IJAMML 17:1 (2023) 27-42 March 2023
ISSN: 2394-2258
Available at http://scientificadvances.co.in
DOI: http://dx.doi.org/10.18642/ijamml_7100122264

RESEARCH ON COMMON LUNG DISEASES BASED ON MACHINE LEARNING

Linxuan Li, Zhendi Qin, Xuan Chen, Lie Chen, Fanzhi Kong and Fanbin Meng

School of Medical Information Engineering, Jining Medical University, Rizhao Shandong 276826, P. R. China

Abstract

Covid-19 epidemic have greatly increased the number of patients with lung disease, and physicians have difficulty assessing patients' lung imaging with only personal experience and effort. To guarantee the efficiency of identification, it is necessary to establish a complete system for auxiliary lung disease identification. In response to the above problems, this paper will describe the process and results of a convolutional neural network (CNN) - based framework for lung disease image recognition. We randomly input image data using image data iterator and randomly selected a certain size of sample data for training in each batch. The system can identify chest X-ray images and lung CT images, and the identified lung diseases are Novel coronaviruses, Community-acquired pneumonia (CAP), and Viral pneumonia. The experimental test results of the classification system for

*Corresponding author.
E-mail address: fanbinmeng@qq.com (Fanbin Meng).

2020 Mathematics Subject Classification: 68T10, 68U10.
Submitted by Jianqiang Gao.
Received September 30, 2022

image recognition in lung diseases have a high correct rate of 98.9%. From the experimental results, it is suggested that this system can assist physicians to complete the evaluation of lung imaging.

Keywords: Covid-19, convolutional neural networks, deep learning, lung diseases image recognition classification system.

---

1. Introduction

About 40 years have passed since 1980 when machine learning began to be known as an independent direction. Machine learning techniques are well recognized and effective tools for signal processing and image processing by medical researchers. Zoabi et al. [1] studied the prediction of Covid-19 diagnosis based on basic symptoms, but novel coronavirus is a single-stranded RNA virus [2], prone to mutation, and the corresponding symptoms will change. Zoabi et al. method has too many uncertain factors, not suitable for large-scale promotion. Yuki et al. [3] found that in addition to Covid-19 has respiratory symptoms, thrombosis and pulmonary embolism are also observed in severe diseases, so they can be detected by PCR technology or CT or X-ray forms. Nucleic acid testing has become the norm, but clinically there is no lack of "false negative" patients with severe CT or X-ray examination but negative nucleic acid testing, Ai et al. already compared with Reverse Transcription-Polymerase Chain Reaction (RT-PCR) [4], lung CT imaging is a more reliable, practical and rapid method to diagnose Covid-19. Compared to RT-PCR. The combination of lung CT and chest X-ray for comprehensive analysis and diagnosis can greatly improve the accuracy. For manual diagnosis, most doctors use manual Region of Interest (ROI) delineation. This approach is less efficient, high labor cost. Against the identification of images of a large number of "false-negative" patients. Xu et al. [5] found that models could be trained through machine learning, which helped the doctors quickly to detect and identify pneumonia lesions. To provide information about the diagnosis of the disease and accelerate the

identification of patient symptoms, Oulefki et al. [6] introduces the value of an automated tool for the measurement of Covid-19 pulmonary infection using chest CT image segmentation and recognition system to improve the diagnostic and evaluation efficiency of primary physicians. Yang et al. [7] used CNN to analyze 152 CT images of Covid-19 patients and achieved 89% accuracy in Covid-19 patient prediction. Wang et al. [8] used 1065 CT images (740 Covid-19 negative and 325 Covid-19 positive) as a dataset for training and achieved 85.2% accuracy in correctly predicting patients with Covid-19. Afshar et al. [10] included chest radiographs from Covid-19 patients as a positive dataset and chest radiographs from common pneumonia and normal people as a negative dataset, the accuracy of the training using the capsule network [9] is about 95%. Roberts et al. [11] studied 62 papers based on machine learning to validate COVID-19, and the review found that most of the models were at high risk of bias, probably because the dataset contains a mixture of medical images from children and adults. All of the datasets in our study were medical images of adults used. However, only lung CT images or chest radiographs could be identified in these studies, and it is impossible to achieve multiple imaging types of lung disease identification from lung CT images and chest X-rays simultaneously. While our study can detect both lung CT images and chest radiograph at the same time and correctly identify lung diseases, which can better reduce the burden on physicians and elevate the correct rate of assessment. Even primary hospitals with limited medical resources can diagnose multiple detection images with the aid of auxiliary pre diagnostic tools, further improving the diagnosis and evaluation efficiency of primary physicians.

