergence of COVID-19, the planet has been plunged into a state of complete turmoil. The World Health Organisation (WHO) has proclaimed the virus to be a pandemic, with each nation taking measures in conjunction with the epidemic by implementing a lock-down to deter the spread of COVID-19 and offering immediate medical services to persons with symptoms or positive results. Owing to the size of the population affected and undergoing monitoring, the small amount of facilities and services accessible to hospitals is not adequate. The world is still lagging behind in terms of the time needed for the arrival of the test results, which not only take hours, but often days. This may be quite dangerous, because the suspected patients will have transmitted the infection to other individuals during that period, because they do not realize they are infected and would be self-isolating. Our motivation is to use deep learning Algorithms to be able to train a model which will detect COVID patients faster and more accurately than the RT-PCR being used today. CNN is used on X-ray scans for image processing and pattern detection to be able to detect COVID X-rays. The accuracy of 96% was achieved during the research which could be very helpful in these dire times.

Keywords: COVID-19, Classification, COVID-19 Detection, Convolutional Neural Network, CNN, Deep Learning, Image processing, Pattern Recognition, Chest X-rays

1 Introduction

The most chaotic transformation of the 21st century came with the origin of the COVID-19 in December of 2019 in Wuhan, China. Someone infected by the coronavirus could take 5-6 days to develop symptoms like fever, dry cough, tiredness and also up to 14 days to develop Serious symptoms like difficulty breathing or shortness of breath, chest pain or pressure, loss of speech or movement which ultimately could lead to death[1]. The Coronavirus is a highly contagious disease, it can spread through direct contactor through touching the same surface as well and respiratory droplets, and under certain circumstances airborne transmission is also possible. Since then it spread across the world affecting society and the global economy, it also affected the global environment[2], World Health Organisation (WHO)[3] has proclaimed this outbreak as a pandemic. They also announce the requirement for the implementation of containment strategies as well as a swift action in an effort to control

[1] https://www2.hse.ie/conditions/coronavirus/how-coronavirus-is-spread.html
[2] https://www.sciencedirect.com/science/article/pii/S0048969720323998
[3] https://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/news/news/2020/3/who-announces-covid-19-outbreak-a-pandemic

the Disease, causing schools, colleges and workplaces to go under lockdown and people staying at home, people had to wear masks if they go out, frequent use of Hand Sanitizers and maintaining a distance of 2 meters from people at all times.

The enormous amount of people being affected by the coronavirus has thrown the Healthcare Institutions across the world in a state of humongous lack of Resources, spanning from beds, manpower to machinery. A system called "Flattening the Curve"[4] was introduced by public health officials. As depicted by Figure 1, the two curves show the consequence that U.S.A will face: 1. The tall curve, if measures to prevent the spread of COVID-19 is not undertaken and 2. People won't be sick enough to need a hospital. The Horizontal line is how many people the hospital can treat at the time. So, if the curve is higher than the line the hospitals won't be adequate to accommodate and treat every people. Therefore, flattening the curve is very important.

