A Comparison of Image Classification with Different Activation Functions in Balanced and Imbalanced Datasets

Moqi Zhang

Thesis submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

Master of Science

in

Computer Engineering

Yang Yi, Chair

Haibo Zeng

Jia-Bin Huang

May 10, 2021

Blacksburg, Virginia

Keywords: deep learning, activation function, COVID-19.

A Comparison of Image Classification with Different Activation Functions in Balanced and Imbalanced Datasets

Moqi Zhang

(ABSTRACT)

When the novel coronavirus (COVID-19) outbreak began to ring alarm bells worldwide, rapid, efficient diagnosis was critical to the emergency response. The limited ability of medical systems and the increasing number of daily cases pushed researchers to investigate automated models. The use of deep neural networks to help doctors make the correct diagnosis has dramatically reduced the pressure on the healthcare system. Promoting the improvement of diagnosis networks depends not only on the network structure design but also on the activation function performance. To identify an optimal activation function, this study investigates the correlation between the activation function selection and image classification performance in balanced or imbalanced datasets. Our analysis evaluates various network architectures for both commonly used and novel datasets and presents a comprehensive analysis of ten widely used activation functions. The experimental results show that the swish and softplus functions enhance the classification ability of state-of-the-art networks. Finally, this thesis distinguishes the neural networks using ten activation functions, analyzes their pros and cons, and puts forward detailed suggestions on choosing appropriate activation functions in future work.

A Comparison of Image Classification with Different Activation Functions in Balanced and Imbalanced Datasets

Moqi Zhang

(GENERAL AUDIENCE ABSTRACT)

When the novel coronavirus (COVID-19) outbreak began to ring alarm bells worldwide, the rapid, efficient diagnosis was critical to the emergency response. The manual diagnosis of chest X-rays by radiologists is time and cost-consuming. Compared with traditional diagnostic technology, the artificial intelligence medical system can simultaneously analyze and diagnose hundreds of medical images and speedily obtain high precision and high-efficiency returns. As we all know, machines are brilliant in learning new things and never sleep. Suppose machines can be used to replace human beings in some positions. In that case, it can significantly relieve the pressure on the medical system and buy time for medical practitioners to concentrate more on the research of new technologies. We need to know that the critical decision unit of the intelligent diagnosis system is the activation function. Therefore, this work provides an in-depth evaluation and comparison of the traditional and widely used activation functions with the emerging activation functions, which helps to improve the accuracy of the most advanced diagnostic model on the COVID-19 image dataset. Besides, the results of this study also summarize the cons and pros of using various neural functions and provide many suggestions for future work.

Dedication

To my family

Acknowledgments

First and foremost, I would like to express my deepest appreciation to my advisor, Dr.Yang Yi, for her unwavering support and belief in me. She is not only my academic mentor but also my life mentor. Without her insightful comments and suggestions, this study is impossible.

Much thanks to Dr.Haibo Zeng for his technical support on my study. He has given me valuable advice for my study life and career future. I cannot leave Virginia Tech without mentioning Dr.Jiabin Huang, the professor in ECE 5424: Advanced Machine Learning. He opened the door of Deep Learning to me, and his caring disposition and teaching style inspired me a lot.

I gratefully acknowledge the assistance for all students and faculty members in the MICS lab at Virginia Tech.

Lastly, I also wish to thank my parents for all the support I received, and they are my warmest harbor.

Contents

List of Figures ix

List of Tables xi

1 Introduction 1

1.1 Motivation . . . 1

1.2 Research Objectives . . . 3

2 Review of Literature 4

2.1 Machine Learning and Deep Learning . . . 4

2.2 Neural Networks . . . 5

2.2.1 Deep Neural Networks . . . 5

2.2.2 Convolutional Neural Networks . . . 7

2.2.3 Recurrent Neural Networks . . . 8

2.3 Activation Function . . . 10

2.3.1 Sigmoid . . . 10

2.3.2 Tanh . . . 12

2.3.3 ReLU Family . . . 13

2.3.4 Swish . . . 16

2.3.5 Siren 16

2.4 Related Work 17

3 Experimental Implementation **19**

3.1 Data Sets 19

3.1.1 CIFAR-10 19

3.1.2 COVID-19 RADIOGRAPHY DATABASE 20

3.2 Neural Networks Setup 22

3.2.1 CIFAR-10 DNN Setup 23

3.2.2 CIFAR-10 CNN setup 24

3.2.3 COVID-19 CNN setup 25

4 Results and Analysis **29**

4.1 Evaluation Criteria 29

4.2 Experimental results 31

4.2.1 CIFAR-10 DNN Experimental results 31

4.2.2 CIFAR-10 CNN Experimental results 33

4.2.3 COVID-19 CNN Experimental results 35

5 Conclusion and Discussion **46**

Appendices **49**

Appendix A Confusion Matrix **50**

A.1 Confusion Matrix RMSprop 0.0001 50

A.2 Confusion Matrix RMSprop 0.00001 52

Bibliography **57**

List of Figures

2.1 Images showing difference in various scenes 6

2.2 Sigmoid Activation Function 12

2.3 Tanh Activation Function 13

2.4 ReLU Families Activation Function 14

2.5 Swish Activation Function 16

2.6 Siren Activation Function 17

3.1 CIFAR Databases 20

3.2 COVID-19 RADIOGRAPHY DATABASE 21

3.3 DNN Architecture 23

3.4 CNN Architecture 24

3.5 Alexnet Architecture 25

4.1 DNN Validation Accuracy with Learning Rate 0.001 32

4.2 CNN Validation Accuracy 34

A.1 COVID-19 RADIOGRAPHY Confusion Matrix RMSprop 0.0001 50

A.1 COVID-19 RADIOGRAPHY Confusion Matrix RMSprop 0.0001 51

A.1 COVID-19 RADIOGRAPHY Confusion Matrix RMSprop 0.0001 53

A.2 COVID-19 RADIOGRAPHY Confusion Matrix RMSprop 0.00001 54

A.2 COVID-19 RADIOGRAPHY Confusion Matrix RMSprop 0.00001 55

A.2 COVID-19 RADIOGRAPHY Confusion Matrix RMSprop 0.00001 56

List of Tables

2.1 Activation Function Table 11

3.1 Training Hyper Parameters Four Classes 27

3.2 Training Hyper Parameters Two Classes 27

3.3 Data Augmentation 28

4.1 Confusion Matrix 30

4.2 Confusion Matrix for class Normal 31

4.3 DNN accuracy Table 33

4.4 CNN accuracy Table 35

4.5 Test Accuracy COVID-19 Four Classes 37

4.6 Test Accuracy COVID-19 Binary Classes 38

4.7 Performance COVID-19 Adam Binary Classes 40

4.8 Contrastive Analysis with Recent Studies 40

4.9 Performance COVID-19 RMSprop 0.00001 42

4.10 Performance COVID-19 RMSprop 0.0001 43

4.11 Performance COVID-19 Adam 0.00001 44

4.12 Performance COVID-19 Adam 0.0001 45

List of Abbreviations

AFs Activation Functions

CNNs Convolutional Neural Networks

COVID-19 Coronavirus Disease 2019

CXR Chest X-ray

DNNs Deep Neural Networks

MNIST Modified National Institute of Standards and Technology

RNNs Recurrent Neural Networks

Chapter 1

Introduction

In this thesis, we present a comparison of image classification with different activation functions under balanced and imbalanced datasets. We use the commonly-used and class balanced dataset CIFAR-10 and a stay up-to-date and imbalanced dataset COVID-19 Radiography Database. Chapter 2 gives a brief review of the relevant literature, while chapter 3 introduces the design of experiments in this work; moreover, the assumptions and results had been illustrated in Chapter 4.

1.1 Motivation

"The COVID-19 pandemic, also known as the coronavirus pandemic, is an ongoing global pandemic of coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)" [44]. Since the COVID-19 pandemic in 2019, our human lives and work have been hugely affected. Nevertheless, to fight and prevent this pandemic, artificial intelligence technology has stepped onto the world stage and plays a significant role in various areas. There are many applications of AI in COID-19 disease, including early diagnosis and analysis of cases [22] [7] [2], predictions of mortality [5], and estimate the risks and effects of such an epidemic [27] and so on. The development of vaccines, of course, is a top priority. In this regard, we would like to thank all the frontline medical workers. COVID-19 is a difficult task and challenge for medical workers and many researchers in the AI field.

All people of the world united to overcome the COVID-19. A series of research papers based on computer vision technology has constantly been emerging since this pandemic, and many traditional medical-related enterprises have introduced artificial intelligence talents.

In the field of early detection and diagnosis of coronavirus, medical imaging classification is a hot research direction. The deep learning model builds an intelligent platform for medical practitioners to make fast decisions on infected cases, enabling a timely response to outbreaks. The find-trained models process images in milliseconds, whereas radiologists may need minutes. Besides, due to the variety and uncertainty of coronavirus symptoms, the radiologist's diagnosis requires accumulated experience, but not all radiologists are experts. Therefore, it is significant to develop a high accuracy, high efficiency, and reproducible model. Nowadays, chest X-ray (CXR) and computed tomography (CT) scans are the most widely used medical imaging diagnosis methods. I chose CXR as my dataset because the CXR is more widespread, rapid, cheap, simple, and reliable.

Nowadays, many papers have already achieved noticeable results on CXR diagnosis, including well-known COVID-net [41], COVID-ResNet [6], and COVIDX-Net [14]. It is generally known that medical images are class unbalanced and coarse-grained labels. We can tackle the imbalanced learning problem from three aspects: data level, algorithm level, and hybrid level [19], but the undistinguished features of chest X-rays make it extremely difficult to generate new data for the minority group, and the removal of data from majority groups results in a lack of data for the entire dataset. Consequently, we might as well modify the classification algorithms. Almost all investigations focus on the complexity of network architectures; however, the cumbersome computation of networks will overfit the class with a bigger dataset, which is inevitable.

In this work, the main goal is to find the correlation between the activation functions with the balanced and imbalanced data, especially discern the outperforming activation functions on CXR.

1.2 Research Objectives

There are two research objectives in this study:

1. To assess the performance of neural networks with different activation functions.
2. To find the correlation between the selection of activation functions and the balanced or imbalanced data.
3. To determine the outstanding activation function for unbalanced chest radiography.

Chapter 2

Review of Literature

2.1 Machine Learning and Deep Learning

Artificial intelligence is one of the most popular Computer Science branches and aims to simulate human intelligence by machine. It makes machines perform tasks that a human always performs. However, the difference is that AI does not have emotionality, but human does. This research area was first raised by mathematician Alan Turing, who pointed out that "Can machines think?"in the paper "Computing Machinery and Intelligence." [40].To achieve this proposal, the scientist gives the machine a set of instructions to learn from data instead of hard-coding precise instructions, including every step it needs to do.

Machine learning is a method to realize artificial intelligence through experiential learning. Because of the rapid development of computers in the past 20 years, machine learning has made a breakthrough. It also has made significant achievements in computer vision and natural language processing, ushering in a fresh round of artificial intelligence explosive development. Deep learning is the key to these breakthroughs. As a result, deep learning now pervades all fields, slowly affecting and changing human life, such as self-driving cars, personal assistants (Alexa), and automatic translation.

Compared with traditional machine learning algorithms, which require manual extraction of features, deep learning uses a computational model composed of multiple processing layers to extract features directly from data, reducing the workload of designing feature extractors for each problem.

2.2 Neural Networks

Neural network is a kind of artificial intelligence like machine learning technology that simulates the human brain neural network. The smallest unit in a neural network is a neuron that mimics the neurons of the human brain to perform some complex tasks. Neural network is mainly composed of three layers, the input layer, hidden layer, and output layer. However, the way each neuron is connected plays an essential role in neural networks. Looking back at the development of neural networks, we must start from the invention of the perceptron. The perceptron was created by Frank Rosenblatt in 1958, and he used different weights on connections and in order to address the linear prediction problem[32].

2.2.1 Deep Neural Networks

In recent years, inspired by the perceptron, the multilayer perceptron networks appeared to solve complex nonlinear problems. Neural network is an extension based on perceptron, and Deep Neural Networks(DNNs) can be regarded as a neural network with many hidden layers, sometimes called multilayer perceptron networks. As shown in the figure 2.1, each connection of the neuron has a weight, so the neuron's output is equation 2.2(wi,j are weights, and x1, x2 are the inputs for this neuron, b1, b2 are bias number, and h1, h2, h3 are hidden neurons, o1 is output neuron).

$$h1 = f\left(w1,1 * x1 + w1,4 * x2 + b1\right) \tag{2.1}$$

$$o1 = f\left(w2,1 * h1 + w2,2 * h2 + w2,3 * h3 + b2\right) \tag{2.2}$$

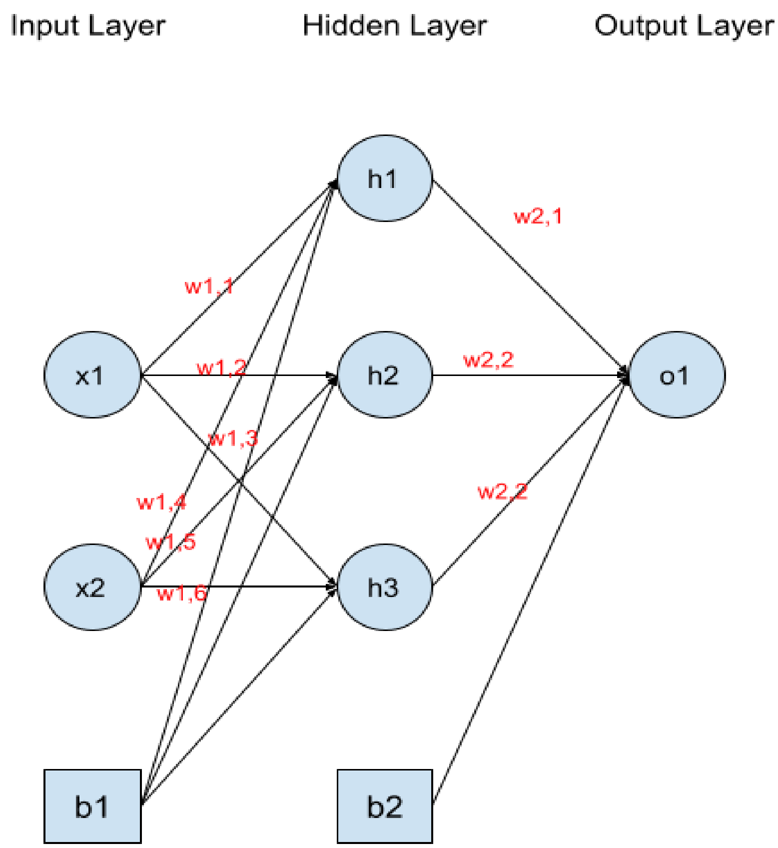


Figure 2.1: Images showing difference in various scenes

The sum of weighted inputs is passed through function f which is a nonlinear function called the activation function. The essence of a neural network is to fit the real functional relationship between features and targets through parameters and activation functions.

The training algorithm of neural networks is to adjust the weight to the best so that the performance of the entire network can reach the best. Base on the standard Artificial Neural Networks, as the name suggests, Deep Neural Networks are going deeper with an increasing number of hidden layers and the number of neurons in every single layer. As long as we provide enough amount of data, the DNNs can mimic the human brain to make many decisions, and predictions [24]. Deep Neural Networks is currently the basis of many AI applications and is used in various disciplines such as text recognition [23], speech recognition[34] and face recognition [38], where DNNs can even surpass human accuracy. However, the superior accuracy of DNNs comes at the cost of high computational complexity, so the rapid growth of it in recent years can be attributed to the rapid development of the Graphics Processing Unit(GPU) field.

2.2.2 Convolutional Neural Networks

Inspired by the human visual nervous system, Convolutional Neural Networks (CNNs) is good at image processing and can be trained end-to-end, including feature extraction, object detection, classification, and prediction. Image contains a tremendous amount of data, while a 256*256 image contains 65536 parameters. First of all, CNNs reduce a considerable number of parameters into a small number of parameters and then process them. Second, this kind of network extracts the features of the image during the process of dimensionality reduction. When the image is flipped, rotated, or changed in position, it can also effectively identify similar images.

A typical CNN is composed of three layers: convolutional layers, pooling layers, and fully connected layers. Convolutional layers capture the spatial and temporal dependencies in an image (extract local features) and reduce its dimension, making it easier to be processed. Simultaneously, convolution layers catch low-level features initially, including edges corners, and then capture higher-level features (shapes and categories) when networks get deeper and deeper. Pooling layers significantly reduce the amount of parameter and computation; it is a form of non-linear down-sampling. It summarizes features extracted from the previous layer and ignores the useless information. Like traditional neural networks, fully connected layers output the desired classification results.

Convolutional neural networks divide the image into regions and process only the pixels within each region, rather than all pixels, during image processing, which can greatly reduce the total number of parameters. Deep Convolutional Neural Networks is a state-of-the-art image processing and speech recognition model and can differentiate images from one to another. It has been proved that CNNs are more potent than traditional deep neural networks, especially on reduced spectral variation in the input signal. Incorporated speaker adaptation named fMLLR into CNNs shows significant improvements in word error rate compares to common DNNs. Taken together, the advantages of image analysis using CNNs constitute a powerful reason for the advancement of speech feature extraction. [34].

