# MACHINE LEARNING- BASED CLASSIFICATION OF COVID- 19 USING CHEST RADIOGRAPHY IMAGES.

**CHAN YU T'NG**

FACULTY OF ENGINEERING
UNIVERSITY OF MALAYA
KUALA LUMPUR

2021

Universiti Malaya

# MACHINE LEARNING- BASED CLASSIFICATION OF COVID- 19 USING CHEST RADIOGRAPHY IMAGES.

**CHAN YU T'NG**

**RESEARCH PROJECT SUBMITTED TO THE FACULTY OF ENGINEERING UNIVERSITI MALAYA, IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF BIOMEDICAL ENGINEERING**

**FACULTY OF ENGINEERING**
**UNIVERSITY OF MALAYA**
**KUALA LUMPUR**

**2021**

Universiti Malaya

**UNIVERSITY OF MALAYA**

**ORIGINAL LITERARY WORK DECLARATION**

Name of Candidate: Chan Yu T'ng

Matric No: S2020132

Name of Degree: Master of Biomedical Engineering

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

Machine Learning- based Classification of COVID- 19 using Chest Radiography Images

Field of Study: Machine Learning, Medical Imaging

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date: 31 August 2021

Subscribed and solemnly declared before,

Witness's Signature Date: 27/9/2021

Name:

Designation:

Universiti Malaya

# MACHINE LEARNING- BASED CLASSIFICATION OF COVID-19 USING CHEST RADIOGRAPHY IMAGES.

## ABSTRACT

The novel coronavirus disease, also known as COVID- 19, was first reported in Wuhan, China and has since spread around the world. Up to July 2021, it has infected over 197 million people and caused over 4 million associated death worldwide. As the number of reported cases escalates, most countries are running out of resources. The scarcity of testing kits, lengthy testing time, and the growing number of daily cases urged researchers around the world to devise alternative methods such as medical imaging to be used in conjunction with computer aided diagnosis systems, to assist radiologists and physicians in detecting COVID- 19 cases more quickly and reliably. The aim of this project is to implement a machine learning- based binary classifier that can detect COVID- 19 positive cases from COVID- 19 negative cases using chest CT images. Transfer learning technique in conjunction with VGG16 and ResNet50 architecture has been adopted in developing the binary classifier. To achieve an optimal performance of the baseline models, many performance improvement strategies such as data augmentation, re- training of weights, fine- tuning of hyperparameters, and 5- fold cross validation have been implemented and incorporated. Thorough experimentation demonstrates that the proposed classification models are computationally less expensive while yielding astoundingly good results where Model 1 based on VGG19 architecture achieved an accuracy of 95.19% and Model 2 based on ResNet50 architecture achieved an accuracy of 98.29%.

Keywords: COVID- 19 Classification, Chest CT, Transfer Learning, VGG16, ResNet50

Universiti Malaya

**KLASIFIKASI COVID- 19 BERASASKAN PEMBELAJARAN MESIN MENGGUNAKAN GAMBAR RADIOGRAFI DADA.**


## ABSTRAK

Penyakit coronavirus, juga dikenalkan sebagai COVID- 19, pertama kali dilaporkan di Wuhan, China dan sejak itu tersebar di seluruh dunia. Sehingga Julai 2021, COVID-19 telah menjangkiti lebih dari 197 juta orang dan menyebabkan lebih dari 4 juta kematian berkaitan di seluruh dunia. Dengan jumlah kes yang dilaporkan semakin meningkat, kebanyakan negara telah menghadapi isu kekurangan sumber. Kekurangan alat ujian, masa ujian yang panjang, dan semakin banyak kes harian telah mendorong para penyelidik untuk merancang kaedah alternatif seperti pengimejan perubatan digunakan bersama dengan sistem diagnosis komputer, untuk membantu ahli radiologi dan doktor dalam mengesan COVID- 19 kes lebih cepat dan boleh dipercayai. Tujuan projek ini ialah mencapai pengkelasan binari berasaskan pembelajaran mesin untuk mengesan kes positif COVID- 19 dari kes negatif COVID- 19 dengan menggunakan gambar CT dada. Teknik "transfer learning" telah digunakan bersama dengan seni bina VGG16 dan ResNet50 dalam implementasi pengkelasan binari. Untuk mencapai prestasi model dasar yang lebih optimistik, pelbagai strategi seperti "data augmentation", "re- training of weights", "fine-tuning of hyperparameters", dan "5- fold cross validation" telah dilaksanakan. Eksperimen projek ini telah menunjukkan bahawa model klasifikasi yang direka cipta adalah lebih menjimatkan masa dan proses komputasi dan dapat memberikan keputusan yang sangat memuaskan, di mana Model 1 berdasarkan seni bina VGG19 telah mencapai ketepatan 95.19% dan Model 2 berdasarkan seni bina ResNet50 telah mencapai ketepatan 98.29%.



Kata kunci: Pengkelasan COVID- 19, CT dada, Transfer Learning, VGG16, ResNet50



Universiti Malaya

ACKNOWLEDGEMENTS

I would like to thank everyone who had contributed to the successful completion of this project. I would like to express my gratitude to my research supervisor, Ir. Dr. Lai Khin Wee for his invaluable advice, guidance, and patience throughout the development of the research.

In addition, I would also like to express my gratitude to my loving parents and friends who had helped and given me encouragement.

Universiti Malaya

**TABLE OF CONTENTS**



Universiti Malaya

Universiti Malaya

Universiti Malaya

## LIST OF FIGURES



Universiti Malaya

Universiti Malaya

## LIST OF TABLES



Universiti Malaya

## LIST OF SYMBOLS AND ABBREVIATIONS

| | | |
|---|---|---|
| ADRS | : | Acute Respiratory Distress Syndrome |
| AI | : | Artificial intelligence |
| ANN | : | Artificial neural network |
| CAD | : | Computer- aided diagnosis |
| COVID- 19 | | Coronavirus Disease |
| CNN | : | Convolutional neural network |
| CT | : | Computed Tomography |
| CV | : | Cross validation |
| FPN | : | Feature pyramid network |
| FN | : | False Negative |
| FP | : | False Positive |
| FC | : | Fully connected |
| GPU | : | Graphics Processing Unit |
| GGO | : | Ground- glass opacity |
| ILSVRC | : | ImageNet Large Scale Visual Recognition Challenge |
| MV | : | Majority voting |
| MADE | : | Multi- objective adaptive differential evaluation |
| MERS | : | Middle East Respiratory Syndrome |
| MSE | : | Mean Squared Error |
| MSSP | : | Multi- scale spatial pyramid |
| MSCNN | : | Multi- scale convolutional neural network |
| RT- PCR | : | Real- time reverse transcription polymerase chain reaction |
| RNA | : | Ribonucleic acid |
| RSR | : | Russian Society of Radiology |

Universiti Malaya

| | | |
|---|---|---|
| SARS | : | Severe Acute Respiratory Syndrome |
| SGD | : | Stochastics Gradient Decent |
| TN | : | True Negative |
| TP | : | True Positive |
| WHO | : | World Health Organization |

Universiti Malaya

**LIST OF APPENDICES**



Universiti Malaya

# CHAPTER 1: INTRODUCTION

## 1.1 General Introduction

On December 31, 2019 a cluster of pneumonia cases was reported in Wuhan, China by Wuhan Municipal Health Commission. Soon, in January 2020, the World Health Organization (WHO) proclaimed the outbreak of this new infectious disease to be a global health issue and labeled this new coronavirus as COVID- 19 (WHO, 2021).

COVID- 19 can cause different effects on different individuals. The virus causes a variety of respiratory diseases, ranging from minor pneumonia to the more serious and deadly Acute Respiratory Distress Syndrome (ADRS) (Yang et al., 2020). The majority of people suffered common symptoms such as fever, cough, tiredness etc. Some people may experience additional symptoms such as headache, sore throat, and loss of smell and taste. In severe cases, patients have reported chest pain, shortness of breath, and inability to move or talk (Rahimzadeh, Attar & Sakhaei, 2021).

The most significant characteristic of the ongoing pandemic disease is its rapid and wide transmission capability. Up to July 2021, COVID- 19 had infected more than 197 million people and associated more than four million deaths across the globe (WHO, 2021). It is due to the fact that COVID- 19 virus can be transmitted through direct or indirect contact with infected person. The virus is most commonly transmitted directly from infected people to others; it can also be spread indirectly through the air or surfaces within the environment that infected person had contacted with.

Currently, the most common screening method used for identifying the disease is known as real-time reverse transcription polymerase chain reaction (RT- PCR). It is considered the "gold standard" in COVID- 19 infection detection, it detects ribonucleic acid (RNA) or genetic material that is specific to the COVID- 19 virus. Due to its time-consuming procedure, the PCR diagnosing method was unable to cope with the growing



Universiti Malaya

demand. Collected respiratory samples are required to be processed in specialized machine and it would normally take two to three hours before the results can be retrieved. Besides, the shortage of the PCR test kits in most infected regions worldwide is prompting researchers to devise innovative and simpler methods of COVID- 19 diagnosis.

Before the pandemic, medical imaging is one of the most popular techniques utilized in diagnosis of pneumonia due to the accessibility and availability of imaging device in most medical facilities. In most of the COVID- 19 infected patients, the new coronavirus causes infections in their lungs, which aroused researchers to use chest medical imaging as a tool for diagnosing COVID- 19. With the approval and ability to use chest medical images to diagnose COVID- 19, several ways have been suggested to use these images.

Studies have shown that, in clinical practice, the easily accessible chest X- rays and computed tomography (CT) are competent in facilitating triage, diagnosis and severity assessment of the COVID- 19 patients (Abelaira et al., 2021). Both X- rays and CT can reveal abnormalities in lung structures manifested by COVID- 19 patient. Despite being the most widespread and common medical imaging modality, chest X- rays suffers in term of sensitive compared to that of CT. There are chances of early or mild diseased patients appear to be normal on chest X- ray images.  CT has been proven to be more sensitive in terms of detecting abnormalities even before the infection become detectable by the PCR test. CT also has a higher resolution, which allows it to detect CT characteristics that appear at a fine scale, such as ground- glass opacity (GGO), which acts as a hazy opacity on top of pulmonary vessels (Saygılı, 2021).

Making an accurate diagnosis from those chest medical images, on the other hand, necessitates expert knowledge and extensive experience. Therefore, incorporating computer- aided diagnosis (CAD) approaches into radiologist diagnosis frameworks will greatly minimize doctors' workload while further improving the analysis' quantity and



Universiti Malaya

reliability. For instance, the Google's notable study of diagnostic classification of diabetic retinopathy with artificial intelligence (AI) and machine vision has demonstrated prominent performance beyond specialist's ability in enhancing diagnostic accuracy and decrease the burden on healthcare system (Gulshan et al., 2016).

Therefore, the proposed research stresses to develop a machine learning algorithm to detect the presence of COVID- 19 pneumonia using medical images and the specific objectives of this project include:

(1) To study the role of medical imaging and machine learning in the ongoing COVID-19 pandemic.

(2) To implement a binary machine learning- based classifier that can classify COVID-19 positive cases from COVID- 19 negative cases using chest CT images.

(3) To compare the classification performance of the implemented models with previous literature.

## 1.2 Report Outline

The rest of this report is organized as follows: The role of chest CT and machine learning in the ongoing pandemic is explored and presented in *Chapter 2*. The dataset description, methodology for classifying COVID- 19 chest CT images, as well as the experimental setup and performance metrics are given in detail in *Chapter 3*. Result analysis and discussion are provided in *Chapter 4* and *Chapter 5* respectively, and finally, the conclusion and future work recommendations are summarized in *Chapter 6*.

Universiti Malaya

# CHAPTER 2: LITERATURE REVIEW

AI implies the modeling of human intelligence in machines in order to solve complex problems with the ability to learn from experience (or given data) and improve performance over time. In recent years, the rise of AI has inspired researchers around the world to apply AI techniques in a variety of fields, especially medical detection. It has benefited the healthcare sector by allowing for more accurate diagnosis and also reducing diagnosis time. In this ongoing COVID- 19 pandemic, AI combining with medical imaging provides an alternative approach to COVID- 19 detection. Machine learning and deep learning models have been used extensively by many researchers for detection and classification of COVID- 19.

This part of the report will first describe the events that take place in the chest of the patient after COVID- 19 infection. The review will then continue by exploring the role of machine learning in medical detection as well as compare and discuss the previous machine learning or deep learning model proposed by other researchers in classifying COVID- 19.

## 2.1 CT Findings of COVID- 19

COVID- 19, similar to other pneumonias, causes increase in lung density. This can be seen on the film as increased whiteness in the lungs, proportional to the severity of the pneumonia. However, in COVID- 19, initial CT observations include bilateral, multilobed GGO with a peripheral or posterior distribution, mostly in the lower lobes and less often in the middle lobes (Yang et al., 2020). GGO is caused by increased in whiteness, but not enough to completely obscure lung markings, thus giving a ground glass appearance. Thickened interlobular and intralobular lines or so- called crazy paving are sometimes seen in conjunction with GGO. As the disease progresses, these marks will become invisible or "white- out", this phenomenon is known as consolidation.



Universiti Malaya

Consolidation is found in a smaller percentage of cases, mostly in the older adults. *Table 2.1* and *Table 2.2* summarises the common patterns of initial CT images and the CT changes over time with COVID- 19.

**Table 2.1: Common features and distribution of 919 COVID- 19 patients on initial chest CT images.**

| CT Features | Distribution |
|---|---|
| GGO | 88% |
| Bilateral Involvement | 88% |
| Posterior distribution | 80% |
| Mutilobar involvement | 79% |
| Peripheral distribution | 76% |
| Consolidation | 32% |

*Note*. From "Clinical characteristics and imaging manifestations of the 2019 novel coronavirus disease (COVID-19): a multi-center study in Wenzhou city, Zhejiang, China," by W. Yang, Q. Cao, L. E. Qin, X. Wang, Z. Cheng, A. Pan, ... & F. Yan, 2020, *Journal of Infection*, *80*(4), 388-393.

**Table 2.2: CT changes over time.**

| Stages | CT Findings |
|---|---|
| Early stage (0- 4 days) | GGO, partial crazy paving, lower number of lobes involved |
| Progressive stage (5- 8 days) | Extension of GGO, increased crazy paving pattern |
| Peak stage (10- 13 days) | Consolidation |
| Absorption stage (≥ 14 days) | Gradual resolution |

*Note*. From "Clinical characteristics and imaging manifestations of the 2019 novel coronavirus disease (COVID-19): a multi-center study in Wenzhou city, Zhejiang, China," by W. Yang, Q. Cao, L. E. Qin, X. Wang, Z. Cheng, A. Pan, ... & F. Yan, 2020, *Journal of Infection*, *80*(4), 388-393.