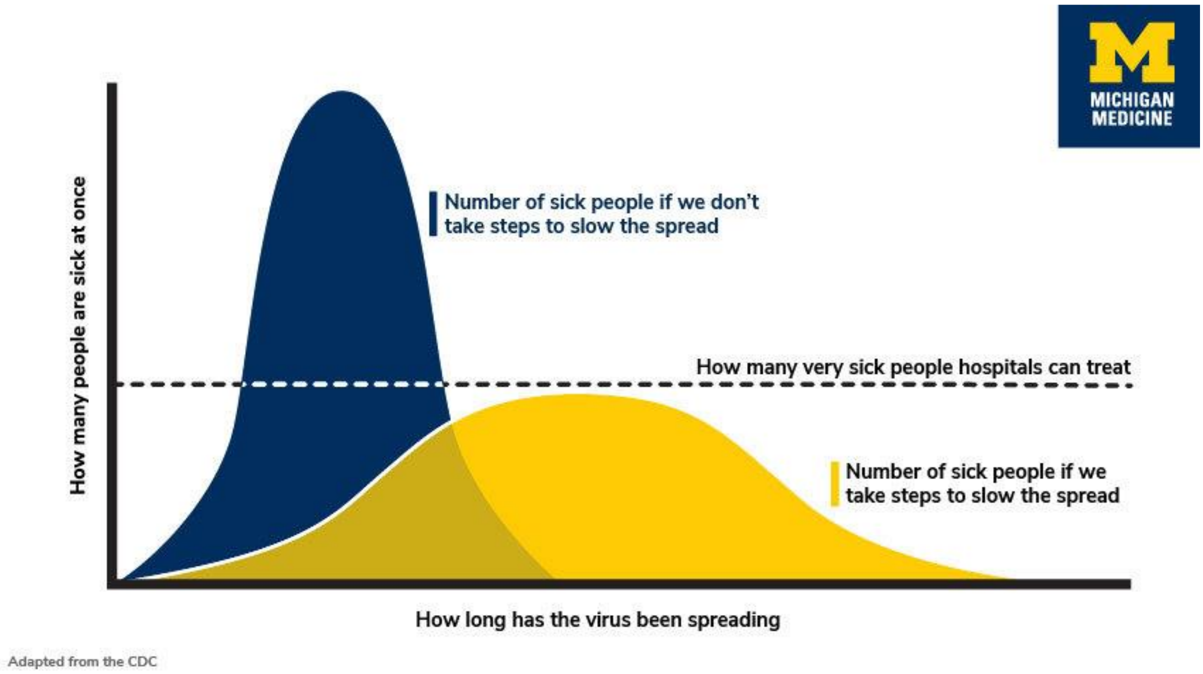

Figure 1: The Curve [4]

The Current COVID-19 Test system used by the Healthcare facilities are Genetic PCR Tests[5] where a swab sample of the patient is taken and sent to centralized labs in the country for testing just the transit time for the samples to reach the lab can take up to 24 hours depending on the distance between the hospitals and Labs, the process itself might take 5 – 6 hours, on average it may take up to 72 hours to get the results back but it is not uncommon for the turnaround time for the results to take up to a week. Currently, Irish Laboratories can only process 2000 - 2500 tests per day[6] and with a target of 15,000 tests per

---

[4] https://healthblog.uofmhealth.org/wellness-prevention/flattening-curve-for-covid-19-what-does-it-mean-and-how-can-you-help
[5] https://www.npr.org/sections/health-shots/2020/03/28/822869504/why-it-takes-so-long-to-get-most-covid-19-test-results?t=1597583698765
[6] https://www.con-telegraph.ie/2020/04/05/plans-to-conduct-up-to-4500-covid-19-lab-tests-per-day-in-ireland/

day[7], it could be very hard to cope. Although after all this progress the COVID-19 PCR Tests can give up to 30% False Negative Results[8] i.e. Even if the patient is affected by COVID-19, the results show negative. This means the RT PCR is only 80% – 85% accurate. This could become a huge issue as a patient with the false-negative result could socialize with people as he tested negative, and this could give rise to more spread of the infection.

How can deep learning help identify COVID-19 infection from the Chest X-rays with high accuracy and help improve the current COVID-19 detection systems used by the hospitals?

This paper aims at reducing these kinds of mistakes and give faster means of testing for coronavirus easily and with high accuracy which could benefit the Healthcare in this fight against the pandemic and in turn could help the world become a better place. Through the implementation of Deep Learning and pattern recognition, we make a system to Classify between a Covid-19 patient and a normal person through their Chest X-rays.

The report is divided into sections taking you through the whole workflow of the topic. The Literature Survey section deals with previous works of researchers that this project takes inspiration from or is somehow connected. The third section is the Research Methodology this section describes in detail the methods for activities necessary for the project. Then comes Design and implementation, Evaluation, Conclusion, and discussion or future work respectively. The performance accuracies are 97.97% foe

2 Literature Survey

Research on COVID-19 is a relatively new subject in the field of science. Many researchers have done a lot of work in the field. This section is a concise in-depth critical analysis about the work done in the field of COVID-19, image processing and pattern recognition.

The study (Hammoudi, et al., 2020) uses deep learning model are proposed to detect Pneumonia infection during the time of corona virus. A tailored CNN architecture and other classification model They Tested the Bacterial, viral and normal dataset. The DenseNet169 had best performance with an average accuracy of 95.72% and had accuracies 97.97% for bacteria, 96.62% for Normal and 92.57 for the virus class. The idea behind the workflow of the project where if the outcome is bacterial or viral, it will do a deep sub image analysis and if the subject still has bacterial or viral infection would be directly recommended to a health professional. Although the tailored CNN-based architecture model is extremely good at finding pneumonia with a 99.3% accuracy, their model for finding other types of virus is extremely bad even the one that has the highest accuracy at 60.64%. This paper seems to focus on too many different classes, and this could be the reason for them to get low

[7] https://www.irishtimes.com/news/ireland/irish-news/coronavirus-target-of-15-000-tests-per-day-never-hit-by-irish-labs-1.4297173

[8] https://www.bostonglobe.com/2020/04/02/nation/how-accurate-are-coronavirus-tests-doctors-raise-concern-about-false-negative-results/

accuracy. Having a primary focus could maybe increase the accuracy that they are currently able to achieve. The negative aspects from this paper could teach to narrow down your focus just like it is in the current paper where there are only 2 classes COVID-19 and Normal.

