Detection and Classification of Pneumonia from Lung Ultrasound Images

Jiaqi Zhang
Department of Mechanical Engineering
National University of Singapore
Singapore, Singapore
e0343933@u.nus.edu

Chin-Boon Chng
Department of Mechanical Engineering
National University of Singapore
Singapore, Singapore
mpeccbo@nus.edu.sg

Xuan Chen
BGI Research
National University of Singapore
Shenzhen, China
chenxuan@genomics.cn

Chunshuang Wu
Department of Emergency Medicine
2nd Affiliated Hospital, Zhejiang University School of Medicine
Hangzhou, China
21518193@zju.edu.cn

Mao Zhang
Department of Emergency Medicine
2nd Affiliated Hospital, Zhejiang University School of Medicine
Hangzhou, China
z2jzk@zju.edu.cn

Yuqi Xue
Research Institute of Surgery
Daping Hospital, Army Medical University
Chongqing, China
xueyuqi77@126.com

Jianxin Jiang
Research Institute of Surgery
Daping Hospital, Army Medical University
Chongqing, China
hellojjx@126.com

Chee-Kong Chui
Department of Mechanical Engineering
National University of Singapore
Singapore, Singapore
mpecck@nus.edu.sg

***Abstract*— The lungs are the primary organs of the respiratory system in humans. Meanwhile, lungs are also vulnerable and are easily damaged by inflammation or impact lesions during the course of our daily lives. Due to the epidemic of COVID-19 pneumonia, the confirmed and suspected cases often grow rapidly beyond the capabilities of medical institutions, rapid and accurate diagnosis for patients have become the first priority. Hence, ultrasound images have started to be adopted in lung diagnosis as they are more convenient, flexible, cheaper, and without ionizing radiation as compared with CT and CXR. This paper aims to use VGG, ResNet and EfficientNet networks to accurately classify Lung Ultrasound images of pneumonia according to different clinical stages based on self-made LUS datasets. The hyperparameters of the three networks were tuned and their performances were carefully compared. Our results indicate that the EfficientNet model outperformed the others, providing the best classification accuracies for 3 and 4 clinical stages of pneumonia are 94.62% and 91.18%, respectively. The best classification accuracy of 8 imagological features of pneumonia is 82.75%. This result is a proof of the promising potential of the LUS device to be used in pneumonia diagnosis and prove the viability of deep learning for LUS classification of pneumonia.**

Keywords—lung ultrasound image, deep learning, medical image classification, diagnosis of pneumonia.

I. INTRODUCTION

The lung structure contains lobes, bronchus and pulmonary alveoli which are connected to important veins. All components need to function well for the efficient exchange of oxygen and carbon dioxide. Due to the softer and more fragile structure than other organs, the lungs are more easily injured by external shocks and internal inflammation. The injuries can result in lung consolidation, and even pulmonary effusion, which will further lead to respiratory failure and finally cause apnoea and multiple organ failure (MOF). Due to the symptoms during early stage are mild, the development of lung lesions is extremely fast. Finding a rapid and efficient way of diagnosing respiratory system diseases at the earlies stage is thus a critical area of research that bears investigation.

The advancement of medical technology has led to numerous medical examination technologies such as Chest X-ray (CXR), Computed Tomography (CT) and Ultrasound (US). For the diagnosis of pneumonia, CT scan is regarded as gold standard. However, in the context of the rapid escalation of COVID-19 cases, the low speed and throughput due to limited number of CT examinations, high costs and higher doses of radiation, restricts the feasibility of frequent imagological examinations with CT devices to closely monitor the development of COVID-19 pneumonia. Although CXR devices are smaller, cheaper, with less ionizing radiation. However, it is difficult for CXR to distinguish between pneumonia, pulmonary embolism, acute respiratory distress syndrome and pulmonary fibrosis, with low sensitivity especially for detecting tiny and mild lesions in early stages of pneumonia in practical clinical situation [1]. Compared with CXR and CT, Lung ultrasound (LUS) is a versatile, low cost, radiation-free and convenient imaging modality that is widely available in most modern healthcare systems. Because Ultrasound imaging is based on the pulse-echo principle of sound wave, which led that clinicians underestimated the potential of LUS in pneumonia diagnosis over a long period in the past. This coupled with the detection capability of CT and CXR could adequately meet the daily examination requirements, the acceptance of LUS device in pneumonia diagnosis was limited in clinical practise [2].

