Distinction of COVID-19 and Analysis on Symptoms and Hospitalization Time

Xiran Gao[1, †] Guang Ni[2, †] Jingyuan Zhang[3, *, †] Xiaoning Zhao[4, †]

[1]*Maple Leaf International high school, Zhenjiang, Jiangsu Province, 212002, China*
[2] *Business administration college, University of Pittsburgh, Pittsburgh, Pennsylvania, 15260, in USA*
[3]*Mathematical science college, Jiangsu University, Zhenjiang, Jiangsu Province, 212013, China*
[4]*St.John college, Durham University, Durham, DH1 3RJ, UK*
[*]Corresponding author email E-mail: *3190113023@stmail.ujs.edu.cn*
[†] *These authors contributed equally.*

ABSTRACT
This paper aims to use the machine learning model to distinguish more precisely whether the patients get COVID-19 or not and analyze symptoms and hospitalization time of the patients. We use CNN to test the hypothesis: we can find from their X-rays that whether the patients get COVID-19. The result showed a 95 percent accuracy indicates that it can be found who are infected with COVID-19 from the model easily. It suggests that X-ray is an important and accurate indicator to find COVID-19 since respectively, X-rays results from patients with COVID-19 and normal people differ significantly. In addition, after the analysis of symptoms and time of staying in hospital, we found that patients were not likely to had no symptoms or experiencing and who had one of the symptoms accounted for the largest group of patients. The symptoms or experiencing they behaved had exact combinations rather than randomly combined, like someone may have fever, tiredness and dry-cough at the same time but he cannot have fever, dry-cough and difficulty-in-breathing simultaneously. What is more, the result also showed that the elder the patients, the longer they stayed in hospitals. The CNN model used in this study has higher accuracy. In addition, the result can help the hospitals effectively avoid the over concentration of medical resources and allocate them reasonably.

Keywords: *COVID-19, Distinction, Analysis, Machine learning*

1. INTRODUCTION

COVID-19 is the disease caused by SARS-CoV-2, first reported in December 2019, and has caused a global pandemic. The highly contagious disease is pummeling the global health system with strained hospital capacity and caused the death of 765,722 in Unite State alone as of October 29 2021. Globally, as of 4:40pm 29 October 2021, there have been 245,373,039 confirmed cases of COVID-19, including 4,979,421 deaths. In addition, the pandemic has brought a series of unintended consequences, such as restriction of human interaction created socioeconomic hardship and poverty caused by economic recession in the long term [1]. The unemployment rate rocketed very fast in the first half of 2020 and the highest rate was about 15% in the U.S. [2].

This pandemic continuous challenging the global medical system, in terms of insufficient medical resource and equipment, low efficient of resource allocation and infections of medical stuff [3,4]. Thus, accurate and efficiently identify infections by machine and effectively allocate medical resource is curial. Machine learning is an effective method to apply in this kind of tasks [5].

As the patient cases continues accumulating, patient data become an available and crucial resource for researcher to predict and identify disease through machine learning. Machine learning approach already applied in predicting the number of new cases will rise [4], screening infections [6], stratify the mortality risk of patient [7].

We try to implementing machine learning to predicted covid patient hospital stay time to help hospital optimize their resource allocation. We sperate this problem into two steps: first step is use convolution neural network (CNN) to analysis chest x-ray image; second step is use

traditional analysis to give general analysis of result. The reason we can't implement machine learning is the trail of patient is highly homogenization which means normal classification method like logistic regression, CatBoost and neural network is unable to separate data. By using CNN we can sperate covid patient and normal people with very good precision. Our paper is use to reduce wasting hospital resources and optimize it which can let more covid patient to get treatment they needed.

The rest of this paper is organized as follows. Section 2 introduce the data and variables we used in build model of distinguish X-ray and predict hospital time respectively. Section 3 presents the deep learning method used in this paper. Section 4 discusses the experimental results. Conclusion are given in Section 5.