2. Materials and Methods

This paper trains three models (including Model 1, Model 2, Model 3) using TensorFlow Deep Learning Frameworks [12] and PaddlePaddle [13].

2.1. Model 1 for identifying chest X-ray images and lung CT images

Data sets a total of 3300 images were acquired, 1650 each for chest radiographs and CT images, and 40% of the images were randomly selected as test set data by custom functions. The remaining 60% of the data set was used as the training set. After classification, images were preprocessed using imagedatagenerator with batch size of 16 as best, not too large or too small, batch size too large makes training time too long and loss function value difficult to drop. Too small a batch size would make individual features repeated, prone to overfitting. The number of images with batch size 16 was randomly selected, followed by normalization and rotation before transferring the image data to the convolutional neural network for model training.

Convolutional neural networks are constructed, after thousands of experiments, as shown in Figure 1, the model takes the form of convolutional layer maximum pooling layer full connected layer. Interpolation of images was calculated to image data for 128 × 128 and divided by 255 to normalize to input data.

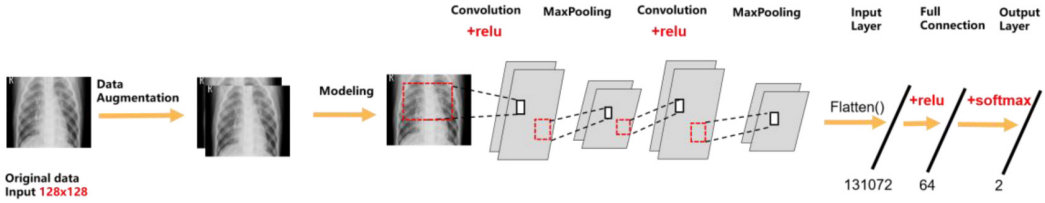

Figure 1. Basic structure of the CNN in Model 1. Image dataset from PaddlePaddle. The convolutional layer was two layers (convolution core size is 3 × 3), two layers of maximum pooling layer (pooling core size is 2 × 2), and the 64 groups of neurons in the fully connected layer had the shortest training time and the highest accuracy.

Besides, overfitting phenomenon can easily happen during training when dataset data is small, so we use keras.dropout(). When designing convolutional neural networks, each neuron is a single feature learned by the machine, and each layer of neurons is a combination of features learned by the machine. When the data set is small, there are many repetitions and redundancies among the single features learned by each neuron, resulting in overfitting, while the dropout function directly reduces the number of single features, thus reducing feature repetition and redundancy, preventing overfitting, and improving model generalization.