While LUS has yet to become the recommended diagnosis device for pneumonia, this paper reviews the recent literature of LUS in pneumonia diagnosis and summarizes the specific features corresponding to CT images in the past 20 years. In one of the earlier studies by Lichtenstein et al. [2], the accuracy of

LUS for the diagnosis of alveolar consolidation was evaluated for critically ill patients in the emergency department (ED). The feasibility of the method was found to be 99%, with a sensitivity of 90% and a specificity of 98%, showing the potential of LUS as a beside non-invasive diagnostic method. In 2006, Volpicelli et al. [3] studied the use of LUS for the diagnosis of alveolar interstitial syndrome (AIS) in mild cases of pneumonia, reporting the sensitivity and specificity to be 85.3% and 96.8% respectively. Their findings showed that the comet-tailed artifacts (B-lines) could be used to accurately diagnose AIS, meanwhile, exclude pneumothorax and pulmonary edema. In 2009, Parlamento et al. [4] compared LUS against CXR and CT, reporting accuracy = 96.9%.

For accurately identifying the clinical features with LUS, Xirouchaki et al. [5] identified four clinical signs - pulmonary consolidation, positive air bronchogram, abnormal pleural line and pleural effusion. Pagano et al. [6] further added subpleural lung consolidation, alveolar syndrome with dynamic air bronchograms and interstitial syndrome with 3 or more B lines. With the newer criteria, they achieved a positive predictive value (PPV) of 0.838, negative predictive value (NPV) of 0.960, a sensitivity of 0.985 and a specificity of 0.649, outperforming CXR. Besides the accuracy, time efficiency and feasibility were also focused. Seyedhosseini et al. [7] compared the diagnostic time, with the mean admission-treatment time of LUS with BLUE protocol was 17 mins compared to 38 mins of CT control group. Copetti et al. [8] investigated the use of LUS in paediatric departments. Since the ionizing radiation has negative effects on children's growth and health. Their results suggested that the performance of LUS was comparable to CT and exceeded CXR in evaluating children. Correspondingly, LUS modality is extremely attractive for resource-limited remote areas due to the ease of use, portability and low cost.

COVID-19 has brought the need for rapid detection of pneumonia to the forefront. In the earlier period of the outbreak, before diagnostic test kits like real-time reverse transcription polymerase chain reactions (RT-PCR) were developed, pneumonia symptoms were used to identify patients. According to previous work, the Handbook of COVID-19 Prevention and Treatment and the COVID-19 diagnosis and treatment plan (7th Edition) [9], the clinical stages and diagnostic signs of LUS and CT are highly consistent and are summarized in Table I. The development of pneumonia from the initial onset of symptoms such as fever and coughs were often sudden and accompanied by major complications such as ARDS. Thus, meticulous monitoring of the progression of symptoms, would enable early treatment and can minimize complications and costs of follow-up treatments. However, with the rate of transmission and the number of infections rising to unprecedented levels, the diagnostic efficiency of CT devices could no longer meet the testing needs for so many confirmed and suspected patients. Similarly, the amount of manpower would still be substantial even with the LUS devices.

Deep learning method such as convolutional neural networks have demonstrated comparable performances to that of humans on a range of image classification tasks. A precise and fast LUS image classification system based on deep learning method could potentially assist clinicians and ease their workload. Hence, with LUS being a rapid, convenient, accurate, radiation-free and easy to implement bedside method for the visualization of pulmonary diseases, this paper focus on the investigation of the feasibility of computer-assisted ultrasound diagnosis with three CNN-based deep learning models - VGG, ResNet and EfficientNet for the detection and classification of pneumonia on a self-made lung ultrasound image dataset.