This study (Khana, et al., 2020) propose CoroNet a Deep Convolutional Neural Network based on a pre-trained architecture called the Xception architecture for the objective of aiding radiologist and clinicians to detect COVID-19 from chest X-rays. They use 4 classes for training and testing of the model which are Normal, Pneumonia Bacterial, Pneumonia Viral and COVID-19. An accuracy of 89.5%, 94.59% and 99% has been achieved in 4 classes, 3 classes and between binary classes respectively. It also compares CoroNet with COVID-Net in measuring the accuracy of the 4 classes. The COVID-Net gets an average accuracy of 83.5% and the CoroNet gets an average accuracy of 89.6%. The proposed model has been trained on very less data, this could be limitations to the research, with their future work pointing towards them training CoroNet in the future with a lot more data is a positive sign. But comparing CoroNet to CovidNet seems a bit unfair as CovidNet is an old algorithm at this point. Everyday with the increase in the COVID situation across the more and more novel and better Deep CNN are coming to light as well as huge amounts of data have started coming in from trusted sources like hospitals and radiologist, The CoroNet can now be trained even better than before.

The COVID-Net (Wang, et al., 2020) is a tailored Deep Convolutional Neural Network Design for detection of COVID-19 cases from Chest X-rays. It is trained using the COVIDx Dataset. The average accuracy of the COVID-Net is 93.3%. The study compares COVID-Net with bunch of traditionally used methods but still gets beaten by VGG-19 in sensitivity at COVID-Net being at 95 and VGG-19 being at 98. COVID-Net might need an upgrade as it got beaten by CoroNet (Khana, et al., 2020).

DeTraC is a Robust Deep Convolutional Network is being adapted in this paper (Abbas, et al., 2020) for COVID-19 detection from chest X-ray images. It is very good at dealing with irregularities in the image dataset. DeTraC uses transfer Learning and has an accuracy of 95.12%

3 Research Methodology

The research has been performed on the COVID-19 dataset which is open source and free for everyone to use accessible at: 1)The COVID-19 infected Dataset are available here - https://github.com/ieee8023/covidchestxray-dataset and the Normal Chest X-ray images are available here - https://www.qmenta.com/covid-19-kaggle-chest-x-ray-normal/.

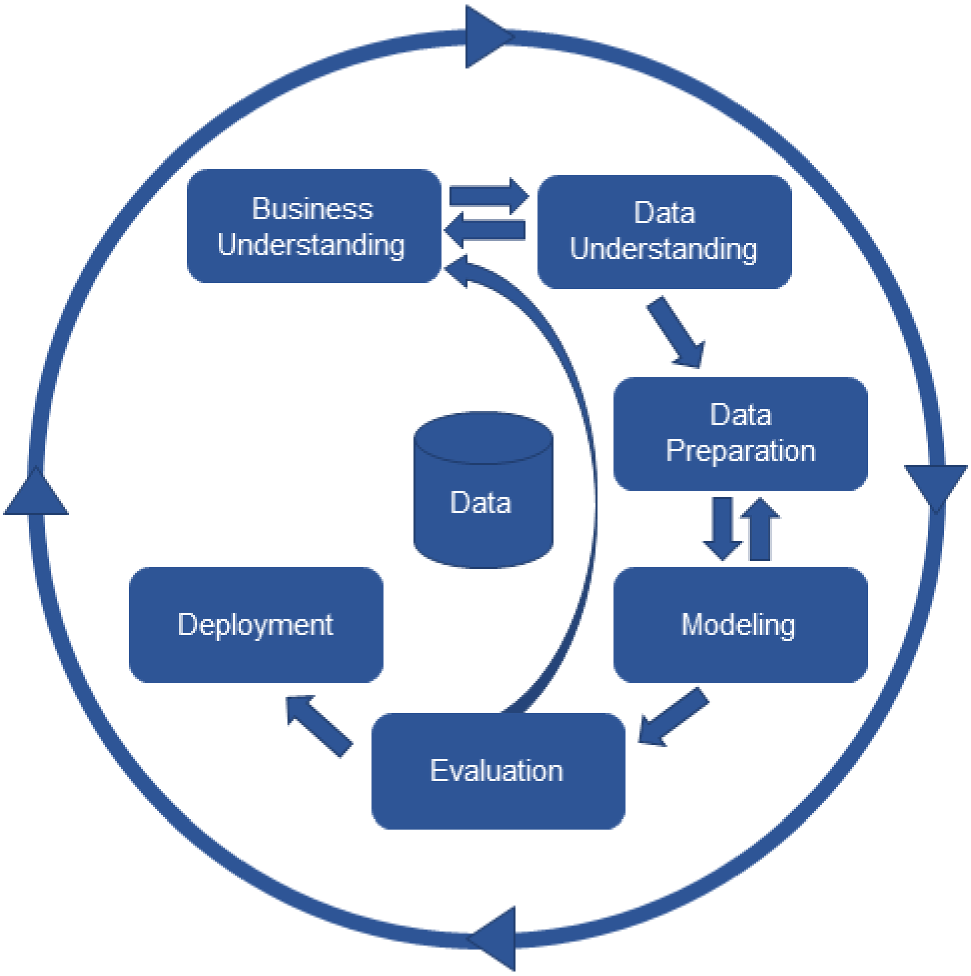

Figure 2: The Stages of Crisp-DM

The Methodology used is the CRISP-DM. The input data are images of chest X-rays. Data preparation is done in 3 stages. First the data is prepared for implementation in the model. In this stage we remove the corrupt data and gather the highest quality of images. We resize the image and process the image to remove noise and make sure that it fits our criteria. Eventually, the processed pictures are used for implementing Machine Learning.