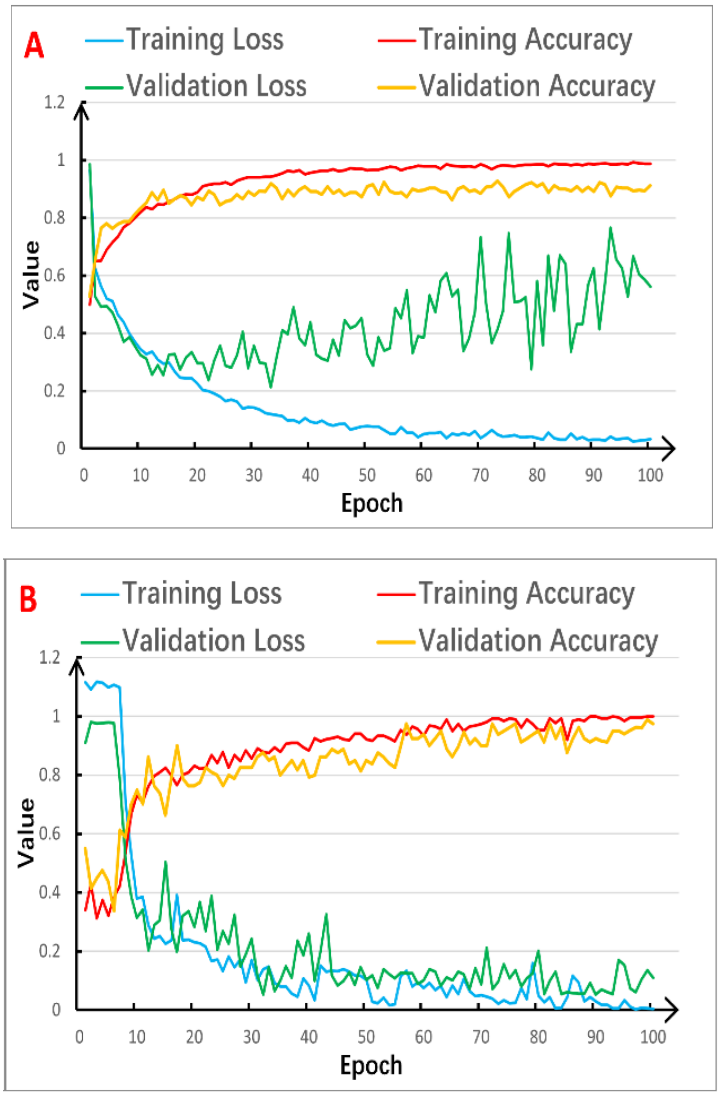

Figure 2. In the case of the same learning rate. The image A is the model training process without the dropout function. There is overfitting in the image A. On the image B is a model with a dropout function with a value of 0.4, and the overfitting phenomenon in the image B has been significantly alleviated.

In this study, the optimizers used categorical crossentropy to match the softmax activation function. Loss uses adagrad, whose full name is adaptive gradient. The adagrad algorithm adapts the learning rate of parameters automatically, the direction of parameters with large gradients slows the learning rate, the direction of parameters with small gradients accelerates the learning rate, and adagrad can greatly improve the robustness of SGD (a system or organization has the performance to resist or overcome unfavourable conditions). Metrics was assessed using the accuracy function. Results as shown in Figure 3, the model loss has successfully converged, the training set and test set accuracy rates both reached over 99%, and the model data were saved. Training on this model was completed.

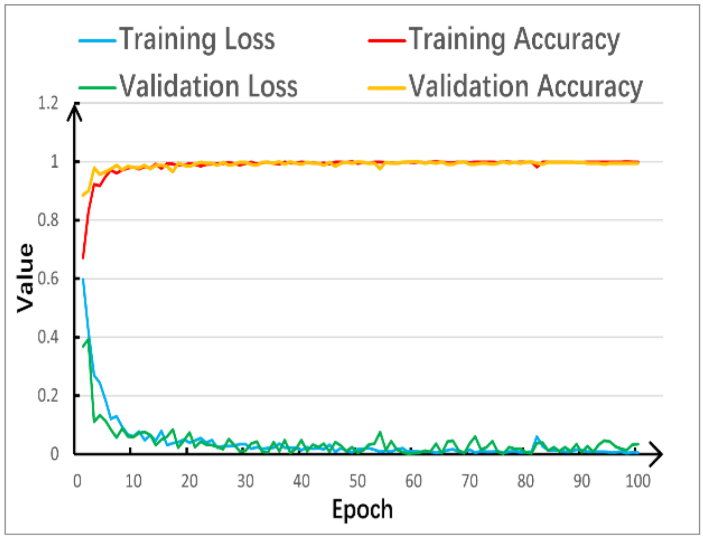

Figure 3. Training process of Model 1. The initial learning rate was 0.03, the number of trainings was 50, and the model was saved after the completion of training, the learning rate was changed to 0.01, and the number of trainings was 50.