TABLE I. DIAGNOSTIC SIGNS OF PNEUMONIA AT DIFFERENT CLINICAL STAGES

Sign of Pneumonia	Pattern in LUS	Pattern in CT	Stage
AIS	3 or more B lines	ground-glass opacity	Early or middle
Consolidation (with air bronchogram)	Pieces and hepatization	white lungs	Middle and severe
Plueural Effusion	Sinusoid	Effusion shadows	Severe

II. MATERIAL AND METHODS

A. Dataset Construction

A total of 10350 LUS images were used in the construction of the dataset. As these images were collected from multiple sources with different file naming rules and sizes, each image was preprocessed to homogenize the filenames, remove all private information. Besides, due to lack the LUS data at early stage of COVID-19, some of the LUS images are collected from the patients of lung impact lesion which have similar imagological features as pneumonia. Each image was manually classified and relabeled into 8 classes according to clinical features of pneumonia (Fig.1). The clinical features of pneumonia are summarized in Table II. While there are 8 clinical features which can be identified in LUS images of pneumonia patients, clinicians often need to combine several features based on pathology in order to accurately diagnose the stages of pneumonia. Hence we grouped multiple features together according to clinical stages, a) 3-classes, b) 4-classes to further differentiate the severe cases, and c) 8-classes based on all clinical features.

The allocation of images into training and testing sets is shown in Table III. The proportions of training and testing sets are 90% and 10%, respectively. Due to the limitation of the imbalance number of 8 clinical feature images, allocating more images for testing set would severely reduce of the training data, especially for class 2 and 7. With the new grouping into 3 and 4-classes, the distribution of images is more balance and able to provide better performance for the classification models.

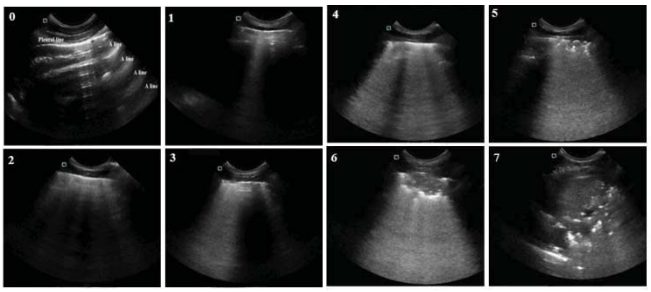

Fig. 1. Processed images of 8 clinical features.

TABLE II. EIGHT CLINICAL FEATURES OF PNEUMONIA IN LUS IMAGES

#	Clinical Feature
0	Normal
1	B lines are less than 3
2	B lines are more than 3
3	Area of merging B line is less than half
4	Area of merging B line is more than half
5	Depth of Pieces is less than 1cm
6	Air bronchogram and depth of Hepatization is less than 3 cm
7	Pleural effusion and depth of Hepatization is more than 3 cm

TABLE III. ALLOCATION OF IMAGES FOR EACH DATASET INTO TRAINING AND TEST SETS

# Class	Categories (Train/Test)							
	0	*1*	*2*	*3*	*4*	*5*	*6*	*7*
3	2608 /288	4044/363				2654/393		
4	2608 /288	4044/363				2305/126		349/ 267
8	2608 /288	1958 /200	184/ 32	1014 /58	888/ 73	1625 /51	679/ 75	349/ 267

B. Models Construction and Training

The aim of this paper is to rapidly verify the accuracy of the deep learning method in the classification task of pneumonia LUS images. This would pave the way for wide adoption of LUS device for the diagnosis of pneumonia. Thus, this paper does not focus on constructing a brand new ConvNet structure, apart from adding some steps to overcome data imbalance. The performance of three state of the art models were investigated in this work – VGG-19, ResNet-101 and EfficientNet-B5 [10]. The reason is that VGG and ResNet belong to the typical single-dimension scaling methods (by depth) and EfficientNet belongs to latest compound scaling method (by depth, width and resolution together). All models utilized transfer learning to accelerate the training process of the models. Pre-trained weights of VGG-19 and Resnet-101 were trained on the ImageNet and CIFAR-100 dataset. The pre-trained EfficientNet model was trained on 8 large datasets and 1 private breast cancer ultrasound dataset. For our application, the models were used as feature extractors, with their original output layers reconnected to provide an output shape according to the number of classes in our LUS datatsets. All models were fine-tuned on a desktop computer with an i5-4570 processor with a Nvidia 1080Ti 11Gb GPU and with 16Gb of RAM. The software framework used was Python 3.7 with PyTorch 1.4. Image augmentation was also performed through random cropping and horizontal flipping in order to reduce the imbalance between classes.