3.1 Business Understanding

The first and the most important step of CRISP-DM approach is Business Understanding. It is intended to gain full knowledge on the subject and research from a business perspective. It requires close study and understanding of the project objectives. Research goals are established and are specifically identified at this point, providing an in-depth knowledge required for study and analysis of the business. Business Understanding consists of three parts Business Goals, Business Plan and Plan Evaluation.

3.2 Data Understanding

Another Important part of CRISP-DM is Data understanding. To be able to understand all the properties and usability of data can give us the ability to be able to analyse it with depth and precision. First, we download and study the dataset. The COVID-19 X-rays dataset has been used in Figure 3. There are actually 2 classes of data that needs to be implemented. COVID-19 X-rays (Figure 3) and the chest X-rays of normal people.

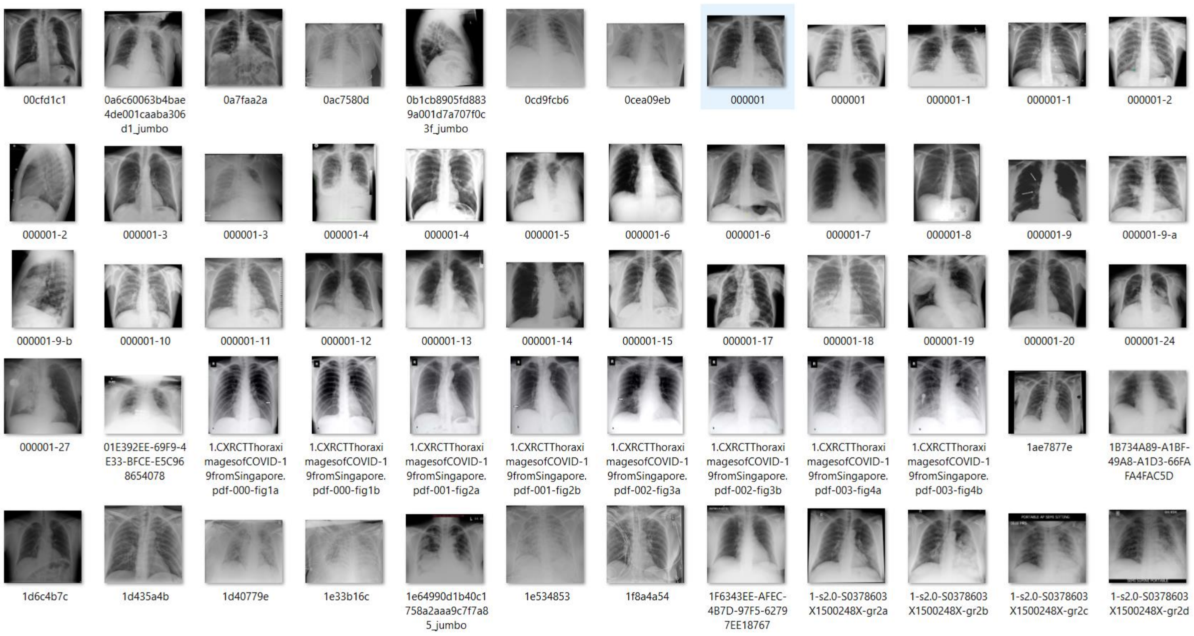

Figure 3: Data Exploration

3.3 Data Preparation & Pre-processing

The data downloaded from the COVID-19 dataset required to be Pre-processed before it can be used in the model. The Images have been Pre-processed by rescaling, changing the zoom range and flipping it horizontally. The Images also needed brightness adjustment because some of the pictures had contrast problems ad were very bright. The Pictures were then separated into Training and Testing sets, which were then again separated into COVID & Normal X-ray images.

3.4 Modelling

To build a machine learning model which can precisely detect and classify the Chest X-rays, deep Learning Image Processing Algorithms were necessary. Therefore, Convolutional Neural Network (CNN) is Implemented.

Convolutional Neural Network

A Convolutional Neural Network[9] is a type of Neural Network which has multiple layers and they are connected to each other. The first layer of the CNN takes the image as an input. The output of one layer is the input for the next layer. Each Layer of the CNN has a filter. Generally, a filter is a 3x3 matric that is put on top of the image and slides as it makes matric calculations on the image. These calculations are done then then feature extraction takes place with the help of pooling. Most commonly used pooling method is the Max pooling which extracts the highest value from each sliding calculation of the filter. This is called feature extraction. On the lower levels the CNN might be able to extract and detect features like curves and edges but at a higher level it can also detect eyes, nose, mouth etc. Each node has an activation function, we have used relu as our activation function. At the end there is an output Node which outputs the feature map of the image.