2.2.3 Recurrent Neural Networks

The layers are fully connected in the traditional neural network model, but the neurons in every single layer do not share information with each other. Although this kind of neural network already had some success in prediction, it is ineffective for some temporal problems and ordered data. For example, if you want to predict the weather tomorrow, you usually need to know the weather before because the weather can not change entirely independently.

Based on the traditional neural network, the Recurrent Neural Networks (RNNs) adds additional weights between neurons in each hidden layer to realize the internal circulation, which is used to retain information and maintain internal states. In other words, RNNs remember the data from previous states and apply it to the computation of the current output.

As early as the 1990s, Recurrent Neural Networks were used in a variety of fields, including financial forecasting[10], music analysis[36], and stock-trend alerts[33]. In order to address temporal complexity problems, the architecture range of RNNs from fully connected recurrent networks to partially connected networks. The issue of vanishing gradients in standard RNNs, nevertheless, makes them hard to train, which means that it is tough for Recurrent Neural Networks to remember long-term history[18]. Gradient vanishing or gradient explosion refers to the fact that the gradient decreases or increases during calculation and back-propagation, and after a certain period, the gradient will converge to zero (vanishing) or diverge to infinity (explosion). Simply put, the problem with long-term dependency is that as each time interval increases, standard RNNs lose their ability to connect to long-term information. To overcome these problems, researchers have proposed several solutions, such as Echo State Network, Leaky Units. One of the most successful and widely used solutions is the gated RNN, designed by Sepp Hochreiter and Jürgen Schmidhuber in 1997, called Long Short-Term Memory (LSTM) [17]. The reason why LSTM can solve the long-term dependency problem is that it includes three gates to control the flow and loss of features. They are forget gate, input gate, and output gate, respectively. RNNs learn long-term dependence from the dataset; however, not all previous information needs to be remembered. LSTM elaborately creates cell states process gates responsible for removing or adding elements to cell states to decide if the information is pertinent.

Sometimes the prediction may need to be determined by the combination of previous and subsequent inputs to be more accurate. Therefore, Bidirectional Recurrent Neural Networks (BRNNs) are proposed [35]. Bidirectional networks train the model in both forward and backward directions simultaneously, and every cell state is comprehensively determined by past and future. Meanwhile, BRNNs theoretically provide a model computing by posterior probabilities. Most of the time, bidirectional networks work better than unidirectional networks on relation classification and speech translation [39].

2.3 Activation Function

Whether it is the traditional neural network model or the state-of-the-art deep learning algorithm, we can see that activation functions https://www.overleaf.com/project/603d9f13ab2a7637a3749cdc(play a significant role. The so-called activation function is a function that runs on neurons and is responsible for mapping input to output.

In neural networks, each neuron in the hidden layer accepts the neuron's output value in the upper layer as the input value of its own and passes an output value to the next layer. The input layer directly transfers the attribute input value to the hidden layer. In a multilayer neural network, there is a functional relationship between the output of the upper neurons and the input of the lower neurons, which is called the activation functions. In the equation, we used character f to represent an activation function. In the network without AFs, the relationship between the input and output of the hidden layer is linear, so its fitting ability is very limited. Therefore, to model complex real-world problems, the concept of AFs needs to be invoked. Besides the benefits discussed earlier, it also helps networks to limit the output of the neuron to a certain range as required. If the output value is not defined within a specific range, it can become huge, especially in a deep neural network with millions of parameters, leading to excessive computation.

2.3.1 Sigmoid

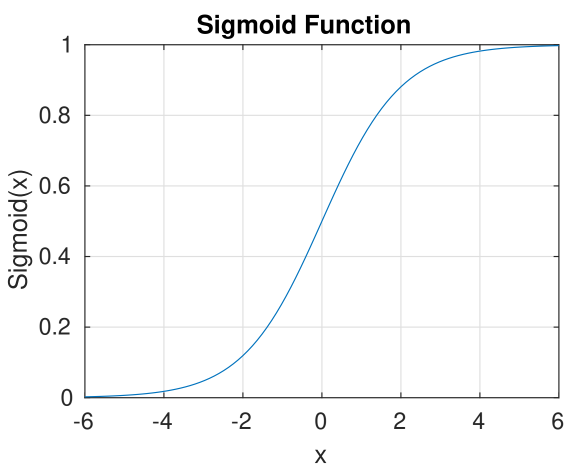
Sigmoid function is the most commonly used nonlinear activation function, which has the exponential shape closest to biological neurons. It compresses a real number in the range of 0 to 1. When the input number is large, the result is close to 1. Otherwise, when input is a large negative number, the result will reach zero. This function gives a good explanation of the extent to which neurons are activated when stimulated: 0 represents no activation at

Name	Activation functions	Derivative function	Equation number
Sigmoid	$f(x) = \frac{1}{1+e^{-x}}$	$y = \frac{e^{-x}}{(1+e^{-x})^2}$	1
ReLU	$f(x) = max\,(0, x)$	f'(x) = $\begin{cases} 0 & \text{for} \quad x < 0 \\ 1 & \text{for} \quad x \geq 0 \end{cases}$	2
LeakyReLU	$f(x) = max\,(0.01x, x)$	$f'(x) = \begin{cases} 0.01 & \text{for} \quad x < 0 \\ 1 & \text{for} \quad x \geq 0 \end{cases}$	3
ELU	$f(x) = \begin{cases} \alpha(exp(x)–1) & \text{for} \quad x \leq 0 \\ x & \text{for} \quad x > 0 \end{cases}$	$f'(x) = \begin{cases} f(x) + \alpha & \text{for} \quad x \leq 0 \\ 1 & \text{for} \quad x > 0 \end{cases}$	4
GELU	$f(x) = 0.5x(1 + tanh(\sqrt{\frac{2}{\pi}}(x + 0.044715x^3)))$	$f'(x) = 0.5tanh(0.0356774x^3 + 0.797885x) +$ $(0.0535161\mathrm{x}^3 + 0.398942x)sech^2(0.0356774x^3 + 0.797885x) + 0.5$	5
PReLU	$f(x) = \begin{cases} a_i x & \text{for} \quad x < 0 \\ x & \text{for} \quad x \geq 0 \end{cases}$	$f'(x) = \begin{cases} a_i & \text{for} \quad x < 0 \\ 1 & \text{for} \quad x \geq 0 \end{cases}$	6
Softplus	$f(x) = ln(1 + e^x)$	$f'(x) = \frac{1}{1+e^{-x}}$	7
Tanh	$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$	$f'(x) = 1 - (\frac{e^x - e^{-x}}{e^x + e^{-x}})^2$	8
Siren	$f(x) = sin(x)$	$f'(x) = cos(x)$	9
Swish	$f(x) = \frac{\mathrm{x}}{1+e^{-x}}$	$\frac{x(1+e^{-x})+(1+e^{-x}-x)}{(1+e^{-x})^2}$	10

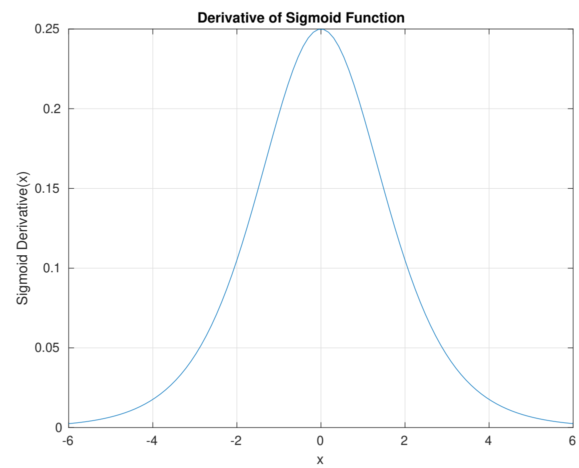
Table 2.1: Activation Function Table

all; for full activation. However, the output of the activation function should be symmetric to zero so that the gradient does not move in a particular direction during gradient descent. This activation function, furthermore, is rarely seen in deep learning in recent years because Sigmoid function is likely to lead to gradient vanishing or exploding problems.

During back-propagation, we need to take the derivative of the S-type activation function. In the figure 2.2, the derivative of Sigmoid is always less than 0.25; if we update weights by multiplying their gradient, these parameters will persistently decrease, causing the gradient vanishing problem.



(a) $y = Sigmoid(x)$

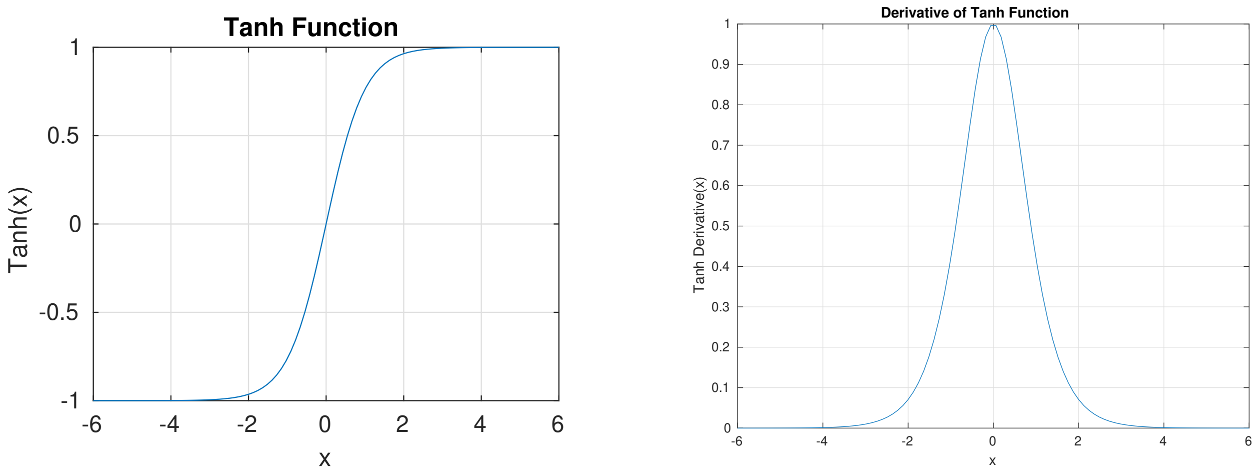


(b) $y = Sigmoid'(x)$

Figure 2.2: Sigmoid Activation Function

2.3.2 Tanh

Tanh is also known as Hyperbolic Tangent Function, is an alternative to the sigmoid function. From the figure 2.3, it is obvious that Hyperbolic Tangent Function shares a similar trend with Sigmoid. However, Tanh is almost always preferable to using sigmoid because Tanh is a zero-centered activation function that helps in centering the data by bringing mean close to 0. Moreover, the derivative of Hyperbolic Tangent Function is larger than Sigmoid, so it converges dramatically faster during gradient descent.



(a) $y = Tanh(x)$ (b) $y = Tanh'(x)$

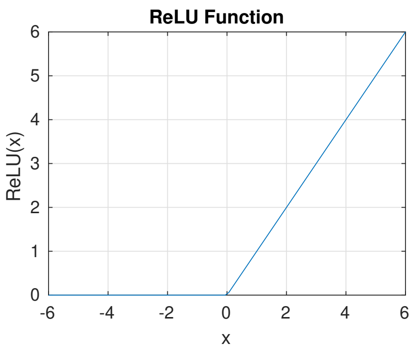
Figure 2.3: Tanh Activation Function

2.3.3 ReLU Family

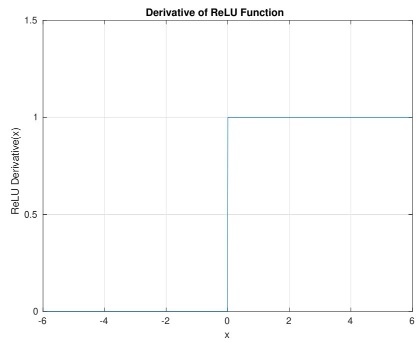
The rectified linear activation function (ReLU) is the most common function used for hidden layers, especially in Convolutional Neural Networks, and introduced by this paper [11]. This function replaces Sigmoid and Tanh and takes the throne as the default activation function in Deep Learning. In a nutshell, ReLU is a piecewise linear function, i.e. equation 2 in table 2.1. In most cases, it can suffer from vanishing gradients problems and often achieves better performance. It is worth noting; however, Malay Haldar mentioned a problem of ReLU in paper [12], when there is abnormal input, the ReLU unit may produce large gradients during the process of back-propagation. These outliers could permanently shut down the ReLU activation function, killing the neurons. Obviously, the reason of Dying ReLU is that when the input is negative, the output of ReLU is always zero.

Leaky Relu (LReLU) [25], Exponential Linear Units (ELU) [4] and Parametric Rectified Linear Unit (PReLU) [13] have been developed to address this problem to a certain extent.

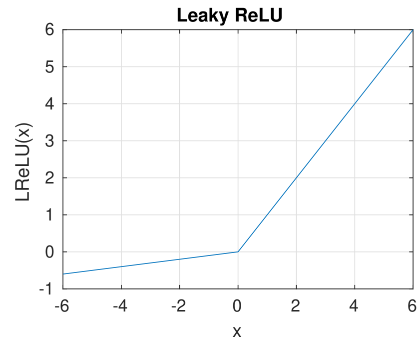
From the figure 2.4, all of them modified the negative part of ReLU function slightly.



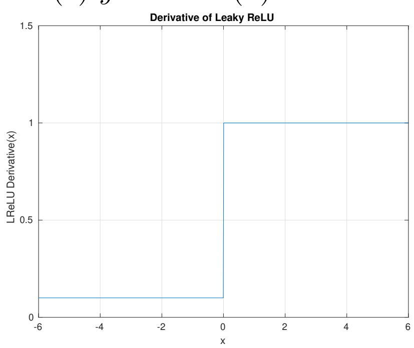
(a) $y = ReLU(x)$



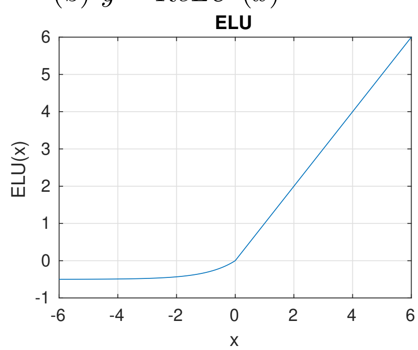
(b) $y = ReLU'(x)$



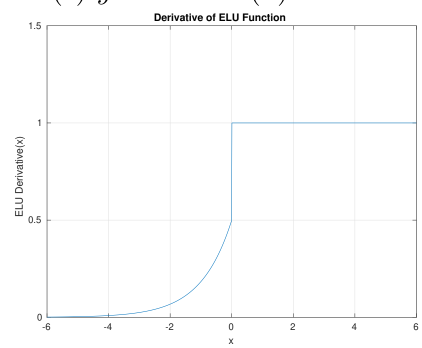
(c) $y = LReLU(x)$



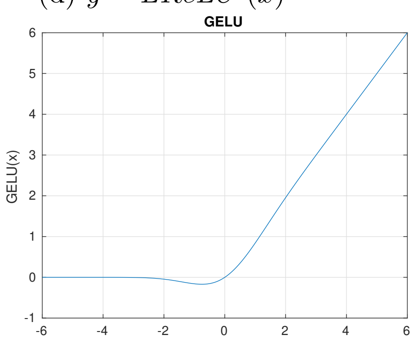
(d) $y = LReLU'(x)$



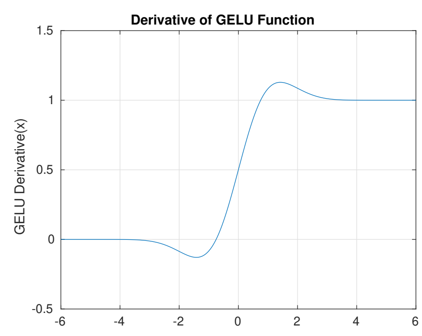
(e) $y = ELU(x)$



(f) $y = ELU'(x)$



(g) $y = GELU(x)$



(h) $y = GELU'(x)$

Figure 2.4: ReLU Families Activation Function

Leaky Relu (LReLU)

As shown in the table 2.1, Leaky Relu (LReLU) modified the equation of ReLU by making the coefficient of leakage.

Exponential Linear Units (ELU)

Nevertheless, ELU combines the idea from sigmoid and ReLU, the positive part same as ReLU, and the negative part same as sigmoid. It refers to the concept of regularization, which pushes the mean value of the function to zero so that the neural network has fewer iterations, faster learning speed, and higher accuracy.

Parametric Rectified Linear Unit (PReLU)

As shown in the table 2.1, the expression of PReLU is consistent with that of LReLU. However, the parameter α is fixed in the former and learnable in the latter. In other words, the slope value learns through backward propagation, which is flexible and variable.