2.2. Model 2 is able to identify three types of CT images of Covid-19, community-acquired pneumonia (CAP) and normal people

A total of 2700 images were present in the dataset. The CT images of Covid-19, CAP, and normal people were all 900. The remaining 60% of the data set was used as the training set. We randomly selected 40% of the images as test set data by a custom function. The remaining 60% of the data set was used as the training set. Once classification is complete, images are preprocessed again using the imagedatagenerator, which automatically generates a label value (one for each folder) for the training data. Batch sizes have been tested many times and either batch sizes too large or too small can affect training, as shown in Figure 4. It was experimentally found that the best training learning rate is best when the batch size is 32, which is discussed below.

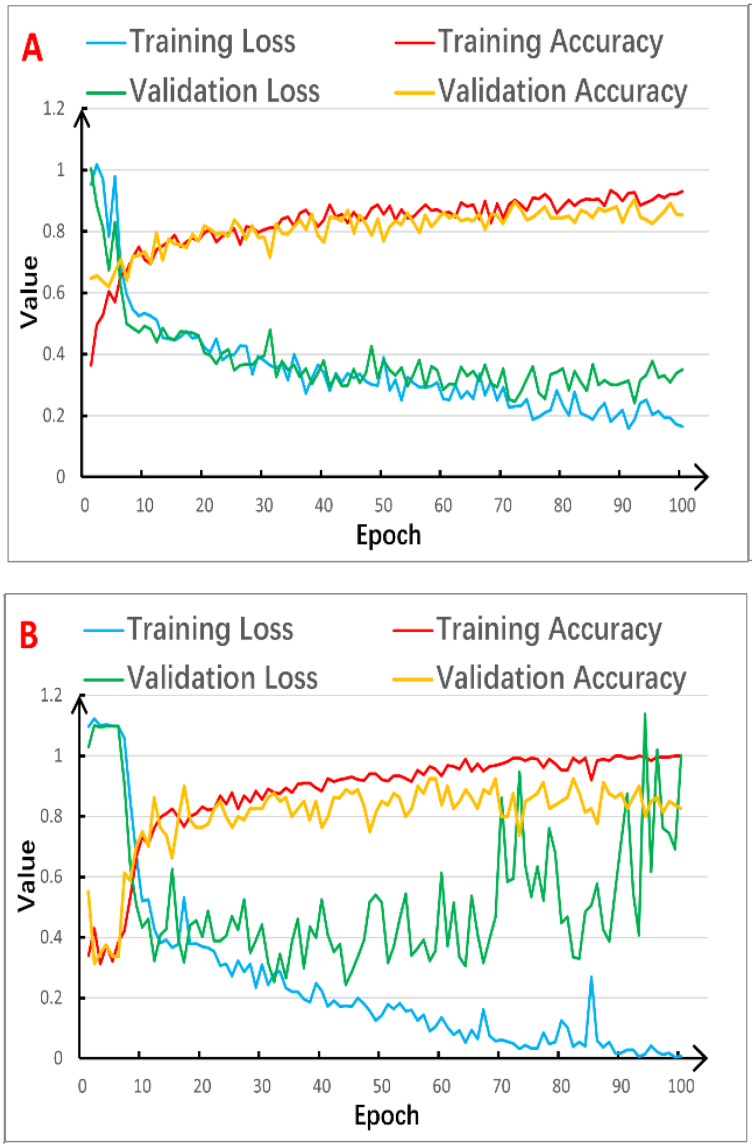

Figure 4. Image A is a case where the batch size is too large, the resulting image features are too many, the training time is too long, and it is difficult to converge. Image B is a case where the batch size is too small and individual features too repetitive, overfitting has occurred.

Construction of convolutional neural networks. Interpolation of images was calculated to image data for 128 × 128 and divided by 255 to normalize to input data. After thousands of experiments, the convolutional neural network structure shown in Figure 5 was found to have the shortest training time and the best effect.