3.5 Evaluation Parameters

After the proposed model is trained using the training dataset, we use the testing dataset to test on them. It is important to check if the model was able to classify the images correctly. We choose Precision, Accuracy, Recall and F1 score. Higher value of accuracy implies higher ability of the model to better detect and classify the X-ray images. Sensitivity is also called the true positive rate and Recall implies the ability to accurately identify the positive class. F1 score gives the balance between the precision and the recall and also decides the overall accuracy of the model.

3.5.1 Classification Report

Confusion classification[10] report summarizes all prediction results of a classifier.
The summary of all the prediction is given in a way which is simple to understand. The predictions are divided into individual classes. It gives a deep understanding into the errors of the classifier. The classification report showing the result for a binary classifier is known as confusion matrix.
The Confusion Matrix is divided into 4 classes**:**

True Positive: These represent the data that are positive and have also been correctly classified as positive by the classifier.

False Positive: These represent the data that are positive and have also been incorrectly classified as negative by the classifier.

True Negative: These represent the data that are negative and have also been correctly classified as negative by the classifier.

[9] https://towardsdatascience.com/simple-introduction-to-convolutional-neural-networks-cdf8d3077bac
[10] https://medium.com/@kennymiyasato/classification-report-precision-recall-f1-score-accuracy-16a245a437a5

False Negative: These represent the data that are negative and have also been correctly classified as negative by the classifier.

3.5.2 Precision

Precision is evaluated by dividing total number correctly classified X-rays by number of classified X-rays. Higher precision indicates that the model is very good. The formula of precision is as follows:

$$Precision = \frac{Number\ of\ correctly\ identified\ Xrays}{Number\ of\ identified\ Xrays}$$

Equation 1: Precision

3.5.3 Recall

High Recall implies that the X-ray class has been correctly identified. Recall is calculated by total number of correctly identified X-ray images divided by Total number of X-ray images. The formula is as follows:

$$Recall = \frac{Number\ of\ correctly\ classified\ Xrays}{Number\ of\ classified\ Xrays}$$

Equation 2: Recall

3.5.4 F-Measure

F-Measure has a value very near Recall and Precision. This can be used to get a measurement to very nicely represent Precision and Recall. This formula of F-Measure is:

$$F\ Measure = \frac{2\ x\ Recall\ x\ Precision}{Recall + Precision}$$

Equation 3: F-Measure

4 Design Specification

This section goes into details regarding the specifications. The Figure 4 below shows the design specification of the CRISP-DM workflow. The first step is to download the Dataset from the COVID-19 dataset. The perform data exploration and pre-processing on the data and make it suitable for data processing. Then splitting the data into training set and test set.

Figure 4: Design flowchart

Then building the CNN model using TensorFlow and Keras to be able to properly process the image and perform pattern recognition. The next step is to train the model using the training set of data. After the model has been trained, the test data is passed through the model for classification. Then the result of the classification is Evaluated using Classification reports and gives us the final accuracy of the model.

5 Implementation Specification

Model implementation is the most important part of the project. Figure 5 shows the architectural design and methodology that we will be following. This depicts the whole from data extraction to results.

