APPLY DEEP LEARNING FOR DETECTING COVID-19 PATIENT FROM A LARGE DATASET OF LUNG CT IMAGES

Tran Dai Chi
Computer Science Student
VNUHCM - University of Science
18127070@student.hcmus.edu.vn
ORCID: 0000-0001-6345-8658

Phan Tan Dat
Computer Science Student
VNUHCM - University of Science
18127078@student.hcmus.edu.vn
ORCID: 0000-0002-1967-3884

April 26, 2021

ABSTRACT

Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus. It is a serious global problem and as of today, more than 161 million infections and 3 million deaths worldwide. One of the main ways we can reduce the damage caused by disease is early detection of an infected person and AI can play a key role that could help many hospitals increase number of patients can be tested quickly to avoid overload. In this paper, we want to propose a method based on neural network to help detect COVID-19 from lung CT scan of patient quickly without having to go through too many other original tests. Our dataset contains 48260 CT scan images which is collected from 282 normal persons and 15589 images from 95 patients with COVID-19 infection. Our model will takes the entire CT scan of a patient as input to determine if the patient is positive or not. At an early stage, because the number of images we use for training is large, our model will run an image processing algorithm to remove the images that cannot be seen in detail. For example, a lung CT image usually consists of two black slices and if a patient is positive with COVID-19, the most common manifestation we can see on CT chest is a Ground glass opacity, Consolidation, or Crazy paving pattern...[1] However, if the two black slices in our CT scans are too small it will be difficult to recognize, so we must ignore these images when training the model to reduces the number of recognized images, thereby reducing processing time. In the next phase, we will use the ResNet50V2 model, an improved form of ResNet50 enhanced by a feature pyramid network (FPN) to increase the accuracy of the classification and our model has improved its accuracy up to more than 98%, a rather high result when compared to just applying the model and classification as usual.

Keywords Deep learning, COVID-19 diagnosis, lung CT scan, Medical image analysis

1 Introduction

The first patient of this virus originated in Wuhan city, China. The cause was initially thought to originate in the wildlife trade and bats were the first suspect species, suggesting that the possibility of transmitting the virus from animals to humans is still a very dangerous problem [2]. This outbreak quickly spread to other provinces in China and then globally. It disrupted the world economy, affect everyone's lives and caused a lot of other serious damage.

The most important feature of the virus is its ability to spread rapidly through the air, even on the fingerprints of an infected person [3]. According to WHO statistics, the death rate caused by this virus is not high, but due to its widespread spread, identifying at-risk patients to quickly quarantine play an important role in preventing the spread [4].

The common symptoms of the virus are: fever, dry cough, fatigue. There are also other serious symptoms such as: difficulty breathing, pain or tightness in the chest, loss of ability to speak or move [5].

The most popular method of rapid test results today is Realtime-PCR [6], but the accuracy of this method varies widely depending on the sampling technique and the expertise of the investigator [7]. In addition, the lack of rapid test kits in many parts of the world is making researchers give more effective ways to diagnose this diseases.

In most infected patients, lung damage from CT scans can help diagnose the disease and is due to the availability of medical imaging devices in most medical centers, hospitals, many other methods has been proposed by scientists based on the use of these images [8, 9].

Most patients with COVID-19 can be detected within 15 days [10], however it can take shorter or longer, so using these CT scans helps to make an early diagnosis of whether a patient has a positive or not. In addition, the Realtime-PCR method may give a negative result once or twice for the first time, but later times it may give a positive or we call weak positive result [11], so the prediction use CT scans can be seen as an effective method for early diagnosis.

One of the benefit of using medical imaging is that we can take the advantages in the field of machine learning and deep learning methods have also been widely applied in many different fields previous to increase productivity and minimize human errors [12].

In this article, we will introduce a deep learning-based approach to quickly detect whether a patient is positive or not based on a series of lung CT scans of a patient from which help doctors quickly test a patient and limit errors when compared to manual inspection.

The summary of all work in this paper is shown in fig. 1.

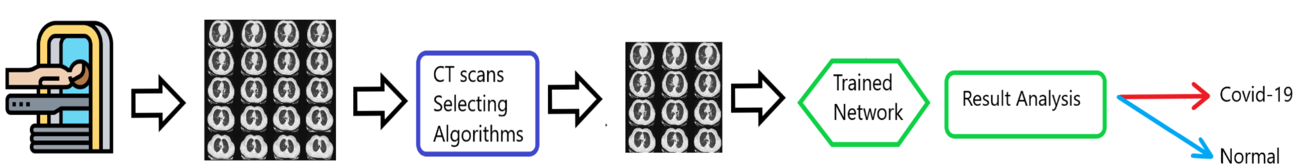

Figure 1: Our proposed method for the classification.

The remainder of the paper will cover: In section 2, we will talk in detail about the dataset, the deep learning model that has been used for classification. In section 3, we will discuss the experimental results, In section 4, the paper is discussed. Section 5 is the conclusion of the paper and finally, the path to the code and the dataset is provided as well as reference links.