Gaussian Error Linerar Units(GELU)

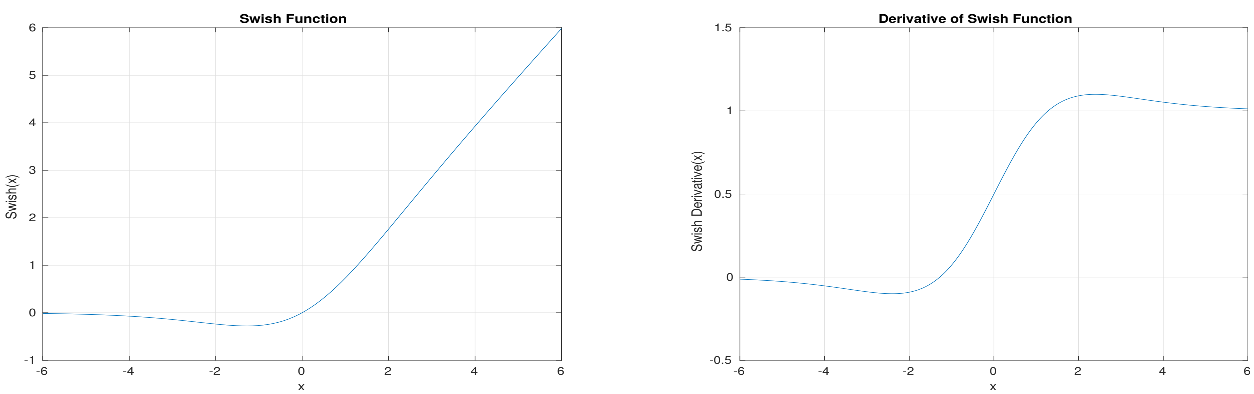
GELU is a combination of dropout, zone-out, and rectifying neurons, and it has been proved that it surpasses the accuracy of the rest of the version of the ReLU family [15]. According to equation 5 in table 2.1, When the input x decreases, the input will have a higher probability of being dropout so that the activation transformation will be randomly dependent on the input. The plot of it shown in the figure 2.4g. It is a probabilistic representation of neuron input, which is more in line with the natural way of nerve activation intuitively.

Softplus

Softplus is a much smoother version of ReLU activation function, which might give us unexpected results when the outputs of neurons need to be smoother and more continuous. It is worth trying to use novel AFs in this study.

2.3.4 Swish

The advent of ReLU function and its variants is a breakthrough in the history of deep learning, but the invention of Swish by Prajit Ramachandran in 2017, has enabled a state-of-the-art performance of image recognition and machine translation on [30]. Swish and ReLU share a similar trend, and the equation of Swish is equation 10 in table 2.1. It has a series of advantages, including unsaturated, smooth, and non-monotonic properties. Figure 2.5 shows that the output decreases first, then increases, which means the swish activation function does not have continually positive or negative derivatives, avoids Dying ReLU problem when gradient nears to zero. Meanwhile, It does not change suddenly at a certain point, It does not change suddenly at a certain point, making it easier to converge during training.

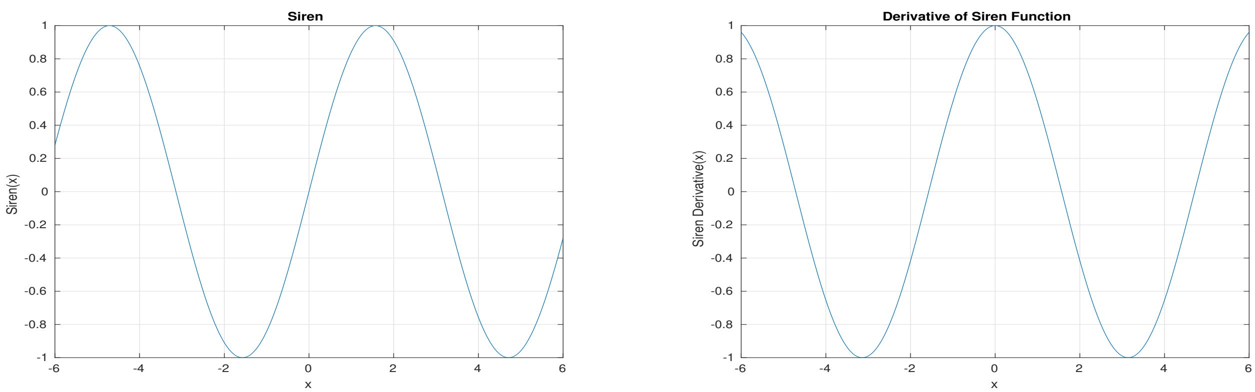


(a) $y = Swish(x)$ (b) $y = Swish'(x)$

Figure 2.5: Swish Activation Function

2.3.5 Siren

Stanford University researchers introduce sinusoidal representation networks (SIRENs) as a method to represent signals in a paper in 2020 [37]. Since the ReLU network is piecewise linear and its second derivative is zero everywhere, it is impossible to describe the higher derivative of the natural signal. However, the sinusoid function has a continuous derivative. It is obvious from the figure 2.6 that the cosine curve is the shifted form of the sine curve. Meanwhile, The first derivative and the second derivative are both shifted sinusoids. The researchers evaluate the performance of SIRENs against other classical algorithms; the SIRENs not only converge rapidly during the training but also show more details. The performance of signal representation has been enhanced significantly, so I implement the siren activation function on traditional artificial intelligence problems.



(a) $y = Siren(x)$ (b) $y = Siren'(x)$

Figure 2.6: Siren Activation Function

2.4 Related Work

Admittedly, many published papers demonstrate comparative analyses of trend, performance, and efficiency of activation function applications in all aspects. Nevertheless, few precedents investigate research on activation function in imbalanced datasets, which got me interested.

From the literature of state-of-the-art image classification architectures, most of them still using ReLU as the AFs. To enhance the generalization ability of classifer, researches refers to dropout, regularization, normalization. Despite the fact that these neural networks have shown colossal achievement in computer vision success, we can still improve their performance by replacing the AFs. In paper [26], the author compared trends in recent AFs applications and pointed out that it is worth observing if there would be improved performance results on outperforming architectures with state-of-the-art functions. Tomasz Szanda evaluated the time cost and accuracy of 11 commonly used Functions in image classification and concluded that AFs depend on the dataset and the position in networks, but no ultimate decision for choicest AFs [1]. Both of them [31] [48] shows that the alternatives of ReLU function play a crucial role in helping to optimize the models further.

Chapter 3

Experimental Implementation

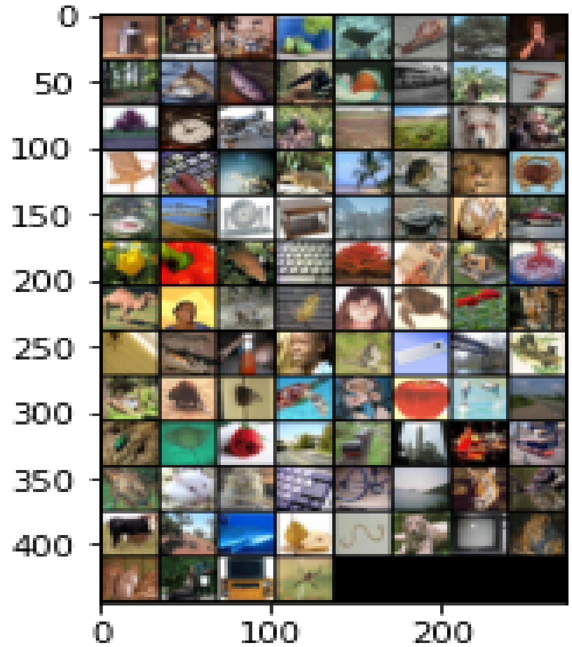
3.1 Data Sets

In this article, we chose two datasets. The first is the classic CIFAR-10 dataset [42], the number of samples is perfectly balanced between each class, and every category has 6000 images. The second is COVID-19 RADIOGRAPHY DATABASE [28], which has a severe class imbalance problem.

3.1.1 CIFAR-10

Unlike traditional gray color MNIST Database(Modified National Institute of Standards and Technology Database) [45] and Fashion-MNIST dataset[46], the CIFAR-10 dataset in figure 3.1 consists of 60,000 color images spread over ten classes [42]. CIFAR (Canadian Institute For Advanced Research) collected and organized CIFAR-10 and CIFAR-100 for complexity machine learning problems and a benchmark for state-of-the-art image recognition and classification algorithms. As their name, CIFAR-100 consists of 60,000 color images spread over hundred classes. Recently, CIFAR-10 has become one of the most traditional image classification datasets and has also significantly contributed to deep learning.

(a) CIFAR-10

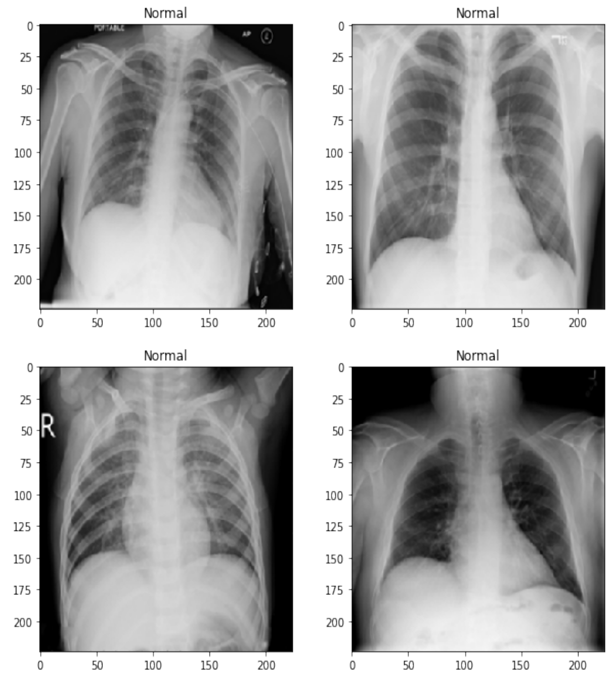


(b) CIFAR-100

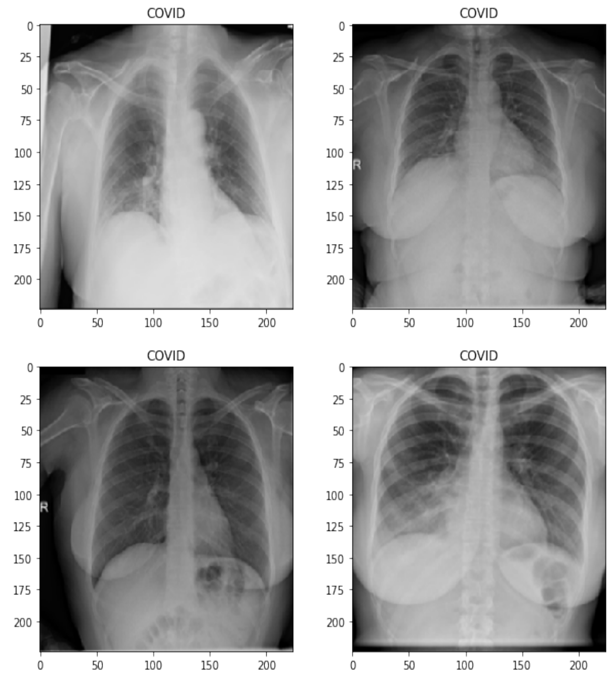
Figure 3.1: CIFAR Databases

3.1.2 COVID-19 RADIOGRAPHY DATABASE

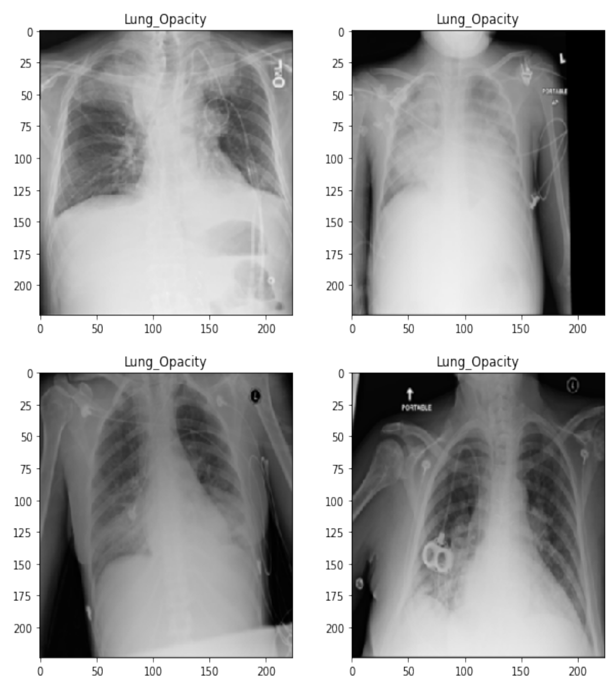
Coronavirus disease 2019 (COVID-19) is a contagious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [43]. On December 31, 2019, the World Health Organization (WHO) was informed of a case of pneumonia caused by an unknown microorganism in Wuhan, Hubei Province, China. WHO subsequently announced that they had detected a novel coronavirus in samples from this group of patients. Then the epidemic escalated and quickly spread around the world. With the rapid increase in the number of cases, there is an increasing need for AI techniques to help doctors, such as the detection and classification of chest X-rays, the analysis of CT scans, and the prediction and analysis of the vaccination progression. Especially in countries with large populations, such as China and India, the ratio of doctors to patients is far from sufficient. Therefore, medical researchers quickly established a real-time update COVID-19 database, enabling more researchers to



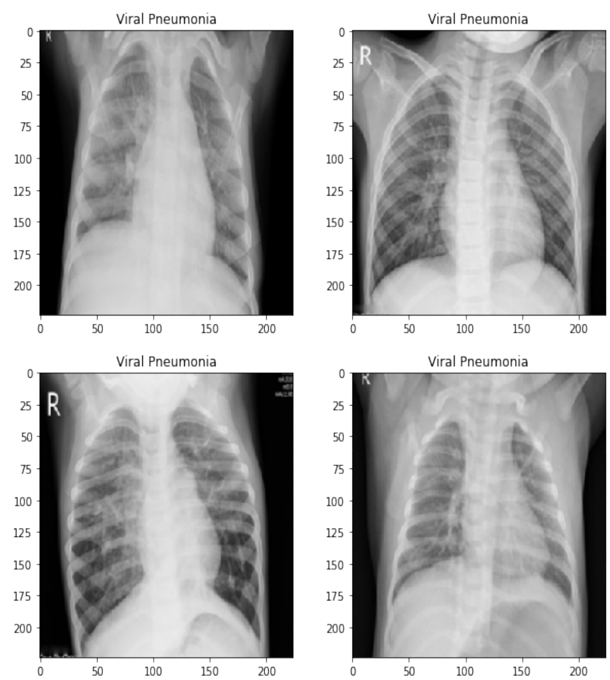
(a) Normal



(b) COVID



(c) Lung Opacity



(d) Viral Pneumonia

Figure 3.2: COVID-19 RADIOGRAPHY DATABASE

take part in this challenge.

The latest COVID-19 Radiography Database (Kaggle) consists of chest X-rays of 3616 COVID-19 positive images, 10192 normal images, 6012 Lung Opacity images and 1345 viral pneumonia images[28][29][3]. This dataset got the winner of COVID-19 dataset award by the Kaggle community, and it only has 261 COVID-19 chest X-rays in the first release but updated to 3616 after the second update.

Because of the rapid spread of COVID-19, clinicians worldwide face a challenge relying on laboratory tests to confirm COVID-19 cases is time-consuming and may delay diagnosis and treatment. The diagnosis by X-ray must be made with the help of a specialist doctor. As you can see in the figure 3.2 below, it is impossible for someone who is not in a medical major to tell the difference in X-ray images between the Normal and COVID-19 patients. Therefore, using deep learning to train an artificial intelligence classifier can dramatically speed up diagnosis and reduce the workload of specialist doctors.

3.2 Neural Networks Setup

In the section 2.2, we provide a brief description about neural networks. There are many variants of Neural Networks, such as Back Propagation Neural Networks, Convolutional Neural Networks(CNNs), and Long Short-term Memory Networks (LSTM). Three sets of experiments are designed on different neural network structures with varying datasets to analyze the performance of different activation functions.

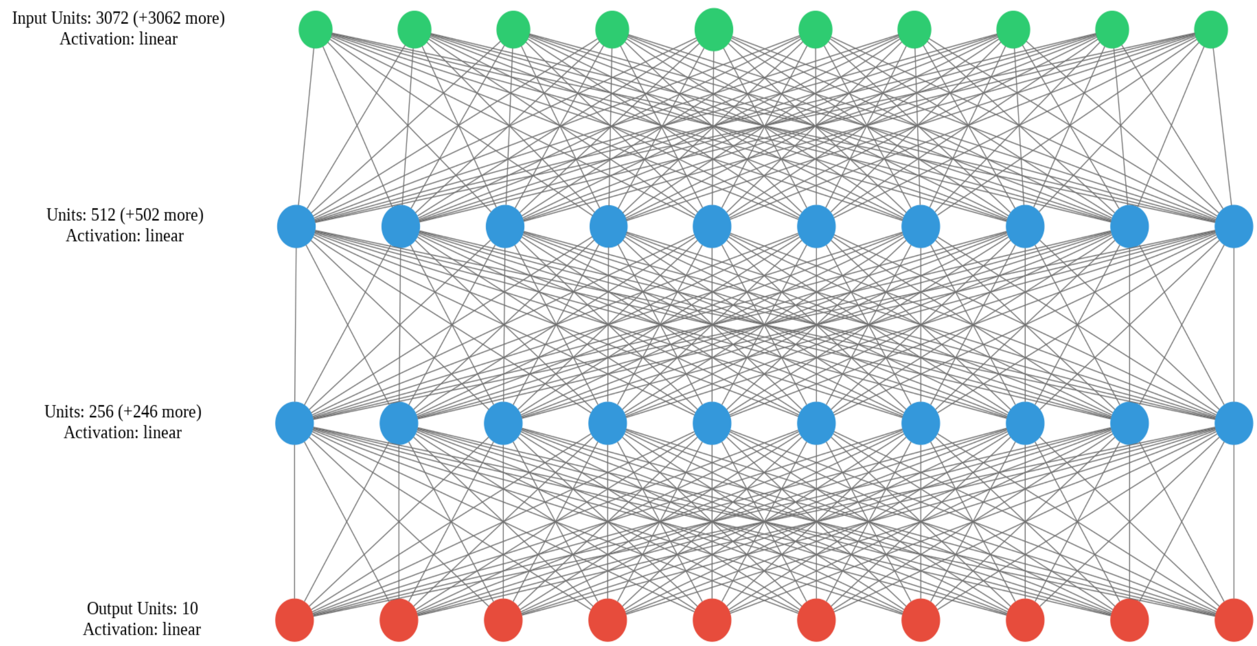


Figure 3.3: DNN Architecture

3.2.1 CIFAR-10 DNN Setup

However, only by understanding the original of the classics can we better understand the more powerful modern variant. The classic neural network, which is the most fundamental neural network, is called multilayer perceptron networks (MLP). In order to better evaluate the effect of different activation functions on image classification, we adapted a fixed number of layers DNN model to get a benchmark for the proposed solution on the CIFAR-10 dataset.This step aims to get a basic understanding of the effects of different activation functions, not to achieve the highest accuracy. We have shown the architecture of the network in the figure 3.3; it had one input layer with 3072 input units (CIFAR-10 has 32 by 32 images with RGB three-channel), two hidden layers of 512 units and 256 units, respectively, and one output layer with ten output units (CIFAR-10 has ten classes). And it used Adam optimizer with learning rates 0.001, 0.0001, and 0.00001 for 50 epochs in the training step.