Take one example, *Figure 2.1* below shows an axial and coronal planes chest CT images of a 29- year- old man. The patient was tested positive with RT-PCR on February 5, 2020. At the onset (*Column A*), a normal chest CT was obtained in both axial and coronal planes. On January 26, 2020 (*Column B*), the chest CT of the patient shows minimal GGO in the lower areas of both lungs (yellow arrows). Two days later (*Column C*), the chest CT shows increased GGOs (yellow arrowheads). As the disease progresses

Universiti Malaya

to severe COVID- 19 pneumonia, on February 3, 2020 (*Column D*), the chest CT shows mixed GGOs and linear opacities in the subpleural region of the lungs. Six days later (*Column E*), the chest CT shows the absorption of GGOs and organizing pneumonia.

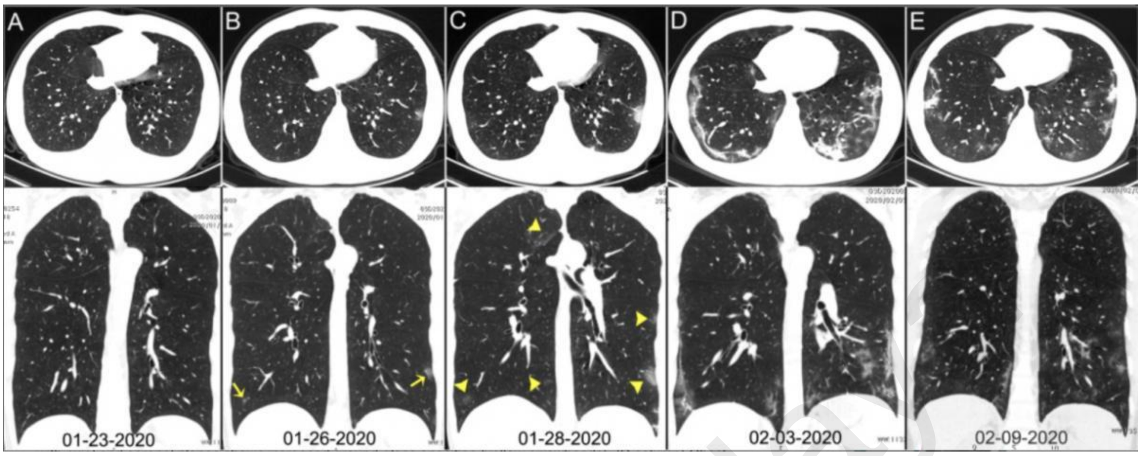


**Figure 2.1: Chest CT images of a 29-year-old man infected with COVID- 19.**

*Note*. From "Correlation of chest CT and RT-PCR testing for coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases," by T. Ai, Z. Yang, H. Hou, C. Zhan, C. Chen, W. Lv, ... & L. Xia, 2020, *Radiology*, *296*(2), E32-E40.

Bronchiectasis, septal thickening, pleural thickening, and subpleural involvement are amongst the less typical findings, which often appear only in the later stages of the disease. Some other rare but probable symptoms as the development of the disease are cavitation, CT halo sign, lymphadenopathy, pleural effusion, pericardial effusion, and pneumothorax (Yang et al., 2020).

### 2.1.1 Overlapping of Characteristics with other Atypical Pneumonia

COVID- 19 is classified as one of the atypical pneumonia due to the radiographical presence of GGO, linear opacities, and consolidation. Some other atypical pneumonias such as SARS and MERS also show similar characteristics. Although the imaging characteristics are similar to those of MERS and SARS, COVID- 19 is more likely to show activity in both lungs (bilateral) on initial chest imaging; initial imaging anomalies in SARS and MERS are more often unilateral (Hosseiny et al., 2020). *Table 2.3* below

Universiti Malaya

summarize the difference between radiographical features of SARS, MERS and COVID-19 pneumonia.

**Table 2.3: Comparison between radiographical features of SARS, MERS and COVID- 19 pneumonia.**

| Feature | SARS | MERS | COVID-19 |
|---|---|---|---|
| Imaging finding | | | |
| Acute phase | | | |
| Initial imaging | | | |
| Normal | 15–20% of patients | 17% of patients | 15–20% of patients |
| Abnormalities | | | |
| Common | Peripheral multifocal airspace opacities (GGO, consolidation, or both) on chest radiography and CT | Peripheral multifocal airspace opacities (GGO, consolidation, or both) on chest radiography and CT | Peripheral multifocal airspace opacities (GGO, consolidation, or both) on chest radiography and CT |
| Rare | Pneumothorax | Pneumothorax | Pneumothorax |
| Not seen | Cavitation or lymphadenopathy | Cavitation or lymphadenopathy | Cavitation or lymphadenopathy |
| Appearance | Unilateral, focal (50%); multifocal (40%); diffuse (10%) | Bilateral, multifocal basal airspace on chest radiography or CT (80%); isolated unilateral (20%) | Bilateral, multifocal, basal airspace; normal chest radiography findings (15%) |
| Follow-up imaging appearance | Unilateral, focal (25%); progressive (most common, can be unilateral and multifocal or bilateral with multifocal consolidation) | Extension into upper lobes or perihilar areas, pleural effusion (33%), interlobular septal thickening (26%) | Persistent or progressive airspace opacities |
| Indications of poor prognosis | Bilateral (like ARDS), four or more lung zones, progressive involvement after 12 d | Greater involvement of the lungs, pleural effusion, pneumothorax | Consolidation (vs GGO) |
| Chronic phase | | | Unknown, but pleural effusion and interlobar septal thickening have not yet been reported |
| Transient reticular opacities[a] | Yes | Yes | |
| Airtrapping | Common (usually persistent) | | |
| Fibrosis | Rare | One-third of patients | Not yet reported |

*Note*. From "Radiology perspective of coronavirus disease 2019 (COVID-19): lessons from severe acute respiratory syndrome and Middle East respiratory syndrome.," by M. Hosseiny, S. Kooraki, A. Gholamrezanezhad, S. Reddy, & L. Myers, 2020, *American Journal of Roentgenology, 214*(5), 1078-1082.

## 2.2 Machine Learning in Medical Detection

Machine learning is a branch of AI that emphasize on using data and algorithms to mimic the way humans learn, with the aim of continuously improving accuracy (IBM, 2021). Machine learning algorithms aid classification in the processing of high-dimensional data by learning the complex connections and patterns from input data and automatically predicting or classifying data using the learnt model. A standard machine learning- based classification algorithm involve feature extraction, feature training, and classification.

Universiti Malaya

Deep learning is a subfield of machine learning which hierarchical architecture made up of numerous processing layers which tends to learn high level abstractions of data feeding into the deep learning algorithm (Lecun, Bengio & Hinton, 2015). Unlike machine learning algorithms, deep learning algorithms allow raw data to be used as input and automatically recognize highly discriminative features from it. Furthermore, the performance of deep learning models is not heavily reliant on data pre- processing. Therefore, when deep learning techniques are used, pre- processing steps become less important.

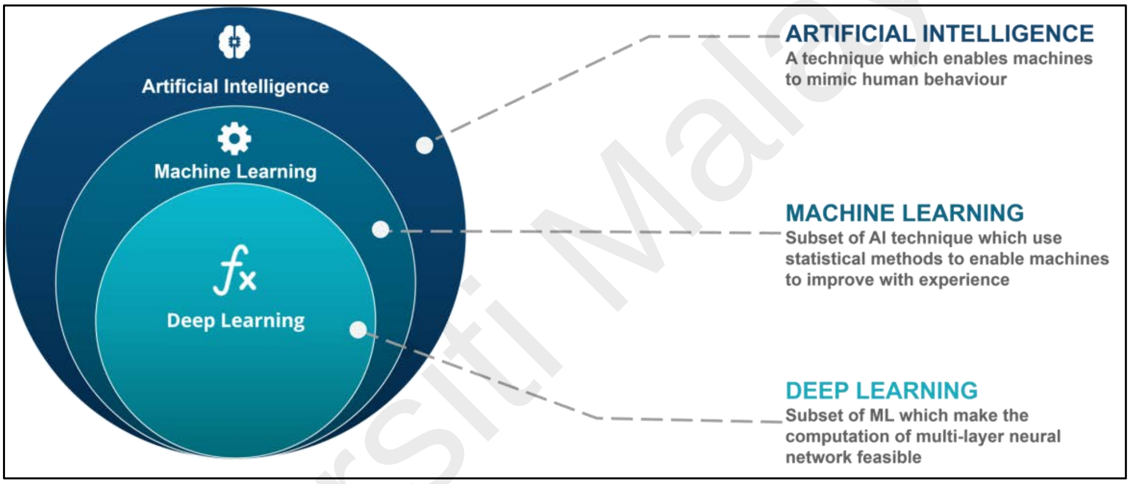


**Figure 2.2: Artificial Intelligence, Machine learning and Deep Learning.**

Over the years, deep learning is gaining popularity worldwide as a means to improve performance and solve problems in a variety of disciplines such as classification, segmentation and detection. Convolutional neural network (CNN) being the most well-known deep learning algorithm, has been proven outstanding results organ segmentation and disease detection (Yamashita et al., 2018).

### 2.2.1 Convolution Neural Network (CNN)

CNN, a form of artificial neural network (ANN) prominent in computer vision, are gaining popularity in a variety of disciplines, including medical imaging. The effectiveness of CNN is mainly due to its architecture to learn spatial hierarchies of



Universiti Malaya

features actively and automatically by backpropagation. A general CNN consists of three major building blocks which are convolution layers, pooling layers and fully connected (FC) layers. The pipeline of CNN architecture is shown in *Figure 2.3*.

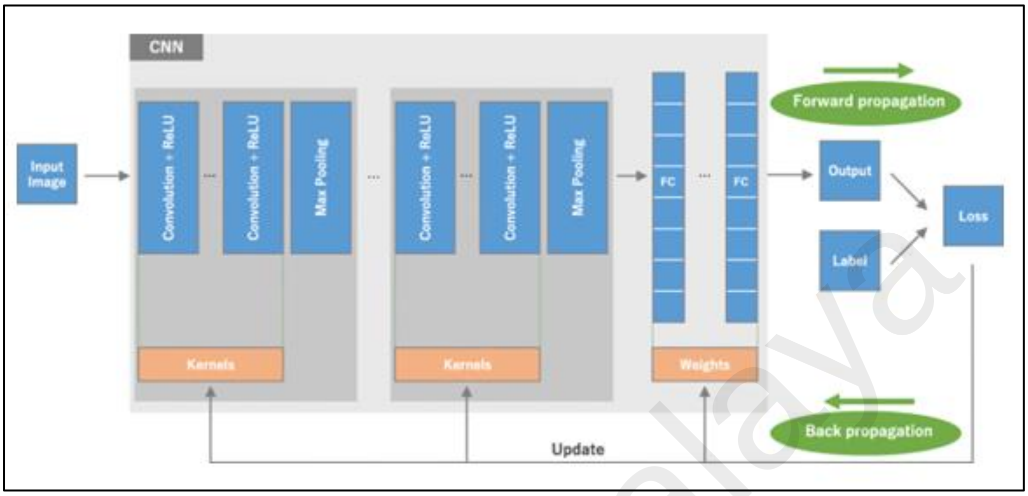


**Figure 2.3: General CNN architecture pipeline.**

In general, two main stages for training CNN which are feed forward propagation and backpropagation. The main objective of feed forward stage is to multiply the input value with the parameters such as kernels and weights in each layer. Then passing the summation of product to a non- linear activation function (ReLU) (Lecun, Bengio and Hinton, 2015). The loss is then calculated with the predicted output with ground truth labels. Next, the backward propagation stage uses derivatives or chain rules to compute gradient. The gradient computed will be used to modify the initial parameters for the next feedforward propagation. The computed gradient is used to minimalize the loss function, hence, creating a linear relationship where a decrease in loss function will increase the model performance of the CNN model (Murugan, 2017). The CNN learning will stop generally after sufficient repetition of feedforward and backpropagation stage.

Universiti Malaya

#### 2.2.1.1 Convolution Layer

The convolutional layer is where the inputs of image undergo kernels or convolution matrix which is a form of filter to convolve the input image to create various feature maps as output as shown in *Figure 2.4* (Szegedy et al., 2014). A kernel is usually a 3×3 or 5×5 shaped matrix that move around the input matrix to capture high- and low-level features in the pattern (LeChun, Bengio & Hinton, 2015). Stride parameter is tuned to adjust the number of steps shifting across the input matrix.

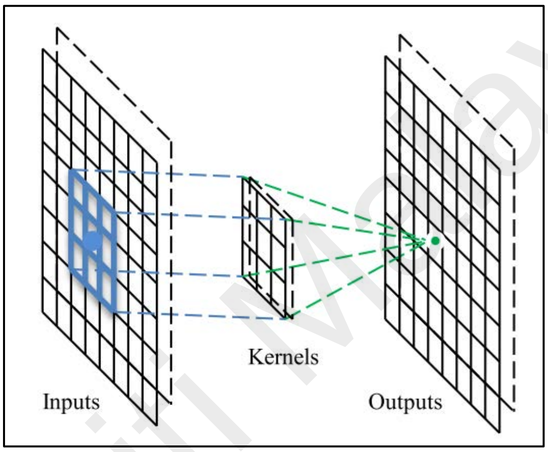


**Figure 2.4: Illustration of convolutional layer.**

*Equation 2.1* below describes the output of convolutional layer (Narin, Kaya & Pamuk, 2020).

$$x_j^l = f(\sum_{a=1}^{N} w_j^{l-1} * y_a^{l-1} + b_j^l) \qquad (2.1)$$

where

$x_j^l$ = $j$-th feature map in layer 1

$w_j^{l-1}$ = $j$-th kernels in layer 1-1

$y_a^{l-1}$ =$a$-th feature map in layer 1-1

$b_j^l$ = bias of the $j$-th feature map in layer 1

$N$ = number of total features in layer 1-1

$(*)$ = vector convolution process

Universiti Malaya

#### 2.2.1.2 Pooling Layer

Pooling layers as shown in *Figure 2.5* generally aim to reduce the size of an image or the parameters if the image size is oversized. Pooling layer is also used to control overfitting as it progressively reduces spatial size of parameters in the network.

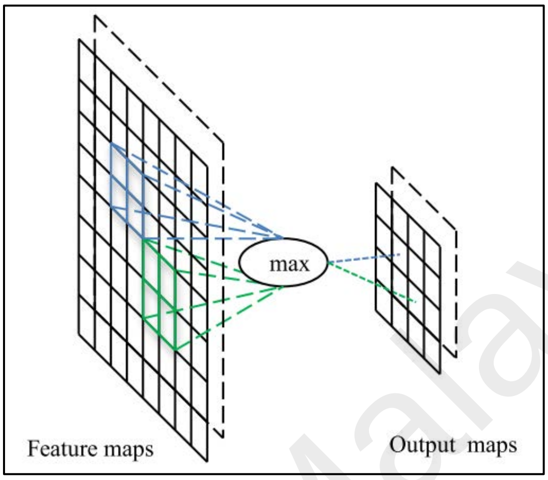


**Figure 2.5: Illustration of pooling layer.**

The most general pooling approach that are being used, are the max pooling and average pooling. According to Boureau and team, the paper has proved a detailed analysis of the performance of max pooling and average pooling (Boureau, Ponce & Lecun, 2010). Scherer and team also performed a study to compare the two pooling approaches and discovered that max pooling excels at generalization of parameters, ability to select advance invariant features and rapid merging of data (Scherer, M¨uller & Behnke, 2010).