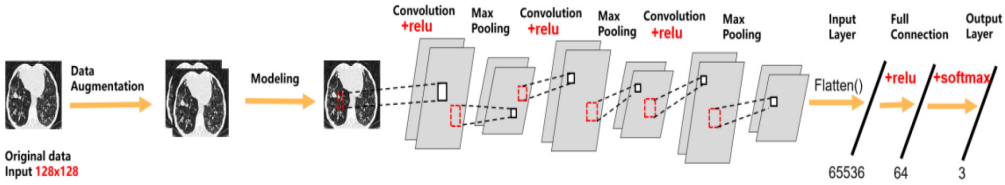

Figure 5. Basic structure of the CNN in Model 2. Image dataset from PaddlePaddle. The convolutional layer is three layers (convolution core size is 3 × 3), three layers of maximum pooling layers (pooling core size is 2 × 2), and the 64 groups of neurons are fully connected layers.

The training process of Model 2 was also prone to overfitting (extremely high accuracy in the training set and poor accuracy in the test set). The dropout function has a value of 0.5 to prevent overfitting. The optimizers used Categorical crossentropy to match the softmax activation function. Loss uses adagrad. Metrics was assessed using the accuracy function. As shown in Figure 6, loss has successfully converged (the average fluctuation value of the loss function value is less than 0.05), the accuracy of the training set and the test set reached more than 98%, the model data were saved, and the model training was completed.

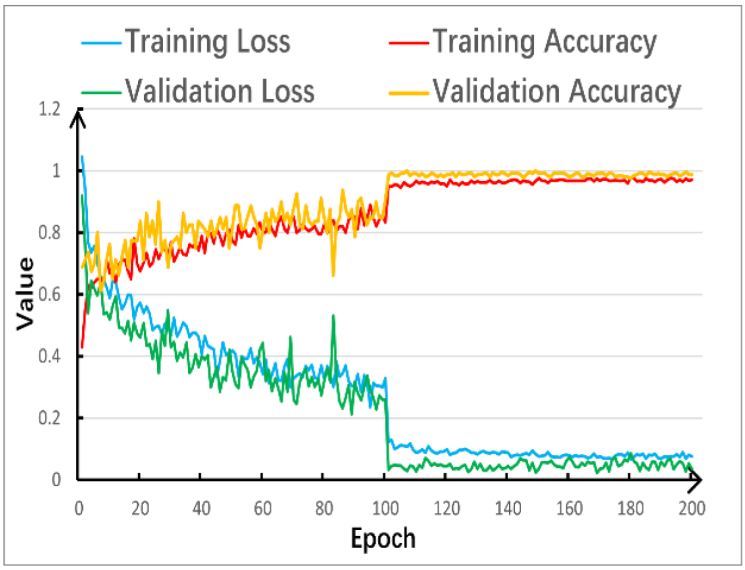

Figure 6. The training process for Model 2. The initial learning rates were 0.05 and 100 training epochs, respectively, and the trained model was saved. The learning rate becomes 0.005 and the number of trainings is 100.

2.3. Model 3 is able to identify three types of chest X-rays of Covid-19, virus pneumonia and normal people

A total of 2400 images were present in the dataset. The chest X-rays images of Covid-19, Viral pneumonia, and normal people were all 800. 40% of the images were randomly selected as test set data and the remaining 60% of the data set was used as training set. Image generators were used to preprocess images. Batch sizes trained on the model were tested multiple times and Model 3 training was found to work best when the batch size was 32.

Construction of convolutional neural networks. Interpolation of images was calculated to image data for 128 × 128 and divided by 255 to normalize to input data. After a hundred experiments and previous experiences, the structure shown in Figure 7 is the most suitable with the shortest training time and the highest accuracy.

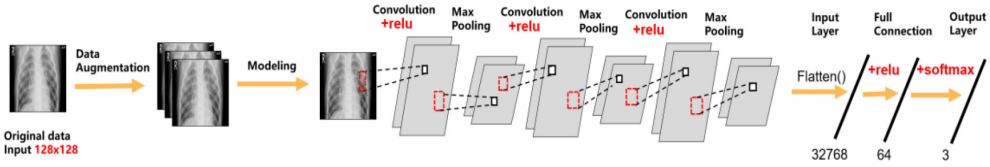

Figure 7. Basic structure of the CNN in Model 3. Image dataset from PaddlePaddle. The convolutional layer is three layers (convolution core size is 3 × 3), three layers of maximum pooling layers (pooling core size is 2 × 2), and the 64 groups of neurons are fully connected layers.

