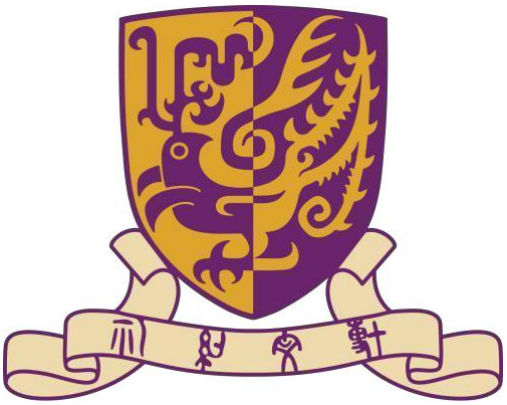

The Chinese University of Hong Kong, Shenzhen

CSC3180

Final Report

Session 1 Team 4

Song Zefang 119010269

Lu Dongzhuyuan 119010216

Song Wenxin 120090625

An Improved COVID-19 Positive Cases Recognition System Based on X-ray Images

1. Introduction

COVID-19 is a relatively new respiratory disease resulting from the infection of the coronavirus SARS-CoV-2 (Queensland Health, 2020). The virus is transported from an infected person to the healthy through coughing or sneezing. As the virus spreads down one's respiratory tract, the patient may become more and more difficult to breathe. Therefore, effective measures to detect and diagnose the infection of COVID-19 as early as possible are in great need and necessary. PCR tests perform one widely applied technique to diagnose the virus. Although doing a PCR test is cheap, it could be tremendously inaccurate, as researchers from John Hopkins Medicine have shown that 20% of such tests might produce false-negative results (Bird, 2020). Some patients with lung diseases can still diagnose with COVID-19 after getting two or even three negative results. Thus, X-ray images are needed to help diagnose COVID-19. However, after the patients have taken their chest X-rays, they still need the doctors to analyze the image, which is a waste of time both for the patients and the doctors. A well-known technique to recognize the features of images and classify them is called deep neural networks. Researchers made one of the existing recognition systems based on St. John's University technology. They used five convolutional layers, with ReLu as the activation function and softmax as a classification modeling tool (Ahmed, 2020).
However, their model has not been applied to the COVID-19 diagnosis in practice yet. When retrying their method, our prediction accuracy is much lower than the data presented in their research paper, which is only about 72%. Besides, their prediction accuracy varies greatly after each training (shown in Figure 1 and Figure 2).

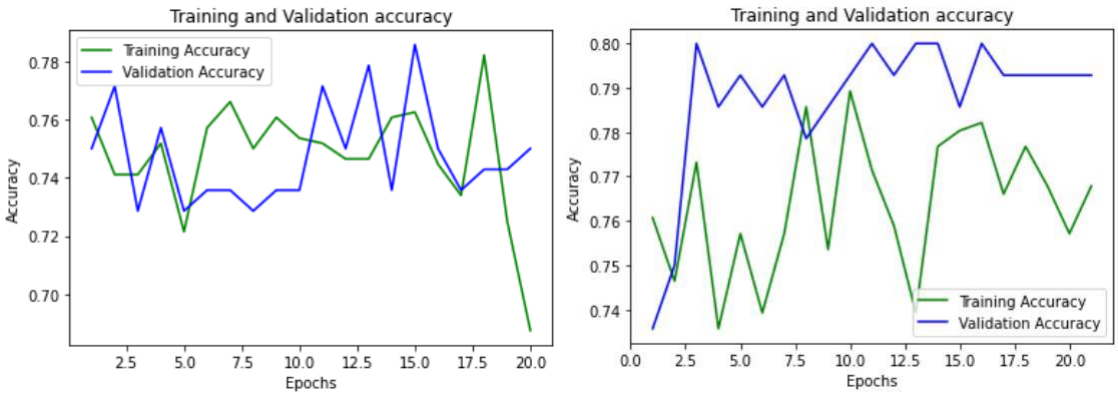

Figure 1. Training and Validation Accuracy Figure 2. Training and Validation Accuracy

These two problems will prevent the model from being applied to COVID-19 diagnosis in practice. By preprocessing the training data and modifying the model used to improve the current recognition system, our project improves its stability and prediction accuracy, making it more possible for the model to be applied to COVID-19 diagnosis in practice. The results of our project can be used to ease the unbalanced of medical resources distribution in China. Doctors can also liberate themselves from heavy repetitive work and focus on more valuable things.

2. **Dataset and images**

In our project, all the images are from three Kaggle datasets in the same database. This database is the winner of the COVID-19 Dataset Award by Kaggle Community, created by researchers from different universities and many medical doctors. It contains 3616 chest X-ray images for COVID-19 positive cases, normal and Viral Pneumonia. Our group did not directly use the collated datasets from this database. When viewing the images from these three datasets, we found that they contain lots of images without prominent characteristics, which may affect the training effect of the model.