Hyperparameter tuning was performed, with a range of learning rates (from 0.01 to 0.0002), batch sizes (from 3 to 24) and optimizers (SGD and ADAM) were explored for their accuracy. The parameters selected for the final models are summarized in Table IV. Choosing 0.0002 as the learning rate of EfficientNet-B5 is based on the experimental results, which has higher accuracy and less training time under lr = 0.0002 than lr = 0.0001. All EfficientNet models converged within 50 epochs, taking an approximate maximum of 2.5 hours while VGG and ResNet models took more than 8 hours for the 8-class dataset and approximately 3 hours for the 3 and 4-class dataset.

TABLE IV. SELECTED HYPERPARAMETERS OF MODELS

Model	Parameters							Results
	# Class	*Learning rate*	*Batch size*	*Epoch*	*Loss function*	*Optimizer*	*Momentum*	*Accuracy*
VGG-19	3	0.0001	12	250	cross entropy	SGD	0.9	0.891
	4	0.0001	12	250	cross entropy	SGD	0.9	0.884
	8	0.0001	12	250	cross entropy	SGD	0.9	0.600
ResNet-101	3	0.0001	24	250	cross entropy	ADAM	-	0.886
	4	0.0001	24	250	cross entropy	ADAM	-	0.875
	8	0.0001	24	250	cross entropy	ADAM	-	0.624
EfficientNet-B5	3	0.0002	16	50	cross entropy	ADAM	-	0.946
	4	0.0002	16	50	cross entropy	ADAM	-	0.912
	8	0.0002	16	50	Weighted Smooth cross entropy	ADAM	-	0.823

Comparing across classification models, the EfficientNet model performed far better across all 3 datasets, with less time needed for fine-tuning. For all models of EfficientNet, training results stabilized fairly quickly within less than 20 epochs. With the speed and accuracy of the model in mind, the EfficientNet-B5 model was selected for use in the final implementation and was retrained with the same parameters for a final evaluation in the next section.

III. RESULTS AND DISCUSSION

The results for EfficientNet-B5 after the final retraining are shown in the confusion matrices in Fig. 2-4 The best hyperparameters used for EfficientNet-B5 were learning rate = 0.0002, batch size = 16, epoch = 50, ADAM optimizer (step-size = 7, gamma = 0.1), cross entropy loss function for the 3 and 4-class dataset. Meanwhile, in order to reduce the class imbalance in 8-class dataset, weighted smooth cross entropy for the 8-class dataset were used, the weights for each class are setting as: (0.7198,0.7896,0.9801,0.8910,0.9046,0.8254,0.9270,0.9625). These weights are calculated according to the image allocation of 8-class dataset which can effectively increase the weights for the classes with less images and suppress the weights of the classes with more images. Besides that, the label smoothing and k-fold cross-validation (k = 10) were also used to further mitigate the class imbalance, the hyperparameter of label smoothing is 0.2. The true probability distribution and loss function were revised in the following equation [11] to add some noise to the weights of true label via soft one-hot tag type, which can increase the loss and effectively avoid overfitting for the training dataset with imbalance class.

$$p_i = \begin{cases} (1-\varepsilon) & if(i=y) \\ \dfrac{\varepsilon}{k-1} & if(1^1 \neq y) \end{cases}$$

$$Loss_i = \begin{cases} (1-\varepsilon) * Loss & if(i=y) \\ \varepsilon * Loss & if(1^1 \neq y) \end{cases}$$

The EfficientNet-B5 model performed the best on the 3-class dataset. As this grouping was based on the diagnostic signs summarized in Table I, the results suggest that the model can achieve a performance in accordance with clinical guidelines for COVID-19. The sensitivity of the 2nd class (clinical features 1-