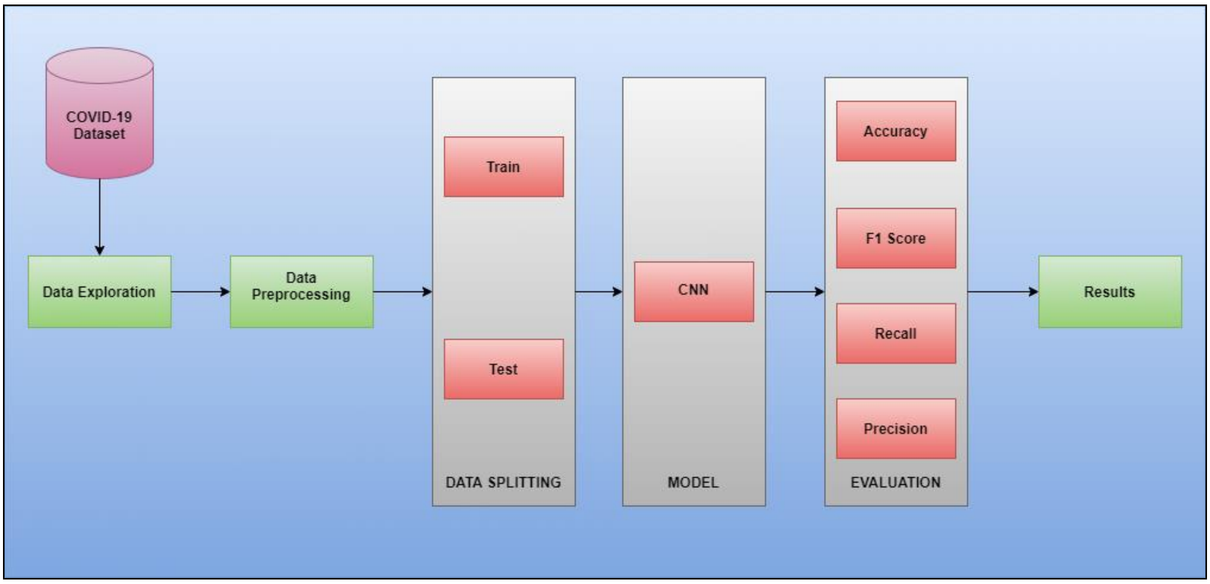

Figure 5: Implementation Details

5.1 Initial platform setup

Initial setup is extremely important before starting anything because lack of initial setup could cause an error later on which could result in starting all over. First thing required is to download and install python 3.8[11]. Then download and install the Anaconda Navigator[12]. Then download necessary packages like TensorFlow, Keras and also other necessary libraries like pandas, matplotlib, scikit-learn and numpy.

---

[11] https://www.python.org/downloads/release/python-385/
[12] https://www.anaconda.com/products/individual

5.2 Data Downloading

Before we start implementing the model, we need data. We can download the required dataset for COVID-19 X-rays from https://github.com/ieee8023/covidchestxray-dataset and the Normal Chest X-ray data from https://www.qmenta.com/covid-19-kaggle-chest-x-ray-normal/.

5.3 Data Exploration and Pre-Processing

Now the downloaded data needs to be explored and pre-processed. The details of the complete exploration and pre-processing has been given in the methodology section. After pre-processing the data, it is divided into train and test folders, both of which further contain Both COVID and normal folder, these folders contain the chest X-rays of the respective folders. The data in the training folder will be used to train the model and the data in the testing folder will be used to test the performance of the model.

5.4 Convolutional Neural Network (CNN) Implementation

Model Construction, training, testing and evaluation are the necessary steps for implementation of Convolutional Neural Network(CNN). We begin by defining the classifier = Sequential(). For defining the layers and adding them to the classifier we use; classifier.add (type_of_layer()). We also gave arguments to the ImageDataGenerator, such as:

Rescaling: Some portion of the height and width are used to rescale the image along its height and width.
Code: rescale=1. /255

Sheer range: Shear the image counterclockwise in degrees.
Code: shear_range=0.2

Zoom range: Range for random zoom.
Code: zoom_range=0.2

Horizontal flip: It is used to flip the image horizontally.
Code: horizontal_flip=True

The code of the implementation of CNN architecture consists of 3 layers. The convolution layers (Conv 2D) with activation function relu and Max Pooling 2D layers. Max 2D and relu are the pooling and nonlinear layers. In the first Conv 2D layer output filter has been defined to be 32. The height and width of the convolutional layer depends on the size of the kernel. The kernel size has been defined as 3, this mean the convolution window is a 3x3 square. The first convolution array takes the input in the form of pixels.

The activation function for the model is relu. A threshold is set by relu and any unnecessary details are eliminated from the channel. Maxpooling2D pools the spatial data. The size of the pool is defined as (2,2). We have successfully made 2 layers which are now connected. Now we make a densely connected layer with the activation function as relu and with a filter size of 64. We then use Dropout to avoid the problem of overfitting. Overfitting happens when the model recognizes samples of the training set but does not do good on the test sample set.

Now we will compile the model. The loss function has been set as binary_crossentropy. Optimizer is used for recurring neural networks. Accuracy signifies the model performance. The training samples are given in a batch size of 32.

The ImageDataGenerator transforms the training and the testing data, this causes data multiplication. The target size and the directory of the train and the test folders are specified, the images will resize to the target size when it interacts with the model. It's time to train the model. The epoch indicates the number of times the training data will repeat. The validation_steps is the batch of samples and steps_per_epoch is calculated by Total number of training samples divided by size of the batch. We train the model and then save it in my_model.h5. Running this code could take quite some time. After the model is finished training, you get validation accuracy i.e. how good the model is when testing new data. At last we check the model's accuracy on the testing data and then we plot the visualization for accuracy, loss and epoch.

6 Evaluation

In this section we will discuss and evaluate the results generated by our classification model. The result of our test set is displayed by the Confusion Matrix.