For example, both Figure 3 and Figure 4 are the chest X-rays for COVID-19 positive cases, while Figure 2 has more obvious features than Figure 3. Then we will choose Figure 2 instead of Figure 3.

Figure 3. COVID-19 Positive case Figure 4. COVID-19 Positive case

When we select the images, we label the normal cases with 0, Covid-19 positive cases with 1, and Viral Pneumonia cases with 2. The dataset for our project contains 560 images for each class as training data and 700 images for each class as testing data.

3. **Methods:**

3.1 Image pre-processing

A sample image that will be trained using CNN is shown below.

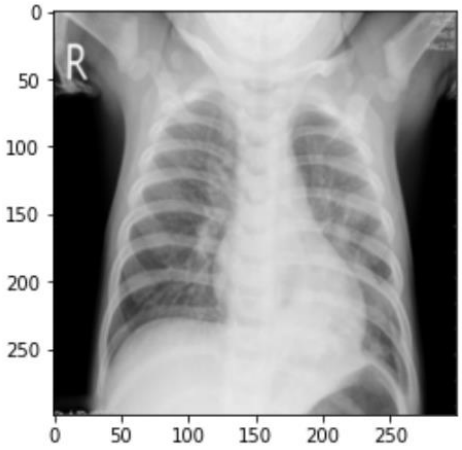

Figure 5. Training sample

The above figure is appropriate enough for the optimizer to train and return the corresponding CNN model. However, in the actual situation, we shall never expect the X-ray figure given to be in a somewhat “ideal” appearance. Therefore, we use specific techniques (shear, rotate...) to modify the training sample. The pre-processed training samples are shown below.

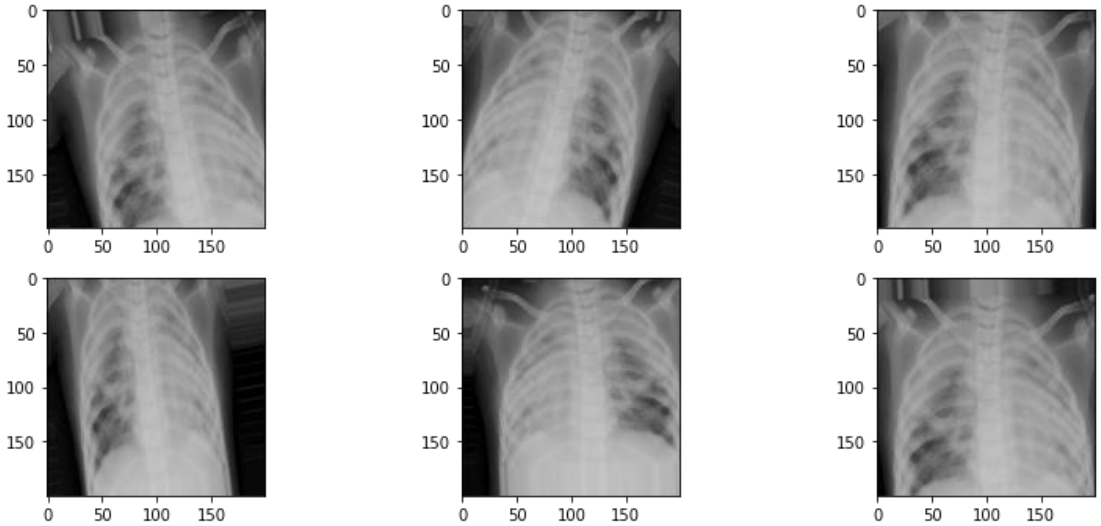

Figure 6. Pre-processed Samples

3.2 Five-layer Convolutional Neural Networks

a. A Single-layer Neural Network

Before we delve into the 5-layer CNN model, we first introduce some fundamental concepts

and mechanisms of a single-layer CNN.

An essential component of the neural network is a neuron, illustrated in the following figure.

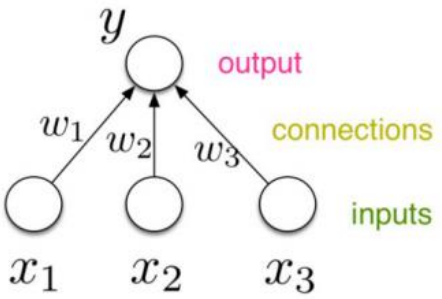

Figure 7. A Single Neuron

In the above graph, x_1, x_2 and x_3 stand for the input 3-dimensional vector, while w_1, w_2 and w_3 represent the relationship between the input data and *y*. For linear connection, we can express *y* in the following way:

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

In our case, *x* is our input image, we want to find parameters so that we can extract specific features of the image.