#### 2.2.1.3 Fully Connected Layer

The operation of FC layer comes right after the pooling layer. In FC layer as shown in *Figure 2.6*, representation of data was accomplished by converting the initially 2D parameters into 1D parameter. Most of the parameters contained in CNN are represented in FC layer, hence, creating disadvantage of having this layer due to heavy workload when training these CNN models (Guo et al., 2016). However, FC layer allows forward



Universiti Malaya

propagation of parameters with a predefined size into various categories for classifying images (Krizhevsky, Sutskever and Hinton, 2012).

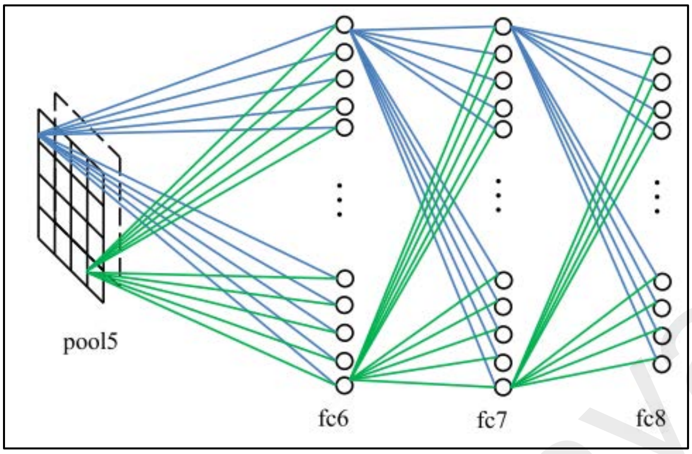


**Figure 2.6: Illustration of FC layer.**

At the last layer of the FC layer, a softmax function is often utilized to normalize the network's output to a probability distribution and then predict the output classes. The *Equation 2.2* below describes the mathematical computation of the standard softmax function (Narin, Kaya & Pamuk, 2020).

$$Softmax(x_i) = \frac{e^{x_i}}{\sum_{y-1}^{m} e^{x_y}} \qquad (2.2)$$

where

$x_i$ =input data

$m$ = number of classes

### 2.2.2 Transfer Learning

Despite being the most well- developed approach for deep learning, the major drawback of CNN is that training CNN models are not only time- consuming but also require high performance equipment (Yu et al., 2020). Therefore, transfer learning was introduced. Transfer learning is the reapplication of a previously learned model to a new problem. In transfer learning, a system uses prior task expertise to enhance generalization

Universiti Malaya

about the new model to be trained. The main benefit of employing the transfer learning approach is that it enables training with less data and reduces calculation costs.

There are two approaches to use transfer learning: (1) Develop model approach and (2) Pre- trained model approach (Brownlee, 2017). The pre- train model approach is an easier approach because it eliminates the need to develop a model from scratch to address a similar problem. In the study of Pham (2021) to investigate the performance of newly developed models and fine- tuned pre- trained model, the results indicate that fine-tuning the network learning parameters can help avoid wasting time and effort constructing more sophisticated new models when current ones can achieve the same or better performance.

### 2.2.3 Pre- trained Model

There is probably more than a dozen of high- performance image recognition models available to be used as the baseline for image classification and other computer vision applications. Some of the most popular models that have accomplished outstanding result in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) are summarize in *Table 2.4* below (Das, 2017). All mentioned models are broadly applied in transfer learning, not just for their outstanding performance, but also for their state- of- art architecture, namely the inception modules (GoogLeNet), consistent and repeating structures (VGG) and residual modules (ResNet).

**Table 2.4: CNN architectures.**

| Year developed | CNN | Developed by | Place | Top-5 error rate | No. of parameters |
|---|---|---|---|---|---|
| 2012 | AlexNet (7) | Alex et al. | 1st | 15.3% | 60 million |
| 2014 | GoogLeNet (19) | Google | 1st | 6.67% | 4 million |
| 2014 | VGG (16) | Simonyan, Zisserman | 2nd | 7.3% | 138 million |
| 2015 | ResNet (152) | Kaiming He et al. | 1st | 3.6% | - |



Universiti Malaya

#### 2.2.3.1 VGG16

VGG, also known as OxfordNet, is a CNN model presented by Karen Simonyan and Andrew Zisserman of the Visual Geometry Group (VGG) (Simonyan & Zisserman, 2014). The VGG architecture outperforms other models with its consistent and repeating structures. Instead of having a massive number of hyper- parameters, VGG was designed with numerous small kernel- sized convolution filters (3×3 with stride 1, the smallest size feasible while still capturing up/down and left/right) and always utilized the same padding and maxpool layer of 2×2 filter with stride 2. This configuration of convolution and maxpool layers is consistent throughout the whole architecture (Thakur, 2019).

VGG is available in two varieties: VGG16 and VGG19, where 16 and 19 refers to the number of layers. VGG16 achieved a 92.7% top- 5 accuracy in ImageNet dataset containing over 14 million pictures classified into 1000 categories. *Figure 2.7* below illustrates the architecture of VGG16 model. As shown in *Figure 2.7*, the VGG16 model contains 5 convolution blocks (conv block-1 to conv block-5) which made up of the following element, therefore giving 16 layers:

- **conv block-1** with 64 different 3×3 kernels with a stride size of 1 repeated 2 times giving **2 layers**, along with a max pooling with stride size of 2.
- **conv block-2** with 128 different 3×3 kernels with a stride size of 1 repeated 2 times, giving **2 layers**, along with a max pooling with stride size of 2.
- **conv block-3** with 256 different 3×3 kernels with a stride size of 1 repeated 3 times, giving **3 layers**, along with a max pooling with stride size of 2.
- **conv block-4** with 512 different 3×3 kernels with a stride size of 1 repeated 3 times, giving **3 layers**, along with a max pooling with stride size of 2.
- **conv block-5** with 512 different 3×3 kernels with a stride size of 1 repeated 3 times, giving **3 layers**, along with a max pooling with stride size of 2.

Universiti Malaya

- Two FC layer with 4096 nodes giving **2 layers**.
- Lastly, FC layer containing 1000 nodes and softmax function in the end giving **1 layer**.

| ConvNet Configuration | | | | | |
|---|---|---|---|---|---|
| A | A-LRN | B | C | D | E |
| 11 weight layers | 11 weight layers | 13 weight layers | 16 weight layers | 16 weight layers | 19 weight layers |
| input (224 × 224 RGB image) | | | | | |
| conv3-64 | conv3-64<br>**LRN** | conv3-64<br>**conv3-64** | conv3-64<br>conv3-64 | conv3-64<br>conv3-64 | conv3-64<br>conv3-64 |
| maxpool | | | | | |
| conv3-128 | conv3-128 | conv3-128<br>**conv3-128** | conv3-128<br>conv3-128 | conv3-128<br>conv3-128 | conv3-128<br>conv3-128 |
| maxpool | | | | | |
| conv3-256<br>conv3-256 | conv3-256<br>conv3-256 | conv3-256<br>conv3-256 | conv3-256<br>conv3-256<br>**conv1-256** | conv3-256<br>conv3-256<br>**conv3-256** | conv3-256<br>conv3-256<br>conv3-256<br>**conv3-256** |
| maxpool | | | | | |
| conv3-512<br>conv3-512 | conv3-512<br>conv3-512 | conv3-512<br>conv3-512 | conv3-512<br>conv3-512<br>**conv1-512** | conv3-512<br>conv3-512<br>**conv3-512** | conv3-512<br>conv3-512<br>conv3-512<br>**conv3-512** |
| maxpool | | | | | |
| conv3-512<br>conv3-512 | conv3-512<br>conv3-512 | conv3-512<br>conv3-512 | conv3-512<br>conv3-512<br>**conv1-512** | conv3-512<br>conv3-512<br>**conv3-512** | conv3-512<br>conv3-512<br>conv3-512<br>**conv3-512** |
| maxpool | | | | | |
| FC-4096 | | | | | |
| FC-4096 | | | | | |
| FC-1000 | | | | | |
| soft-max | | | | | |

**Figure 2.7: Architecture of VGG16.**

*Note*. From "Very deep convolutional networks for large-scale image recognition," by Simonyan, K., & Zisserman, A., 2014, *arXiv preprint arXiv:1409.1556*.

#### 2.2.3.2 ResNet50

ResNet, also known as Residual Neural Network, was designed by Kaiming He and team (He et al., 2016). ResNet was revolutionary when it was initially released because it presented a new solution to a huge challenge for deep neural networks at the time: the vanishing gradient problem (Shi, 2019). Although neural networks are universal function approximators, adding additional layers can cause the training process to become slower as well as the accuracy to saturate at a certain point and then degrades rapidly leading to higher training error as shown in *Figure 2.8*.

Universiti Malaya

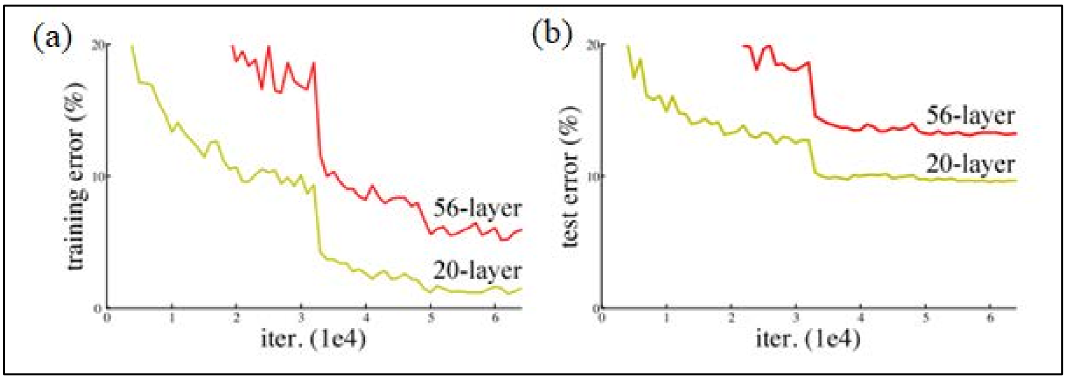


**Figure 2.8: (a) Training error (b) and test error on CIFAR-10 with 20- and 56-layer "plain" network.**

*Note*. From "Deep residual learning for image recognition," by K. He, X. Zhang, S. Ren, & J. Sun, 2016, *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

To tackle this issue, ResNet introduces identity shortcut connections between layers, which essentially skip the training of one or more layers, resulting in a residual block as shown is *Figure 2.9*. Thanks to this, distortion that take place as the network gets deeper are prevented. *Figure 2.10* shows the comparison after applying ResNet of 18 and 34 layers (*Figure 2.10 (b)*) with no additional parameters in relation to the "plain" network (*Figure 2.10 (a)*).

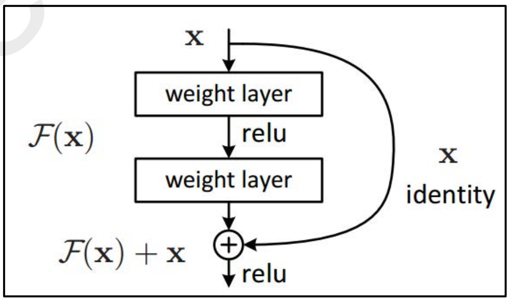


**Figure 2.9: Building block of residual learning.**



Universiti Malaya

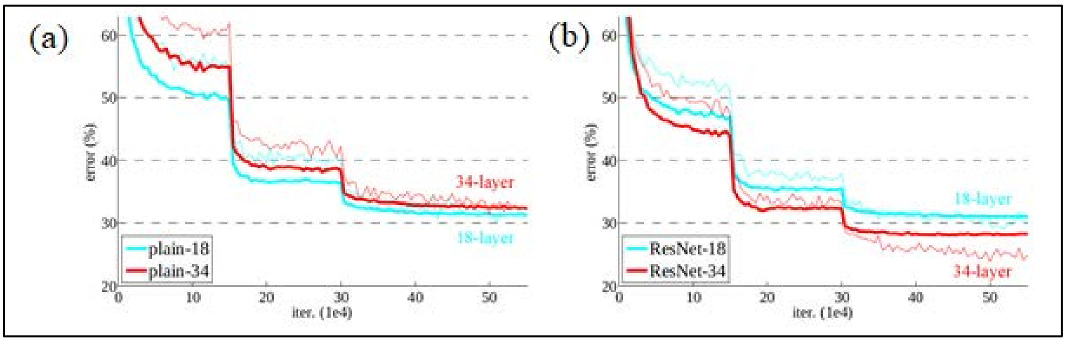


**Figure 2.10: Training (thin curve) and validation (bold curve) error on ImageNet with (a) Plain network (b) ResNet.**

*Note*. From "Deep residual learning for image recognition," by K. He, X. Zhang, S. Ren, & J. Sun, 2016, *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

ResNet50 is a 50- layer CNN trained on the ImageNet dataset. The result of ResNet50 was rather impressive on the ImageNet validation set, where it achieved a top- 5 error rate of 5.25%. *Figure 2.11* below illustrates the architecture of ResNet50 model. As shown in *Figure 2.11*, the ResNet50 model contains five convolution blocks (conv1 to conv5_x) which made up of the following element, therefore giving 50 layers:

- **conv1** with 64 different 7×7 kernels with a stride size of 2 giving **1 layer**.
- Max pooling with stride size of 2.
- **conv2_x** with 1×1, 64 kernels, followed by 3×3, 64 kernels, and lastly 1×1, 256 kernels. These three layers are repeated 3 times giving **9 layers**.
- **conv3_x** with 1×1, 128 kernels, followed by 3×3, 128 kernels, and lastly 1×1, 512 kernels. These three layers are repeated 4 times giving **12 layers**.
- **conv4_x** with 1×1, 256 kernels, followed by 3×3, 256 kernels, and lastly 1×1, 1024 kernel. These three layers are repeated 6 times giving **18 layers**.
- **conv5_x** with 1×1, 512 kernels, followed by 3×3, 512 kernels, and lastly 1×1, 2048 kernel. These three layers are repeated 4 times giving **9 layers**.
- Lastly, average pooling layer connected to the FC layer containing 1000 nodes and softmax function in the end giving **1 layer**.