3.2.2 CIFAR-10 CNN setup

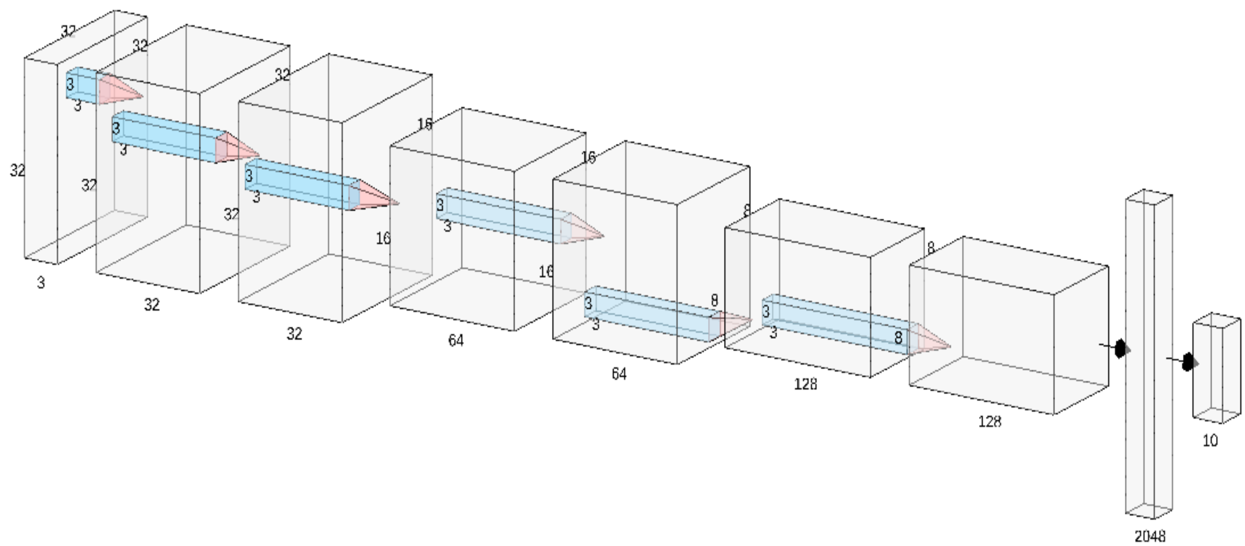


Figure 3.4: CNN Architecture

According to the 2.2.2, convolutional neural networks extract local features from images with fewer parameters than MLP. We adapted a seven-layered convolution neural network followed by a flatten layer as a benchmark for CIFAR-10 classification, created by Abhijeet Kumar [21]. The architecture of our network is summarized in Figure 3.4; a maximum pooling performed after every two convolutional layers to reduce the dimension of data. Like the DNN in the previous section, we use the same number of 50 epochs and Adam optimizer with learning rates 0.001, 0.0001, 0.00001 to compare performance. Our loss function of choice is cross-entropy.

We discarded the last two dropout layers from the original model. Dropout refers to the random deletion of units from the network at a specific dropout rate during the training step. It is a regularization strategy used to tackle overfitting problems and prevent complex co-adaptations on the training data [16]. It has been proved that dropout consistently helps networks achieve lower error and better performance than the networks without dropout in [16]. Nevertheless, it is not popular in convolutional neural networks. Christian Garbin and Xingquan Zhu compared the model training speed and prediction accuracy of networks with or without dropout layers and concluded that adding dropout reduced accuracy significantly [8], so we need to be careful about adding the dropout layer to networks.

3.2.3 COVID-19 CNN setup

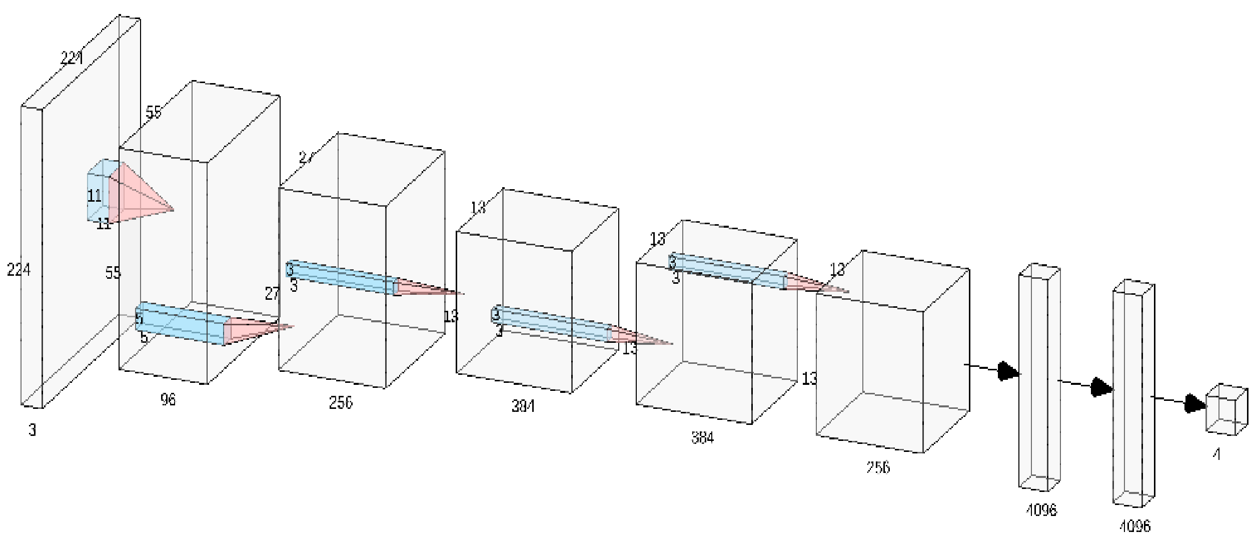


Figure 3.5: Alexnet Architecture

The application of deep learning in medical imaging analysis is a milestone in the development of medicine. Although artificial intelligence cannot wholly replace doctors' diagnoses, it can be used as an auxiliary diagnosis system to relieve the heavy workload of doctors. Unlike the typical image datasets, the grayscale values in the COVID-19 radiography images are almost identical, and it is challenging for even a professional radiologist to tell the difference between pneumonia and coronavirus. Because the tiny features on radiography are indistinguishable by the naked eye.

The application of deep learning in coronavirus diagnosis has become a hot area of research in the past year. Thousands of research groups are working around the clock to improve the accuracy of neural networks. Almost all researchers focus on the complexity of neural network models[2], data augmentation approaches, and the attention theory in image classification and segmentation [9]. Despite the fact that these approaches have achieved significant success in the end-to-end diagnosis on medical imaging, specially COVID-Net created by Linda Wang, Alexander Wong from the University of Waterloo [41], It still would be desirable to compare the activation functions further and, perhaps, achieve the same or even better results using simpler neural networks.

In the field of image classification, AlexNet can be said to be the cornerstone of network structure. Deep learning has been silent for a long time before this. Since AlexNet was born and won the ImageNet Competition in 2012, CNN has become deeper and deeper, more and more complex [20].

In this study, the AlexNet is our proposed baseline CNN, and the architecture of it is presented in figure 3.5. It has a total of 8 layers, the first 5 layers are convolutional layer, the back 3 layers are fully connection layer, the last fully connected layer output is passed to the N path Softmax layer, corresponding to the distribution of the number of class labels.

We used the pre-trained VGG model to prevent occasionality, adding three more trainable dense layers as a controlled experiment. Due to limited computational power, the VGG model parameters are pre-trained and set non-trainable in this work, and also because the activation function in the pre-trained Keras model is ReLU. According to the univariance principle, we only train the activation functions at the outermost three fully connected layers during the backpropagation. Moreover, considering the complexity of the model, VGG has more than 14714688 parameters, which requires much time for training and high demand on GPU performance.

Training Hyper Parameters Four Classes

This experiment is built based on the following experimental parameters in the table 3.2.

	Input Size	Optimizer	Learning Rate	Batch Size	Epochs
Parameter	224*224*3	Adam	0.0001/0.00001	32	30
	224*224*3	RMSprop	0.0001/0.00001	32	30

Table 3.1: Training Hyper Parameters Four Classes

Training Hyper Parameters Two Classes

Many papers have yielded notable results in CXR diagnosis, including well-known COVID-NET [41], COVID-RESNET[6], and COVID-NET[14]. Most of them were tested on binary classification or 3 class classification problems. Therefore, to make a comprehensive comparison between different activation functions, I also test them on binary classification and compare the performance with the previous research. This experiment is built based on the following experimental parameters in the table 3.2.

	Input Size	Optimizer	Learning Rate	Batch Size	Epochs
Parameter	224*224*3	Adam	0.00001	32	30
	224*224*3	RMSprop	0.00001	32	30

Table 3.2: Training Hyper Parameters Two Classes

Data Augmentation

Deep neural networks perform well in many tasks, but these networks typically require large amounts of data to avoid over-fitting. In most cases, however, it can be challenging to obtain adequate data, such as medical image analysis. Data augmentation makes minor changes to an existing dataset to increase the amount of data, including flips, transformations, rotations. Our experiment with the following augmentation types in table 3.3

	Rescale	Shear Range	Zoom Range	Horizontal Flip	Validation Split
Parameters	1 / 255	0.2	0.2	True	0.25

Table 3.3: Data Augmentation

Chapter 4

Results and Analysis

4.1 Evaluation Criteria

Evaluating machine learning algorithms is a cornerstone of any project. Therefore, to assess the performance of the proposed deep learning classifier under different activation functions, we first trained different network models on the classical and balanced data sets CIFAR-10 and compared the accuracy of the models and the time cost when the models reached the best accuracy.

To test whether the selection of activation functions has a distinguished effect on the unbalanced and low contrast grayscale dataset, we used the most basic image classification model, AlexNet, as a benchmark.

As the name implies, image classification is a classification problem; The goal is to classify different images into different categories to achieve the minimum classification error. In machine learning, the confusion matrix 4.1 presents a method of performance visualization. We plot the confusion matrix of COVID-19 experiments for every activation function. It is a table with two rows and two columns, comprising false positives, false negatives, true positives and true negatives. For a better and more in-depth comparison, our experiment on the COVID-19 dataset used the following four key indicators, including accuracy 4.1, precision 4.2, recall 4.3, F1-score 4.4 and specificity 4.5, which could be calculated using the components of the confusion matrix.

<table>
<tr><td colspan="2" rowspan="2"></td><td colspan="2">Actual Class</td></tr>
<tr><td>P</td><td>N</td></tr>
<tr><td rowspan="2">Predicted Class</td><td>P</td><td>TP</td><td>FP</td></tr>
<tr><td>N</td><td>FN</td><td>TN</td></tr>
</table>

Table 4.1: Confusion Matrix

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{4.1}$$

$$Precision = \frac{TP}{TP + FP} \tag{4.2}$$

$$Recall(Sensitivity) = \frac{TP}{TP + FN} \tag{4.3}$$

$$F1 - Score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{4.4}$$

$$Specificity = \frac{TN}{TN + FP} \tag{4.5}$$

Where, TP, FP, TN and FN represent True Positive, False Positive, True Negative and False Negative, respectively.

However, our experiments on COVID-19 RADIOGRAPHY DATABASE is a multiclass classification problem. The table 4.1 and equations 4.1 to 4.4 is focus on the binary classification. Same as the binary classification confusion matrix, the addition of matrix row data is the number of actual classes, the addition of column data is the number of predicted classes.

The correctly classified samples, which are those whose predicted labels are consistent with the actual labels, are arranged diagonally from the upper left corner of the table to the lower right corner. Therefore, we could extend the standard confusion matrix for the multiclass problem, and table 4.2 is for the Normal patient class.

		Actual Class			
		Normal	COVID	Lung Opacity	Viral Pneumonia
Predicted Class	Normal	TP	FP	FP	FP
	COVID	FN	TN	TN	TN
	Lung Opacity	FN	TN	TN	TN
	Viral Pneumonia	FN	TN	TN	TN

Table 4.2: Confusion Matrix for class Normal

4.2 Experimental results

This section is broken down into three parts. The first and second part records the performance of the base DNN and CNN model for the CIFAR-10 database under different activation functions. The last part shows the AlexNet multi-level classification results for chest X-ray images.

4.2.1 CIFAR-10 DNN Experimental results

In the table 4.3, we show the best accuracy of the model under different training parameters and the number of epochs that reached the best accuracy.

As can be seen from the table, it is obvious that the activation function has a significant influence on the accuracy of training data. However, under different learning rates, the performance of the activation function tends to be stable. In other words, no matter how the learning rate changes, the power of an excellent activation function is still outstanding.

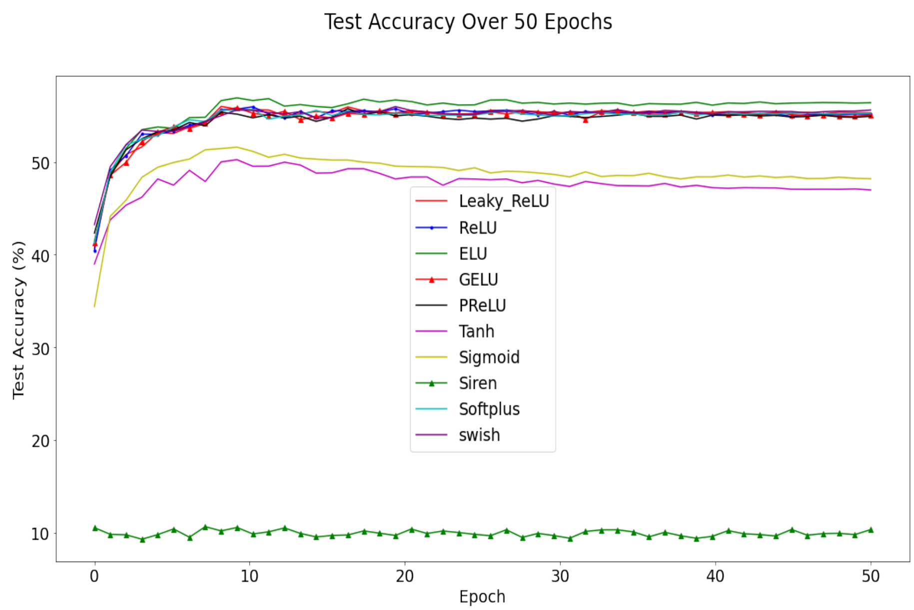


Figure 4.1: DNN Validation Accuracy with Learning Rate 0.001

Among all tested classifiers, except for the classifier using Siren, the accuracy of the model using ELU, PReLU, and Softplus are the highest, which are 56.95%, 56.24%, and 56.19%, respectively, while the performance of the model using Sigmoid is the worst. The version of the activation function SIREN, however, is quite extraordinary. During the training process under a small learning rate (0.00001), the accuracy of other classifiers improves slowly, while Siren can achieve remarkable accuracy expeditiously. On the other hand, in the training process with a higher learning rate, no matter the number of training iterations how to increase, the correctness of the Siren model remains at about 10% and never converges.

Because of the shortcoming of the DNN model structure, it is difficult to process complex spatial information, leading to overfitting. Indeed, according to the figure 4.1, we can observe that the validation accuracy curve for all test cases under a learning rate of 0.001 increases rapidly from Epoch 0 to Epoch 10, then stabilizes or drops off slightly. The accuracy curve of the ELU model is outstanding in the figure above.