Here, we introduce φ as the activation function. There are numerous activation functions that we can apply for the neural networks. Activation function like *sigmoid* (σ) can be used to nonlinearized the model so as to predict much more complicated data. Some common activation functions are shown as follows. Here, we use ReLu as our activation function.

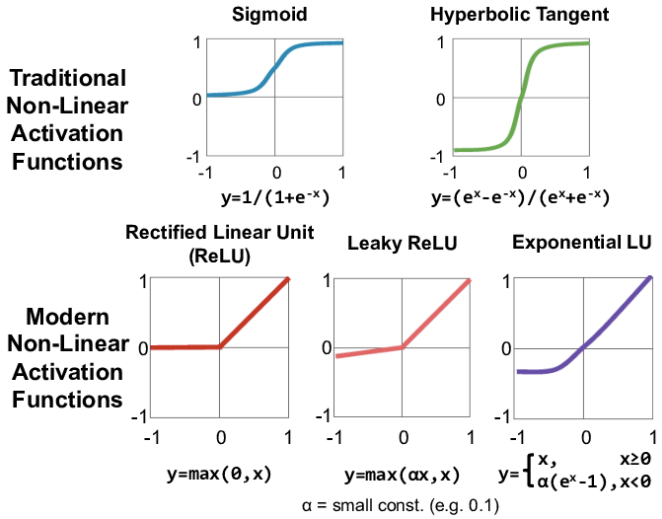

Figure 8. Activation Functions

In supervised learning (image classification), each data *x* is assigned with a definitive label

y. Given the input data, we can predict the label of x to be $\hat{y}$. Then, we are able to formulate the loss function:

$$L(w) = \sum \frac{1}{2}(y - \hat{y})^2$$

b. Multi-layer Neural Networks

In many situations, a single layer is not expressive enough to represent a more complicated model. Hence, we introduce multiple layers to achieve better performance. A double-layer neural network is shown below.

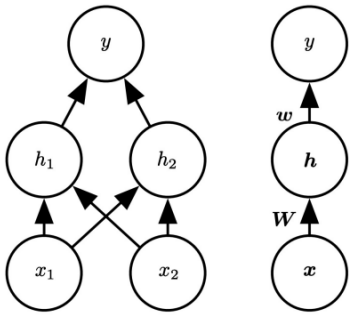

Figure 9. A Double-Layer Neural Network

In this case, y can be formulated as a linear transformation as follows.

$$h = g_2(Wx + c)$$

$$y = g_1(w^T h + b)$$

The loss function is similar as the preceding section.

c. Convolutional Neural Networks (CNN)

In image classification, the input data is usually three dimensional. If we transform the image data into a vector, the parameters trained will be enormous which is time-consuming. To tackle the problem, we introduce the CNN.

A basic step in CNN is to convolve an image using some filter shown below.

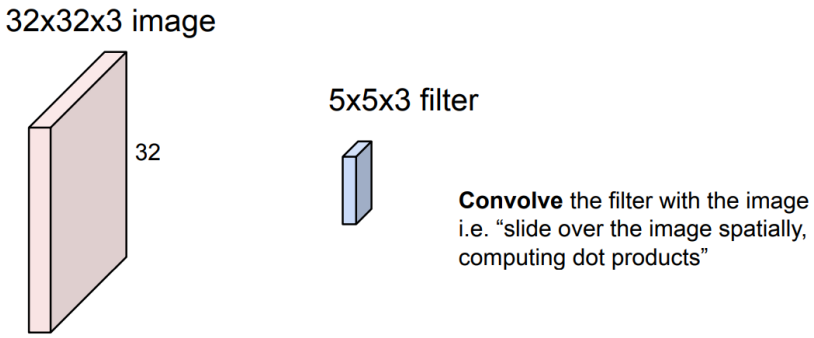

Figure 10. CNN_1

Then we apply dot product in the three-dimensional state space.

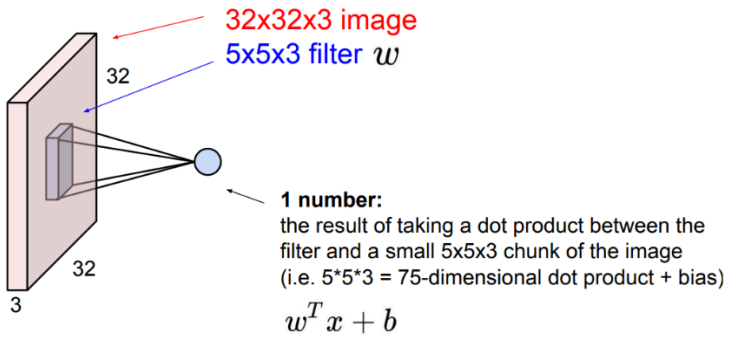

Figure 11. CNN_2

By doing this, we can extract certain features of the original image and reduce the size of the image. However, a single convolution layer sometimes may not satisfy the requirement. Another layer is we should use to extract feature is called max-pooling layer, which is usually applied after each convolution. The mechanism for max-pooling is shown in the following figure. The filter simply selects the maximum value in the specified region to represent the whole region. In this way, we can also reduce the dimensionality of the image.