2 Materials and methods

2.1 COVID-CTset

COVID-CTset is the primary data set we use for training in TIFF format with a resolution of 512 * 512 pixels [13]. One of the special features of this data set is the use of an image in 16bit format instead of 8bit to help improve the method's outcome and because the image we use is a medical one, so it requires for a high degree of precision so we will use images in TIFF format which is lossless instead of JPG or PNG. TIFF images have float values whereas JPG and PNG images have integer values, so it is important to preserve this precision when training the model. Some of the images of our dataset are presented in fig. 2.

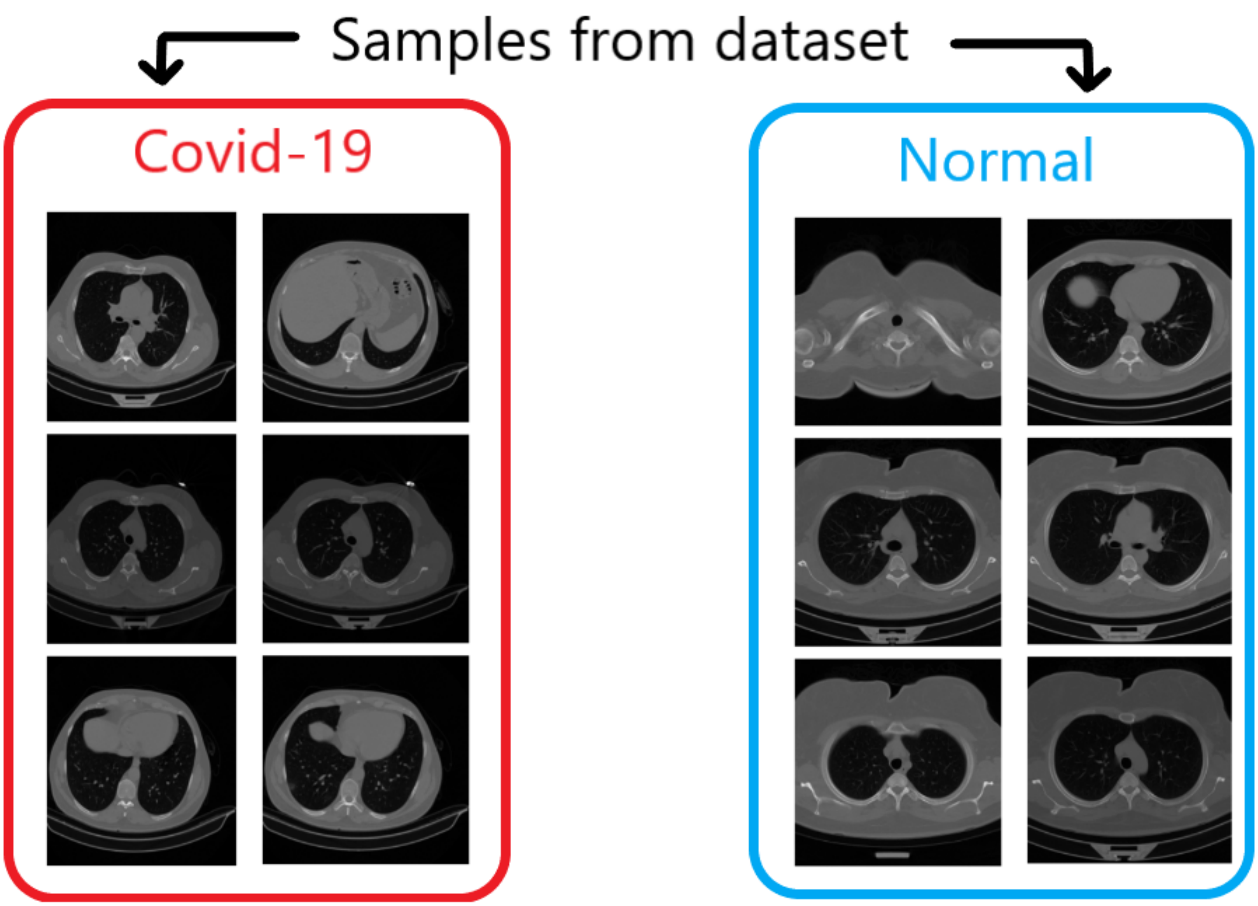

Figure 2: Some of the images in COVID-CTset

COVID-CTset was made up of 15589 images of 95 infected patients and 48260 images of uninfected patients (table 1). The distribution of the patients in COVID-CTset is shown in fig. 3

	COVID-19 Patients	Normal Patients	COVID-19 Images	Normal Images
Number	95	282	15589	48260