The dropout function in keras was also used in Model 3 training. The dropout function has a value of 0.6 to prevent overfitting. The optimizers used categorical crossentropy to match the softmax activation function. Loss uses adagrad. The metrics was assessed using the accuracy function the training process and Model 2 were approximately the same. As shown in Figure 8, loss has converged successfully and the accuracy rate of both the training set and the test set reached more than 98%. Model data was saved and model training completed.

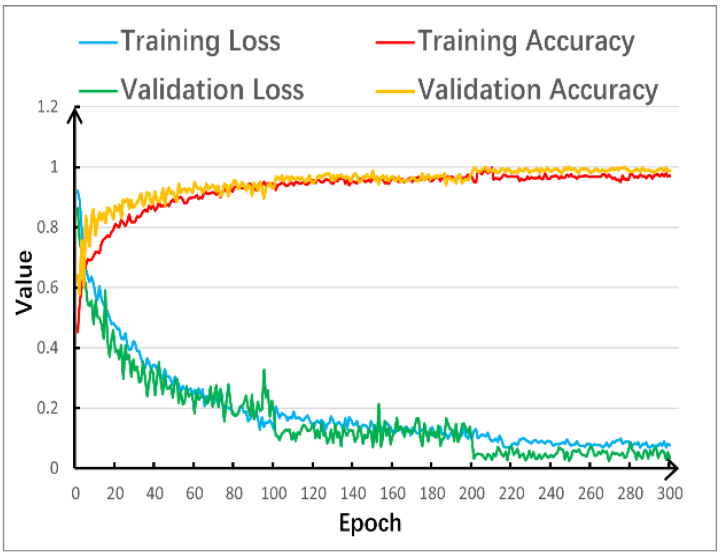

Figure 8. Training process of Model 3. The initial learning rate was 0.03, and the number of trainings was 100. The change learning rate was 0.01, and the number of trainings was 100. The change-over learning rate was 0.001, and the number of trainings was 100.

3. Results and Discussion

After the training of the three models, saved as three H5 files. Our written image detection program, three models were imported for image sample validation, from the experimental results, the classification accuracy of chest X-ray images and lung CT images in Model 1 is 99.6%. The accuracy of the three classifications (Covid-19, cap, normal) of Model 2 lung CT images was 98.6%. Model 3 the accuracy of the three classifications of chest X-ray images (Covid-19, viral pneumonia, normal) reached 98.4%. After integrating the three models by a certain program, and then validating the fusion model, 600 image samples (300 lung CT images, 300 chest X-ray images) were verified. With only seven prediction errors. The combined identification correct rate is up to 98.9%. The lung disease recognition framework has basically met the need to predict these four types of lung diseases.

4. Conclusion

The Covid-19 epidemic is still ravaging globally, and outbreak control efforts should not be loosened. The earlier it is detected and treated, the less pain the patient experiences, and prevention and diagnosis are particularly important. Today, with rapid advances in artificial intelligence, combining Covid-19 diagnosis and prevention with artificial intelligence technologies is key, and using deep learning techniques to predict Covid-19 is central to future prevention and control of Covid-19. In the future, identifying more lung diseases, reducing training time, and simplifying network structure are our next research directions.

Acknowledgements

This work is partly supported by the Doctoral Research Foundation of Jining Medical University under Grant No. 600529001, and the 2022 Innovation and Entrepreneurship Training Program for College Students under Grant Nos. 202210443002, 202210443003, cx2022002z, cx2022009z and cx2022022z.

References

[1] Yazeed Zoabi, Shira Deri-Rozov and Noam Shomron, Machine learning-based prediction of Covid-19 diagnosis based on symptoms, NPJ Digital Medicine 4 (2021), 1-5; Article 3.