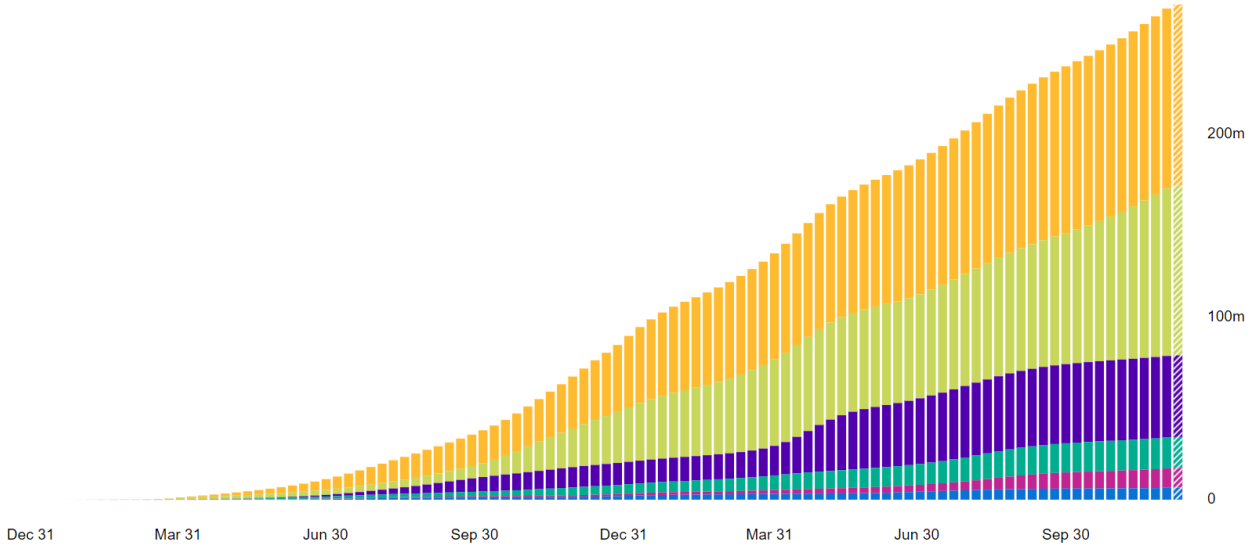

Figure 1 Globally situation of COVID-19

Notes: The source came from World Health Organization. Parts of different colors from top to bottom represents the patients confirmed in America, European, South-East Asia, Eastern Mediterranean, Western Pacific and Africa. The data may be incomplete for the current day or week.

2. DATA AND VARIBLES

2.1. Overview

The datasets that we use include X-rays of normal people and patients, the symptoms that the patients behave and hospitalization time and other data of patients. All the datasets can be found and downloaded at Kaggle (www.kaggle.com).

2.2. X-rays

It shows the pulmonary x-rays of normal people and patients. We downloaded two different datasets about x-rays of both the normal and the infected. The one with less pictures was used to train and the other one was used to test.

When it comes to the features of X-rays from people with COVID-19, it has been divided into 4 different stages. In the early time, chest imaging changes caused by negative or basic diseases. Then it can progress rapidly in a short time (1~3 days), which is characterized by increased and disordered bilateral lung markings, intertwined into grid or honeycomb, especially in the lower lobe of both lungs. After that, in the severe stage, diffuse interstitial lesions in both lungs with patchy and patchy shadows of increased density. Finally at the recovery stage, the scope of lesion is reduced and residual fiber strips [8].

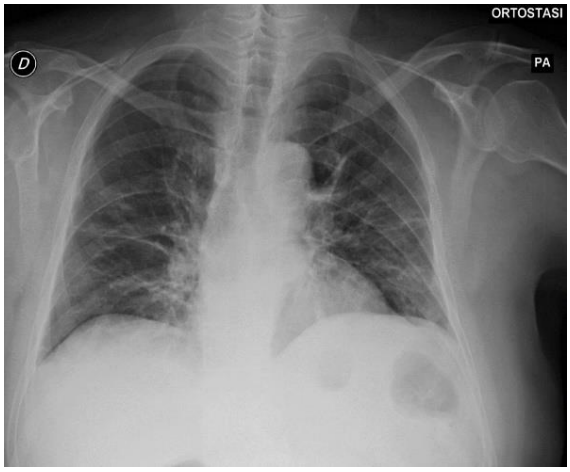

Figure 2 X-ray of People with COVID-19

Figure 3 X-ray of Normal People

Notes: These two figures came from the dataset that could be downloaded at Kaggle.

2.3. Symptoms

This part shows how the patients with COVID-19 behave. There are 316,800 sets of patients' data in the dataset. The symptoms include fever, tiredness, dry-cough, difficulty-in-breathing, sore-throat, pains, nasal-congestion, runny-nose and diarrhea. It also includes those who do not have any symptoms or experiencing. The dataset also gives us the information like the gender, severity and whether he or she had contacted with other patients. This group of data is as of November 7, 2020.