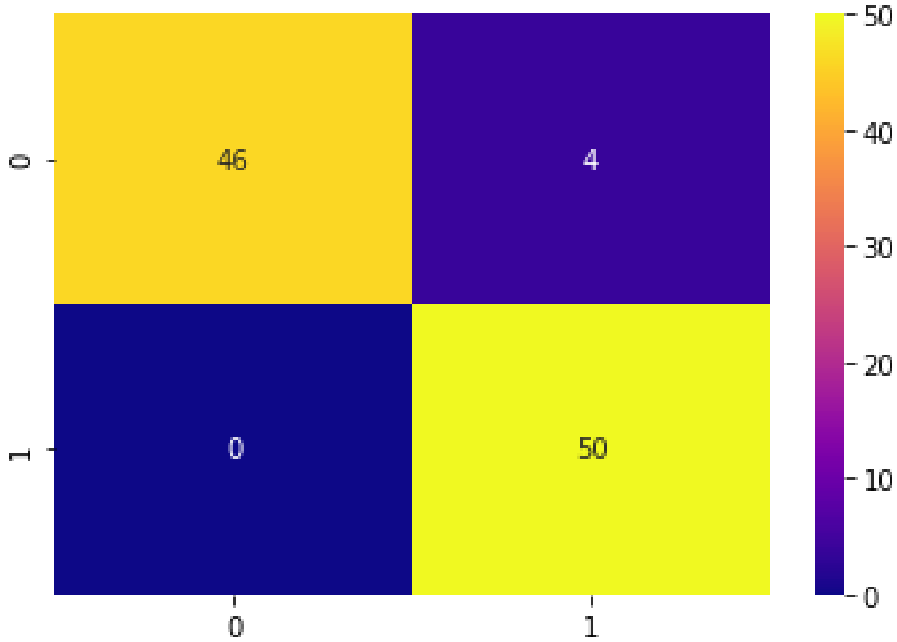

Figure 6: Confusion Matrix

As the confusion matrix reveals the result of the test set classification by the model:
There are 50 True Positive data, 0 False Negative, 46 True Negative and 4 False Negative classification of data.

With **0** being **COVID** positive and **1** being **Normal**. This means that all the Normal X-rays were accurately classified by the model whereas 46 out of 50 COVID X-rays were accurately identified by the model and 4 COVID X-rays were incorrectly identified.

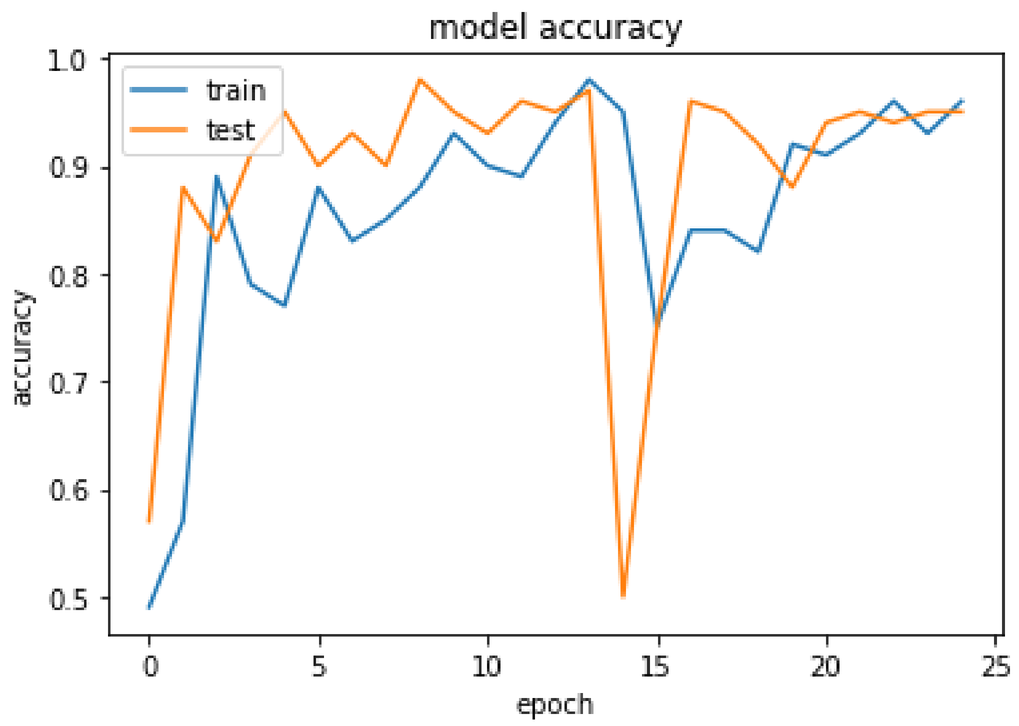

Figure 7: Accuracy Graph

The accuracy graph shows the increase of accuracy of the model with each epoch.