Activation Functions	Learning Rate					
	0.001		0.0001		0.00001	
	Accuracies	Convergence Epochs	Accuracies	Convergence Epochs	Accuracies	Convergence Epochs
Sigmoid	52.02	10/50	48.80	48/50	37.24	46/50
ReLU	55.96	11/50	55.71	32/50	48.80	47/50
LeakyReLU	56.10	10/50	55.99	41/50	48.72	37/50
ELU	**56.95**	10/50	55.63	36/50	47.06	50/50
GELU	56.13	9/50	**56.13**	27/50	48.77	49/50
PReLU	**56.24**	10/50	55.69	25/50	48.82	47/50
Softplus	**56.19**	9/50	54.30	27/50	45.81	47/50
Tanh	50.53	11/50	51.84	28/50	43.96	50/50
Siren	10.63	8/50	19.40	31/50	**54.81**	9/50
Swish	55.99	10/50	**56.02**	33/50	47.64	49/50

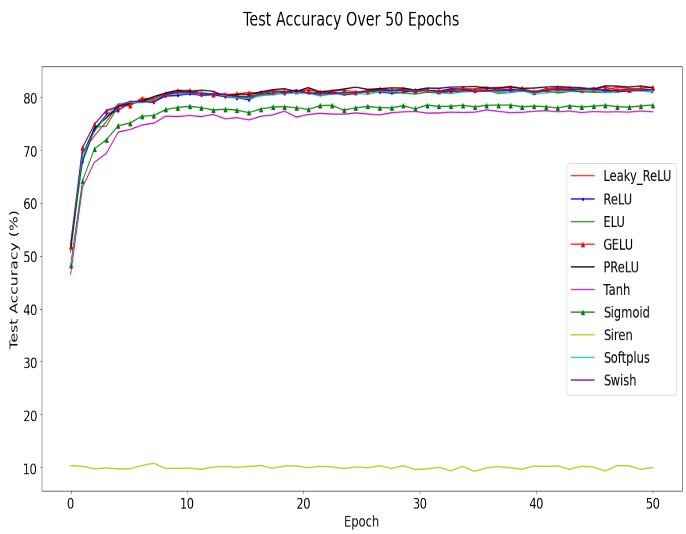
Table 4.3: DNN accuracy Table

4.2.2 CIFAR-10 CNN Experimental results

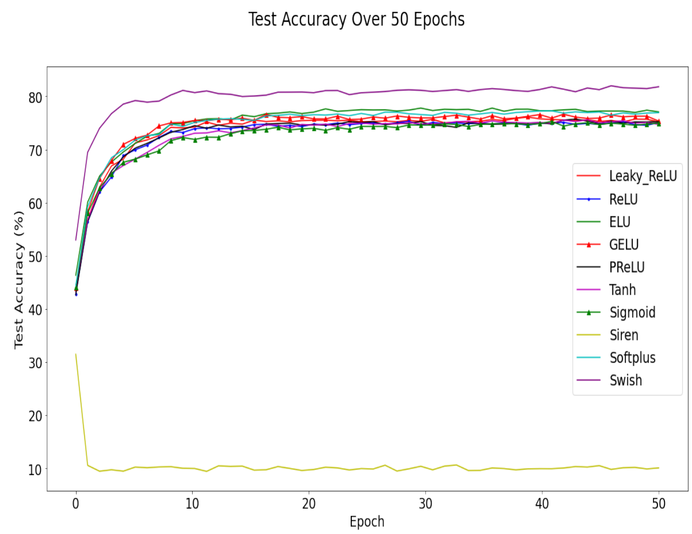
The comparison of CNN with different activation functions is demonstrated in the table 4.4; it is observed that the CNN model with GELU and PReLU outperforms the other comparative cases when the learning rate is 0.001, and its accuracy achieved 81.97% and 82.28%. PReLU, it should be noted, reached the best performance faster than GELU. Equally unsurprisingly, the sigmoid in the rank of the worst performers of the experiments and Siren only show its power when the learning rate is 0.00001. Nevertheless, different from the inference of the DNN experiment, the CNN model with Swish activation function is not affected by the learning rate, and its accuracy is always around 82%.

In the figure 4.2, A rapid increase of accuracy curve can be observed from epoch 0 to epoch 5 for various learning rates and different choices of activation functions. Moreover, the stability of the swish activation function Once again proving in these plots of training accuracy.

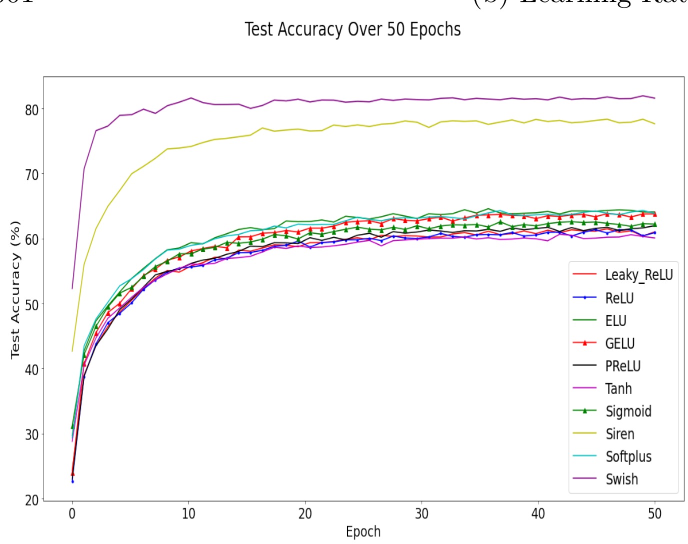
From the foregoing, we can draw the conclusion that when the learning rate range of the training model is uncertain, the Swish activation function should be selected, and the PReLU



(a) Learning Rate 0.001



(b) Learning Rate 0.0001



(c) Learning Rate 0.00001

Figure 4.2: CNN Validation Accuracy

Activation Functions	Learning Rate					
	0.001		0.0001		0.00001	
	Accuracies	Convergence Epochs	Accuracies	Convergence Epochs	Accuracies	Convergence Epochs
Sigmoid	78.58	27/50	75.28	41/50	62.84	42/50
ReLU	81.74	47/50	75.31	29/50	61.52	46/50
LeakyReLU	81.86	36/50	76.10	39/50	61.74	47/50
ELU	81.55	49/50	77.80	30/50	64.90	45/50
GELU	**81.97**	49/50	76.69	33/50	64.04	50/50
PReLU	**82.28**	43/50	75.71	43/50	62.05	50/50
Softplus	81.29	29/50	77.34	43/50	64.31	49/50
Tanh	77.68	40/50	75.59	44/50	60.69	46/50
Siren	10.79	8/50	31.44	1/50	**78.36**	28/50
Swish	81.67	48/50	**81.99**	46/50	**81.95**	49/50

Table 4.4: CNN accuracy Table

activation function is the best choice to obtain the best performance in a short time. Interestingly, the ReLU has been long used for activation function in neural networks but show nothing advantage in this experiment, no matter in terms of speed, stability, or accuracy.

4.2.3 COVID-19 CNN Experimental results

COVID-19 Radiography Database is unbalanced, so we intend to compare the performance of the activation function in various ways, including accuracy, precision, recall, and F1-score, and confusion matrix.

Accuracy

The following table gives the best performance of the base model and pre-trained VGG model along with the four combinations of different hyper-parameters on chest X-ray dataset. In order to evaluate the performance of the model with ten various activation functions, we modified the optimizer and learning rate.

The observation of the table 4.5 demonstrates that it is tricky to find out the relationship between classification accuracy and activation functions under different optimizers and learning rates because of class imbalance. Nevertheless, for base model, Softplus accomplished objective results(91.13%) under all experimental conditions, while Swish achieved the best accuracy of 90.94% (RMSprop) and 90.09%(Adam) at the learning rate is 0.00001. Interestingly, all members of the ReLU family were able to achieve more than 90% accuracy as long as the learning rate or the optimizer were adjusted during the training, but their results were not consistent. Meanwhile, as we expected, the Sigmoid, Siren, and Tanh were not as good as the rest of the activation functions.

Similarly, compared to the results of Alexnet, the pre-trained VGG improved the classification accuracy under adam optimizer, but the softplus function still expressed its outstanding working ability and reached the best performance 91.55% in this study on CXR images. Simultaneity, except for Sigmoid and Siren functions, the accuracy achieved above 90% in all cases, and PReLU and Swish also stand out in the transfer learning model with 91.32% and 91.41% accuracy. The use of Adam optimizer resulted in the highest possible accuracy with the pre-trained model, but not with the Alexnet.

Generally, in transfer learning [47], the researchers combine transferring features, and fine-tuning will generalize surprisingly better results than the model without fine-tuning, even better than the fully trained model on the original dataset. We froze the weights of the pre-trained model in this work and then only fine-tuned the outer fully connected layers. Even so, it still improved the accuracy.

Multiple studies have shown that neural networks, especially convolutional neural networks, can accurately detect the presence of COVID-19 from CXR. However, due to the lack of a COVID-19 public dataset, model training and test results in previous studies have been affected by biases. But in March of this year, a team of researchers from the University of

Qatar in Doha and elsewhere, working with doctors, updated a database of 21,165 chest X-ray images of COVID-19-positive cases and radiography of Normal, Opaque Lung and viral pneumonia. Therefore, this study used this dataset and made a comparative analysis with the previous works. As aforementioned, Since most published research focused on binary classification, we set experiments only with COVID-19 and Normal categories.

The table 4.6 illustrates the performance comparison of different activation functions in COVID-19 RADIOGRAPHY. Interestingly, Gelu and Swish won out in this experiment, achieving 98.12% and 98.70% accuracy, respectively. The particular reason for this is that Gelu is similar to the Swish function; the only difference between them is that a constant factor scales GELU's input. Likewise, the most widely used activation function, ReLU, did not show its power in this case. The best scores of training and validation accuracy were achieved for Swish and GeLU activation function, and the worst case is resulted by the Sigmoid function.

	Alexnet				**VGG(pre-trained)**			
	Adam		**Rmsprop**		**Adam**		**Rmsprop**	
Learning Rate	0.0001	0.00001	0.0001	0.00001	0.0001	0.00001	0.0001	0.00001
Sigmoid	0.8995	0.8872	0.8768	0.8697	-	0.8886	-	0.8858
ReLU	**0.9061**	0.8815	0.8848	0.8679	-	0.9014	-	0.9056
LeakyReLU	0.8797	0.8924	**0.9089**	0.8990	-	0.9075	-	**0.9094**
ELU	0.8759	0.8839	0.8768	**0.9033**	-	0.9099	-	0.9066
GELU	0.8749	0.8977	0.8957	**0.9028**	-	0.9122	-	**0.9094**
PReLU	0.8726	**0.9037**	0.8999	0.8943	-	**0.9132**	-	0.8996
Softplus	**0.9051**	**0.9084**	**0.9032**	**0.9113**	-	**0.9155**	-	**0.9108**
Tanh	0.8754	0.8855	0.8655	0.8858	-	0.9009	-	0.9014
Siren	0.8433	0.8961	0.7697	0.8664	-	0.8924	-	0.8895
Swish	0.8646	**0.9009**	0.8849	**0.9094**	-	**0.9141**	-	0.8990

Table 4.5: Test Accuracy COVID-19 Four Classes

Network	Optimizer	Learning Rate	Activation Functions									
			Sigmoid	ReLU	Leaky ReLU	ELU	GELU	PReLU	Softplus	Tanh	Siren	Swish
Alexnet	**Adam**	0.00001	0.9588	0.9761	0.9725	0.9790	**0.9812**	0.9703	0.9667	0.9696	0.9609	**0.9870**
	Rmsprop	0.00001	0.9711	0.9653	**0.9747**	0.9616	**0.9783**	0.9667	0.9703	0.9725	**0.9747**	**0.9747**

Table 4.6: Test Accuracy COVID-19 Binary Classes

Precision, Recall, F1-Score and Specificity

Precision is the percentage of the predicted positive samples that are truly positive, reflecting the correct prediction rate of a category. The precision of the Lung Opacity category has poor performance, while viral pneumonia and COVID are able to hit the 100% in some cases. In general, those with smaller samples had the worst class precision, but it is puzzling that the number of viral pneumonia was only a quarter of Lung Opacity.

The recall is the ability of the model to recognize the actual positive, that is, how many the actual positives the model predicts it as positive. The F1-score is a balance between recall and precision. Specificity indicates that the model can correctly identify patients without the disease. In COVID-19 diagnosis, it refers to the percentage of non-COVID-19 subjects that are correctly classified as uninfected cases.

However, for the multi-class classification problem; we need to pay attention to different evaluation parameters for each category. For the classification of Normal, we should try to make it as accurate as possible; that is to say, we want the healthy patients predicted by the model to be indeed healthy. As long as there is even a slight risk of the infection, we want to capture and diagnose it. Unlike the Normal class, the value of recall is more crucial for Covid, Lung Opacity, and Viral Pneumonia. In other words, we are able to capture all the actual positives and predict them as positives.

Four Classes Accuracy, however, is generally used to evaluate the global classification ability, which cannot comprehensively assess the performance of a model. It is essential to note that the tables 4.9, 4.10, 4.11, 4.12, demonstrate the detailed comparison of precision, recall, F1-score, and weight average accuracy for four-class classification and The Softplus always performs well in weight average accuracy, but not in precision, recall, F1-score.

To my surprise, with the adam optimizer and a learning rate of 0.0001, tanh function got the results we was hoping for, even though its accuracy was unimpressive. Its precision of Normal is 95%, and the recall for the rest categories are 91%, 94%, 92%, respectively.

For classifiers that only diagnose COVID-19 X-rays, ReLU and Softplus have their distinguished advantages. ReLU achieved 93% precision for Normal and 95% for COVID, while softplus obtained 93%, 94%.

Binary Classes As aforementioned, since the freshest published studies focused on binary classification, we conducted experiments on COVID-19 and Normal to analyze the activation function comprehensively, and the results are included in the table. In this work, we present only one set of evaluation parameters with the best accuracy, as shown in the table 4.7.

Unsurprisingly, the SoftPlus and Swish function were the winners, achieving 99% precision (sensitivity); This positively suggests that if the model predicts a normal patient, then there is a 99 percent chance that the actual label is normal, significantly reduces the probability of misdiagnosis.

Contrastive Analysis with Recent Studies

Due to the limitation in GPU capabilities, we did not test on the fine-tuned transfer learning models and more complex learning models(VGG, DenseNet). Nevertheless, according to the

Network		Alexnet							
Optimizer		Adam							
Evaluation parameters		Precision		Recall		F1-Score		Specificity	
Patient Status		COVID-19	Normal	COVID-19	Normal	COVID-19	Normal	COVID-19	Normal
Activation Functions	Sigmoid	0.90	0.98	0.95	0.96	0.92	0.97	0.96	0.95
	ReLU	0.99	0.97	0.91	1.00	0.95	0.98	1.00	0.91
	Leaky ReLU	0.95	0.98	0.94	0.98	0.95	0.98	0.98	0.94
	ELU	0.97	0.98	0.95	0.99	0.96	0.99	0.99	0.95
	GELU	0.99	0.98	0.94	1.00	0.96	0.99	1.00	0.94
	PReLU	0.95	0.98	0.93	0.98	0.94	0.98	0.98	0.94
	Softplus	0.92	**0.99**	**0.96**	0.97	0.94	0.98	0.97	0.96
	Tanh	0.95	0.97	0.93	0.98	0.94	0.98	0.98	0.93
	Siren	0.92	0.98	0.93	0.97	0.93	0.97	0.97	0.93
	Swish	0.99	**0.99**	**0.96**	1.00	0.97	0.99	1.00	0.96

Table 4.7: Performance COVID-19 Adam Binary Classes

result in table 4.8

An essential aspect of the results is that even when we use the most traditional neural network rather than the most state-of-the-art model, the network shows much stronger performance with Swish function. To sum up everything that has been stated so far, if we replace the activation function, the model can be significantly improved.

Classifier	Optimizer	Number of Parameters	Activation Functions	Accuracy	Patient Status	Precision	Recall	F1-score	Specificity
VGG19 [14]	SGD	138 million	ReLU	0.90	COVID-19	0.83	1.00	0.91	-
					Normal	1.00	0.80	0.89	-
DenseNet201 [14]	SGD	27.2 million	ReLU	0.90	COVID-19	0.83	1.00	0.91	-
					Normal	1.00	0.80	0.89	-
This Work (Alexnet)	Adam	61 million	Swish	0.9870	COVID-19	0.99	0.96	0.97	1.00
					Normal	0.99	1.00	0.99	0.96

Table 4.8: Contrastive Analysis with Recent Studies

Confusion Matrix

The confusion matrixes in most of experiments are presented in the Appendix for your reference. During training, we obtain the history of train and validation accuracy, which in this case, is only an indicator of network convergence. Reflects the overall performance from experimental hyperparameter tuning regarding training and validation accuracy is also presented in the Appendix.