Table 1. Information about Patients

Category	Content
Symptoms	Fever, Tiredness, Dry-cough, Difficulty-in-breathing, Sore-throat, None
Experiencing	Pains, Nasal-congestion, Runny-nose, diarrhea, none
Country	China, France, Germany, Iran, Italy, Other-EUR, Republic of Korean, Spain, UAE
Gender	Male, Female, Transgender
Severity	Minor, Moderate, Severe, Extreme
Contact with other patients	No, Little, Yes

2.4. Hospitalization Time

This part shows the hospitalization time and other extra data of the patients. It includes 318,438 sets of data. The dataset shows the case ID in the hospital, the code of the hospital, the type code of the hospital, the city code of the hospital, the region code of the hospital, the number of extra rooms that the hospital has, the hospitalization time, etc.

Table 2. Information about patients

Category	Content
Severity	Minor, Moderate, Extreme
Age	0~100
Hospitalization time	0~More than 100 days

Notes: Here only shows the part that we use.

3. METHODOLOGY

After preprocessing the image, each image becomes a three-dimensional array, which has shape (1000, 1000, 1) with a single label value in sample space (0, 1). This processed gray scale image can be the input of the function.

The CNN model is a type of artificial Neural Network, most commonly applied to classified visual imagery. CNN were inspired by animal visual cortex, which a single neuron only responds to a restricted region of field of view. CNN model has similar layer that mimic this behavior which called Convolutional layer, kernel (or filter) is filed of view of activation function (neuron). We will use our model as example to tell how this CNN work, and explain why we choose them for our project.

Here is the graph of model we using to distinguish the COVID patient chest x-ray image and normal chest x-ray image.

Graph show below is an overview of the CNN model we using to analysis the x-ray graph. Each node represents different layer in the CNN mode. Node name with Convolutional Layer and Max Pooling is layer doing computation. Then node name with input image, feature maps is use to demonstrate shape of intermediate stage. There is two more layer between Flatting and output, and that will show in next graph. In flattening part, we using two fully connected layer to reduce the parameter size, and first layer has 300 neuron, second layer has 64 neurons. The layer with orange color is output layer which has probability value of yes no.

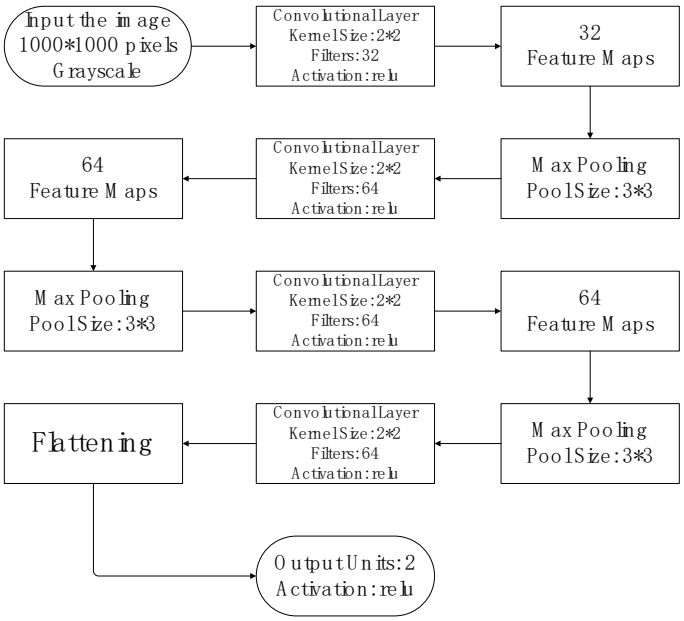

Figure 4 Introduction of the model

Here is the detail of the flattening part.

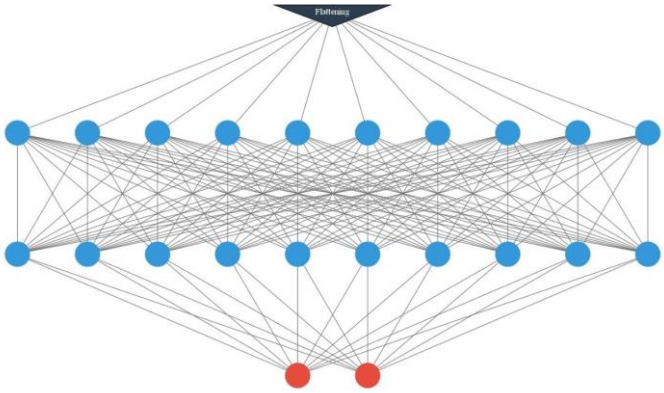

Figure 5 Detail of 'Flattening' in Figure 4

First is the input shape, after preprocess 94 picture in the training set. We turn those pictures into a tensor store in a four-dimensional array (94, 1000, 1000, 1). First represent number of pictures to train, upcoming two number is length and width of the picture, last is channel number. For example, the value at (0, 0, 0, 0) is gray-scale level of the first picture's pixel at (0, 0).