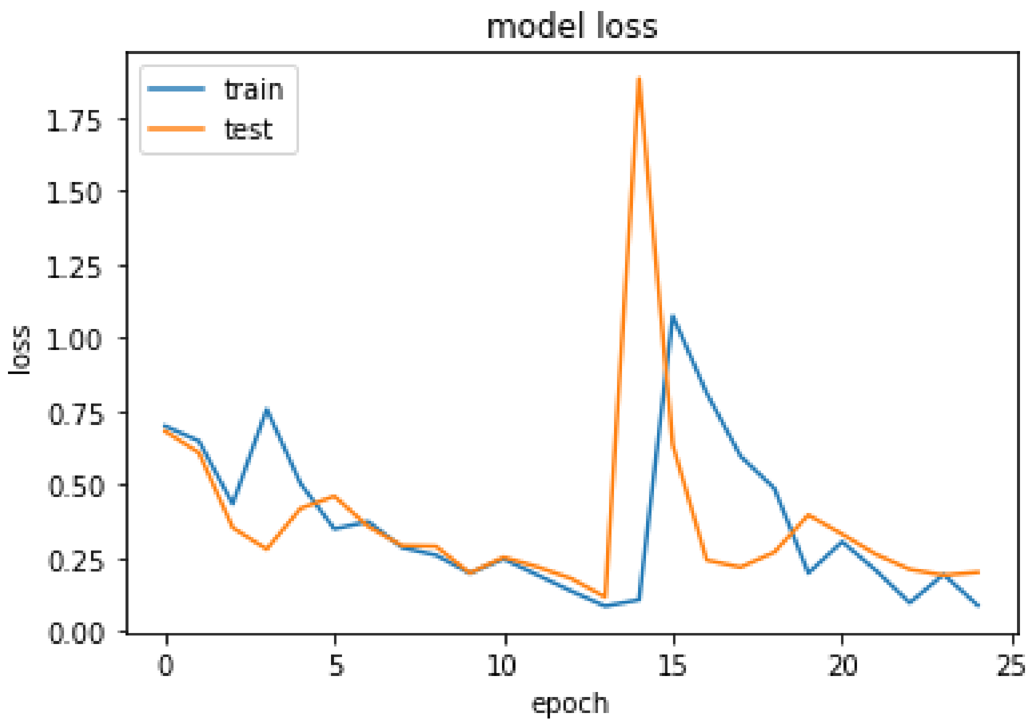

Figure 8: Loss Graph

The Loss graph shows the decrease in Loss of the model with each epoch.

Classification report provides us the detail evaluation of the classifier. The report consists of the 4 critical parameters that we will be basing our performance on: Precision, Recall, F-measure and support. The report also contains accuracy, macro average and weighted average.

	Precision	**Recall**	**F1-score**	**Support**
0	1.00	0.92	0.96	50
1	0.93	1.00	0.96	50
Accuracy	-	-	0.96	100
Macro avg	0.96	0.96	0.96	100
Weighted avg	0.96	0.96	0.96	100

Figure 9: Classification Report

The model after being implemented on the testing data resulted in an accuracy of 96%. That is higher than all the related studies reviewed in this paper. The model has worked very nicely.

7 Conclusion and Discussion

The coronavirus is a pandemic with no solution in sight. Situations due to coronavirus will never go back to normal until a vaccine has been found. But until then we have to support the health care personals putting their life on the line to fight this war. A lot of thigs can be added to this project in the future like trying several different classification models and comparing between them to see who the best is to train or better pre-processing of the data. Also, with maybe more data we can build something more astonishing. By using CNN, we measured the accuracy to be 96% the research meets the qualifications to fulfil the research question. Hopefully this project can be help assist and support the Medical Field in our fight against coronavirus.

Acknowledgement

I am exceptionally appreciative of my supervisor Mr. Hicham Rifai for his unrelenting help support and understanding. He is extremely helpful and knowledgeable. I equally want to acknowledge National College of Ireland for providing me the essential mentoring and relevant skills throughout my academic career while pursuing Masters. Also, I would like to thank my family for always being there for me.

References

Abbas, A., Abdelsamea, M. M. and Gaber, M. M. (2020) ‘Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network’, *arXiv:2003.13815 [cs, eess, stat]*. Available at: http://arxiv.org/abs/2003.13815 (Accessed: 27 September 2020).

Hammoudi, K. *et al.* (2020) 'Deep Learning on Chest X-ray Images to Detect and Evaluate Pneumonia Cases at the Era of COVID-19', *arXiv:2004.03399 [cs, eess]*. Available at: http://arxiv.org/abs/2004.03399 (Accessed: 27 September 2020).

Khan, A. I., Shah, J. L. and Bhat, M. (2020) 'CoroNet: A deep neural network for detection and diagnosis of COVID-19 from chest x-ray images', *Computer Methods and Programs in Biomedicine*, 196, p. 105581. doi: 10.1016/j.cmpb.2020.105581.

Wang, L. and Wong, A. (2020) 'COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images', *arXiv:2003.09871 [cs, eess]*. Available at: http://arxiv.org/abs/2003.09871 (Accessed: 27 September 2020).