Table 1: COVID-CTset data distribution

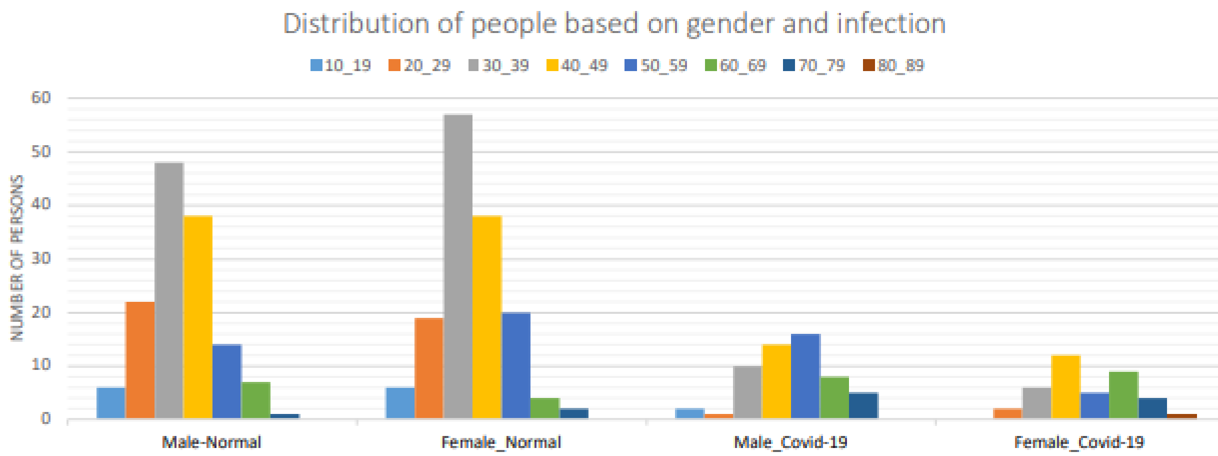

Figure 3: The number of patients based on age, gender and infections.

2.2 CT scans Selection

High-resolution computed tomography (HRCT) is a type of computed tomography (CT) with specific techniques to enhance image resolution [14]. The lung HCRT scan devices will take a series of images from the chest of patient, then check if this patient is infected or not. The doctor will use these images and if he finds on some of them signs of infection, it is likely that the patient is positive with COVID-19.

Here, we will use a trained neural network to classify COVID-19 patients based on lung CT images of that patient selected by our image processing algorithm to make the inside of the image clearly visible [15]. If we were to use the entire image of a patient for training, as depicted in fig. 4, in the early and late stages we would not be able to see the details inside the lungs which could lead to error if we use these images to train the model.

Although we can divide the input data set into three categories: infection-visible, normal and not clearly visible image, if we do, we will incur a fee for re-labeling the entire image, for this reason our image processing algorithm will remove the images inside the lungs that are not clearly visible, which can reduce implementation costs.

Fig. 6 shows the steps of image processing algorithm. As we can see in fig. 5, the main difference between open and closed lungs is that the area of open lung is more black, so we will put an area in the middle to analyze the pixel values in it. Since the lung positions of the patients are different and our picture has a resolution of 512 * 512, so the appropriate position for our region is 120 - 370 pixels in the x axis and 240 - 340 in the y axis [16]. Fig. 7 shows the selected area in different images. In the next step, we will measure pixels in the selected area of each image and if the value is less than 300, they are called dark pixels [17].

For all lung CT image of the patient, we will count the sum of the values in the selected area if the value is less than 300, then we will save the whole count on the images into an array. Next, we will set threshold according to the formula: (max value of array - min value of array) / 1.5. Finally, we will take the index of counter from count array and compare the each index with the threshold, if the index is greater than the threshold, it means that image is considered to have more dark area pixels and such images will be selected for training the model.

In fig. 8, we can look at a series of CT scans of a patient where we can see which images will be removed or selected by the algorithm.

2.3 Deep convolutional neural networks

As we know, computer vision is a long-developed sphear and has many wide applications in many different fields such as biology, medicine, geography...[18] However, its implementation on deep neural networks has yielded an impressive performance [19]. In this research, we will use deep convolutional network to classify a series of lung CT images of a patient to know that patient is normal or infected with COVID-19 based on the ResNet50V2 model with a little modification to deliver higher performance for the classification [20, 21].

Detecting small objects in an image is a problem that should be solved to improve accuracy and Feature Pyramid Networks (FPN) is an neural network model designed based on the pyramid concept to solve this problem [22]. It combines model information in a bottom-up direction combined with top-down to detect an object, while conventional algorithms only use bottom-up. This helps the network generate more semantic features for detecting objects at different scales.