After take the input data, we use a convolution layer to execute the convolution. We choose to use the smallest kernel (2, 2) with 32 layers filter. Use multiple layers to reduce overfitting issue. This step also includes an activation layer with Rectified Linear Unit (ReLU) activation function.

The layer come after convolution is a maxpooling layer. This will choose highest value in the feature map that generated by the convolution layer. Maxpooling means we will down sampleing the feature which can reduce the reasource during the training. This combine of convolution and maxpolling will repeat for three times.

The 'Flattening' layer will convert the feature map to normal neural network. We add two fully connected layer (the blue points in Figure 5) to give a smooth transition to output layer give below. This output layer only have two nodes (the red points in Figure 5), which is COVID and normal.

Table 3. General view of layers in CNN

Layer(type)	Output Shape	Param #
conv2d(Conv2D)	999, 999, 32	160
max_pooling2d	333, 333, 32	0
conv2d_1(Conv2D)	332, 332, 64	8256
max_pooling2d_1	110, 110, 64	0
conv2d_2(Conv2D)	109, 109, 64	16448
max_pooling2d_2	36, 36, 64	0
conv2d_3(Conv2D)	35, 35, 64	16448
flatten(Flatten)	78400	0
dense(Dense)	300	23520300
dense_1(Dense)	64	19264
dense_2(Dense)	2	130

4. RESULT

4.1. Distinction by X-rays

a. Data preprocessing

To analyze the result of the X-Ray, the first step is to determine how the image was processed. At the first sight, our goal was to consider the image's size to be two thousand width times two thousand height pixels, however, the sever that can run this size of image could not be found. For instance, we change the size to one thousand width times one thousand height pixels and since the image should be black and white, it was run through the single tunnel. The final size of the image imputed into the model is 1000 times 1000 times 1, which respectively indicates length, width and gray scale. The output from the model is a one-dimension value which is basically yes or no.

Also, as we have two different datasets of x-rays, we used the one which have less pictures as the training set and the other one as the test set. What's more, we set 10 epochs to do this machine learning.

Table 4. Specific division of datasets

	Normal	COVID-19	Total
Training	25	69	94
Test	28	70	98
Total	53	139	192

b. Distinction Results

After training the program to distinguish the normal people and patients, we tried to let itself to find who are infected and who are not. Actually the result is gratifying. The accuracy in identifying patients is almost 99% at last, which means we can find who are infected with COVID-19 from x-rays easily.

Compared with current researches, most of them use computed tomography (CT) to check whether people have COVID-19 as CT is more advanced and it can give doctors more information. But X-rays are cheaper and convenient [9]. What's more, without machine learning, it requires an experienced doctor to diagnose quickly. It costs a lot of time and labor power. It can be another choice as a result of the greatly high accuracy.

What's more, CNN is a good method to do the distinction as it is more accurate than many other methods. For instance, the accuracy of MobileNet is about 94.90% [10]. This shows the advantage of CNN.

Table 5. Accuracy of machine learning

Epoch	Loss	Accuracy
Epoch1	0.4799	0.7979
Epoch2	0.2338	0.9362
Epoch3	0.1863	0.9468
Epoch4	0.1005	0.9787
Epoch5	0.0779	0.9894
Epoch6	0.1440	0.9574
Epoch7	0.1776	0.9574
Epoch8	0.2274	0.8830
Epoch9	0.0574	0.9894
Epoch10	0.0702	0.9894

4.2. Symptoms Analysis

After analyzing the data of symptoms, we find something useful.

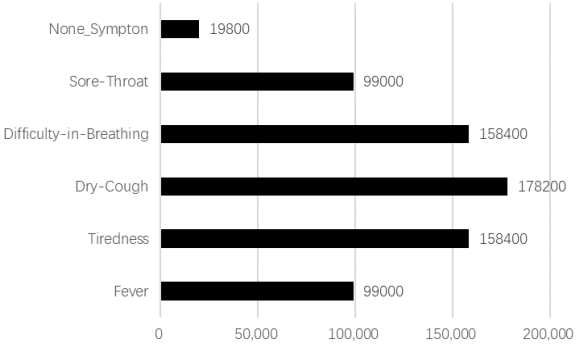

Figure 6 Numbers of the patients with different symptoms

Notes: Symptoms include fever, tiredness, dry cough, difficulty in breathing and sore throat. The unit of the data is 'people'.

From Figure 6, it is easy to see that as there are 316,800 patients, more than half of them had dry cough. Half of them had difficulty in breathing or feel tired. About one in third of the patients had fever or sore throat. Only about 6% of the patients have no symptoms.