Activation functions	**Sigmoid**					**ReLU**					**LeakyReLU**				
Patient Categories	Normal	COVID	Lung Opacity	Viral Pneumonia	**Weighted Average**	Normal	COVID	Lung Opacity	Viral Pneumonia	**Weighted Average**	Normal	COVID	Lung Opacity	Viral Pneumonia	**Weighted Average**
precision	0.92	0.83	0.81	0.86	0.87	0.84	0.85	0.91	0.98	0.87	0.91	0.89	0.86	0.99	0.90
recall	0.87	0.83	0.87	0.96	0.87	0.95	0.86	0.73	0.85	0.87	0.93	0.86	0.86	0.95	0.90
f1-score	0.90	0.83	0.84	0.91	0.87	0.89	0.86	0.81	0.91	0.87	0.92	0.88	0.86	**0.97**	0.90
support	1020	362	602	135		1020	362	602	135		1020	362	602	135	
	PReLU					**Softplus**					**ELU**				
Patient Categories	Normal	COVID	Lung Opacity	Viral Pneumonia	**Weighted Average**	Normal	COVID	Lung Opacity	Viral Pneumonia	**Weighted Average**	Normal	COVID	Lung Opacity	Viral Pneumonia	**Weighted Average**
precision	0.92	0.93	0.83	0.90	0.90	**0.93**	0.96	0.85	0.94	0.91	0.89	**0.98**	0.88	0.91	0.91
recall	0.90	0.84	**0.90**	0.95	0.89	0.93	0.85	**0.90**	0.98	0.91	0.94	0.81	0.87	**0.99**	0.90
f1-score	0.91	0.89	0.86	0.92	0.89	**0.93**	0.90	0.87	0.96	0.91	0.92	0.89	**0.88**	0.95	0.90
support	1020	362	602	135		1020	362	602	135		1020	362	602	135	
	Tanh					**GELU**					**Siren**				
Patient Categories	Normal	COVID	Lung Opacity	Viral Pneumonia	**Weighted Average**	Normal	COVID	Lung Opacity	Viral Pneumonia	**Weighted Average**	Normal	COVID	Lung Opacity	Viral Pneumonia	**Weighted Average**
precision	0.88	0.93	0.85	**1.00**	0.89	0.88	0.90	**0.94**	0.98	0.91	0.85	0.86	0.89	0.99	0.87
recall	0.94	0.82	0.84	0.87	0.89	**0.97**	**0.91**	0.79	0.91	0.90	0.94	0.86	0.78	0.74	0.87
f1-score	0.91	0.88	0.84	0.93	0.89	0.92	0.90	0.86	0.95	0.90	0.89	0.86	0.83	0.85	0.87
support	1020	362	602	135		1020	362	602	135		1020	362	602	135	
	Swish														
Patient Categories	Normal	COVID	Lung Opacity	Viral Pneumonia	**Weighted Average**										
precision	0.92	0.93	0.86	0.97	0.91										
recall	0.92	0.90	0.89	0.96	0.91										
f1-score	0.92	**0.92**	0.87	**0.97**	0.91										
support	1020	362	602	135											

Table 4.9: Performance COVID-19 RMSprop 0.00001

Activation functions	**Sigmoid**					**ReLU**					**LeakyReLU**				
Patient Categories	Normal	COVID	Lung Opacity	Viral Pneumonia	**Weighted Average**	Normal	COVID	Lung Opacity	Viral Pneumonia	**Weighted Average**	Normal	COVID	Lung Opacity	Viral Pneumonia	**Weighted Average**
precision	0.82	0.92	**0.96**	**1.00**	0.89	0.91	0.89	0.83	0.96	0.89	0.93	0.95	0.84	0.96	0.91
recall	**0.99**	0.85	0.76	0.59	0.87	0.91	0.84	0.87	0.91	0.88	0.91	**0.91**	0.90	0.96	0.91
f1-score	0.90	0.88	0.85	0.74	0.87	0.91	0.86	0.85	0.94	0.89	0.92	**0.93**	**0.87**	**0.96**	0.91
support	1020	362	602	135		1020	362	602	135		1020	362	602	135	
	PReLU					**Softplus**					**ELU**				
Patient Categories	Normal	COVID	Lung Opacity	Viral Pneumonia	**Weighted Average**	Normal	COVID	Lung Opacity	Viral Pneumonia	**Weighted Average**	Normal	COVID	Lung Opacity	Viral Pneumonia	**Weighted Average**
precision	0.89	**1.00**	0.86	0.98	0.90	0.91	0.94	0.85	0.99	0.90	0.82	0.97	0.94	0.99	0.89
recall	0.97	0.78	0.85	0.93	0.90	0.93	0.89	0.87	0.91	0.90	**0.99**	0.76	0.75	0.93	0.88
f1-score	**0.93**	0.87	0.86	0.95	0.90	0.92	0.92	0.86	0.95	0.90	0.89	0.85	0.84	**0.96**	0.87
support	1020	362	602	135		1020	362	602	135		1020	362	602	135	
	Tanh					**GELU**					**Siren**				
Patient Categories	Normal	COVID	Lung Opacity	Viral Pneumonia	**Weighted Average**	Normal	COVID	Lung Opacity	Viral Pneumonia	**Weighted Average**	Normal	COVID	Lung Opacity	Viral Pneumonia	**Weighted Average**
precision	0.82	0.93	0.92	0.94	0.87	0.88	0.99	0.86	0.99	0.90	0.83	0.68	0.70	0.89	0.77
recall	0.98	0.76	0.72	**0.97**	0.87	0.96	0.76	0.87	0.90	0.90	0.83	0.57	0.80	0.69	0.77
f1-score	0.89	0.83	0.80	0.95	0.86	0.92	0.86	0.86	0.94	0.89	0.83	0.62	0.75	0.78	0.77
support	1020	362	602	135		1020	362	602	135		1020	362	602	135	
	Swish														
Patient Categories	Normal	COVID	Lung Opacity	Viral Pneumonia	**Weighted Average**										
precision	**0.94**	0.94	0.77	0.98	0.90										
recall	0.86	0.85	**0.94**	0.89	0.88										
f1-score	0.90	0.90	0.84	0.93	0.89										
support	1020	362	602	135											

Table 4.10: Performance COVID-19 RMSprop 0.0001

Activation functions	**Sigmoid**					**ReLU**					**LeakyReLU**				
Patient Categories	Normal	COVID	Lung Opacity	Viral Pneumonia	**Weighted Average**	Normal	COVID	Lung Opacity	Viral Pneumonia	**Weighted Average**	Normal	COVID	Lung Opacity	Viral Pneumonia	**Weighted Average**
precision	0.91	0.87	0.85	0.97	0.89	0.93	0.84	0.82	0.95	0.89	0.93	0.91	0.81	0.98	0.90
recall	0.93	0.81	0.86	0.93	0.89	0.86	**0.95**	0.87	0.93	0.88	0.88	0.91	0.90	0.93	0.89
f1-score	**0.92**	0.84	0.85	0.95	0.89	0.89	0.89	0.84	0.94	0.88	0.90	0.91	0.86	0.95	0.89
support	1020	362	602	135		1020	362	602	135		1020	362	602	135	
	PReLU					**Softplus**					**ELU**				
Patient Categories	Normal	COVID	Lung Opacity	Viral Pneumonia	**Weighted Average**	Normal	COVID	Lung Opacity	Viral Pneumonia	**Weighted Average**	Normal	COVID	Lung Opacity	Viral Pneumonia	**Weighted Average**
precision	0.91	0.97	0.84	**0.99**	0.91	0.92	0.93	0.86	**0.99**	0.91	0.93	0.91	0.79	0.98	0.89
recall	0.92	0.87	0.90	0.91	0.90	0.92	0.89	0.90	0.92	0.91	0.86	0.88	0.92	0.92	0.88
f1-score	0.92	**0.92**	0.87	0.95	0.90	**0.92**	0.91	**0.88**	0.95	0.91	0.89	0.89	0.85	0.95	0.89
support	1020	362	602	135		1020	362	602	135		1020	362	602	135	
	Tanh					**GELU**					**Siren**				
Patient Categories	Normal	COVID	Lung Opacity	Viral Pneumonia	**Weighted Average**	Normal	COVID	Lung Opacity	Viral Pneumonia	**Weighted Average**	Normal	COVID	Lung Opacity	Viral Pneumonia	**Weighted Average**
precision	0.92	0.88	0.82	**0.99**	0.89	0.90	**0.98**	0.84	0.94	0.90	0.89	0.94	**0.88**	0.98	0.90
recall	0.89	0.88	0.89	0.84	0.88	0.93	0.81	0.89	**0.97**	0.90	**0.95**	0.84	0.85	0.90	0.90
f1-score	0.90	0.88	0.85	0.91	0.88	0.91	0.89	0.86	**0.96**	0.90	0.91	0.89	0.86	0.93	0.90
support	1020	362	602	135		1020	362	602	135		1020	362	602	135	
	Swish														
Patient Categories	Normal	COVID	Lung Opacity	Viral Pneumonia	**Weighted Average**										
precision	**0.94**	0.95	0.81	**0.99**	0.91										
recall	0.89	0.89	**0.93**	0.93	0.90										
f1-score	0.91	**0.92**	0.86	**0.96**	0.90										
support	1020	362	602	135											

Table 4.11: Performance COVID-19 Adam 0.00001

Activation functions	**Sigmoid**					**ReLU**					**LeakyReLU**				
Patient Categories	Normal	COVID	Lung Opacity	Viral Pneumonia	**Weighted Average**	Normal	COVID	Lung Opacity	Viral Pneumonia	**Weighted Average**	Normal	COVID	Lung Opacity	Viral Pneumonia	**Weighted Average**
precision	0.87	0.94	0.91	**0.99**	0.90	0.90	0.91	0.89	0.98	**0.91**	0.85	0.97	0.88	0.93	0.88
recall	**0.97**	0.85	0.82	0.87	0.90	0.93	**0.95**	0.83	0.93	**0.91**	0.95	0.75	0.81	**0.97**	0.88
f1-score	**0.92**	0.89	0.86	0.93	0.90	**0.92**	0.93	0.86	**0.95**	**0.91**	0.90	0.85	0.84	**0.95**	0.88
support	1020	362	602	135		1020	362	602	135		1020	362	602	135	
	PReLU					**Softplus**					**ELU**				
Patient Categories	Normal	COVID	Lung Opacity	Viral Pneumonia	**Weighted Average**	Normal	COVID	Lung Opacity	Viral Pneumonia	**Weighted Average**	Normal	COVID	Lung Opacity	Viral Pneumonia	**Weighted Average**
precision	0.86	0.85	0.89	0.93	0.87	0.93	0.84	0.90	0.98	**0.91**	0.86	0.83	**0.91**	0.97	0.88
recall	0.93	0.87	0.75	0.96	0.87	0.92	0.94	0.87	0.87	**0.91**	0.94	0.78	0.81	0.93	0.88
f1-score	0.90	0.86	0.82	**0.95**	0.87	0.92	0.89	**0.88**	0.92	**0.91**	0.90	0.80	0.85	**0.95**	0.87
support	1020	362	602	135		1020	362	602	135		1020	362	602	135	
	Tanh					**GELU**					**Siren**				
Patient Categories	Normal	COVID	Lung Opacity	Viral Pneumonia	**Weighted Average**	Normal	COVID	Lung Opacity	Viral Pneumonia	**Weighted Average**	Normal	COVID	Lung Opacity	Viral Pneumonia	**Weighted Average**
precision	**0.95**	0.87	0.77	**0.99**	0.89	0.87	**0.98**	0.83	0.95	0.88	0.85	0.85	0.81	0.97	0.84
recall	0.82	0.91	**0.94**	0.92	0.88	0.94	0.67	0.87	0.93	0.87	0.92	0.73	0.78	0.87	0.84
f1-score	0.88	0.89	0.84	**0.95**	0.88	0.90	0.80	0.85	0.94	0.87	0.88	0.79	0.79	0.91	0.84
support	1020	362	602	135		1020	362	602	135		1020	362	602	135	
	Swish														
Patient Categories	Normal	COVID	Lung Opacity	Viral Pneumonia	**Weighted Average**										
precision	0.92	0.80	0.80	0.95	0.87										
recall	0.86	0.82	0.88	0.93	0.86										
f1-score	0.89	0.81	0.84	0.94	0.87										
support	1020	362	602	135											

Table 4.12: Performance COVID-19 Adam 0.0001

Chapter 5

Conclusion and Discussion

The complementation of artificial intelligence and medical imaging is a trend of modernization. It has to be said that the COVID epidemic in 2019 pushed the rapid development of this area in recent years. Medical image diagnoses, as fast as possible, can reduce the cost of treatment, reduce the workload of medical staff, and buy time for patients.

In this work, we have successfully evaluated the performance of the model with different activation functions in various aspects. We proposed two models for the balanced dataset, CIFAR-10, and two models for the imbalanced dataset, COVID-19 Radiography Database (Kaggle).

Notwithstanding, we cannot draw a clear conclusion about the correlation between the effect of the activation function and the imbalance of data. In all the cases covered in our research, the sigmoid is the worst function, and Siren is extremely sensitive to the learning rate, while Swish function is stable and is not affected by the learning rate most of the time.

The experimental results show that the chest X-ray classification accuracy reaches 91.55% on the softplus function, and the maximum accuracy is 91.44% on the Swish function. Despite that, the research on softplus is rarely found in the previous literature, so in future work, we could verify that this function is superior to other activation functions in chest X-Ray analysis.

This work illustrates that the comparative performance of neural networks with different activation functions in image classification, especially on the highly imbalanced dataset. The conclusion can be drawn that the Swish function model has a venerable image classification result in general, while the model using Softplus function has a more prominent performance in chest X-ray image classification.

We could start with the Swish function for future work, then move on to alternatives to the ReLU functions (PReLU, GELU) if it does not give satisfactory results, but start with both Softplus and Swish for coronavirus classifier.

In future research, the data augmentation of X-ray images can be further studied. In recently published papers, the data augmentation is not always good for accuracy and sometimes hurts the performance of the model. My assumption is that because the x-rays are gray, low-contrast images and COVID-19 confirmed radiographs showed ground-glass opacity with occasional patchy, peripheral, and bilateral regional consolidation. Meanwhile, the differences between these varieties of Chest X-ray are so subtle that they are hard to tell with the naked eye, and conventional data augmentation might not help. In the future, we can use the GaN model to enhance the data, or we can determine data enhancement methods suitable for medical images.

In addition, because of the update of the COVID-19 database, more and more data are available for model training, and the COVID-19 classifiers in the recent studies have been very effective, fast, and accurate. We can conclude that researchers can add more medical images of various pneumonia infections to these databases to generate a comprehensive X-ray classifier.

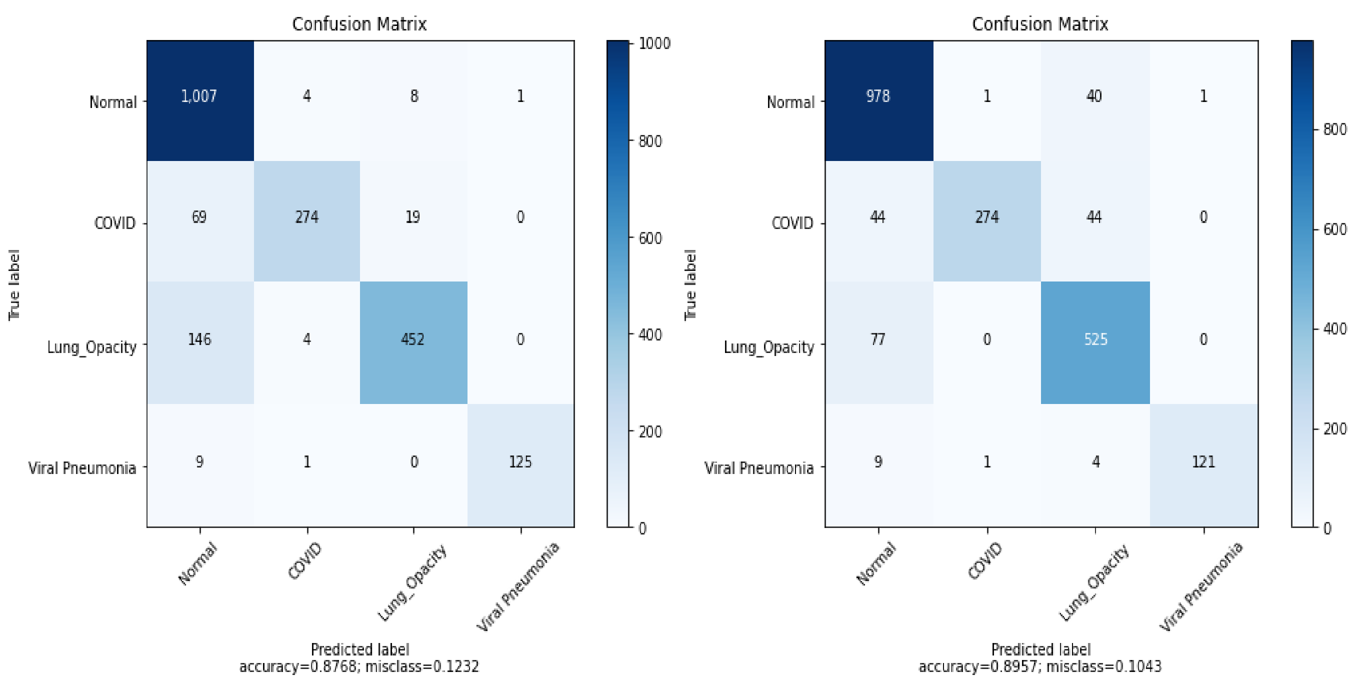
In this work, compared with ten different activation functions, it is found that the Swish function has the characteristics of strong stability and high accuracy. However, we only draw this conclusion in the classification of images. In future studies, the performance of these activation functions can be compared in areas such as speech recognition and image segmentation.