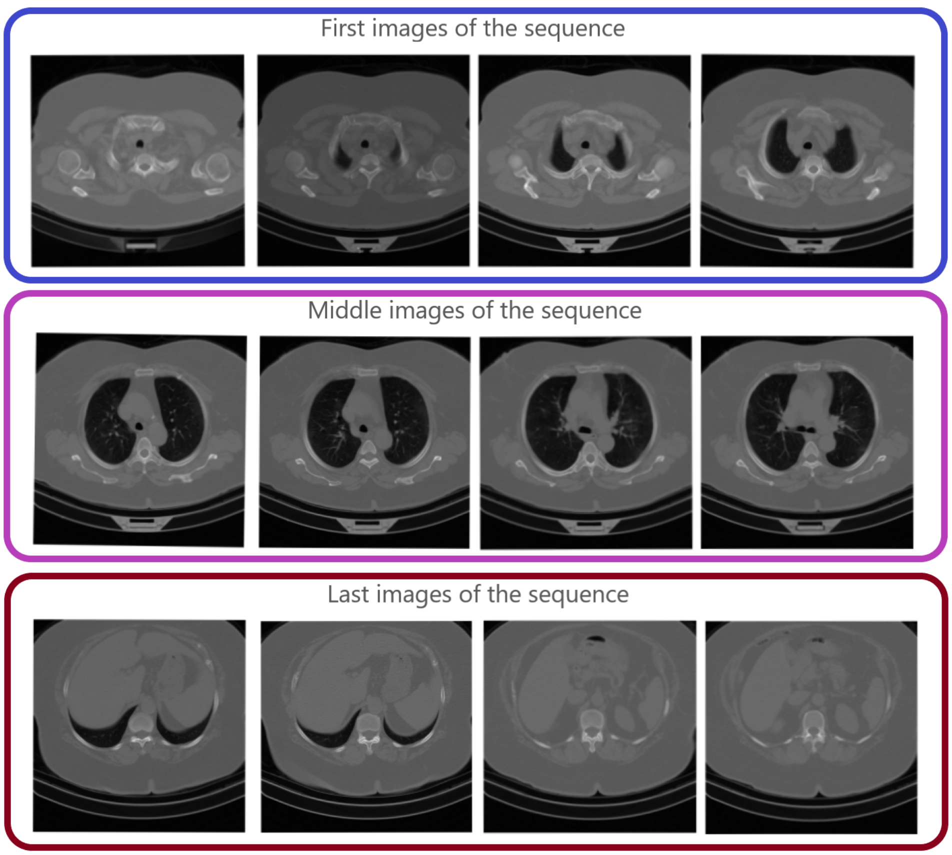

Figure 4: This figure show some of the first, middle, and last images of CT scans sequence.

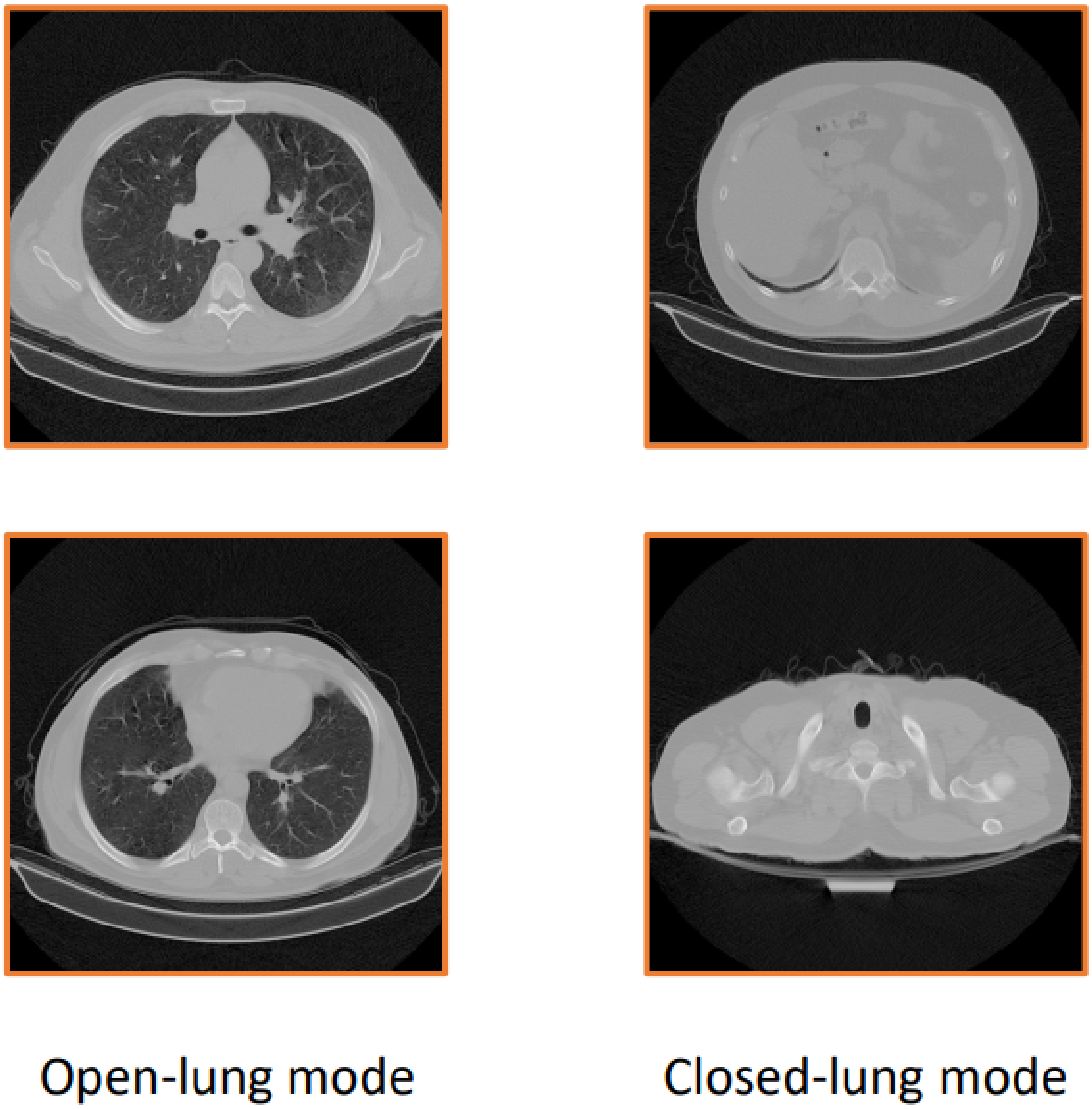

Figure 5: it is clearly visible from the figure that the open-lung has two larger pixel black slices.