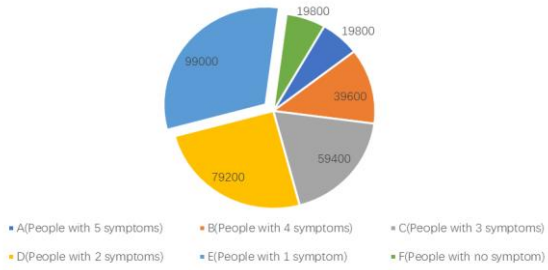

Figure 7 Numbers of patients have different quantities of symptoms

Notes: The unit of the data is 'people'.

Table 6. Possible combinations of symptoms

Number	Combination
4 Symptoms	1. fever, tiredness, dry-cough and difficulty-in-breathing 2. tiredness, dry-cough, difficulty-in-breathing and sore-throat
3 Symptoms	1. fever, tiredness and dry-cough 2. tiredness, dry-cough and difficulty-in-breathing 3. dry-cough, difficulty-in-breathing and sore-throat
2 Symptoms	1. fever, tiredness 2. tiredness, dry-cough 3. dry-cough, sore-throat 4. dry-cough, difficulty-in-breathing

From Figure 7, we can see that those who had only one symptom account for the majority. It is about 31.3%. Those who had 2 symptoms followed and accounted for a quarter. But for those who had all 5 symptoms or had no symptom, they both accounted for the smallest part, which was about 6%. For other parts, C accounted for about 19% and B accounted for 12.5%.

From Table 6, we can see what kind of symptoms the patients with 2 symptoms, 3 symptoms and 4 symptoms might have. It is easy to see that not every combination was possible.

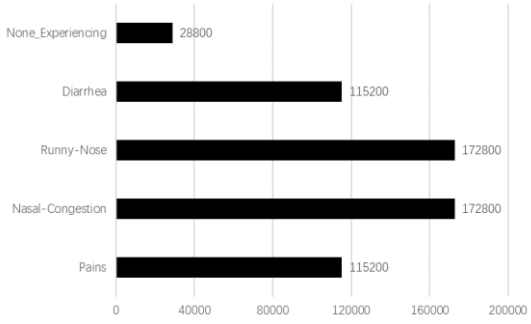

Figure 8 Numbers of the patients with different experiencing

Notes: Experiencing includes diarrhea, runny nose, nasal congestion and pains. The unit of the data is 'people'.

From Figure 8, it is obvious that about 55% of patients had runny nose or nasal congestion. About 36% of the patients had diarrhea or pains. Only less than 10% of the patients had no experiencing.

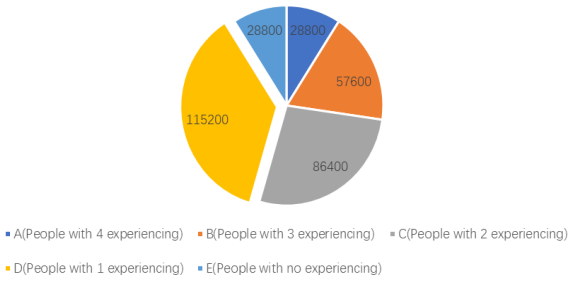

Figure 9 Numbers of patients with different quantities of experiencing

Notes: The unit of the data is 'people'.

Table 7. Possible combination of experiencing

Number	Combination
3 Experiencing	1. pains, nasal-congestion and runny-nose 2. nasal-congestion, runny-nose and diarrhea
2 Experiencing	1. pains and nasal-congestion 2. nasal-congestion and runny-nose 3. runny-nose and diarrhea

From Figure 9, we can see different proportions of patients with different quantities of experiencing. Those who had only one experiencing accounted for about 36%. Similar to symptoms, those who had all 4 experiencing or had no experiencing accounted for the smallest part which was only about 9%. When it comes to other parts, C accounted for about 27% and B accounted for about 18%.

From Table 7, we can see what kind of experiencing the patients with 2 or 3 experiencing might have. Same to symptoms, not all the combinations were possible. Evidently, the number of patients who had no symptoms or experiencing was very small. It might because the data was from the early stage of the epidemic. There were not so many patients who have no symptoms or no experiencing as now. What's more, for both symptoms and experiencing, those who only have one of them accounted for the largest part.

4.3. Hospitalization Time Analysis

At first, we studied about the relationship between ages and the severity.