Universiti Malaya

| layer name | output size | 18-layer | 34-layer | 50-layer | 101-layer | 152-layer |
|---|---|---|---|---|---|---|
| conv1 | $112\times112$ | | | $7\times7$, 64, stride 2 | | |
| conv2_x | $56\times56$ | | | $3\times3$ max pool, stride 2 | | |
| | | $\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$ | $\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times3$ | $\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$ | $\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$ | $\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$ |
| conv3_x | $28\times28$ | $\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$ | $\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times4$ | $\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$ | $\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$ | $\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times8$ |
| conv4_x | $14\times14$ | $\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$ | $\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times6$ | $\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$ | $\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times23$ | $\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times36$ |
| conv5_x | $7\times7$ | $\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$ | $\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times3$ | $\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$ | $\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$ | $\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$ |
| | $1\times1$ | | | average pool, 1000-d fc, softmax | | |
| FLOPs | | $1.8\times10^9$ | $3.6\times10^9$ | $3.8\times10^9$ | $7.6\times10^9$ | $11.3\times10^9$ |

**Figure 2.11: Architecture of ResNet50.**

*Note*. From "Deep residual learning for image recognition," by K. He, X. Zhang, S. Ren, & J. Sun, 2016, *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

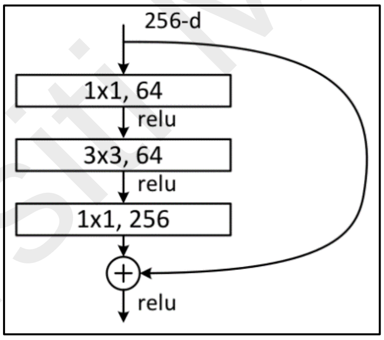


**Figure 2.12: A single residual block in ResNet50.**

## 2.3 Related Works

### 2.3.1 Binary Classification

Binary classification is referred to classification tasks that comprises two different class labels, in this case, COVID-19 infected and clinical normal. Many researchers developed models to focus on only classifying COVID-19 infected patient from clinical normal using CT images. For instance, Yu and team proposed a binary classification model using GoogLeNet as the backbone to predict COVID- 19 infection (Yu et al.,

Universiti Malaya

2020). The model substituted the last two layers of the GoogLeNet architecture with four new layers, including a new dropout layer, two FC layers and a output layer. The result showed by adding a dropout layer to the model can effectively mitigate the problem of overfitting. At slice level, the GoogLeNet- COD model achieved accuracy of 87.50%.

Serte and Demirel (2021) proposed a binary classification model using ResNet18 and ResNet50 CNN to predict COVID- 19 infection for both image- level and scan level CT images. Then, majority voting (MV) technique fuses estimations from the ResNet CNN models to perform classification. The study has shown increasing performance as the number of intermediate scans increases. At scan level, the ResNet50 CNN model together with majority voting achieved accuracy of 96.00%.

In another study, Yan and team (2020) constructed an automated system that can achieve good diagnostic performance with small training dataset. Multi- scale spatial pyramid (MSSP), multi-scale convolutional neural network (MSCNN) and data augmentation are employed together to mitigate the lack of training data while enhancing the model performance. MSSP uses Gaussian pyramid to create three different scans of the CT image, so that important multi- scale information can be captured. Each scale is being used by three different CNNs for training. The study has suggested that pre-processing normalization help in improving prediction of cross-database images. At slice level, the model achieved high per slice accuracy of 97.5% based on a small amount of data.

Rahimzadeh, Attar and Sakhaei (2021) applied a modified feature pyramid network (FPN) to be used together with transfer learning on ResNet50V2 model. FPN serve to improve feature extraction by recognizing COVID- 19 infections at multiple scales, therefore minimizing the percentage of false positives as the model learns better. FPN, on

Universiti Malaya

the other hand, has the disadvantage of increasing computing time and is inefficient. At single image classification stage, the model successfully an achieved accuracy of 98.49%.

Zhu and team (2020) developed a binary classification model based on VGG16 and ResNet50 architectures. The model performed exceptionally well in binary classification, where both the VGG16 and ResNet50 based model reached more than 99% accuracy. The addition of pneumonia images to the model for multiclass classification deteriorates the model's performance, where the accuracy drops to 86.74% and 88.52%.

### 2.3.2 Using X- ray Images for Classification

X- ray imaging is another technique to detect pneumonia and is utilize by many researchers in developing the models to classify COVID- 19 pneumonia during this pandemic period. Singh and team proposed a CNN binary classification model, where the hyperparameters of the CNN model were tuned using multi- objective adaptive differential evaluation (MADE) algorithm (Singh et al., 2020). The MADE- based CNN model achieved an accuracy of 92.55%.

On the other hand, Hemdan and team introduced a model known as COVIDX-Net (Hemdan et al., 2020). The model utilized VGG19 and Dense Convolutional Network (DenseNet) as a backbone to classify the infection status of the patient by analyzing the normalized intensities of chest X- ray images. The VGG19 based model achieve an accuracy of 90%.

Narin and team, on the other hand, compared five different CNN models to classify COVID- 19 X- ray images, namely InceptionV3, ResNet50, ResNet101, ResNet152 and Inception-ResNetV2 (Narin et al., 2020). The result of the study shows the greatest performance by ResNet50 with the accuracy of 99.70%.



Universiti Malaya

From the above discussed studies, it was observed that CNN algorithm is suitable to be used in classifying COVID- 19 pneumonia from chest images to produce reasonably satisfied performance. Additionally, the use of pre- trained models in conjunction with tuning of hyperparameter effectively improves the classification performance while reducing the model complexity. The summary of machine learning algorithm in binary classification of COVID- 19 discussed in the previous subsections is tabulated in *Table 2.5*.

**Table 2.5: Summary of deep learning algorithms used in literature for classifying COVID- 19.**

| Study | Data Type | Architecture | Result | | | |
|---|---|---|---|---|---|---|
| | | | **Accuracy** | **Precision** | **Recall** | **F-measure** |
| Yu et al., 2020 | CT | GoogLeNet-COD | 87.50% | 84.09% | - | 90.91% |
| Serte & Demirel, 2021 | CT | ResNet50+MV | 96.00% | 96.00% | - | 100% |
| Yan et al., 2020 | CT | MSSP-MSCNN | 99.50% | 97.70% | - | 96.20% |
| Rahimzadeh, Attar & Sakhaei, 2021 | CT | FPN- ResNet50V2 | 98.49% | - | - | 94.96% |
| Mishra, Singh & Joshi, 2021 | CT | VGG16<br>ResNet50 | 99.12%<br>99.62% | - | - | - |
| Singh et al., 2020 | X-ray | MADE-CNN | 95.80% | - | 96.16% | 95.60% |
| Hemdan et al., 2020 | X-ray | VGG19<br>DenseNet201 | 90.00%<br>90.00% | - | - | - |
| Narin et al., 2020 | X-ray | InceptionV3<br>ResNet50<br>ResNet101<br>ResNet152<br>Inception-ResNetV2 | 97.70%<br>99.70%<br>97.10%<br>97.00%<br>95.30% | 97.40%<br>99.80%<br>99.10%<br>99.10%<br>98.30% | 82.4%<br>98.30%<br>95.60%<br>95.70%<br>84.00% | 100%<br>98.80%<br>88.30%<br>85.30%<br>70.70% |

## 2.4 Summary

In this chapter of the report, roles of medical imaging and machine learning in the ongoing pandemic were briefly reviewed and summarized as follow:

Universiti Malaya

(1) Similar to other common pneumonia, COVID- 19 induces pathological response in lung of the patient that can be detected using chest CT imaging technique. Although it has overlapping characteristics with other types of atypical pneumonia, there are still notable radiological differences between them. Therefore, making CT imaging technique a useful tool in detecting COVID- 19 cases.

(2) Deep learning is a subset of machine learning that has gain popularity in various computer vision application. CNN being one of the most well- developed deep learning algorithms has been selected to be used in this project due to its outstanding performance in image- related task. However, CNN is not without drawbacks. The major drawback of CNN is that training CNN models are time- consuming and require high performance equipment. Therefore, transfer learning technique was introduced.

(3) Transfer learning is basically the reapplication of a previously learned model to a new problem. Pre- trained model transfer learning approach was selected in this project because it is less complicated to implement while providing similar or better performance than developing model approach.

(4) Lastly, numerous CNN architecture proposed by other researchers were also review and finally VGG16 and ResNet50 architectures were selected as baseline in performing the COVID- 19 classification task due to their state- of- art performance not only in the ImageNet dataset but also in COVID- 19 datasets used by other researchers.

Universiti Malaya

# CHAPTER 3: MATERIALS AND METHODOLOGY

This part of the report describes the methodology of implementing a binary classifier to differentiate between chest CT images of COVID-19 positive patients and chest CT images from COVID- 19 negative patients. The proposed deep learning models were implemented, trained and evaluated using Keras library on Tensorflow backend on Google Collaboratory (Colab) Notebooks which allocates an online cloud service with 12GB NVIDIA Tesla K80 Graphics Processing Unit (GPU) for free.

First of all, CT images of COVID- 19 positive and negative were collected as the input. Then, CNN architecture was used as a baseline to perform classification to classify those images into infected and non- infected categories. Rather than training CNN from scratch, transfer learning was utilized by transferring VGG16 and ResNet50 architecture learned on the large ImageNet dataset to the classification task. Then, tuning of the hyperparameters were performed in order to obtain the best possible outcome. Lastly, k-fold validation will be used to prevent over fitting problem. *Figure 3.1* illustrates the architecture of the proposed model.

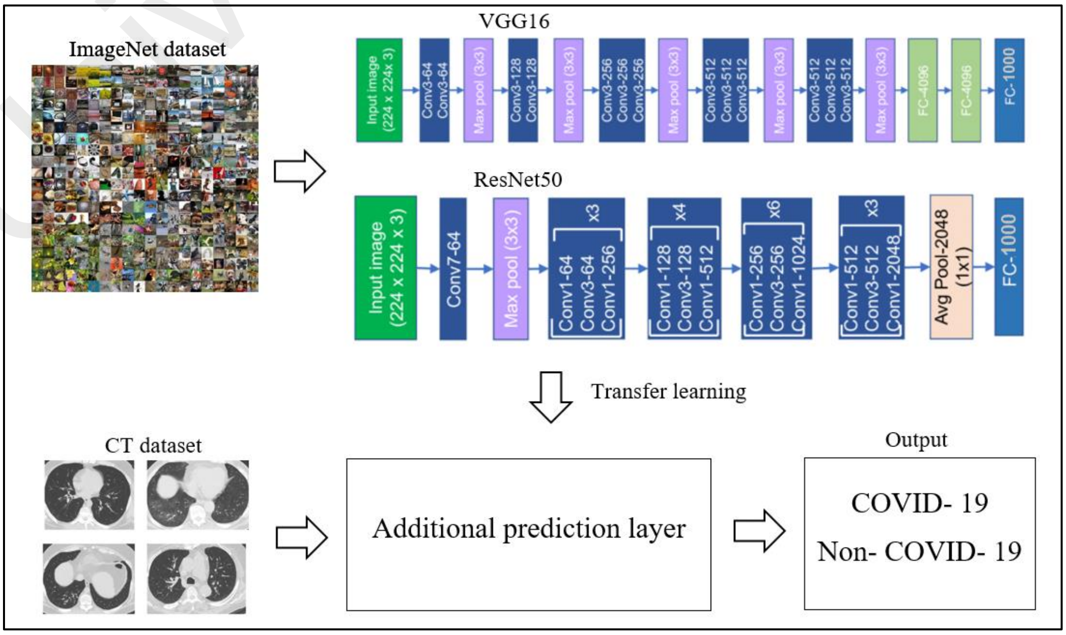


**Figure 3.1: Architecture of proposed model.**

Universiti Malaya

## 3.1 Dataset

In this study, two publicly available dataset was used together as the experimental data to increase the robustness of the implemented models: SARS-COV-2 Ct-Scan Dataset (available at: https://www.kaggle.com/plameneduardo/sarscov2-ctscan-dataset?select=non-COVID) and COVID-CT-Dataset (available at: https://github.com/UCSD-AI4H/COVID-CT). The choice of these dataset for experiment was influenced by the fact that it is an open source, and accessible to the research community and public.

The SARS-COV-2 Ct-Scan Dataset comprises a total of 2482 chest CT scans in the axial view obtained from 160 actual patients in hospitals from Sao Paulo, Brazil. On the other hand, the COVID-CT-Dataset comprises a total of 746 chest CT scans in the axial view obtained from 307 patients from China during the early outbreak of COVID- 19. The distribution of the dataset is shown in *Table 3.1*.

**Table 3.1: Details of data distribution.**

| COVID- 19 | Number of Images (Number of Patients) | | |
|---|---|---|---|
| | SARS- COV- 2 Ct-Scan Dataset | COVID-CT-Dataset | Total |
| **Positive** | 1252 (80) | 349 (216) | **1601 (296)** |
| **Negative** | 1230 (80) | 397 (91) | **1627 (171)** |
| **Total** | **2482 (160)** | **746 (307)** | **3228 (467)** |

*Figure 3.2* shows some of the images from the two datasets. All images in those datasets are already in open- lung mode and are already represented in the 2D PNG format, therefore no image selection algorithm is required to discard the closed- lung images as well as no image format conversion required.

Universiti Malaya

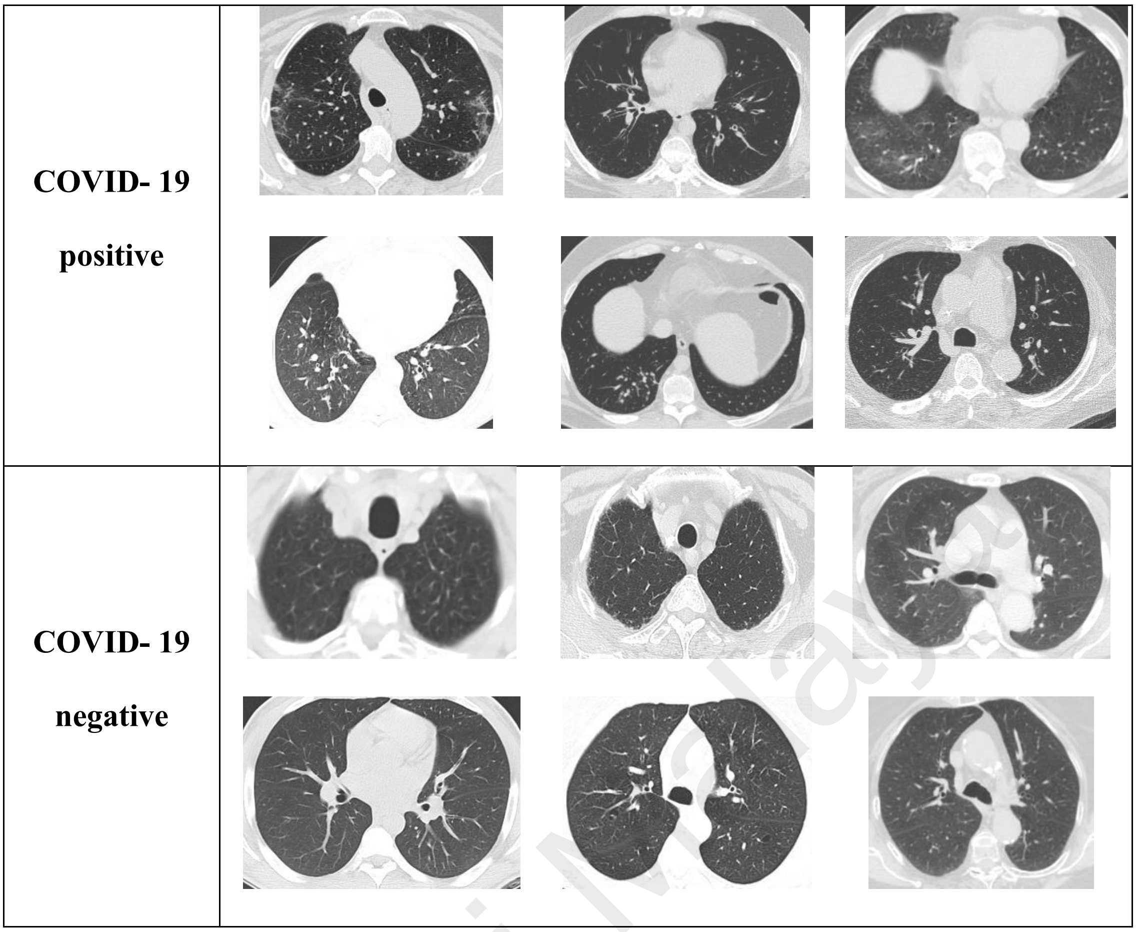


**Figure 3.2: Sample images from datasets.**

## 3.2 Data Preparation

In machine learning, data preparation is the process of organizing raw data to make it suitable to create and train the machine learning model. It is essential to enhance the data quality, therefore facilitating extraction of meaningful information from the data. In the first step of data preparation, the CT images were resized to have the same scale (100 × 100) in order to visualize and facilitate the training process of the CNN. The next step is to split the dataset into training, validation, and testing set. First, the complete dataset was randomly split into the training and testing set with 80:20 ratio, which means 80% of the data used for the training model and 20% for testing the model. Then, the training set was further divided in a ratio of 80:20 for validation purpose. *Figure 3.3* and *Table 3.2* below shows the folder structure and the data distribution of the dataset after data splitting.