Although our task here is to classify images, our model also needs to be learned to detect the infection points. For this reason, FPN can be used to help us better classify the image. In fig. 9, you can see the architecture of the network we have proposed.

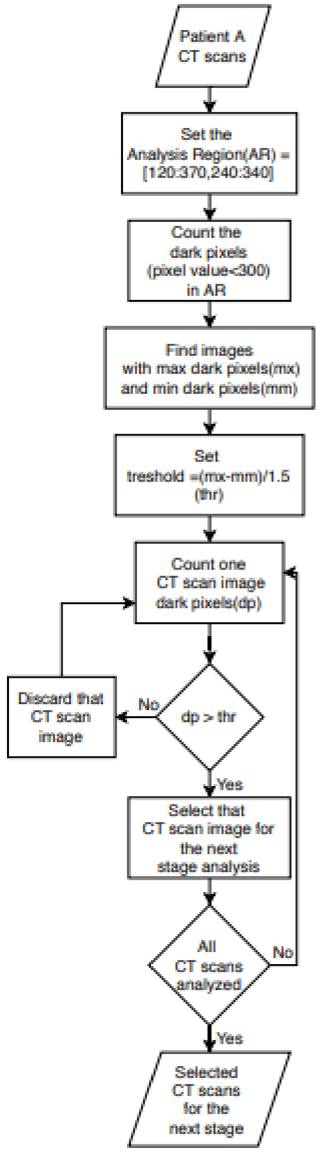

Figure 6: The flowchart of image selection algorithm for training

2.4 Training Phase

Our dataset consists of two parts: the first part is the set of lung CT images as introduced in section 2.1, since these images are all raw which is transferred from the original image when the CT scan passed, so part two will include the csv files for train and validation have been labeled covid and normal for these images. To increase accuracy during training and avoid over-fitting, we split the dataset into five folds for training and validation [23]. Accuracy of each model in 5 folds after training are reported in table 2.