Table 8. Number of patients

Age	Minor	Moderate	Extreme
0-10	2530	3066	658
11-20	8027	7383	1358

21-30	15406	21200	4237
31-40	16462	36958	10219
41-50	12933	37423	13393
51-60	10513	27534	10467
61-70	8199	18273	7215
71-80	9686	19068	7038
81-90	1840	4208	1842
91-100	276	730	296
Total	85872	175843	56723

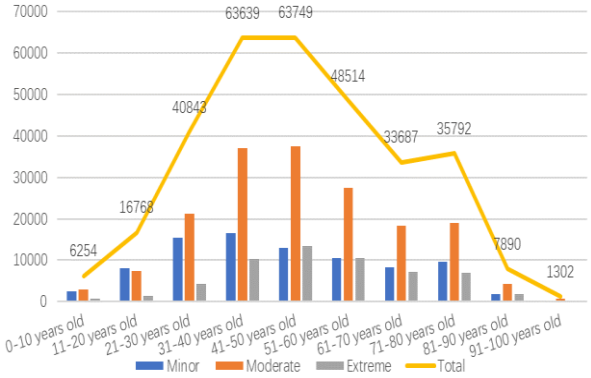

Figure 10 Numbers of the patients at different ages

Notes: The unit of the data in Table 8 and Figure 10 is 'people'.

From Table 8, we can see that patients with moderate severity accounted for the largest part, which is more than 55%. Those with minor severity second only to them and accounted about 27%. In addition, it is easy to find that with the increase of age, the number of patients increased first and then decreased, which is showed at Figure 10. People aged 31 to 50 were more easily to be infected.

Then we continued to learn about the relationship between severity and time of staying in hospital by statistical methods.

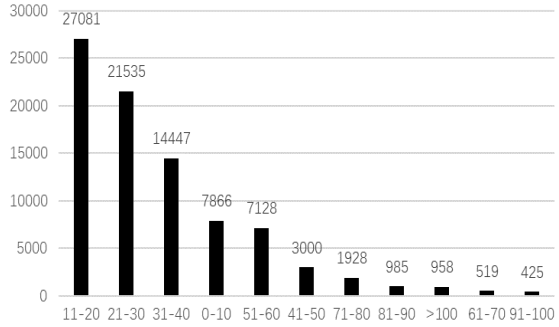

Figure 11 Hospitalization time of minor severity

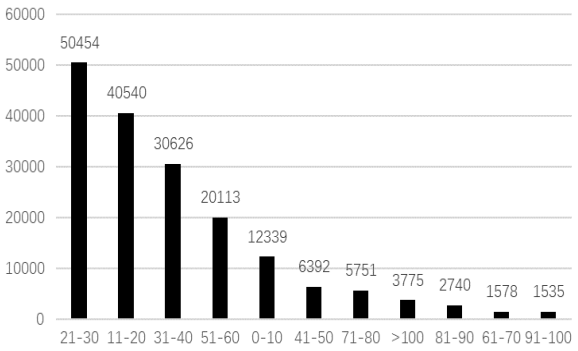

Figure 12 Hospitalization time of moderate severity

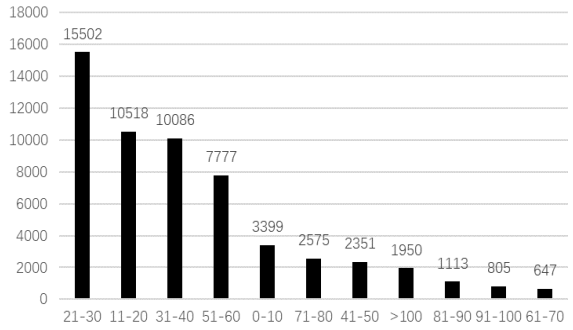

Figure 13 Hospitalization time of extreme severity

Notes: The unit of data in figure 10, figure 11 and figure 12 is 'people'. The transverse axes of the figures represent hospitalization time.

From the three figures above, we can see that as the severity got higher, the time that the patients stayed in hospital become longer. Moreover, most of the patients stayed in hospital less than 60 days.

Also, we did some study about the relationship between different ages and hospitalization time.

After throwing the data of 'More than 100 days', we calculated expectations of hospitalization time. Then we get the result below.

Table 9. Hospitalization time of patients of different ages

	0-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100	>100
0-10	615	1959	1489	1014	187	582	26	153	84	35	110
11-20	1152	5343	4312	2681	510	1429	89	350	223	71	208
21-30	3467	11272	11394	6912	1398	1026	263	1026	546	231	541
31-40	4916	15792	18550	10912	2373	6517	509	1807	801	484	978
41-50	4727	14959	17906	10983	2507	7189	562	2146	885	578	1307
51-60	3428	11346	13508	8569	1735	5739	448	1710	784	499	1199
61-70	2194	7870	9033	5930	1205	4081	325	1230	600	330	889
71-80	2201	7958	9534	6420	1383	4433	378	1367	670	386	1062
81-90	422	1392	1920	1504	379	1082	115	402	216	132	326
91-100	83	248	295	234	66	173	29	63	29	19	63

Notes: The transverse axis represents hospitalization time and the vertical axis represents age. The unit of the data is 'people'.