Appendices

Appendix A

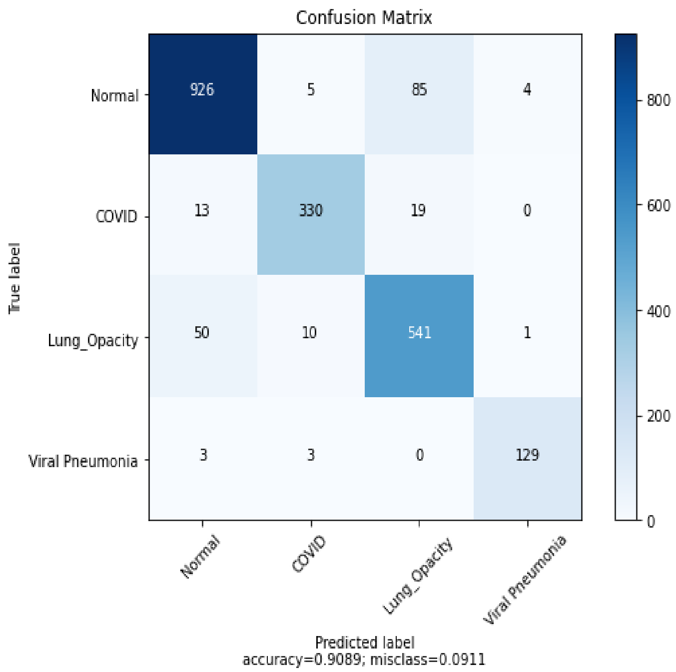
Confusion Matrix

A.1 Confusion Matrix RMSprop 0.0001

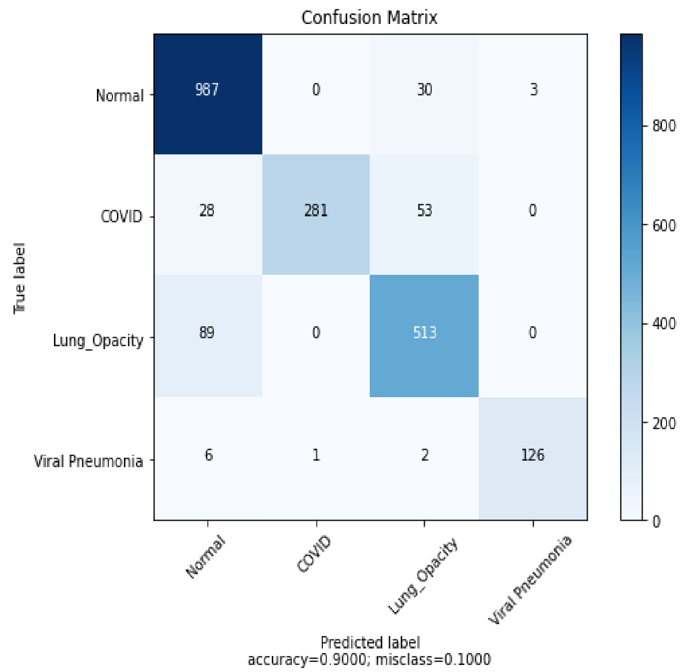


(a) ELU (b) GELU

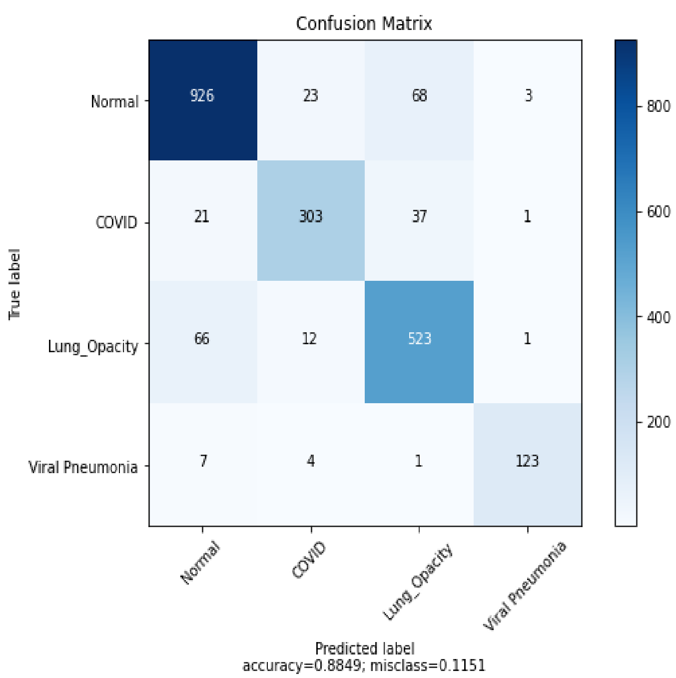
Figure A.1: COVID-19 RADIOGRAPHY Confusion Matrix RMSprop 0.0001



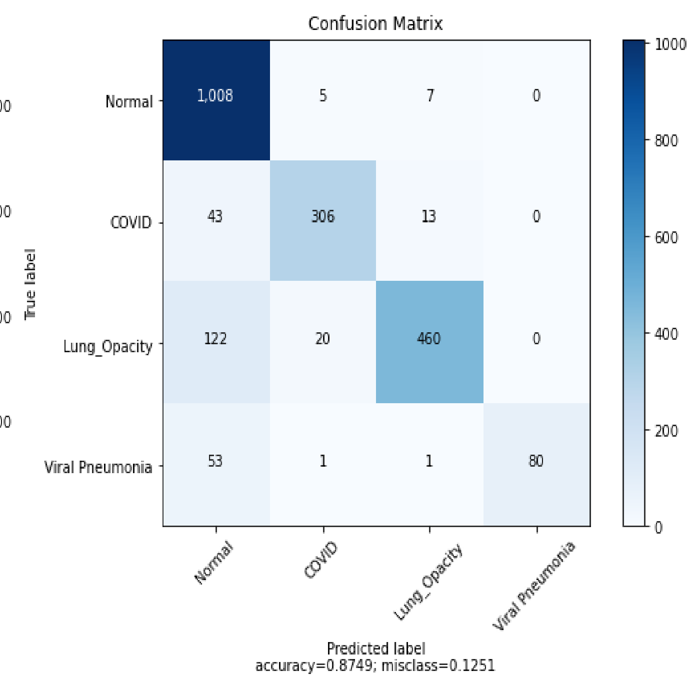
(c) LeakyReLU



(d) PReLU



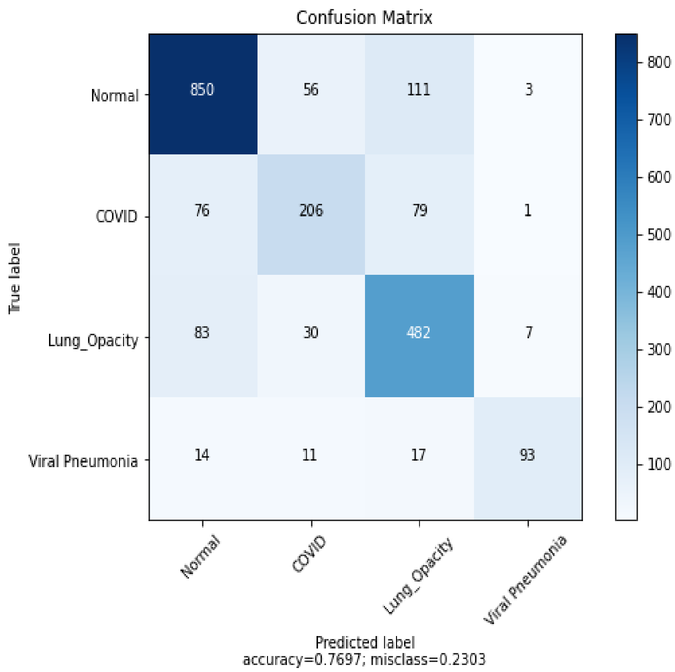
(e) ReLU



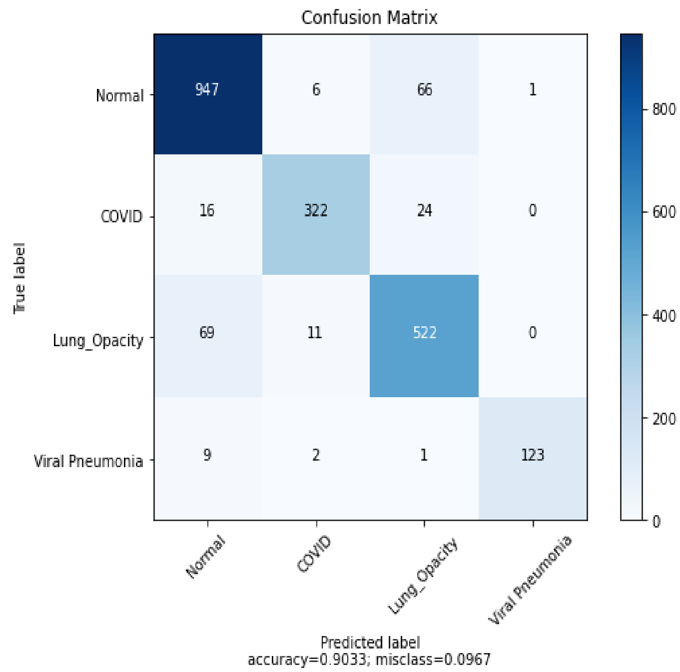
(f) Sigmoid

Figure A.1: COVID-19 RADIOGRAPHY Confusion Matrix RMSprop 0.0001

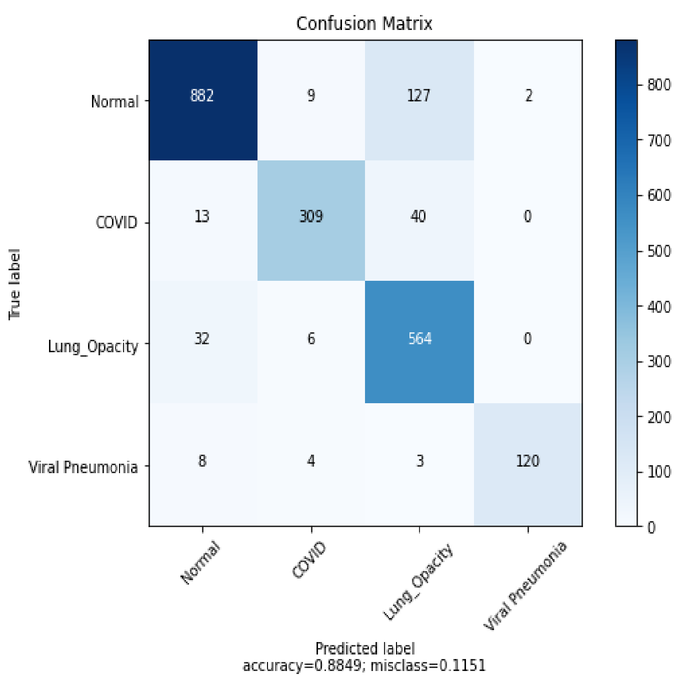
A.2 Confusion Matrix RMSprop 0.00001



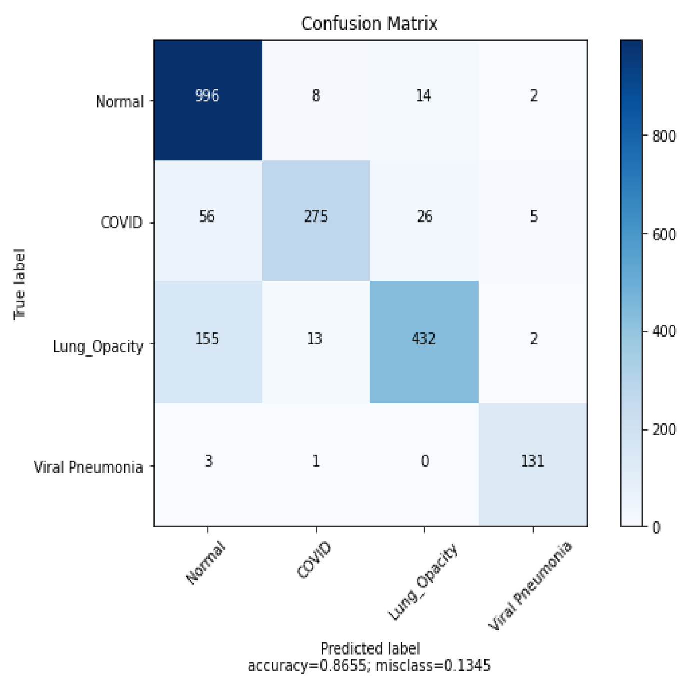
(g) Siren



(h) Softplus

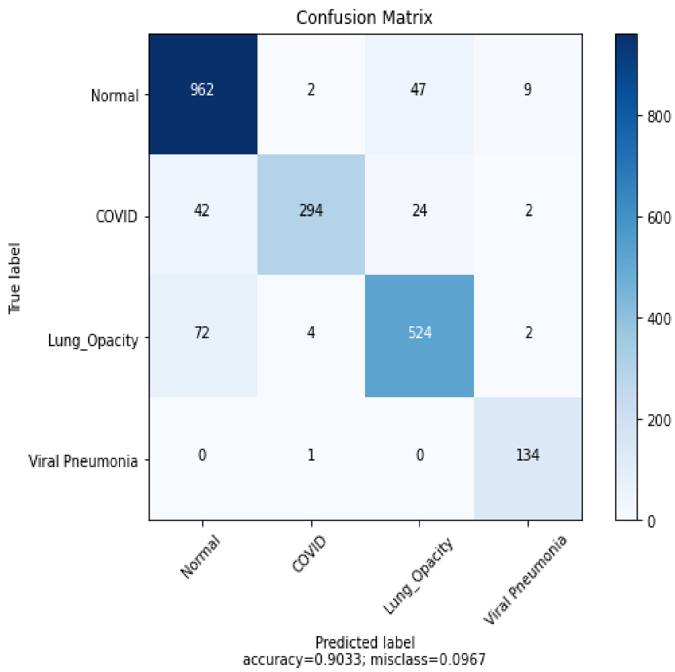


(i) Swish

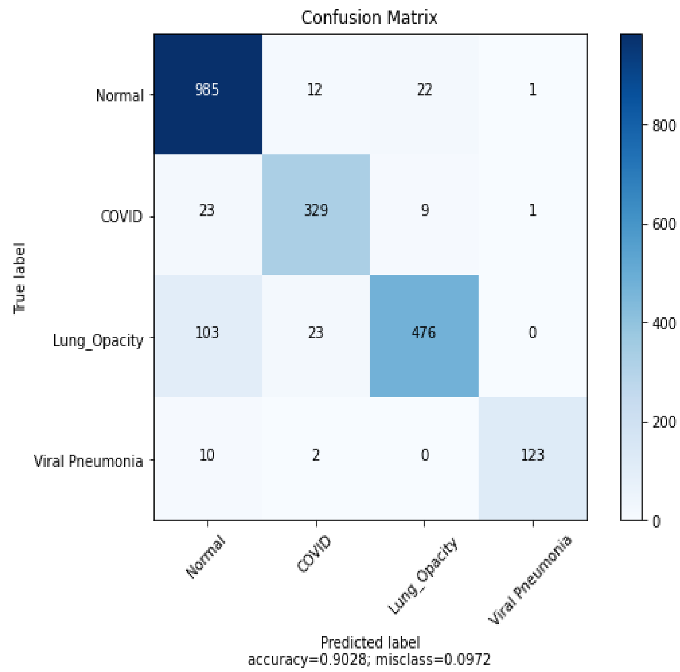


(j) Tanh

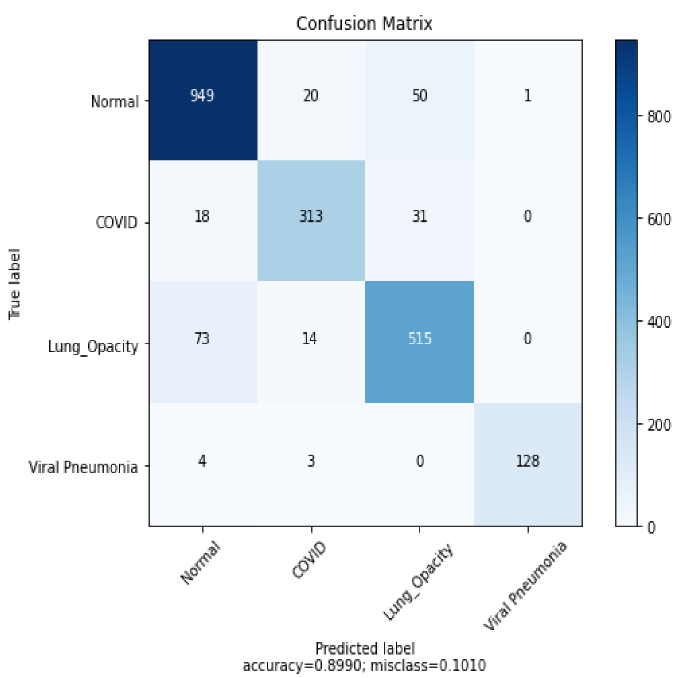
Figure A.1: COVID-19 RADIOGRAPHY Confusion Matrix RMSprop 0.0001