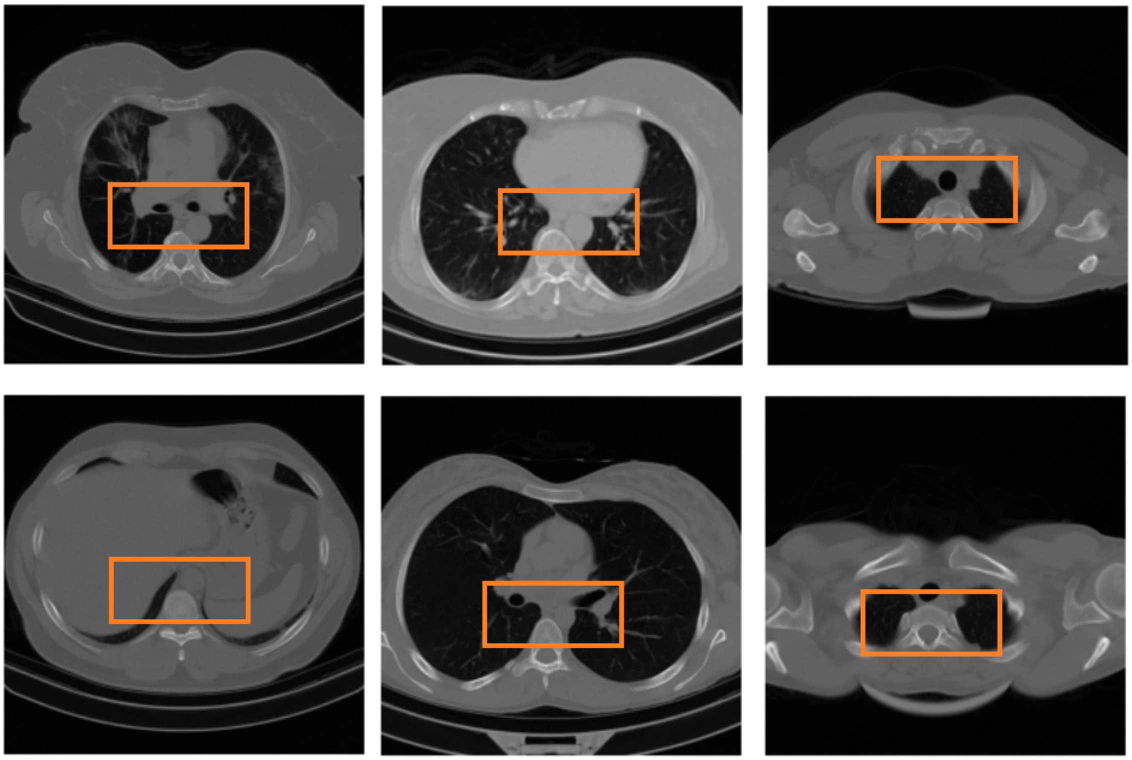

Figure 7: The selected region in different images

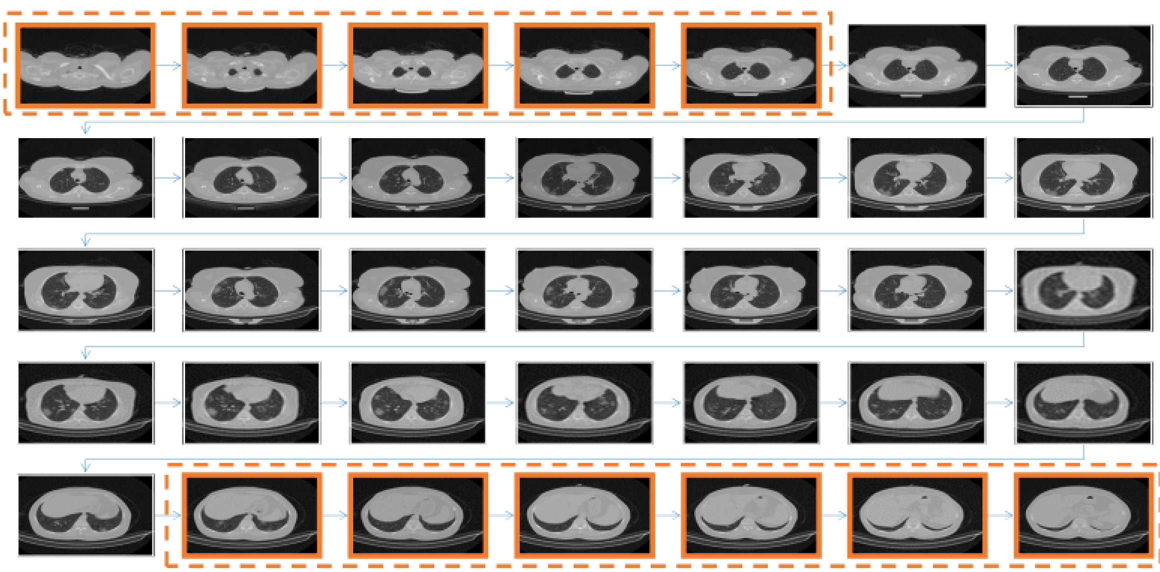

Figure 8: The figure shows the highlight images are the ones that will be ignored by the algorithm during the selection process.

From table 2, the problem arises here is why the number of images of patients with COVID-19 is less than the number of images of the normal person, the reason is that in the early stage before training, we used the image processing algorithm to select lung CT images that are clearly visible, so the number of images is reduced.

We trained our dataset on ResNet50V2 and the modified version of ResNet50V2 with 20 epochs and we used transfer learning from the ImageNet pre-trained weights to improve the accuracy of the model [24]. We also make some configurations to avoid network from over-fitting.

There is a small note that we did not reduce the size of the image while training to keep the details in the image, so the training will take place longer but it will bring more accuracy. The training parameters for our model are listed in table 3.

3 Experimental results

In this section, we will report the experimental results of the model after testing it on the test set. All of our models are trained on Google Colaboratory Notebooks and we also use the Keras library [25] for implementing deep learning models. Patient identification results on the normal and COVID-19 infected test set for each model at each fold are available in table 4.

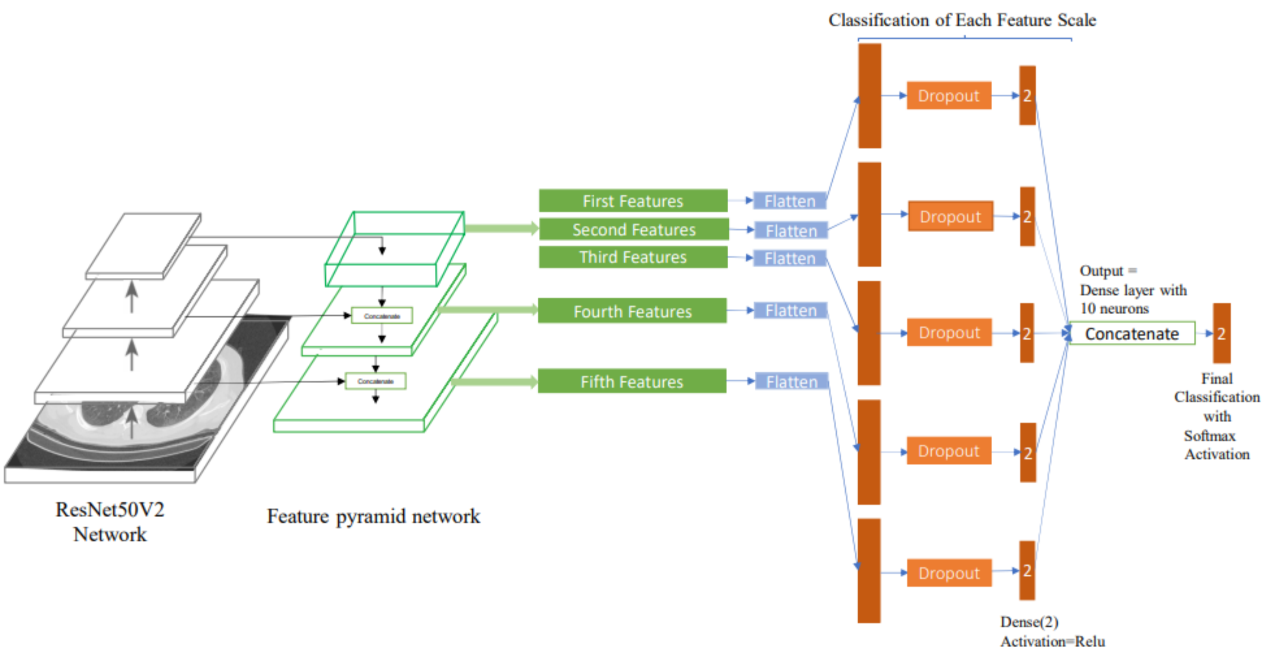

Figure 9: This figure shows the combination of ResNet50V2 with the feature pyramid network for the classification.