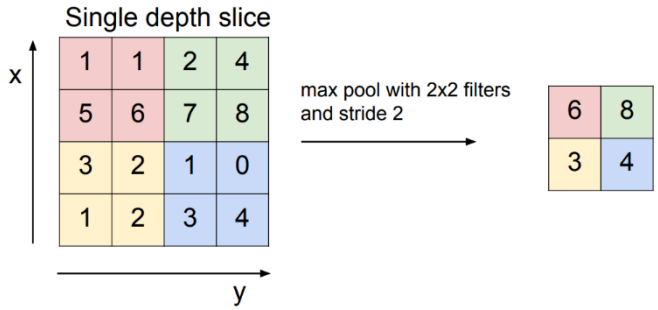

Figure 12. Max-pooling

Pooling layers are used to reduce the dimensions of the feature maps. Thus, it reduces the number of parameters to learn and the amount of computation performed in the network.

Going through the 5-layer CNN, our image will be transformed to a 512 by 1 feature vector which can be used for supervised learning training model.

3.3.1 Softmax

Softmax is used to solve multi-class classification problem. For example, we can choose a Covid-19 positive cases (y) as positive class and put it on the top. After normalize it, we get the Softmax function

$$p(y|x) = \frac{\exp(W_y.x)}{\sum_{c=1}^{C} \exp(W_c.x)}$$

with C being the number of classes. To solve the problem, we minimize the cost function

$$J(\mathbf{W}) = -\frac{1}{m}\sum_{i}^{m}\sum_{j}^{C}\left[\mathbb{I}(y_i = j)\log(f_{\mathbf{w}_j,b_j}(\mathbf{x}_i)\right],$$

for each data. We can use gradient descent to get the solution.

$$\mathbf{w}_j \leftarrow \mathbf{w}_j - \alpha\frac{\partial J(\mathbf{W})}{\partial \mathbf{w}_j},$$

$$\frac{\partial J(\mathbf{W})}{\partial \mathbf{w}_j} = -\frac{1}{m}\sum_{j}^{m}\mathbb{I}(y_i = j)\cdot\frac{1}{f_{\mathbf{w}_j,b_j}(\mathbf{x}_i))}\cdot\frac{\nabla f_{\mathbf{w}_j,b_j}(\mathbf{x}_i)}{\nabla \mathbf{w}_j},$$

$$\frac{\nabla f_{\mathbf{w}_j,b_j}(\mathbf{x}_i)}{\nabla \mathbf{w}_j} = f_{\mathbf{w}_j,b_j}(\mathbf{x}_i)\cdot(1 - f_{\mathbf{w}_j,b_j}(\mathbf{x}_i))\cdot\mathbf{x}_i.$$

In prediction, we will put the new data in the class with highest probability. To improve the accuracy of our prediction, our group explored other models such as SVM and Random Forest.

3.3.2 Cross Validation

Cross-validation is used to choose hyper-parameters so that the model can work best on the data. In K-fold Cross-validation, we pick the hyper-parameters following the four steps:

1. Split the train data into *K* folds
2. Try each fold as validation, while others as train
3. Train the model on the train folds, and evaluate the performance on the validation fold
4. Calculate the average results on all validation folds across all trials, pick the hyper-parameters with the best average result.

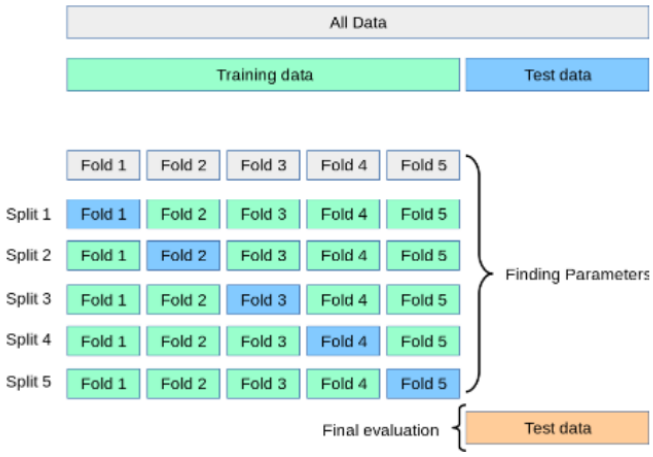

Figure 13. 5-fold Cross Validation

Figure 13 is an example of 5-fold cross validation which we use in this project.