DOI: https://doi.org/10.1038/s41746-020-00372-6

[2] Yen-Chin Liu, Rei-Lin Kuo and Shin-Ru Shih, Covid-19: The first documented coronavirus pandemic in history, Biomedical Journal 43(4) (2020), 328-333.

DOI: https://doi.org/10.1016/j.bj.2020.04.007

[3] Koichi Yuki, Miho Fujiogi and Sophia Koutsogiannaki, Covid-19 pathophysiology: A review, Clinical Immunology 215 (2020); Article 108427.

DOI: https://doi.org/10.1016/j.clim.2020.108427

[4] Tao Ai, Zhenlu Yang, Hongyan Hou, Chenao Zhan, Chong Chen, Wenzhi Lv, Qian Tao, Ziyong Sun and Liming Xia, Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (Covid-19) in China: A report of 1014 cases, Radiology 296(2) (2020), e32-e40.

DOI: https://doi.org/10.1148/radiol.2020200642

[5] Xiaowei Xu, Xiangao Jiang, Chunlian Ma, Peng Du, Xukun Li, Shuangzhi Lv, Liang Yu, Qin Ni, Yanfei Chen, Junwei Su, Guanjing Lang, Yongtao Li, Hong Zhao, Jun Liu, Kaijin Xu, Lingxiang Ruan, Jifang Sheng, Yunqing Qiu and Lanjuan Li, A deep learning system to screen novel coronavirus disease 2019 pneumonia, Engineering 6(10) (2020), 1122-1129.

DOI: https://doi.org/10.1016/j.eng.2020.04.010

[6] Adel Oulefki, Sos Agaian, Thaweesak Trongtirakul and Azzeddine Kassah Laouar, Automatic Covid-19 lung infected region segmentation and measurement using CT-scans images, Pattern Recognition 114 (2021); Article 107747.

DOI: https://doi.org/10.1016/j.patcog.2020.107747

[7] X. Yang, X. He, J. Zhao, Y. Zhang, S. Zhang and P. Xie, COVID-CT-Dataset: A CT Image Dataset About Covid-19 (2020).

[8] Shuai Wang, Bo Kang, Jinlu Ma, Xianjun Zeng, Mingming Xiao, Jia Guo, Mengjiao Cai, Jingyi Yang, Yaodong Li, Xiangfei Meng and Bo Xu, A deep learning algorithm using CT images to screen for Corona Virus Disease (Covid-19), European Radiology 31(8) (2021), 6096-6104.

DOI: https://doi.org/10.1007/s00330-021-07715-1

[9] Jathushan Rajasegaran, Vinoj Jayasundara, Sandaru Jayasekara, Hirunima Jayasekara, Suranga Seneviratne and Ranga Rodrigo, Deepcaps: Going deeper with capsule networks, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019).

DOI: https://doi.org/10.1109/CVPR.2019.01098

[10] Parnian Afshar, Shahin Heidarian, Farnoosh Naderkhani, Anastasia Oikonomou, Konstantinos N. Plataniotis and Arash Mohammadi, Covid-caps: A capsule network-based framework for identification of Covid-19 cases from X-ray images, Pattern Recognition Letters 138 (2020), 638-643.

DOI: https://doi.org/10.1016/j.patrec.2020.09.010

[11] Michael Roberts, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I. Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, Jonathan R. Weir-McCall, Zhongzhao Teng, Effrossyni Gkrania-Klotsas, AIX-COVNET, James H. F. Rudd, Evis Sala and Carola-Bibiane Schönlieb, Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans, Nature Machine Intelligence 3(3) (2021), 199-217.

DOI: https://doi.org/10.1038/s42256-021-00307-0

[12] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu and Xiaoqiang Zheng, TensorFlow: A system for large-scale machine learning, 12th USENIX Symposium on Operating Systems Design and Implementation (2016).

[13] Yanjun Ma, Dianhai Yu, Tian Wu and Haifeng Wang, PaddlePaddle: An open-source deep learning platform from industrial practice, Frontiers of Data and Computing 1(1) (2019), 105-115.

DOI: https://doi.org/10.11871/jfdc.issn.2096.742X.2019.01.011

■