Universiti Malaya

```
/dataset
        /train
                /Covid
                /Non-covid
        /test
                /Covid
                /Non-covid
        /validation
                /Covid
                /Non-Covid
```

**Figure 3.3: Folder structure of dataset.**

**Table 3.2: Details of training, validation and testing dataset.**

| Label | Number of Images | | | |
|---|---|---|---|---|
| | **Training** | **Validation** | **Testing** | **Total** |
| **Covid** | 1016 | 260 | 325 | **1061** |
| **Non- Covid** | 1051 | 256 | 320 | **1627** |
| **Total** | **2067** | **516** | **645** | **3228** |

## 3.3 Proposed CNN Architecture

Based on literature, VGG16 and ResNet50 architectures has been selected as the baseline CNN models to be used in this study due to its outstanding performance compared to the other available models.

Initially, the performance of the baseline models was evaluated. The validation and testing result revealed that the model performed badly on the COVID- 19 dataset. Based on the results, it was expected that the models exhibit signs of overfitting where the model shows good performance on the training data but poor generalization on the validation and test data. Therefore, various performance enhancement techniques were implemented to improve the performance of the baseline models. The details are enumerated in the following subsections.

Universiti Malaya

### 3.3.1 Data Augmentation

Overfitting occurs when the model tends to learn patterns directly from the training data and is unable to generalized, hence fail to perform on test data. In order to prevent over- fitting issue, one of the solutions is to train with more data as more training data will tend to generalize the model, hence, able to perform well on test data (Koehrsen, 2018). Alternatively, overfitting can also be fixed by performing data augmentation to the training set. Data augmentation increases the amount of dataset by artificially creating more training images by introducing random variations from the existing ones, hence effectively reduces overfitting issue. In this project, data augmentation technique was implemented carefully to augment the CT images while maintaining its quality. *Table 3.3* below shows the data augmentation techniques that were used, and they were selected based on references.

**Table 3.3: Augmentation parameters.**

| Technique | Function | Description |
|---|---|---|
| Re- scaling | rescale = 1./255 | Transform every pixel value from range [0,255] to [0,1]. |
| Flipping | vertical_flip = True<br>horizontal_flip = True | Randomly flipping the image vertically and horizontally. |
| Rotation | rotation_range = 10 | Randomly rotating the image in the range of 20 degree. |
| Shifting | height_shift_range = 0.2<br>width_shift_range = 0.2 | Randomly shifting the image vertically and horizontally with a scale of 0.2. |
| Zooming | zoom_range = 0.1 | Randomly zooming the image with a scale of 0.1. |

*Note.* Adapted from *A Comprehensive Guide to Transfer Learning*, by R. Mehrotra, 2020 (https://www.kaggle.com/rajmehra03/a-comprehensive-guide-to-transfer-learning).

### 3.3.2 Re- training of Weights

VGG16 and ResNet50 model pre- trained on the ImageNet dataset were primarily designed to classify millions of pictures into 1000 categories. In order to make the standard pre- trained VGG16 and ResNet50 model more specific, relevant and efficient



Universiti Malaya

for the COVID- 19 classification task, the models' convolutional blocks have been re-trained end- to- end on the CT dataset to enhance the model weights.

### 3.3.3 Tuning of Hyperparameters

An important step in training a classification model is to analyze and fine tune the model to a preferred level of performance required for the specific task. Hence, in transfer learning approach applied in this project, manipulating the hyperparameters such as number of layers, optimization algorithm, learning rates, number of steps, epoch values and batch sizes in the pre- trained models will induce an effect on the performance of the training model.

A good model for classification needs to have an optimum learning rate where the convergence occurs optimally without being underfitting or overfitting, as illustrated in *Figure 3.4*. Underfitting occurs when the model is over regularized and improvement on the test data is still required. The main reason behind underfitting is due to insufficient training time and the model is unable to learn the features of the training data. This is caused by too large value of learning rate where it is hard to converge. On the other hand, overfitting is the opposite of underfitting, it will show a rapid convergence of learning rate. To efficiently counter overfitting, a control over the number of epochs or number of steps for training, will efficiently control the duration of training, hence preventing overfitting.

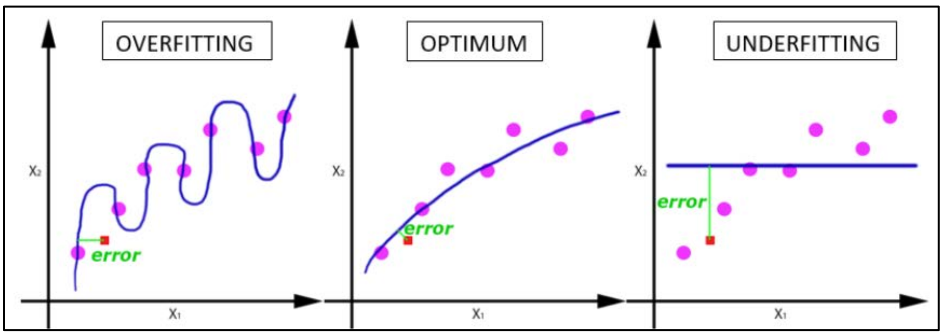


**Figure 3.4: Illustration of overfitting, optimum and underfitting.**

The alteration of batch size, epoch and step size are closely related. Batch size refers to amount of training data used in an iteration while epoch is related to the entire data passing through the neural network backward and forward once. Iteration, also known as step size, is the number of times where an amount of batch size to reach a specific amount of epoch. For instance, if a training dataset of 2000 images are used, a batch size of 10 will require 200 iterations to complete an epoch (Sharma, 2017). The number of epoch or step size is to determine the duration or steps required to complete a training, where an adequate amount is able to obtain an optimize model that is generalized. Whereas a large amount of step size will likely to benefit underfitting model while reduces the performance of an overfitted model. Therefore, a suitable value of the hyperparameters needed to be modified and monitor constantly to obtain the optimum level of performance.

Suitable number of layers should be added to the model because small number of layers would lead to underfitting while large numbers of layers increase model complexity. Various intermediate layer such as Convolutional layers, Batch Normalization, Max Pooling, and Dropout layers can be added to the model.

### 3.3.4 Five- fold Cross Validation

Cross- validation (CV) was utilized to evaluate a model's performance and robustness on unseen data with the goal of minimizing generalization error before selecting the final model (Mishra, Singh & Joshi, 2021). As a CV method, k- fold was chosen in this study and 5- fold CV, as illustrated in *Figure 3.5* has been used. In 5- fold CV, the dataset is separated into five folds, where in each experiment, four folds will be used as the training set and the remaining fold as validation set. The result of the binary class classification for the implemented model will then be recorded for each fold in the confusion matrix (*Figure 3.6*) and the average score will be calculated, as illustrated in *Figure 3.5*.

Universiti Malaya

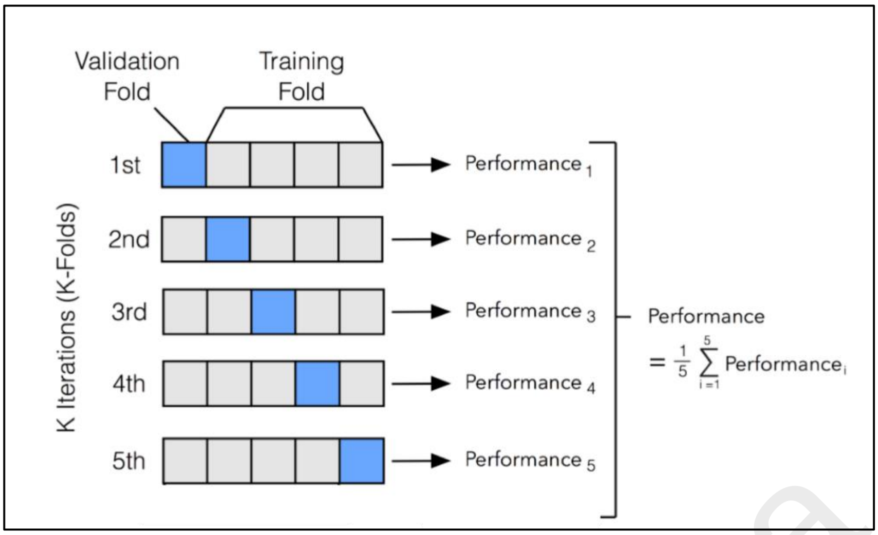


**Figure 3.5: Illustration of 5- fold cross validation.**

| | |
|---|---|
| True Negative (TN)<br>Predicted Negative<br>Actual Negative | False Positive (FP)<br>Predicted Positive<br>Actual Negative |
| False Negative (FN)<br>Predicted Negative<br>Actual Positive | True Positive (TP)<br>Predicted Positive<br>Actual Positive |

**Figure 3.6: Confusion Matrix.**

*Note*. Adapted from *Understanding the Confusion Matrix from Scikit learn*, by S. Agrawal, 2020 (https://towardsdatascience.com/understanding-the-confusion-matrix-from-scikit-learn-c51d88929c79).

## 3.4 Performance Evaluation

The proposed models were evaluated using the training- cross validation- test method, which implies that the model is trained on the training set, tuned on the validation set, and final testing is done on the testing set. It is important to note that the testing set shall never be used to pick between two or more networks, to ensure the error on the testing set provides an unbiased estimate of the generalization error. In other words, after the final model has been evaluated on the testing set, the model shall not be tuned any further.

Universiti Malaya

Four criteria utilized to evaluate the model performance were accuracy, precision, recall and F- measure as summarized in *Table 3.4* below. The selected performance metrics are the top metrics used in measuring the performance of classification models (Khan, Shah & Bhat, 2020). Accuracy is the most straightforward metric; it is simply the fraction of successfully predicted observations to the entire sample size. On the other hand, precision refers to the number of the predicted positives are actually positive, therefore it is particularly useful when the cost of false positive are large. Contrastingly, when there is a significant cost associated with false negative, recall shall be used as the metrics to choose the best model, it is the fraction of correctly predicted positive observations to the actual class. Lastly, F- measure is the measure of model accuracy using precision and recall. The main difference between accuracy and F- measure is that accuracy is measures the true positive and true negative while F- measure measures the false positive and false negative. Therefore, when the class distribution is consistent, accuracy is preferable while F- measure is more commonly used with unbalanced classes (Joshi, 2016).

**Table 3.4: Performance metrics.**

| Performance Metrics | Equation |
|---|---|
| Accuracy | $\frac{(TN+TP)}{(TN+TP+FN+FP)}$ *(3.1)* |
| Precision | $\frac{TP}{(TP+FP)}$ *(3.2)* |
| Recall | $\frac{TP}{(TP+FN)}$ *(3.3)* |
| F- measure | $\frac{(2 \times Precision \times Recall)}{(Precision + Recall)}$ *(3.4)* |

*Note*. Adapted from "CoroNet: A deep neural network for detection and diagnosis of COVID-19 from chest x-ray images," by A.I. Khan, J. L. Shah, & M. M. Bhat, 2020, *Computer Methods and Programs in Biomedicine*, *196*, 105581.

# CHAPTER 4: RESULTS

Numerous efforts have been made to obtain the best possible performance of the pre-trained transfer learning model. The results of the finely- tuned transfer learning models will be discussed in this chapter.

## 4.1 Final Proposed Model

In this project, two different binary classification models were implemented using VGG16 and ResNet50 as baseline. For each model, 5- fold CV was incorporated to assess the effectiveness and robustness of the model. The details of the final proposed models are as shown in *Figure 4.1, Figure 4.2,* and *Table 4.1* below.

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
vgg16 (Functional)           (None, 512)               14714688
_________________________________________________________________
dense (Dense)                (None, 2)                 1026
=================================================================
Total params: 14,715,714
Trainable params: 14,715,714
Non-trainable params: 0
_________________________________________________________________
```

**Figure 4.1: Details of proposed transfer learning Model 1, based on VGG16.**

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
resnet50 (Functional)        (None, 2048)              23587712
_________________________________________________________________
dense (Dense)                (None, 2)                 4098
=================================================================
Total params: 23,591,810
Trainable params: 23,538,690
Non-trainable params: 53,120
_________________________________________________________________
```

**Figure 4.2: Details of proposed transfer learning Model 2, based on ResNet50.**

Universiti Malaya

**Table 4.1: Details of hyperparameters used in proposed deep learning models.**

| Hyperparameters | Model 1 | Model 2 |
|---|---|---|
| Additional layers | Dense(2,activation='softmax') | Dense(2,activation='softmax') |
| Optimizer | Adam | Adam |
| Loss function | Binary cross entropy | Binary cross entropy |
| Metric | Accuracy | Accuracy |
| Learning rate | 0.0001 | 0.00001 |
| Batch size | 128 | 32 |
| Epochs | 100 | 100 |

## 4.2 Performance Analysis

The details of the confusion matrix and performance metrics of the proposed models are presented in *Table 4.2*. In addition, the validation metrics for the binary classifier over 5- fold CV are tabulated in *Table 4.3*.