3.4 SVM

SVM is the abbreviation of support vector machine. When classify two classes, we will choose the classifier with largest margin, which is the distance from the closest point of positive and negative classes to the decision boundary. As shown in the figure below.

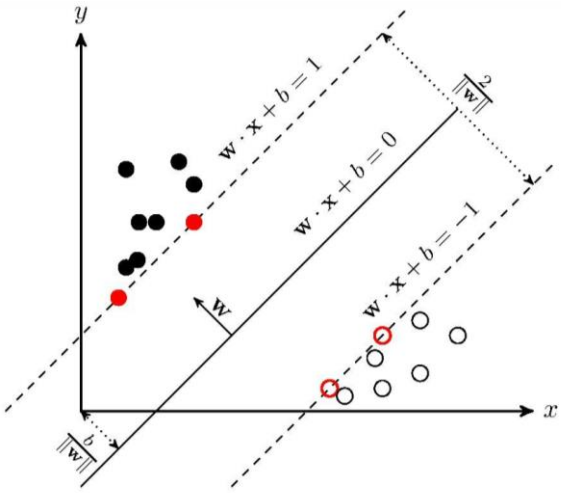

Figure 14. Support vector machine

To find the large margin classifier we need to solve the optimization problem

$$\min_{\mathbf{w},\mathrm{b}} \frac{1}{2}\|\mathbf{w}\|^2$$

$$s.t.\ y_i(\mathbf{w}^T\mathbf{x}_i + \mathrm{b}) \geq 1, \forall i$$

where $\mathbf{x}_i$ is the i_{th} input data, y_i is the ground truth, $\mathbf{w}^T\mathbf{x} + \mathrm{b}$ determines the position of the classifier.

KKT conditions can be used to solve the problem. We first get the Lagrange function

$$\mathcal{L}(\mathbf{w}, b, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 + \sum_i^m \alpha_i(1 - y_i(\mathbf{w}^T\mathbf{x}_i + \mathrm{b})),$$

then utilizing KKT conditions we get the dual problem. Solving the dual problem we get the solution $\mathbf{w} = \sum_i^m \alpha_i y_i \mathbf{x}_i$ and $\mathrm{b} = \frac{1}{|S|}\sum_{j\in S}(y_i - \sum_i^m \alpha_i y_i \mathbf{x}_i^\top \mathbf{x}_j)$, where S is the support set.

3.5 Random Forest

Random Forest produce many unique trees for learning, which can alleviate overfitting and increase accuracy. It first samples the original data with replacement (aka “bootstrap” the

training data), to obtain several diverse training data sets. Then fits an over-grown tree to each resampled training data set, each time a split is to be performed, the search for the split attribute is limited to a random subset of m of the N attributes. Finally, we aggregate the predictions of all single trees using majority or average method.

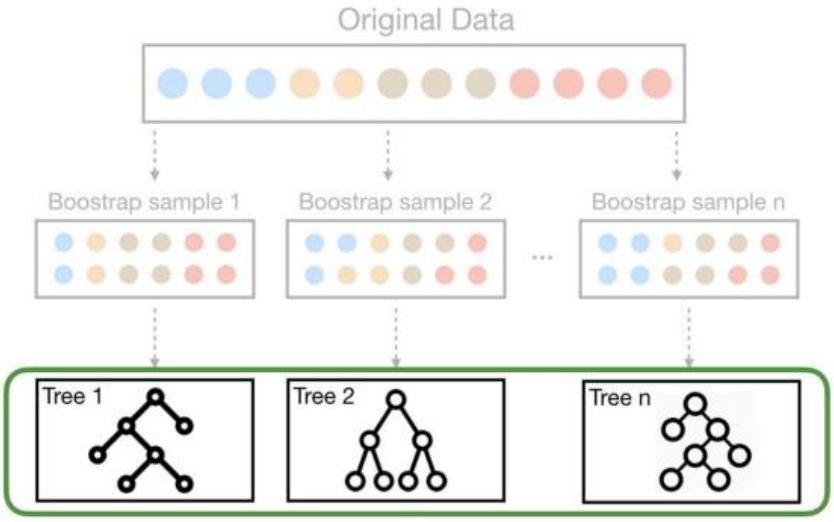

Figure 15. Random Forest

4. Results

The following set of images are the predictions of the softmax model on the testing data. As shown in the figure, the left-most figure is a normal case, and the prediction is 0 which is consistent with our ground-truth label.