	ResNet50	Resnet50V2 + FPN
Fold 1	97.48%	96.49%
Fold 2	98.36%	97.16%
Fold 3	97.77%	98.66%
Fold 4	95.27%	97.80%
Fold 5	95.44%	97.89%

Table 2: Accuracy of each models for 5 folds

Training Parameters	ResNet50	Resnet50V2 + FPN
Learning Rate	1e-4	1e-4
Batch Size	14	14
Optimizer	Nadam	Nadam
Loss Function	Categorical Cross entropy	Categorical Cross entropy
Epoch	20	20
Horizontal/Vertical Flipping	Yes	Yes
Zoom Range	5%	5%
Rotation Range	0° - 360°	0° - 360°
Width/Height Shifting	5%	5%
Shear Range	5%	5%
Early Stopping	Yes	Yes

Table 3: Training Parameters

4 Discussion

Based on the results from table 4, we can see that the average accuracy in all 5 folds of the ResNet50 model when testing on normal patients is about 90%, on covid patients it is 96%. Similarlly, for the ResNet50V2 FPN model is about 93% and 96%. From this result, we can understand that the combination of ResNet50V2 with feature pyramid network make better overall accuracy than ResNet50.

Through this paper, we also hope that the shared dataset can help scientists as well as researchers to come up with new methods to improve the accuracy of the model for advanced medical diagnosis.

Fold	Model	Normal Patient Cases	Correct Identified Patients	Wrong Identified Patients	COVID Patient Cases	COVID Correct Identified	COVID Wrong Identified
1	ResNet50	79	73	6	95	91	4
	ResNet50V2 FPN	79	74	5	95	92	3
2	ResNet50	79	72	7	95	92	3
	ResNet50V2 FPN	79	67	12	95	92	3
3	ResNet50	79	71	8	95	92	3
	ResNet50V2 FPN	79	76	3	95	91	4
4	ResNet50	79	74	5	95	91	4
	ResNet50V2 FPN	79	75	4	95	89	6
5	ResNet50	79	66	13	95	90	5
	ResNet50V2 FPN	79	74	5	95	93	2

Table 4: Patients identification results

5 Conclusion

In this paper, we have proposed an effective method for detecting patients infected COVID-19 from lung HCRT scans. Initially, we used an image processing algorithm to select clearly visible lung CT images, then we applied FPN to the ResNet50V2 model to be able to identify details small in the lung, thereby increasing the accuracy of the model and all the data sets we used in this research were transferred from the CT scan device to the 16bit grayscale and all were in TIFF format which were lossless and let the image retain full details when converting.

In the next stage, we train two convolutional neural networks, ResNet50V2 and ResNet50V2 combining FPN to classify images of normal and patients infected COVID-19. After the training, our model's accuracy is pretty good at over 98%.

In the final stage, we put two models that were trained above to test and the results showed that the ResNet50V2 model using FPN gave quite outstanding results compared to the other model. From there, it can be seen that the method that we propose has increased the detection accuracy of COVID-19 patients. We hope that this approach can help doctors save time and effort in diagnosing early stage COVID-19 patients as well as help researchers devise a variety of methods that can be more efficient than our current approach.

6 Data availability

The public data set we used in this article is available here: https://tinyurl.com/59nm4pxs. The data set is divided into two parts, in which Train&Validation.zip includes lung CT images of all patients which were unlabeled, the rest in CSV folder includes 5 sets train.csv and validation.csv are the data set was labeled corresponding to these images above. We hope this data set can help not only us but other researchers can use it to create more efficient methods of detecting COVID-19.

7 Code availability

All codes we used for training and testing are available in `https://tinyurl.com/3xpvwtrr` and `https://tinyurl.com/te24s6tp`.

Acknowledgment

We would like to thank two authors: Mohammad Rahimzadeh and Abolfazl Attar for providing us the data set as well as helping us in completing this article. We are also committed to this article being created for the purpose of our scientific research methodology course and not for any other issue.

Ethics Statement

We declare that this article has been accomplished with the help of two authors mentioned above and is intended for scientific research methodology course only and for no other purpose. We promise to work for a better world where we can bring all of our knowledge to serve humanity and not to cause harm to others. Although the requirements for an author's ethical standards when writing a scientific paper are great, we write these in my paper because we recognize personal responsibility as the first step head on the road to success.

References

[1] Central Highlands Regional General Hospital. (2020, April 16). Symptoms and indications of chest CT in COVID-19. Benhviendaklak.Org.Vn. `http://benhviendaklak.org.vn/vi/news/thong-cao-bao-chi/CAC-BIEU-HIEN-VA-CHI-DINH-CUA-CT-NGUC-TRONG-COVID-19-510/`.