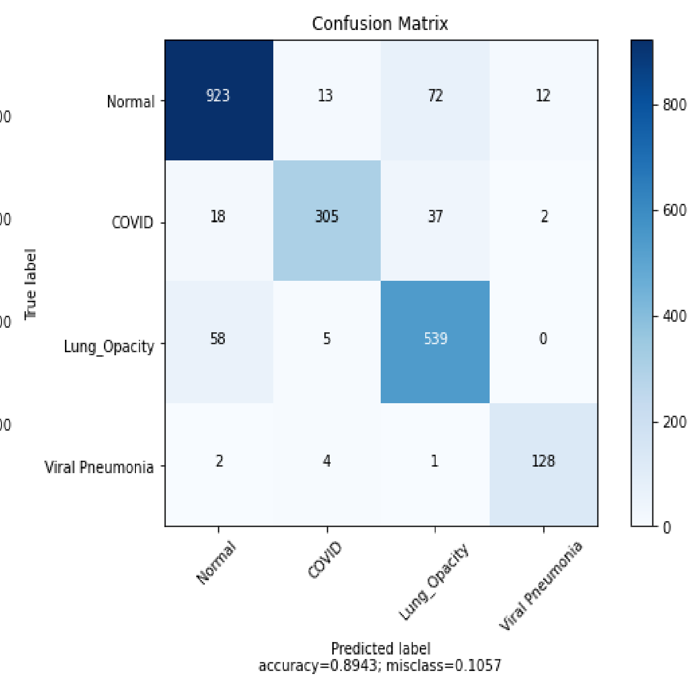
(a) ELU



(b) GELU

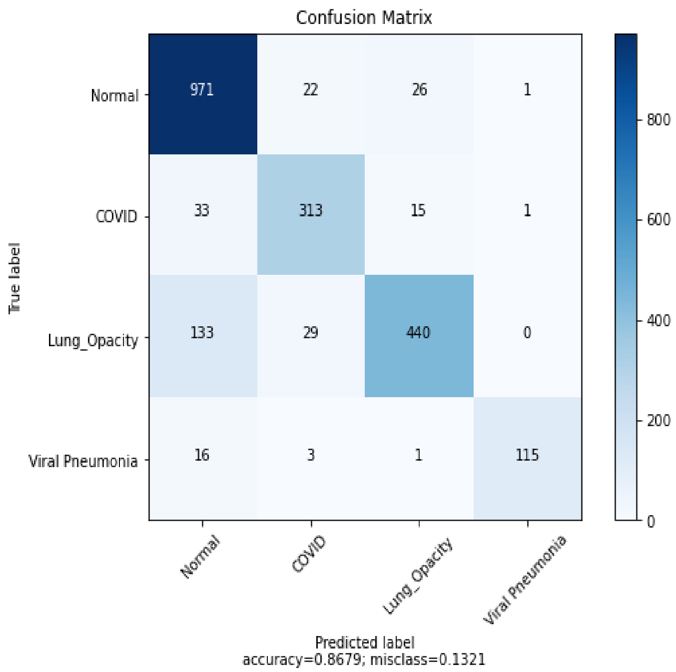


(c) LeakyReLU

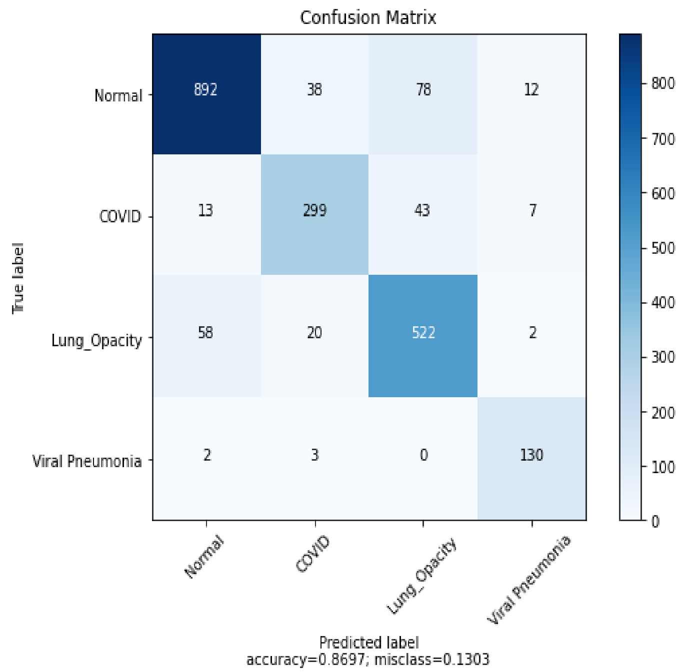


(d) PReLU

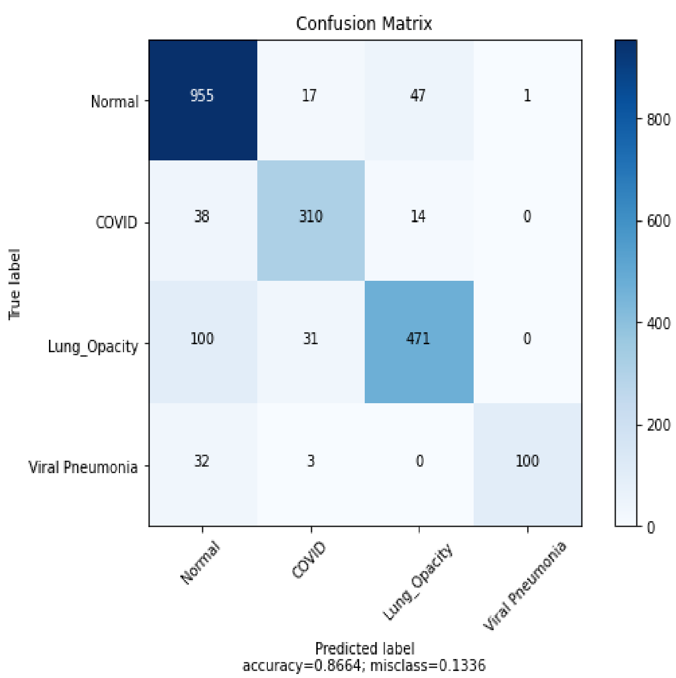
Figure A.2: COVID-19 RADIOGRAPHY Confusion Matrix RMSprop 0.00001



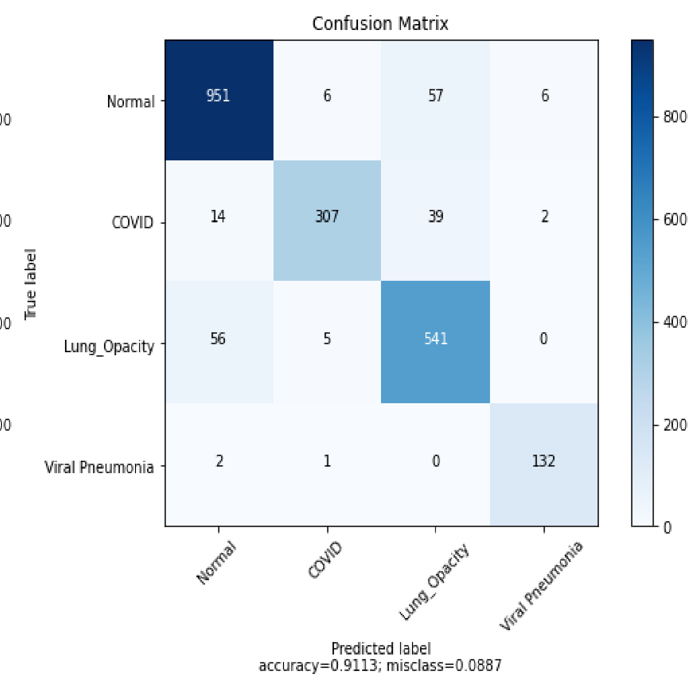
(e) ReLU



(f) Sigmoid

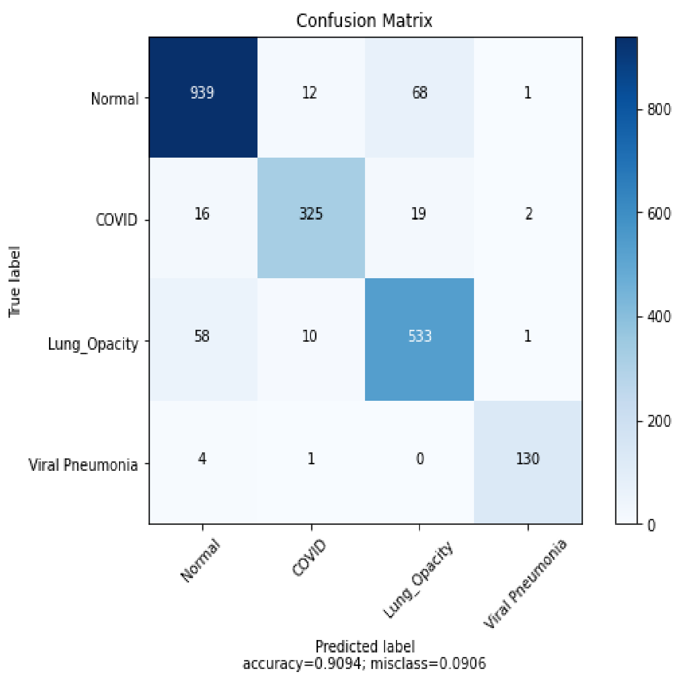


(g) Siren

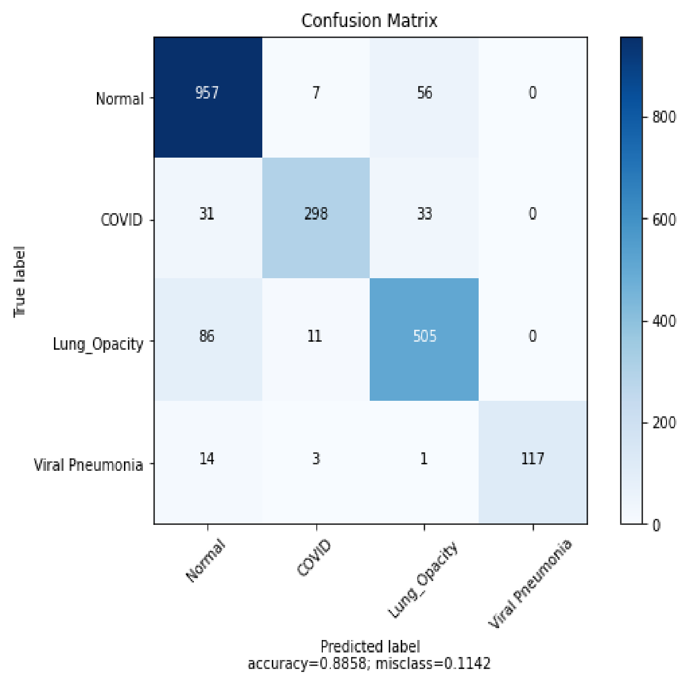


(h) Softplus

Figure A.2: COVID-19 RADIOGRAPHY Confusion Matrix RMSprop 0.00001



(i) Swish



(j) Tanh

Figure A.2: COVID-19 RADIOGRAPHY Confusion Matrix RMSprop 0.00001

Bibliography

[1] Bio-inspired neurocomputing. *Studies in Computational Intelligence*, 2021.

[2] Khalid El Asnaoui, Youness Chawki, and Ali Idri. Automated methods for detection and classification pneumonia based on x-ray images using deep learning, 2020.

[3] M. E. H. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M. A. Kadir, Z. B. Mahbub, K. R. Islam, M. S. Khan, A. Iqbal, N. A. Emadi, M. B. I. Reaz, and M. T. Islam. Can ai help in screening viral and covid-19 pneumonia? *IEEE Access*, 8:132665–132676, 2020.

[4] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus), 2016.

[5] Mohammad A. Dabbah, Angus B. Reed, Adam T. C. Booth, Arrash Yassaee, Alex Despotovic, Benjamin Klasmer, Emily Binning, Mert Aral, David Plans, Alain B. Labrique, and Diwakar Mohan. Machine learning approach to dynamic risk modeling of mortality in covid-19: a uk biobank study, 2021.

[6] Muhammad Farooq and Abdul Hafeez. Covid-resnet: A deep learning framework for screening of covid19 from radiographs, 2020.

[7] Gusztáv Gaál, Balázs Maga, and András Lukács. Attention u-net based adversarial architectures for chest x-ray lung segmentation. *arXiv preprint arXiv:2003.10304*, 2020.

[8] Christian Garbin, Xingquan Zhu, and O. Marques. Dropout vs. batch normalization: an empirical study of their impact to deep learning. *Multimedia Tools and Applications*, 79:12777 – 12815, 2020.

[9] Gusztáv Gaál, Balázs Maga, and András Lukács. Attention u-net based adversarial architectures for chest x-ray lung segmentation, 2020.

[10] C. L. Giles, S. Lawrence, and Ah Chung Tsoi. Rule inference for financial prediction using recurrent neural networks. In *Proceedings of the IEEE/IAFE 1997 Computational Intelligence for Financial Engineering (CIFEr)*, pages 253–259, 1997.

[11] Richard Hahnloser, Rahul Sarpeshkar, Misha Mahowald, Rodney Douglas, and H. Seung. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*, 405:947–51, 07 2000.

[12] Malay Haldar, Mustafa Abdool, Prashant Ramanathan, Tao Xu, Shulin Yang, Huizhong Duan, Qing Zhang, Nick Barrow-Williams, Bradley C. Turnbull, Brendan M. Collins, and et al. Applying deep learning to airbnb search. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, Jul 2019.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[14] Ezz El-Din Hemdan, Marwa A. Shouman, and Mohamed Esmail Karar. Covidx-net: A framework of deep learning classifiers to diagnose covid-19 in x-ray images, 2020.

[15] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus), 2020.

[16] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors, 2012.

[17] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[18] J. F. Kolen and S. C. Kremer. *Gradient Flow in Recurrent Nets: The Difficulty of Learning LongTerm Dependencies*, pages 237–243. 2001.

[19] Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016.

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[21] Abhijeet Kumar. Achieving 90% accuracy in object recognition task on cifar-10 dataset with keras: Convolutional neural networks, 03 2018.

[22] Lin Li, Lixin Qin, Zeguo Xu, Youbing Yin, Xin Wang, Bin Kong, Junjie Bai, Yi Lu, Zhenghan Fang, Qi Song, et al. Artificial intelligence distinguishes covid-19 from community acquired pneumonia on chest ct. *Radiology*, 2020.

[23] Minghui Liao, Baoguang Shi, Xiang Bai, Xinggang Wang, and Wenyu Liu. Textboxes: A fast text detector with a single deep neural network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

[24] Weibo Liu, Zidong Wang, Xiaohui Liu, Nianyin Zeng, Yurong Liu, and Fuad E Alsaadi. A survey of deep neural network architectures and their applications. *Neurocomputing*, 234:11–26, 2017.

[25] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3. Citeseer, 2013.

[26] Chigozie Nwankpa, Winifred Ijomah, Anthony Gachagan, and Stephen Marshall. Activation functions: Comparison of trends in practice and research for deep learning. *arXiv preprint arXiv:1811.03378*, 2018.

[27] Ratnabali Pal, Arif Ahmed Sekh, Samarjit Kar, and Dilip K. Prasad. Neural network based country wise risk prediction of covid-19. *Applied Sciences*, 10(18):6448, Sep 2020.

[28] Tawsifur Rahman. Covid-19 radiography database, 03 2020. data accessed 1-April-2021, https://www.kaggle.com/tawsifurrahman/covid19-radiography-database.

[29] Tawsifur Rahman, Amith Khandakar, Yazan Qiblawey, Anas Tahir, Serkan Kiranyaz, Saad Bin Abul Kashem, Mohammad Tariqul Islam, Somaya Al Maadeed, Susu M Zughaier, Muhammad Salman Khan, and Muhammad E. H. Chowdhury. Exploring the effect of image enhancement techniques on covid-19 detection using chest x-rays images, 2020.

[30] Prajit Ramachandran, Barret Zoph, and Quoc V. Le. Searching for activation functions, 2017.

[31] Dian Eka Ratnawati and Anam Syaiful Marjono, Widodo. Comparison of activation function on extreme learning machine (ELM) performance for classifying the active compound. In *American Institute of Physics Conference Series*, volume 2264 of *American Institute of Physics Conference Series*, page 140001, September 2020.

[32] F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65 6:386–408, 1958.

[33] E. W. Saad, D. V. Prokhorov, and D. C. Wunsch. Advanced neural network training methods for low false alarm stock trend prediction. In *Proceedings of International Conference on Neural Networks (ICNN'96)*, volume 4, pages 2021–2026 vol.4, 1996.

[34] T. N. Sainath, A. Mohamed, B. Kingsbury, and B. Ramabhadran. Deep convolutional neural networks for lvcsr. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8614–8618, 2013.

[35] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[36] Sheng-Fu Liang, A. W. Y. Su, and Cheng-Teng Lin. A new recurrent-network-based music synthesis method for chinese plucked-string instruments - pipa and qin. In *IJCNN'99. International Joint Conference on Neural Networks. Proceedings (Cat. No.99CH36339)*, volume 4, pages 2564–2569 vol.4, 1999.

[37] Vincent Sitzmann, Julien N. P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions, 2020.

[38] Yi Sun, Ding Liang, Xiaogang Wang, and Xiaoou Tang. Deepid3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873*, 2015.

[39] Martin Sundermeyer, Tamer Alkhouli, Joern Wuebker, and Hermann Ney. Translation modeling with bidirectional recurrent neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 14–25, 2014.

[40] A. M. TURING. I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, LIX(236):433–460, 10 1950.

[41] Linda Wang and Alexander Wong. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images, 2020.

[42] Wikipedia contributors. Cifar-10 — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=CIFAR-10&oldid=1014793182, 2021. [Online; accessed 3-April-2021].

[43] Wikipedia contributors. Covid-19 — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=COVID-19&oldid=1015476995, 2021. [Online; accessed 1-April-2021].

[44] Wikipedia contributors. Covid-19 pandemic — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=COVID-19_pandemic&oldid=1018919330, 2021. [Online; accessed 20-April-2021].

[45] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017.

[46] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

[47] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *arXiv preprint arXiv:1411.1792*, 2014.

[48] Jianhao Zhu and Zong Chen. Comparative analysis of various new activation functions based on convolutional neural network. *Journal of Physics: Conference Series*, 1676:012228, nov 2020.