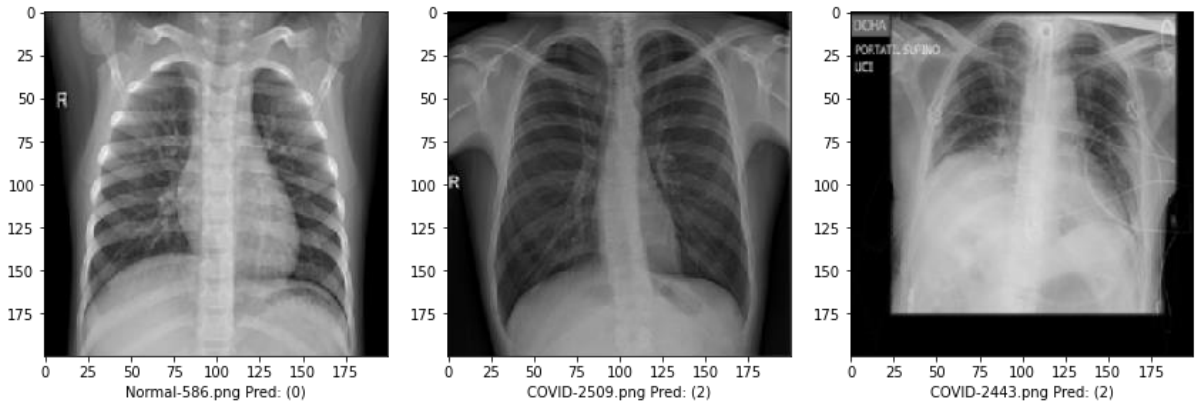

Figure 16. Predictions

Combining Cross-validation with Softmax, we get a more stable result. The accuracy also improved as shown below.