[2] Wikipedia contributors. (2021, April 28). COVID-19. Wikipedia. `https://en.wikipedia.org/wiki/COVID-19`.

[3] U.S. Department of Health & Human Services. (2020, October 5). SARS-CoV-2 and Potential Airborne Transmission. Centers for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/science/science-briefs/scientific-brief-sars-cov-2.html.

[4] Lovelace, B., & Higgins-Dunn, N. (2020, March 4). WHO says coronavirus death rate is 3.4% globally, higher than previously thought. CNBC. https://www.cnbc.com/2020/03/03/who-says-coronavirus-death-rate-is-3point4percent-globally-higher-than-previously-thought.html.

[5] U.S. Department of Health & Human Services. (2021, February 22). Coronavirus Disease 2019 (COVID-19) – Symptoms. Centers for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html.

[6] Wikipedia contributors. (2021a, April 24). Real-time polymerase chain reaction. Wikipedia. https://en.wikipedia.org/wiki/Real-time_polymerase_chain_reaction.

[7] Tahamtana, A., & Ardebili, A. (2020, April 22). Real-time RT-PCR in COVID-19 detection: issues affecting the results. PubMed Central (PMC). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7189409/.

[8] Thomas, L. (2020, March 15). Chest CT shows COVID-19 damage to the lungs. News-Medical.Net. https://www.news-medical.net/news/20200315/Chest-CT-shows-COVID-19-damage-to-the-lungs.aspx.

[9] Bollineni, V. R., Nieboer, K. H., Döring, S., Buls, N., & Mey, J. D. (2021, January 29). The role of CT imaging for management of COVID-19 in epidemic area: early experience from a University Hospital. Insights into Imaging. https://insightsimaging.springeropen.com/articles/10.1186/s13244-020-00957-5.

[10] U.S. Department of Health & Human Services. (2021, February 13). Interim Guidance on Duration of Isolation and Precautions for Adults with COVID-19. Centers for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/hcp/duration-isolation.html.

[11] Lascarrou, J. , Colin, G. , Thuaut, A. L. , Serck, N. , Ohana, M. , Sauneuf, B. , Geri, G. , Mesland, J. , Ribeyre, G. , Hussenet, C. , Boureau, A. S. & Gille, T. (2021, January 27). Predictors of negative first SARS-CoV-2 RT-PCR despite final diagnosis of COVID-19 and association with outcome. Scientific Reports. https://www.nature.com/articles/s41598-021-82192-6.

[12] Chatterjee, M. (2019, February 19). Top 20 Applications of Deep Learning in 2021 Across Industries. Mygreatlearning.Com. https://www.mygreatlearning.com/blog/deep-learning-applications/.

[13] Wikipedia contributors. (2021, April 23). TIFF. Wikipedia. https://en.wikipedia.org/wiki/TIFF.

[14] Wikipedia contributors. (2021a, March 19). High-resolution computed tomography. Wikipedia. https://en.wikipedia.org/wiki/High-resolution_computed_tomography.

[15] ScienceDirect ®. (n.d.). Trained Neural Network - an overview — ScienceDirect Topics. ScienceDirect. https://www.sciencedirect.com/topics/computer-science/trained-neural-network.

[16] Hui, J. (2018, March 28). What do we learn from region based object detectors (Faster R-CNN, R-FCN, FPN)? Medium. https://jonathan-hui.medium.com/what-do-we-learn-from-region-based-object-detectors-faster-r-cnn-r-fcn-fpn-7e354377a7c9.

[17] Suzuki, K. (2012, February 28). Pixel-Based Machine Learning in Medical Imaging. Hindawi. https://www.hindawi.com/journals/ijbi/2012/792079/.

[18] Meel, V. (2021, March 9). 56 Most Popular Computer Vision Applications in 2021. Viso.Ai. https://viso.ai/applications/computer-vision-applications/.

[19] Brownlee, J. (2019, March 26). A Gentle Introduction to the Promise of Deep Learning for Computer Vision. Machine Learning Mastery. https://machinelearningmastery.com/promise-of-deep-learning-for-computer-vision/.

[20] Wikipedia contributors. (2021c, April 27). Convolutional neural network. Wikipedia. https://en.wikipedia.org/wiki/Convolutional_neural_network.

[21] Sachan, A. (2019, September 17). Detailed Guide to Understand and Implement ResNets. CV-Tricks.Com. https://cv-tricks.com/keras/understand-implement-resnets/.

[22] Tsang, S. (2019, January 17). Review: FPN — Feature Pyramid Network (Object Detection). Medium. https://towardsdatascience.com/review-fpn-feature-pyramid-network-object-detection-262fc7482610.

[23] Brownlee, J. (2020, August 3). A Gentle Introduction to k-fold Cross-Validation. Machine Learning Mastery. https://machinelearningmastery.com/k-fold-cross-validation/.

[24] Wikipedia contributors. (2021b, March 29). ImageNet. Wikipedia. https://en.wikipedia.org/wiki/ImageNet.

[25] Team, K. (n.d.). Keras documentation: Models API. Keras.Io. https://keras.io/api/models/