**Table 4.2: Performance on test set.**

| **Models** | **Results** | | | | |
|---|---|---|---|---|---|
| | **Confusion Matrix** | **Accuracy** | **Precision** | **Recall** | **F- measure** |
| Model 1 | [309 11]<br>[ 20 305] | 95.19% | 96.51% | 93.84% | 95.16% |
| Model 2 | [316 4]<br>[ 7 318] | 98.29% | 98.75% | 97.84% | 98.29% |

Universiti Malaya

**Table 4.3: Performance on validation set.**

| Models | Fold | Results | | | | |
|---|---|---|---|---|---|---|
| | | Confusion Matrix | Accuracy | Precision | Recall | F- measure |
| Model 1 | Fold 1 | [229 28]<br>[ 19 241] | 90.90% | 89.59% | 92.69% | 91.11% |
| | Fold 2 | [253 3]<br>[ 16 245] | 96.32% | 98.79% | 93.86% | 96.26% |
| | Fold 3 | [248 8]<br>[ 9 251] | 96.70% | 96.91% | 96.53% | 96.72% |
| | Fold 4 | [255 4]<br>[ 4 256] | 98.45% | 98.46% | 98.46% | 98.46% |
| | Fold 5 | [248 11]<br>[ 3 258] | 97.30% | 95.91% | 98.85% | 97.35% |
| | **Average** | **-** | **95.94%** | **95.93%** | **96.08%** | **95.98%** |
| Model 2 | Fold 1 | [238 19]<br>[ 33 227] | 89.94% | 92.27% | 87.30% | 89.72% |
| | Fold 2 | [247 9]<br>[ 10 251] | 96.32% | 96.53% | 96.16% | 96.35% |
| | Fold 3 | [237 19]<br>[ 23 237] | 97.28% | 99.59% | 95.00% | 97.24% |
| | Fold 4 | [253 3]<br>[ 1 259] | 99.22% | 98.85% | 99.61% | 99.23% |
| | Fold 5 | [253 3]<br>[ 4 258] | 98.64% | 98.85% | 98.47% | 98.66% |
| | **Average** | **-** | **96.28%** | **97.21%** | **97.30%** | **96.24%** |

Training and validation learning curves of classification accuracy and cross- entropy loss was also assessed during the training process. Parameters were tuned based on the cross- entropy loss learning curves, and the models are evaluated based on the classification accuracy learning curves. Overfitting and underfitting issues were also assessed, and fine tuning of the hyperparameters was performed based on these findings to obtain optimal performance. *Figure 4.3* and *Figure 4.4* shows the curves of classification accuracy and cross- entropy loss for Model 1 and Model 2 respectively over 5- fold CV.

Universiti Malaya

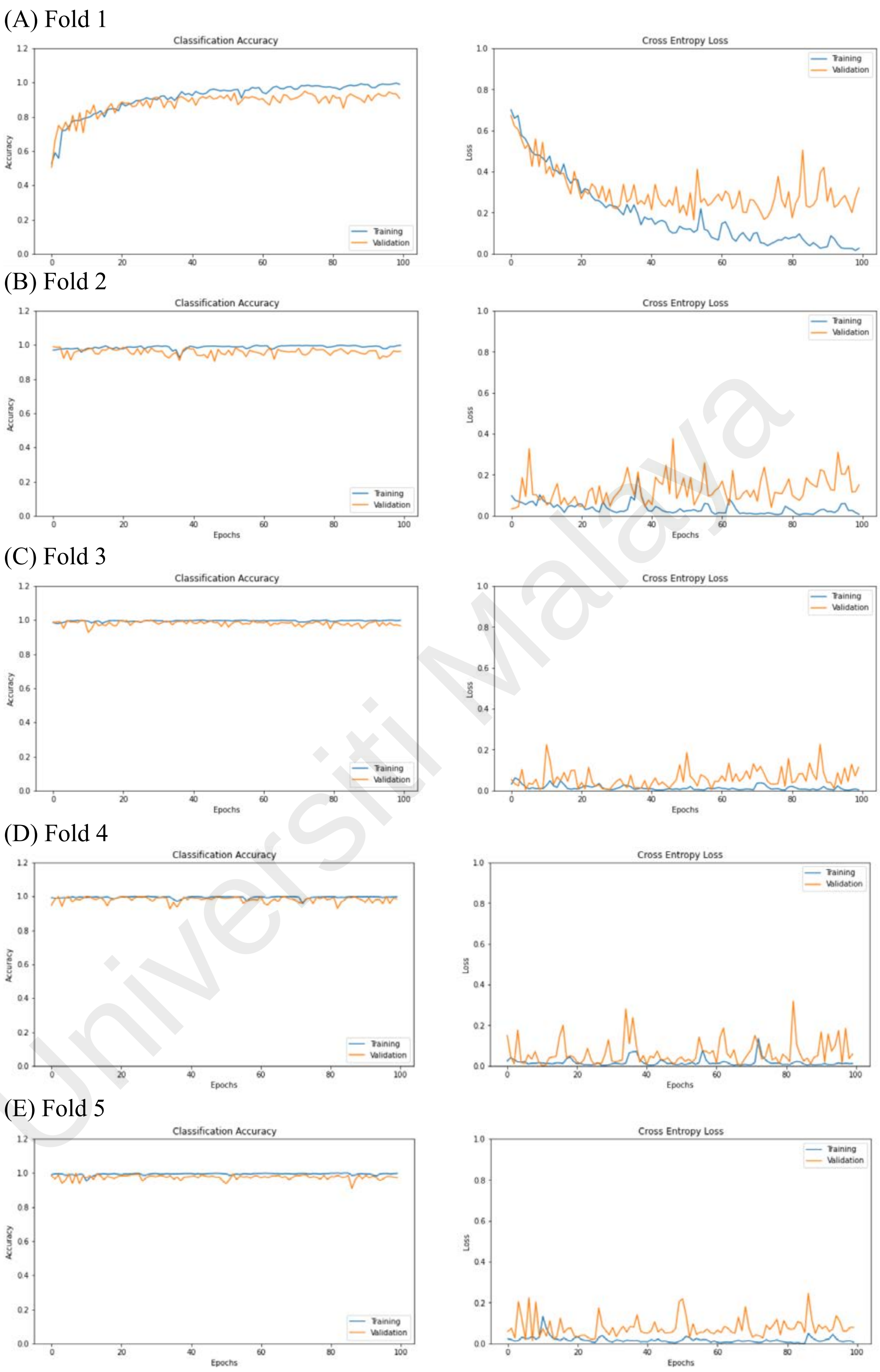


**Figure 4.3: Learning curves of Model 1.**

Universiti Malaya

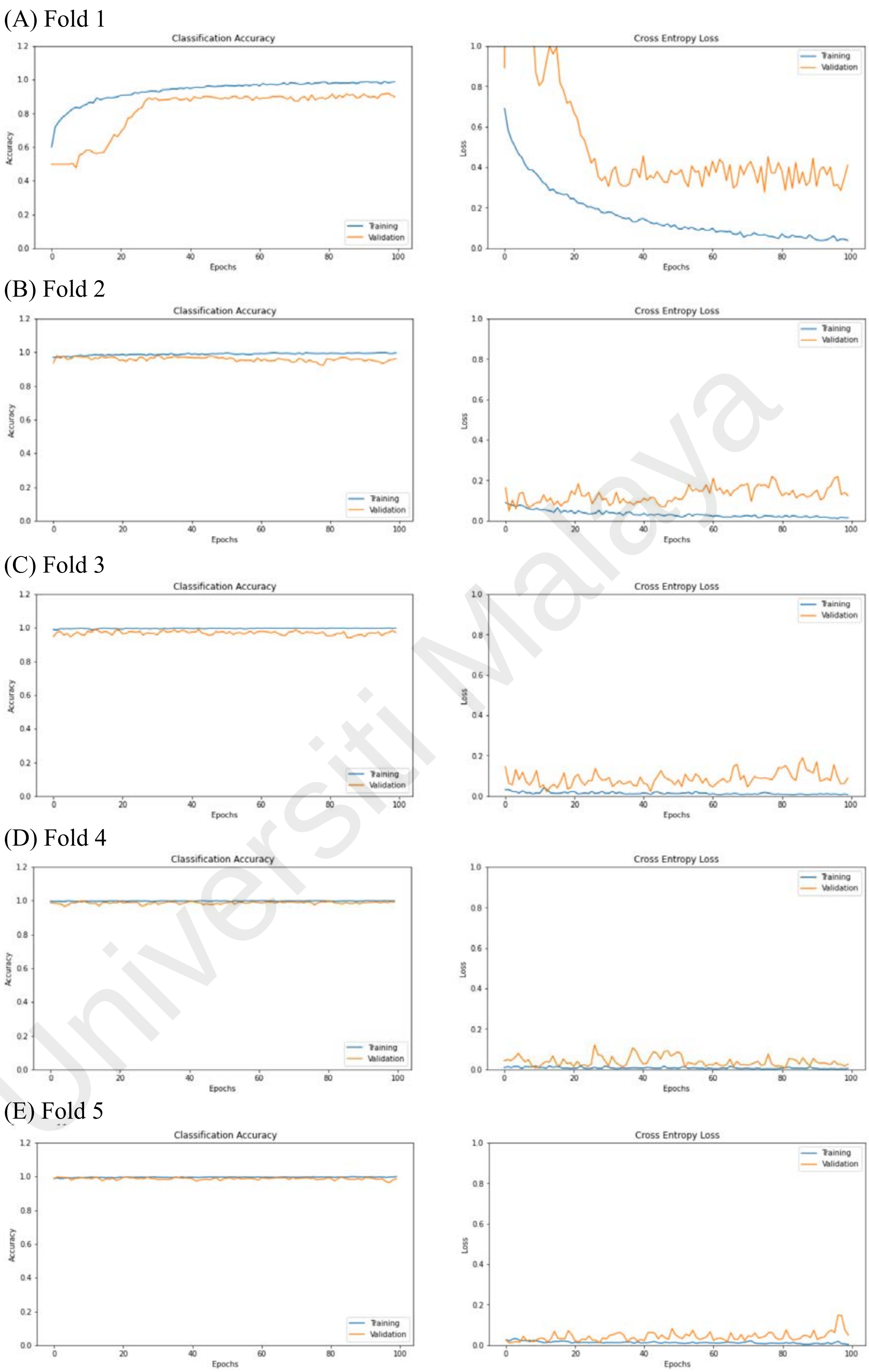


**Figure 4.4: Learning curves of Model 2.**

Universiti Malaya

## 4.3 Tuning of Hyperparameters

To achieve the greatest possible model performance, numerous attempts have been made to tune the hyperparameters of the pre- trained model. The sub- sections that follow described how the hyperparameters were chosen and how would they affect the performance of the model.

### 4.3.1 Additional Layers

In the proposed models, only a single convolution layer with softmax classification function was added at the bottom of the classifier to decrease the output category to fit the number of classes (Covid and Non-Covid). The reason for this is that the experiment demonstrated that adding more layers to pre- trained models does not improve their accuracy rather decreasing accuracy. The architecture and complexity of VGG16 and ResNet50 models were well- designed to require minimal layer customization to perform effectively on a variety of datasets. *Table 4.4* below shows the comparison of the performance of Model 1 and Model 2 with and without additional layers.

**Table 4.4: Multiple attempts on deciding number of additional layers.**

| Models | Additional Layers | Validation Accuracy |
|---|---|---|
| Model 1 | Dense, 256<br>Batch Normalization<br>Dropout, 0.5<br>Dense, 2 | 94.72% |
| | **Dense, 2** | **95.94%** |
| Model 2 | [ Dense, 256<br>Batch Normalization<br>Dropout, 0.5 ] × 2<br>Dense, 2 | 93.05% |
| | **Dense, 2** | **96.28%** |

Universiti Malaya

### 4.3.2 Optimizer

Optimizer is one of the most critical parameters in training a deep learning model to minimize the cost function to provide the best possible accuracy. Cost function represents the difference between the true value of the predicted parameter and the predictions of the model. Therefore, minimizing the cost function improves the accuracy of the model. There are various types of optimization algorithms, namely Adam, stochastic gradient descent (SGD), RMSprop etc. The optimization algorithm that was selected to be used in this project is the Adam optimizer due to its outstanding performance in minimizing loss function as discussed below.

Adam is an extension of SGD introduced by Diederik Kingma and Jimmy Ba in the paper titled "Adam: A Method for Stochastic Optimization 2014" (Kingma & Ba, 2014). In the paper, the authors described some appealing features of adopting Adam compared to the classical SGD, as follow:

(1) It is straightforward to implement.

(2) It is computationally efficient.

(3) It requires little memory.

(4) The gradients are invariant to diagonal rescale.

(5) It is excellent for problems with a significant amount of data and parameters.

Therefore, Adam is currently regarded as the default algorithm to use in machine learning problems, and often outperforms other optimization algorithms, as illustrated in *Figure 4.5* below.

Universiti Malaya

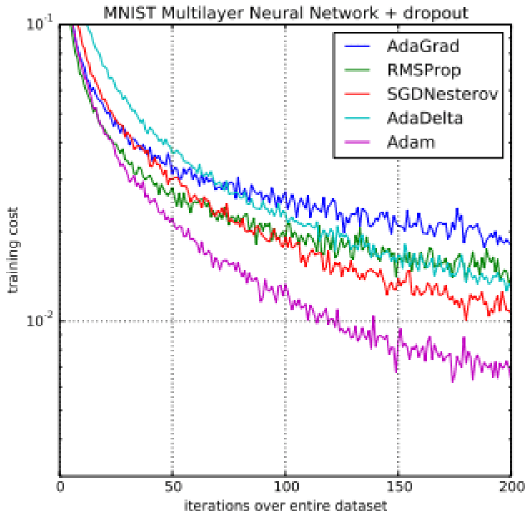


**Figure 4.5: Adam vs. other optimization algorithms in training a multilayer neural network.**

*Note*. From "Adam: A method for stochastic optimization," by D. P. Kingma, & J. Ba, 2014, *arXiv preprint arXiv:1412.6980*.

### 4.3.3 Loss Function

When SGD is used to train deep learning models, it is necessary to pick a loss function while implementing the model. The three most frequently used lost functions used in deep learning models are known as the mean square error (MSE) loss function, which is preferable in regression tasks, binary cross entropy loss function for binary classification tasks, and categorical cross entropy loss function for multi- class classification tasks. Therefore, for the binary classification task in this project, binary cross entropy loss function was selected.

### 4.3.4 Metric

The metric selected to track the result for each step was accuracy which is simply the fraction of successfully predicted observations to the entire sample size. The metric was chosen based on the fact where the class distribution of the dataset is consistent with a total of 1601 'Covid' images and 1627 'Non- Covid' images. As mentioned in *Section 3.4*, when the class distribution is consistent, accuracy is preferable.

Universiti Malaya

### 4.3.5 Learning Rate

Learning rate in an optimizer determines the step size at each iteration to achieve minimum loss function. While too low learning rate stuck the learning process, too high learning rate result in undesired divergent behaviour in the loss function as well as increases the likelihood of overfitting, as illustrated in *Figure 4.6*. After several attempts, the final learning rate were set to 0.0001 for Model 1 and 0.00001 for Model 2.