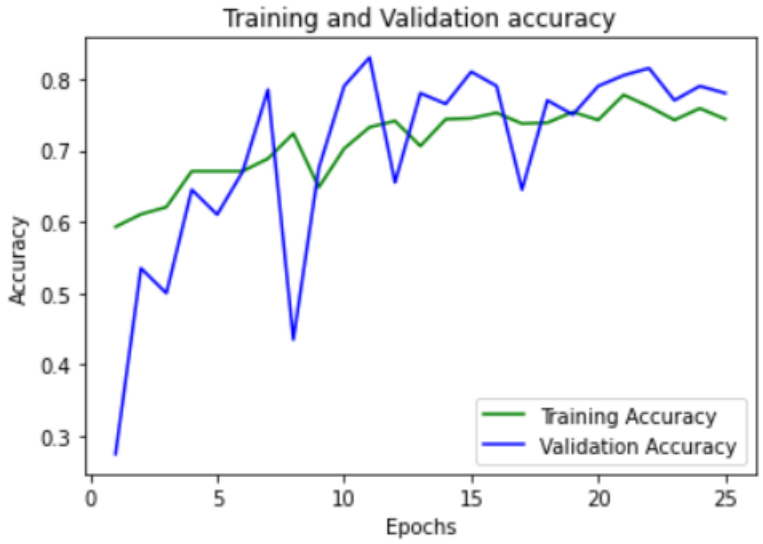

Figure 16. Accuracy after using Cross-validation

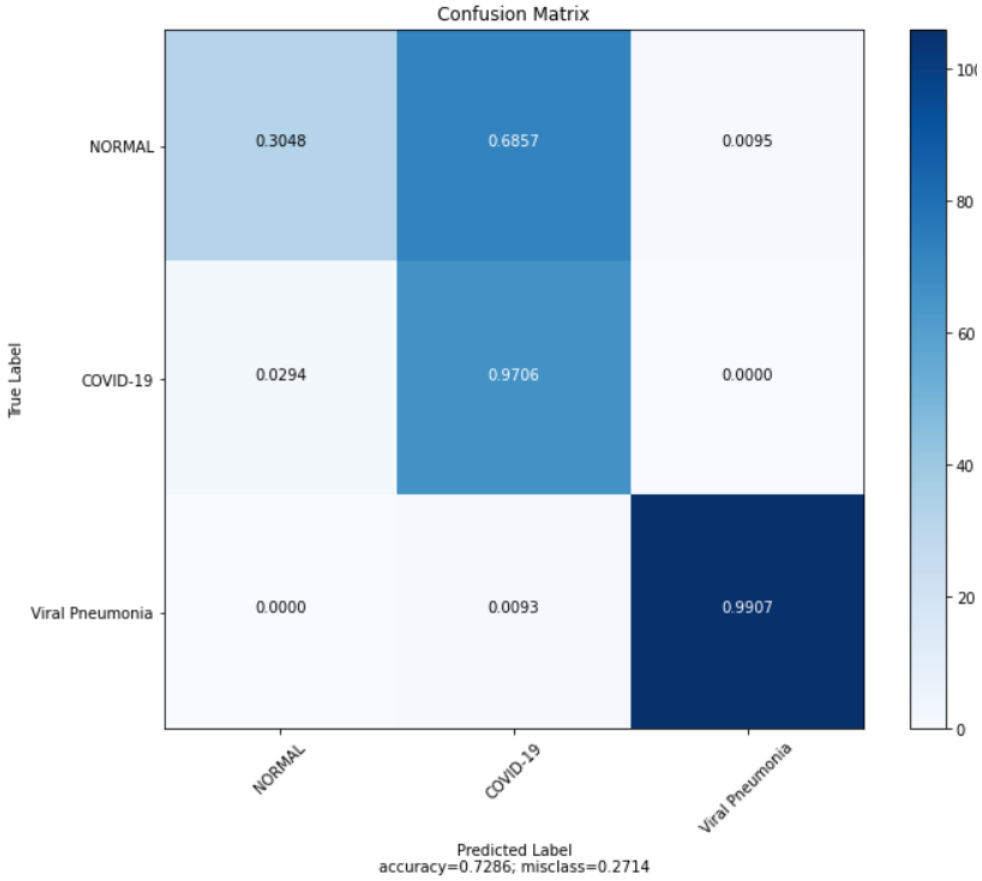

Figure 17. Confusion matrix of Cross-validation

The accuracy of predicting the normal cases is still too low. However, the value apparently increased in Random Forest and SVM model.

```
RF training accuracy: 0.9285714285714286
RF testing accuracy: 0.8324786324786325

array([[163,  12,  25],
       [  9, 191,  11],
       [  7,  34, 133]], dtype=int64)
```

Figure 18. Accuracy of the Random Forest

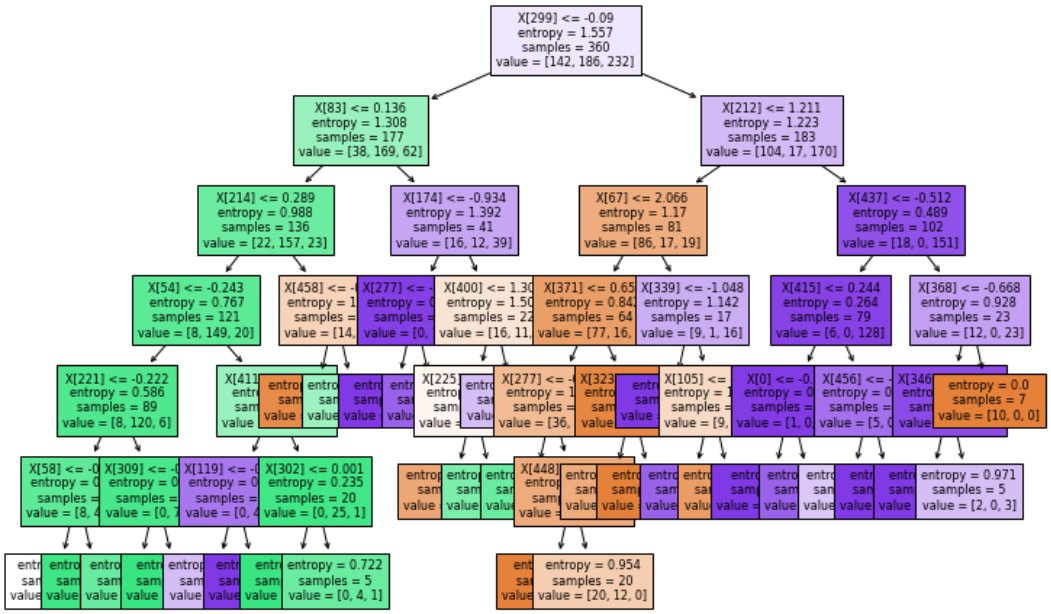

Figure 19. One decision tree in the random forest with 100 trees

```
SVM training accuracy: 1.0
SVM testing accuracy: 0.8461538461538461

array([[164,  16,  20],
       [  3, 201,   7],
       [  2,  42, 130]], dtype=int64)
```

Figure 20. Accuracy of the SVM

5. Discussion

5.1 Softmax with Cross Validation

Cross-validation can reduce the data dependence of the model as it iteratively uses four-folds out of the five as the training set. It can also choose the best hyper-parameters so that the model can fit the data well and thus make the result more stable.

From Figure 17, we can find that the Softmax model with the Cross-validation method has a high prediction accuracy for Covid-19 positive cases while having a low prediction accuracy for normal cases. We suppose that it was because the training database is not good enough, which makes some of the normal cases being mispredicted as Covid-19 positive cases.