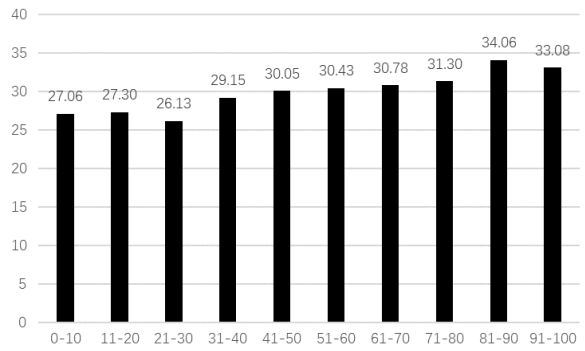

Figure 14 Expectation of patients' hospitalization time of different ages

Notes: The unit of the data is 'day'.

It is clear that with patients' ages get larger, the time they stayed in hospital gets longer.

Then here might be the process of going to the hospital for check. First you need to do an X-ray. If the result is that you are not infected with COVID-19, then you go back home. But if you are infected with COVID-19, then your information will be checked. For example, your age, your symptoms and much other information will be asked to provide to the doctor and the hospital to estimate how long you will stay in the hospital and how many medical resources need to be provided to you. With the age becomes larger and the number of the symptoms becomes bigger, hospitalization time may become longer and the medical resources required may become more. For example, if there is a patient with COVID-19 who is 25 years old and with 3 different symptoms, then he may stay in hospital for about 26 days and need more medicine care than those who have less symptoms. The result may also depend on other factors like the severity, something about the patient himself (time from infection to onset, autoimmunity, etc) [11] and so on.

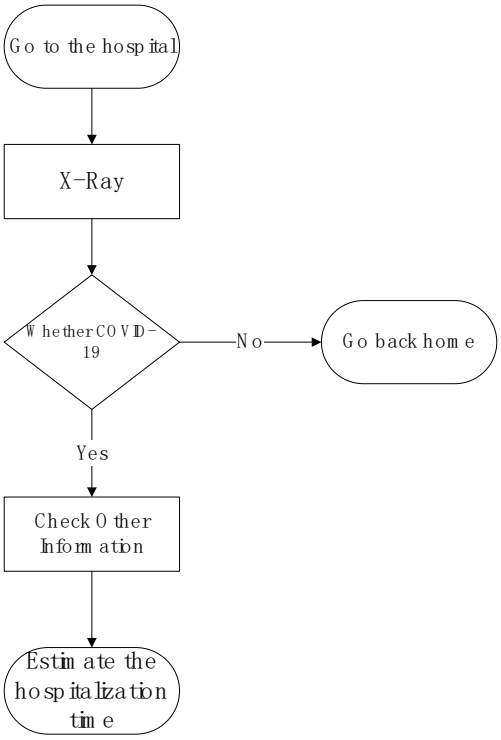

Figure 15 The process of going to the hospital for check

5. CONCLUSION

In this paper, we want to distinguish whether a person is infected by COVID-19 and analyze the symptoms and hospitalization time. We use different method to pursuit the result we looking for. We use CNN model to efficiently identify the COVID patient. We try to use machine learning COVID patient symptom to find out the

time they will stay in the hospital, but because the symptom of the COVID patient is highly homogenization, which means there is hardly to find a viable machine learning model to make distinction. Thus, we choose to use traditional data analysis method to give a general stay time for different level of severity. To sum up, we used three different datasets which were X-rays, symptoms and hospitalization time from different people and patients. X-rays shows the pulmonary x-rays of normal people and patients, symptoms show how the patients with COVID-19 behave and hospitalization time shows the time that patients stayed at the hospital. We input these datasets into the machine learning model, and the accuracy we got in identifying patients is higher than 95 percent which means that we can find who are infected with COVID-19 from X-rays easily. Moreover, when it turns to the symptoms, we concluded that patients were not likely to had no symptoms or experiencing and who had one of the symptoms accounted for the largest group of patients. Finally, for hospitalization time, we summarized that the elder the patients, the longer they stayed in hospitals.

According to the study, we could distinguish more precisely whether the patients get the COVID-19 or not since the accuracy in our study is higher than 95 percent. Secondly, since we exam the different patients with different symptoms and experiencing, we could allocate medical resources to the patients better. Thirdly, the results from the study could improve the efficiency when the doctor treat the patients.