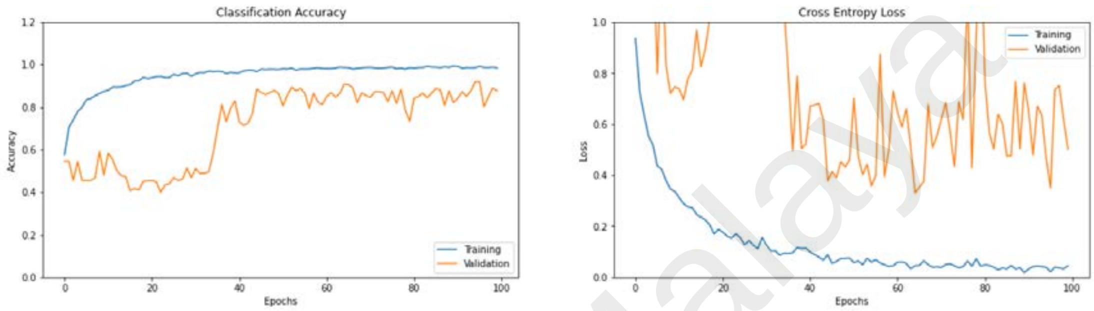


**Figure 4.6: Learning curve for Model 2 (learning rate = 0.0001).**

### 4.3.6 Batch Size

When using SGD to train deep learning models, it is essential to choose a gradient descent variant. Typically, there are three basic types of gradient descent namely batch mode, mini- batch mode, and stochastic mode, as tabulated in *Table 4.5* below.

**Table 4.5: Summary of various gradient descent technique.**

| Gradient Descent | Batch size | Characteristic |
|---|---|---|
| Batch mode | Batch size = Total dataset | Long iteration times |
| Mini- batch mode | 1 < Batch size > Dataset | Faster learning |
| Stochastic mode | Batch size = 1 | Loss of speed due to vectorization |

Among the three mentioned gradient descent, mini- batch mode is regarded as the most recommended gradient descent technique for most machine learning applications, particularly in deep learning, therefore selected to be used in this project (Brownlee, 2017). Typical mini- batch sizes are 32, 64, 128, 256 or 512 where a smaller batch size

Universiti Malaya

guarantees that each training iteration is quick, while a higher batch size provides more exact estimation of the gradients. Therefore, in the experiment, different batch sizes starting from a smaller value were experimented until no significant improvements in performance were observed, as illustrated in *Table 4.6*. After several attempts, the final batch size was set to 128 for Model 1 and 32 for Model 2.

**Table 4.6: Multiple attempts on deciding batch size.**

| Models | Batch Size | Validation Accuracy (Fold 1) |
|---|---|---|
| Model 1 | 64 | 86.07% |
| | **128** | **90.90%** |
| | 256 | 90.52% |
| Model 2 | **32** | **89.94%** |
| | 64 | 88.91% |

### 4.3.7 Epochs

The number of epochs was determined based on the training and validation accuracy and cross entropy loss. Learning curves of the accuracy and cross entropy loss were plotted over epochs to track their values. In the beginning of the experiment, a large number of epochs was selected to observe where the learning curves start saturating to have an idea about the effective number of epochs, as illustrated in *Figure 4.7* and *Figure 4.8*. Finally, the epoch number was set to 100 for both Model 1 and Model 2.

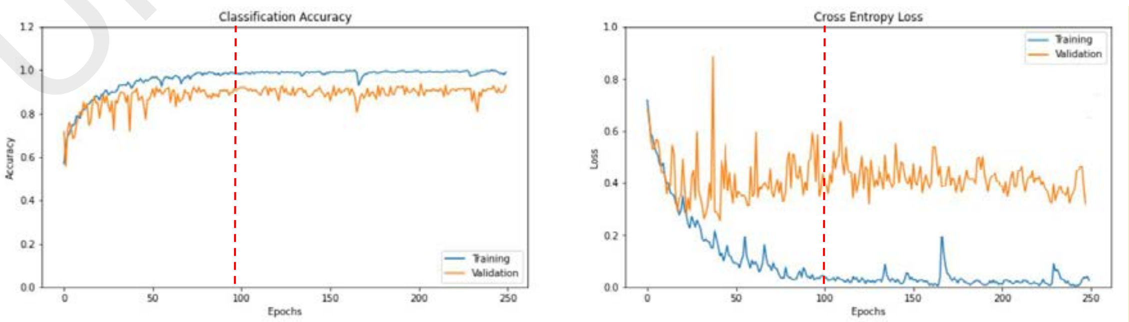


**Figure 4.7: Learning Curve for Model 1 (epochs = 250).**

Universiti Malaya

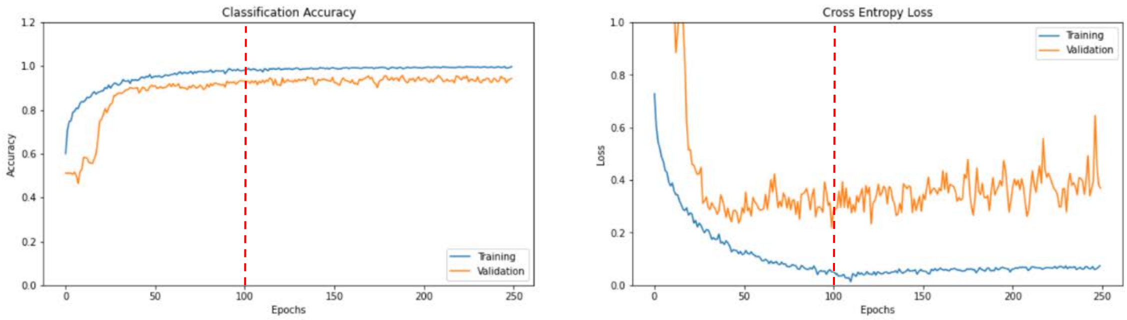


**Figure 4.8: Learning curve for Model 2 (epochs = 250).**

## 4.4 Summary

Some noteworthy observations have been established based on the comprehensive experiment and thorough result analysis of the proposed transfer learning models in this project, and are summarized as follows:

(1) In Model 1, the VGG as baseline model achieved relatively high validation accuracies ranging from 90.90% to 98.45% over the 5- fold CV with an average score of 95.94%. On top of that, the accuracy score on the test set also achieved quite similar result of 95.19%. Moreover, the training and validation learning curves are converging pretty well with no significant deviations, indicating a good fit of the model.

(2) In Model 2, the ResNet50 as baseline model achieved an average accuracy score of 96.28% with each fold ranging from 89.94% to 99.22% over the 5- fold CV. In addition to that, the model performance on the test set was outstanding with an accuracy of 98.29%. Furthermore, the training and validation learning curves also converged nicely with no major deviations, indicating a good model fit.

(3) It was further observed that the performance of Model 2 that utilized ResNet50 model as baseline is better than Model 1 that utilized VGG16 model. This observation is so much in line with the literature where ResNet50 model is expected

to have a higher training performance than a VGG16 model due to the increased depth of the ResNet50 model as mentioned in *Section 2.2.3.2*.

(4) The performance of transfer learning models is highly reliant on the hyperparameter settings. Optimal hyperparameters aid in avoiding model overfitting and underfitting. However, optimal hyperparameters for different models and datasets are often different. As a result, selecting the best optimum hyperparameter to balance the underfitting and overfitting issues, as well as the computing efficiency can be tedious.

(5) K- fold CV plays a crucial role in evaluating the robustness and effectiveness of the proposed model. While using a simple train/ validation split, the model may have an excellent performance if the data split is not reflective of the underlying data distribution. As a result, the model may perform excellently on some training folds while performing relatively poor on others. By using k- fold CV technique, the data is divided into k folds and each fold were trained k times, with one fold serving as the validation set and the others as the training set, and the final accuracy will be the mean accuracy from each fold. Therefore, it averages the biases towards the large variations in the test dataset, thus give better insight on the model and eventually reduces overfitting.

Universiti Malaya

# CHAPTER 5: DISCUSSIONS

## 5.1 Comparative Analysis

As mentioned earlier, numerous efforts have been made by researchers around the world in developing machine learning- based COVID- 19 classification models. However, due to lack of uniform training dataset, direct comparison with other machine learning models in the literature could not be conducted. Therefore, comparisons are only made based on proposed methodologies. The proposed model is compared to some existing state- of- the- art methodologies COVID- 19 classification from chest CT and X-ray images. The comparative analysis is as tabulated in *Table 5.1* below.

**Table 5.1: Comparative analysis.**

| Study | Data Type | Architecture | Result | | | |
|---|---|---|---|---|---|---|
| | | | Accuracy | Precision | Recall | F-measure |
| Yu et al., 2020 | CT | GoogLeNet-COD | 87.50% | 84.09% | - | 90.91% |
| Serte & Demirel, 2021 | CT | ResNet50+MV | 96.00% | 96.00% | - | 100% |
| Yan et al., 2020 | CT | MSSP-MSCNN | 99.50% | 97.70% | - | 96.20% |
| Rahimzadeh, Attar & Sakhaei, 2021 | CT | FPN- ResNet50V2 | 98.49% | - | - | 94.96% |
| Mishra, Singh & Joshi, 2021 | CT | VGG16<br>ResNet50 | 99.12%<br>99.62% | - | - | - |
| Singh et al., 2020 | X- ray | MADE-CNN | 95.80% | - | 96.16% | 95.60% |
| Hemdan et al., 2020 | X- ray | VGG19<br>DenseNet201 | 90.00%<br>90.00% | - | - | - |
| Narin et al., 2020 | X- ray | InceptionV3<br>ResNet50<br>ResNet101<br>ResNet152<br>Inception-ResNetV2 | 97.70%<br>99.70%<br>97.10%<br>97.00%<br>95.30% | 97.40%<br>99.80%<br>99.10%<br>99.10%<br>98.30% | 82.40%<br>98.30%<br>95.60%<br>95.70%<br>84.00% | 100%<br>98.80%<br>88.30%<br>85.30%<br>70.70% |
| Proposed Model | CT | VGG16<br>ResNet50 | 95.19%<br>98.29% | 96.51%<br>98.75% | 93.84%<br>97.84% | 95.16%<br>98.29% |

Universiti Malaya

Although the results obtained by the proposed models do not achieve the highest accuracy compared to other studies in the literature, nevertheless, the key contributions of this work are as follows:

(1) Two transfer learning- based architecture (VGG16 and ResNet50) were extensively studied and evaluated to classify between COVID- 19 positive and negative from the input CT images.

(2) Two different datasets consisting of COVID- 19 positive and negative CT images were used to increase the robustness in training the proposed model.

(3) The proposed method does not include any manual feature extraction or feature selection, it was performed from beginning to end directly using only raw data to increase the model generalization ability.

(4) Multiple performance improvement techniques such as data augmentation, re-training of weights, additional of new layers, fine- tuning of hyperparameters and 5- fold CV were implemented and incorporated to improve the performance of the baseline models.

(5) Extensive experiment demonstrates that the proposed classification model is computationally less expensive while achieving astonishingly good results.

## 5.2 Proposed Model Outperforms Radiologists

Javaheri and team conducted an experiment to compare the classification accuracy of their framework with four qualified radiologists (Javaheri et al., 2021). In the experiment, each radiologist was given a dataset consisting of 20 cases mixed of COVID- 19 positive and negative cases. The average reading performance for the four independent radiologists showed an accuracy, sensitivity and specificity of 81%, 79.00% and 82.14% respectively. The proposed transfer learning classification model, however outperformed

Universiti Malaya

the radiologists and achieved a test accuracy of 95.19% and 98.29% for Model 1 and Model 2 respectively.

This has to do with the fact that, in order for a radiologist to make accurate diagnosis with only radiography images, it requires expert knowledge and great experience. Furthermore, in the ongoing pandemic where there is lack of human resources, radiology experts may need to work overtime. Increase in workload may increase doctor fatigue which can lead to decreased effectiveness and worse, increase in errors in medical diagnosis. For this reason, incorporating CAD approach in the radiologist framework can ideally reduce the workload of the radiologists while increasing the quantitative and reliability of the analysis.

### 5.3 Chest CT versus X- ray for COVID- 19 Detection

Based on *Table 4.4*, it can be observed that deep learning models that utilizes chest CT images as input generally perform better than those using X- ray images. This is owing to the fact that a CT image generally offers a much higher level of detail as compared to a X- ray image. The issue is made worse by the fact that the changes in the lung parenchyma were not apparent when the imaging is taken during early stages of viral infection.

According to Stephanie and team (2020), the sensitivity of chest X- ray detection increased as the severity of COVID- 19 infection shown on the chest X- ray image increased (Stephanie et al., 2020). *Figure 5.1 (a)* below shows the chest X- ray severity progression of COVID- 19 pneumonia infected patients (red line) over time. The severity scores increased from mean of 1.11 before Day 2 of symptom onset to mean of 1.97 after Day 11. *Figure 5.1 (b)* shows the sensitivity (red line) of chest X- ray COVID- 19 detection. It was reported that chest X- ray sensitivity for COVID- 19 detection rises from

Universiti Malaya

55% before Day 2 to 79% after Day 11, with sensitivity noticeably increasing after Day 6.

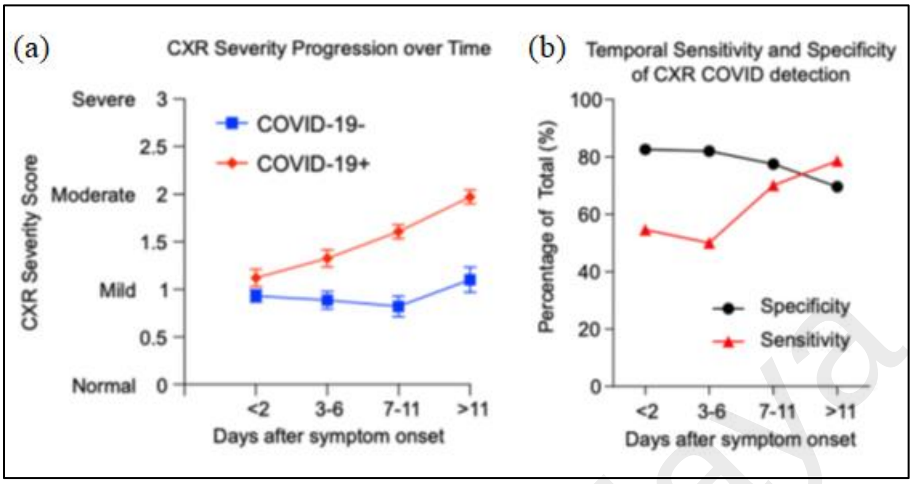


**Figure 5.1: (a) Graph of COVID- 19 chest X- ray severity progression over time. (b) Graph of sensitivity and specificity of chest X- ray COVID- 19 detection.**

*Note*. From "Determinants of chest x-ray sensitivity for covid-19: A multi-institutional study in the united states.," by S. Stephanie, T. Shum, H. Cleveland, S. R. Challa, A. Herring, F. L. Jacobson, ... & M. M. Hammer, 2020, *Radiology: Cardiothoracic Imaging*, *2*(5), e200337.