5.2 Random Forest

Random forests in our training consist of 100 decision trees. The decision tree is a nonparametric machine learning method that does not strongly depend on the assumptions regarding the shape of the relationship between the variables. Hence, one benefit of this model is reducing the influence and bias of the pre-defined model. Moreover, the decision tree considerably increases the interpretability of how the algorithm makes decisions. However, one immediate disadvantage of a single decision tree is that the model overfits easily. To deal with this problem, we introduce random forest, an ensemble model, to improve the performance.

5.3 Support Vector Machine (SVM)

From Figure 20, we find that the average prediction accuracy for the SVM model is higher than that of Softmax. According to Andreas et al., SMV is more suitable for a small dataset, which is the case in our project (2017). It may help to explain why the SVM performs better than Softmax.

5.4 Limitations and future study

The limitations of our project are apparent. First of all, all the images are artificially selected without quantitative criteria. For example, whether the pathological features for the training data are transparent and representative. Besides, the dataset size is still tiny compared with some mature recognition systems, which may use thousands of data to train the model. When comparing the performance of the Softmax model with and without the Cross-validation method, we find that Cross-validation can improve the prediction accuracy to a certain extent. However, we did not combine the other two models, SVM and the Random Forest model, as time is not permitted. In the future study, researchers can try to cooperate with Covid-19 experts and set standard criteria to select the training data. Besides, they can also focus on combining the SVM and Random Forest model to improve the recognition system's performance further.

6. Significance

Our project improved the existing COVID-19 Positive Cases Recognition System in two aspects: improving the model's stability and prediction accuracy for Covid-19 positive cases.

The model often needs to be retrained with new data to optimize it further. The new dataset may contain some improper data, which will lead to sample excursion and reduce the model's performance. Higher stability can reduce the adverse effects of some improper samples. The previous model has not been applied to the COVID-19 diagnosis because its average prediction accuracy is about 72%. However, our project improves the prediction accuracy for Covid-19 positive cases to 84%, making the model more possible to apply COVID-19 diagnosis in practice. Thus, first of all, the results of our project may help to ease the unbalanced of medical resources distribution. Remote areas with insufficient medical experts can have a better medical diagnosis and decision-making ability. Besides, for the area with the difficult epidemic situation, our project may also help reduce the breakdown of the health system by liberating the doctors from repetitive diagnosis work and making it easier for the patients to receive care from the medical workers.

7. Conclusion

Compared to the model by Ahmed et al., our model has a significant improvement in stability and accuracy. By increasing the sample size and ignoring the improper data, the prediction accuracy of our model is stable. By modifying the model, our model gets a higher prediction accuracy, with an average prediction rate of 84% (the average prediction rate for Covid-19 positive cases, normal cases, and Viral Pneumonia cases). This improvement makes it more possible for the diagnosis model to be applied to COVID-19 diagnosis in practice. Thus, the improved COVID-19 Positive Cases Recognition System in this project may help ease the unbalanced in medical resources distribution and help reduce the breakdown of the health system.

References

Ahmed, F., Bukhari, S. A. C. & Keshtkar, F. (2020). A Deep Learning Approach for COVID-19 & Viral Pneumonia Screening with X-ray Images. Digit. Gov.: Res. Pract. 2, 2, Article 18 (December 2020), 12 pages. https://doi.org/10.1145/3431804

Andreas C. Muller and Sarah Guido, "Introduction to Machine Learning with Python: A Guide for Data Scientists", O'Reilly Media, Inc., 2017

Bird, E. 2020. Tests may miss more than 1 in 5 COVID-19 cases. https://www.medicalnewstoday.com/articles/tests-may-miss-more-than-1-in-5-covid-19- cases.

Chowdhury, M. E. H., Rahman, T., Khandakar, A., Mazhar, R., Kadir, M. A., Mahbub, Z. B. Islam, K. R., Khan, M. S., Iqbal, A., Al-Emadi, N., Reaz, M. B. I. & Islam, T. I. 2020. Can AI help in screening Viral and COVID-19 pneumonia? https://www.kaggle.com/tawsifurrahman/covid19-radiography-database.

Hilary Guite. 2020. COVID-19: What happens inside the body? https://www.medicalnewstoday.com/articles/

Nair, A. et al. (2020). A British Society of Thoracic Imaging statement: Considerations in designing local imaging diagnostic algorithms for the COVID-19 pandemic. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7156903/.

Nabeel Sajid. 2020. COVID-19 Patients Lungs X Ray Images 10000. https://www.kaggle.com/nabeelsajid917/covid-19- x-ray-10000-images.

Queensland Health. 2020. How does COVID-19 spread and how can I stop myself from getting it? https://www.health. qld.gov.au/news-events/news/novel-coronavirus-covid-19-how-it-spreads-transmission-infection-prevention-protection.

Wu, Bao yuan, "DDA2020 Machine Learning", The Chinese University of Hong Kong, Shenzhen, 2022