However, there are still some disadvantages of our study. One deficiency is the accuracy of information from data collected since the yes or no symptoms cannot be used efficiently in machine learning. However, if the hospital can score the collecting symptoms in precise number, then, we can use the data to predict. Moreover, it would be useful if there is a machine learning algorithm model that can classify this homogenized feature. Another deficiency is the problem from the CNN model. Due to the inadequate computing resources, we chose to down sample some of the images from two thousand width times two thousand height pixels to one thousand width times on thousand height pixels which may leads to lack of some information that may affects the stability of the model. However, it could not be discovered at the moment. The third deficiency is that we are unable to categorize the symptoms and the reason is that the homogeneity of patients' symptoms is critical.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHORS' CONTRIBUTIONS

Xiran Gao and Jingyuan Zhang collected the data; Jingyuan Zhang and Xiaoning Zhao analyzed the data and did the research. Guang Ni summarized the results. Guang Ni, Jingyuan Zhang, Xiaoning Zhao and Xiran Gao wrote and revised the paper all authors had approved the final version.

REFERENCES

[1] N. Diffenbaugh, C. Field, E. Appel, et al., "The COVID-19 lockdowns: a window into the Earth System," *Nature Reviews Earth & Environment*, vol. 1, pp. 470-481, 29 July 2020.

[2] W. Gui and J. Zhao, "Impact of COVID-19 on unemployment rate of China and US," *The World of Survey and Research*, no. 11, pp. 15-26, November 2021.

[3] I. Miller, A. Becker, B. Grenfell and C. Metcalf, "Disease and healthcare burden of COVID-19 in the United States," *Nature Medicine*, vol. 26, pp. 1212-1217, 16 June 2020.

[4] A. Khakharia, V. Shah, S. Jain, et al., "Outbreak Prediction of COVID-19 for Dense and Populated Countries Using Machine Learning," *Annals of Data Science*, vol. 8, pp. 1-19, 16 October 2020.

[5] W. Sun, B, Zheng and W. Qian, "Automatic feature learning using multichannel ROI based on deep structured algorithms for computerized lung cancer diagnosis," *Computers in Biology and Medicine*, vol. 89, pp. 530-539, 1 October 2017.

[6] Y. Zoabi, S. Deri-Rozov and N. Shomron, "Machine learning-based prediction of COVID-19 diagnosis based on symptoms," *npj Digit. Med*, vol. 4, pp. 1-5, 4 January 2021.

[7] Y. Gao, G. Cai, F. Wei, et al., "Machine learning based early warning system enables accurate mortality risk prediction for COVID-19," *Nature Communication*, vol. 11, pp. 1-10, 6 October 2020.

[8] L. Wang, Z. Qiao, W. Liu, et al., "Guidelines for imaging diagnosis of New Coronavirus pneumonia (Second Edition 2020 Edition)," *Journal of Capital Medical University*, no. 2, pp. 168-173, 2020.

[9] F. Wu, S. Zhao, B. Yu, et al., "A new coronavirus associated with human respiratory disease in China," *Nature*, vol. 579, pp. 265-269, February 2020.

[10] X. Zhao, C. Shi, T. Huang and X. Ji, "Lung image classification based on attention mechanism and convolutional neural network," *Journal of Qiqihar University(Natural Science Edition)*, no. 4, pp. 27-32, 2021.

[11] P. Jia, S. Gao, P. Liu, S. Zheng, Z. Cai and X. Du, "Analysis of duration and its influencing factors about COVID-19," *Journal of Modern Medicine & Health*, no. 11, pp. 1800-1803, 2021.

BIOGRAPHY

Guang Ni was born in Wuhan, China on October, 10, 2000. He currently majors in Finance at University of Pittsburgh which locates at Pittsburgh, Pennsylvania, United States, and the expected graduation date is in April 2022.
He used to work as a CALCULUS TUTOR at University of Pittsburgh at Johnstown in 2019 for 3 months.
His research interest is on quantitative trading field which uses finance, statistics, mathematics modeling and software engineering to trade on stocks or cryptocurrency.

Jingyuan Zhang was born in Changzhou, Jiangsu Province, China on 21 April, 2001. He is now studying at Jiangsu University in Zhenjiang. He majors at applied mathematics.
His research interest is about statistics, mathematics and business analysis.

Xiaoning Zhao was born in Taiyuan, China on April 22, 1998. He graduated for Temple University with Bachelor of Science in computer science, and currently is master student of Durham University in scientific computing and data analysis.
He focuses on study of implementing machine learning on financial time series model like stock or option.

Xiran Gao was born in Nanjing, Jiangsu Province in China on April 29, 2003. Now she is studying in Mapleleaf International high school in Zhenjiang, Jiangsu Province.
She focuses on the study of statistics, economics, and financial analysis.