Therefore, for early diagnosis purposes, X- ray based image test in detecting COVID-19 pneumonia may have limited value. Various studies have suggested that COVID-19 pneumonia diagnosis with CT offers a greater sensitivity and specificity results than radiography method especially in the initial assessment of the patients (Blažić, Brkljačić, & Frija, 2021).

## 5.4 Imaging Test versus Laboratory Test for COVID- 19 Detection

Currently, there are two broad categories of laboratory test: Antigen test and PCR test. PCR test is deemed the "gold standard" in COVID- 19 infection detection. In many countries, PCR test is the first line of diagnosis in patients with COVID- 19 infection. Research have suggested that an initial PCR test demonstrated sensitivity of 91% in identifying COVID- 19 virus (Wong et al., 2020). On the other hand, imaging test in the

Universiti Malaya

initial stage demonstrates a lower sensitivity compared to the initial PCR testing (Herpe et al., 2021).

Beside lower detection rate in the initial stage, the most important factor that limits the application image test in detection of COVID- 19 is that not all people with COVID- 19 infection develop pneumonia, this is known as asymptomatic infections. Imaging tests for asymptomatic infected patient are often negative, leaving laboratory- based detection as the only option. Therefore, image test may have low value in identifying the potential sporadic cases.

On the contrary, although the laboratory test generally shows a more promising detection rate, however, detection rate is influenced by a number of factors, including sampling, storage, and assay performance. In the application of COVID- 19 detection, the typically done throat swabs are influenced by various factors that would reduce detection rates, such as the disease progression, body's immunological state, and the specificity of the infected organs (Meng & Liu, 2020). According to published research, upper respiratory tract PCR test has only sensitivity of 70- 80% when compared to a composite reference standard that includes clinical, radiological, and microbiology data (Fang et al., 2020). Therefore, some researchers support the use of image test to scan the suspected COVID- 19 patient who have clinical and epidemiologic characteristics that are consistent with COVID- 19 infection, especially when the PCR findings are negative (Fang et al., 2020).

The Fleischner Society, an international multidisciplinary medical society devoted to thoracic radiology, does not recommend the use of imaging as a COVID- 19 screening test in asymptomatic individuals; however, for symptomatic patients, specific guidelines are proposed based on the severity of symptoms, pre- test likelihood, and COVID- 19 test results (Rubin et al., 2020). At the same time, the Russian Society of Radiology (RSR) also warns against the use of chest X- ray and CT in asymptomatic patients (Sinitsyn, Tyurin & Mitkov, 2020).

In short, the diagnostic capability of radiology- based image test was not convincing enough to replace the PCR test in the management of COVID- 19, however it can be used as a second diagnosis tool for the symptomatic patients with negative PCR findings or in conjunction with the PCR test as triage purposes. The latter mean to be used as an early detection tool to help determine if a patient should be held in isolation before laboratory test results arrive. This early prediction from the image test stops the illness from spreading quickly to others during the gap when the patient is waiting for the laboratory test result. The image test result is a discriminating factor: If the image is normal, the patient will return home and wait for the laboratory test results.

## 5.5 Limitations and Possible Solution

One limitation in COVID- 19 classification task is the low quality of dataset. Most publicly accessible datasets are either unorganized or with unbalance classes which would affect the training and prediction performances of machine learning- based models. This is understandable as the ongoing pandemic has forced government to allocate more resources in saving lives rather than research initiatives. Therefore, it is suggested that researchers and non- governmental organizations could take the initiative to gather and organize public available data.

Owing to the same fact, studies with promising classification performance were uncertain whether their model could generalize to bigger datasets. While waiting for bigger datasets to be available, studies (including this study) have implemented data augmentation or combined multiple datasets to ensure robustness and reliability of model and prevent negative impacts of small dataset that can cause overfitting issue.

Universiti Malaya

Furthermore, as mentioned in *Section 5.4*, the use of CT scans alone may not be convincing enough for COVID- 19 detection. Therefore, in future work, it is suggested to take into account additional information or data such as the clinical indicators of patients, combined with the imaging scans as input for the CAD system. It is believed with additional information would increase the performance of a CAD system.

Lastly, despite machine learning techniques aided the diagnosis and classification of COVID- 19, nevertheless, choosing the most suitable or high- performance machine learning techniques can be a challenging task. There is a spectrum of options of machine learning techniques creating dilemma in determining which of them could best suit the development of system for COVID- 19 diagnosis and classification. Therefore, instead of testing only one type of model, studies have implemented classification models based on different architectures to compare the performance. For instance, this study utilized both VGG16 and ResNet50 architectures for comparison of results.

Universiti Malaya

# CHAPTER 6: CONCLUSION

## 6.1 Conclusion

To summarize, review on role of medical imaging and machine learning in the ongoing pandemic was discussed. Medical imaging, particularly the CT technique together with machine learning can provide a more efficient way of detecting COVID- 19 while greatly minimizing impact of human error. In this project, two CNN models were proposed to classify those images into normal or COVID- 19 infected categories. VGG16 and ResNet50 models as baseline were used together with transfer learning technique to implement the classification models. Various performance improvement techniques such as data augmentation, re- training of weights, additional of new layers, fine- tuning of hyperparameters and 5- fold CV were also implemented and incorporated to improve the performance of the baseline models. The models have achieved reasonably satisfied performance with test accuracy score of 95.19% and 98.29%, and has outperformed the radiologist in classifying COVID- 19 infected CT scans from normal chest CT scans. In addition, no extensive pre- processing techniques were applied to the input training data therefore increases models' generalization ability, making them readily applied for real time testing for patient using raw images. Yet, it is debatable if an accuracy and specific diagnosis can be made solely on the basis of imaging.

## 6.2 Recommendation for Future Work

In future work, it is suggested that the proposed machine- based classification models can be expanded to perform multi- class classification to differentiate COVID- 19 pneumonia from other atypical pneumonia. However, it can be a challenging task due to the subtle discrepancies between both. Therefore, in order to enhance the performance, it is recommended to perform feature extraction to allow better interpretation and presentation of COVID- 19 lesion that have significant contributions to the classification

Universiti Malaya

results and the overall model performance. In addition to that, expanding the size of dataset by incorporating additional inputs such as ultrasound or X- ray images, clinical indicators of patient, and patient information such as age and gender may also contribute to the overall classification performance.

Universiti Malaya

# REFERENCES


Abelaira, M. D. C., Abelaira, F. C., Ruano-Ravina, A., & Fernández-Villar, A. (2021). Use of conventional chest imaging and artificial intelligence in COVID-19 infection. A review of the literature. *Open Respiratory Archives*, 100078.

Agrawal, S. (2021). Understanding the Confusion Matrix from Scikit learn. Retrieved 31 June 2021, from https://towardsdatascience.com/understanding-the-confusion-matrix-from-scikit-learn-c51d88929c79

Ai, T., Yang, Z., Hou, H., Zhan, C., Chen, C., Lv, W., ... & Xia, L. (2020). Correlation of chest CT and RT-PCR testing for coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases. *Radiology*, *296*(2), E32-E40.

Boureau, Y. L., Ponce, J., & LeCun, Y. (2010). A theoretical analysis of feature pooling in visual recognition. In *Proceedings of the 27th international conference on machine learning (ICML-10)* (pp. 111-118).

Brownlee, J. (2017). A Gentle Introduction to Transfer Learning for Deep Learning. Retrieved 8 June 2021, from https://machinelearningmastery.com/transfer-learning-for-deep-learning/

Das, S. (2017). CNN Architectures: LeNet, AlexNet, VGG, GoogLeNet, ResNet and more. Retrieved 8 June 2021, from https://medium.com/analytics-vidhya/cnns-architectures-lenet-alexnet-vgg-googlenet-resnet-and-more-666091488df5

Fang, Y., Zhang, H., Xie, J., Lin, M., Ying, L., Pang, P., & Ji, W. (2020). Sensitivity of chest CT for COVID-19: comparison to RT-PCR. *Radiology*, *296*(2), E115-E117.

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., ... & Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, *316*(22), 2402-2410.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

Hemdan, E. E. D., Shouman, M. A., & Karar, M. E. (2020). Covidx-net: A framework of deep learning classifiers to diagnose covid-19 in x-ray images. *arXiv preprint arXiv:2003.11055*.


Universiti Malaya

Herpe, G., Lederlin, M., Naudin, M., Ohana, M., Chaumoitre, K., Gregory, J., ... & Tasu, J. P. (2021). Efficacy of Chest CT for COVID-19 Pneumonia Diagnosis in France. *Radiology*, *298*(2), E81-E87.

Hosseiny, M., Kooraki, S., Gholamrezanezhad, A., Reddy, S., & Myers, L. (2020). Radiology perspective of coronavirus disease 2019 (COVID-19): lessons from severe acute respiratory syndrome and Middle East respiratory syndrome. *American Journal of Roentgenology*, *214*(5), 1078-1082.

IBM. (2021). What is Machine Learning?. Retrieved 18 May 2021, from https://www.ibm.com/cloud/learn/machine-learning

Jamil, M., & Hussain, I. (2020). Automatic detection of COVID-19 infection from chest X-ray using deep learning. medRxiv.

Joshi, R. (2016). Accuracy, Precision, Recall & F1 Score: Interpretation of Performance Measures - Exsilio Blog. Retrieved 26 August 2021, from https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/

Khan, A. I., Shah, J. L., & Bhat, M. 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*, 105581.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*, 1097-1105.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436-444.

Mehrotra, R. (2019). A Comprehensive Guide to Transfer Learning. Retrieved 8 June 2021, from https://www.kaggle.com/rajmehra03/a-comprehensive-guide-to-transfer-learning

Meng, X., & Liu, Y. (2020). Chest Imaging Tests versus RT-PCR Testing for COVID-19 Pneumonia: There Is No Best, Only a Better Fit. *Radiology*, *297*(3), E345-E345.

Universiti Malaya

Mishra, N. K., Singh, P., & Joshi, S. D. (2021). Automated detection of COVID-19 from CT scan using convolutional neural network. *Biocybernetics and Biomedical Engineering*, *41*(2), 572-588.

Murugan, P. (2017). Feed forward and backward run in deep convolution neural network. *arXiv preprint arXiv:1711.03278*.

Narin, A., Kaya, C., & Pamuk, Z. (2021). Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks. *Pattern Analysis and Applications*, 1-14.

Ozturk, T., Talo, M., Yildirim, E. A., Baloglu, U. B., Yildirim, O., & Acharya, U. R. (2020). Automated detection of COVID-19 cases using deep neural networks with X-ray images. Computers in biology and medicine, 121, 103792.

Pham, T. D. (2021). Classification of COVID-19 chest X-rays with deep learning: new models or fine tuning?. *Health Information Science and Systems*, *9*(1), 1-11.

Rahimzadeh, M., Attar, A., & Sakhaei, S. M. (2021). A fully automated deep learning-based network for detecting covid-19 from a new and large lung ct scan dataset. *Biomedical Signal Processing and Control*, *68*, 102588.

Rubin, G. D., Ryerson, C. J., Haramati, L. B., Sverzellati, N., Kanne, J. P., Raoof, S., ... & Leung, A. N. (2020). The role of chest imaging in patient management during the COVID-19 pandemic: a multinational consensus statement from the Fleischner Society. *Chest*, *158*(1), 106-116.

Salehi, S., Abedi, A., Balakrishnan, S., & Gholamrezanezhad, A. (2020). Coronavirus disease 2019 (COVID-19): a systematic review of imaging findings in 919 patients. *American Journal of Roentgenology*, *215*(1), 87-93.

Saygılı, A. (2021). A new approach for computer-aided detection of coronavirus (COVID-19) from CT and X-ray images using machine learning methods. *Applied Soft Computing*, *105*, 107323.

Scherer, D., Müller, A., & Behnke, S. (2010, September). Evaluation of pooling operations in convolutional architectures for object recognition. In *International conference on artificial neural networks* (pp. 92-101). Springer, Berlin, Heidelberg.

Serte, S., & Demirel, H. (2021). Deep learning for diagnosis of COVID-19 using 3D CT scans. *Computers in Biology and Medicine*, *132*, 104306.

Universiti Malaya

Sharma, S. (2017). Epoch vs Batch Size vs Iterations. Retrieved 18 August 2021, from https://towardsdatascience.com/epoch-vs-iterations-vs-batch-size-4dfb9c7ce9c9

Shi, G. (2019). Implementing a ResNet model from scratch. Retrieved 8 June 2021, from https://towardsdatascience.com/implementing-a-resnet-model-from-scratch-971be7193718

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Singh, K. K., Siddhartha, M., & Singh, A. (2020). Diagnosis of Coronavirus Disease (COVID-19) from Chest X-ray images using modified XceptionNet. *Romanian Journal of Information Science and Technology*, *23*(657), 91-115.

Sinitsyn, V. E., Tyurin, I. E., & Mitkov, V. V. (2020). Consensus guidelines of Russian Society of Radiology (RSR) and Russian Association of Specialists in Ultrasound Diagnostics in Medicine (RASUDM) role of imaging (X-ray, CT and US) in diagnosis of COVID-19 pneumonia (version 2). *Journal of radiology and nuclear medicine*, *101*(2), 72-89.

Stephanie, S., Shum, T., Cleveland, H., Challa, S. R., Herring, A., Jacobson, F. L., ... & Hammer, M. M. (2020). Determinants of chest x-ray sensitivity for covid-19: A multi-institutional study in the united states. *Radiology: Cardiothoracic Imaging*, *2*(5), e200337.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1-9).

WHO. (2021). Coronavirus disease (COVID-19) – World Health Organization. Retrieved 9 June 2021, from https://www.who.int/emergencies/diseases/novel-coronavirus-2019

Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional neural networks: an overview and application in radiology. *Insights into imaging*, *9*(4), 611-629.

Yan, T., Wong, P. K., Ren, H., Wang, H., Wang, J., & Li, Y. (2020). Automatic distinction between covid-19 and common pneumonia using multi-scale convolutional neural network on chest ct scans. *Chaos, Solitons & Fractals*, *140*, 110153.

Yang, W., Cao, Q., Qin, L. E., Wang, X., Cheng, Z., Pan, A., ... & Yan, F. (2020). Clinical characteristics and imaging manifestations of the 2019 novel coronavirus disease

Universiti Malaya

(COVID-19): a multi-center study in Wenzhou city, Zhejiang, China. *Journal of Infection*, *80*(4), 388-393.

Yu, X., Wang, S. H., Zhang, X., & Zhang, Y. D. (2020). Detection of COVID-19 by GoogLeNet-COD. In *International Conference on Intelligent Computing* (pp. 499-509). Springer, Cham.

Zhang, K., Liu, X., Shen, J., Li, Z., Sang, Y., Wu, X., ... & Wang, G. (2020). Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography. *Cell*, *181*(6), 1423-1433.

Universiti Malaya