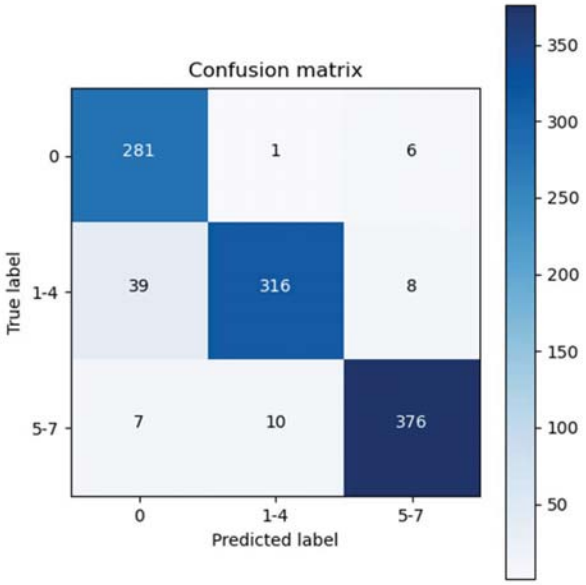

Fig. 2. Confusion Matrix for EfficientNet-B5 based network with 3 class dataset

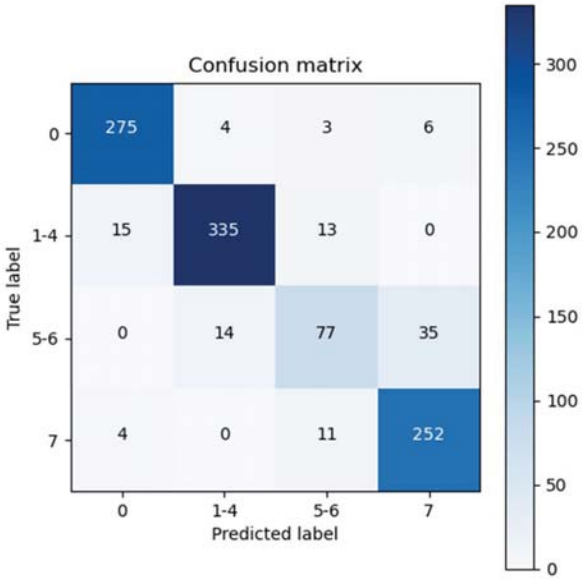

Fig. 3. Confusion Matrix for EfficientNet-B5 based network with 4 class dataset

TABLE V. STATISTIC ANALYSIS OF BEST EFFICIENTNET-B5 NETWORK ON 3 CLINICAL STAGES DATASET

Metric	Clinical Features (Class)			
	0	*1-4*	*5-7*	*Micro-Average*
f1	0.914	0.916	0.960	**0.932**
Accuracy	0.949	0.944	0.970	**0.955**
Sensitivity	0.976	0.871	0.957	**0.932**
Specificity	0.939	0.984	0.978	0.966
Precision	0.859	0.966	0.964	**0.932**
False Postive rate	0.061	0.016	0.022	0.034
False Negative rate	0.024	0.129	0.043	**0.068**
Negative Predictive Value	0.990	0.934	0.974	0.966

TABLE VI. STATISTIC ANALYSIS OF BEST EFFICIENTNET-B5 NETWORK ON 4 CLINICAL STAGES DATASET

Metric	Clinical Features (Class)				
	0	*1-4*	*5-6*	*7*	*Micro-Average*
f1	0.945	0.936	0.670	0.900	0.899
Accuracy	0.969	0.956	0.927	0.946	0.950
Sensitivity	0.955	0.923	0.611	0.944	0.899
Specificity	0.975	0.974	0.971	0.947	0.966
Precision	0.935	0.949	0.740	0.860	0.899
False Postive rate	0.025	0.026	0.029	0.053	0.034
False Negative rate	0.045	0.077	0.389	0.056	0.101
Negative Predictive Value	0.983	0.959	0.948	0.980	0.966

4) for the model is low, with a substantial number of images mistakenly labelled as class 0. This may be indicative that some images of the 2nd class share similar features to class 0 and are not well separable. Comparatively for the 4-class dataset, the sensitivity for the 3rd class (clinical features 5-6) was 0.611, with almost a third misclassified as the most severe class. This could be due to both the 3rd and 4th class representing varying severity of consolidation, which share clinical features of pieces and hepatization. In addition, the specification of a 3cm length as a clinical feature may not be suitably strong as it depends on the how the LUS device is positioned and oriented. Also, the class imbalance may have caused the overfitting of class 7. Finally, for the 8-class dataset, the EfficientNet model performed poorly for classes 2,3 and 6. Even with the image augmentation, weighted loss function, k-fold cross-validation and label smoothing, many images of classes 2 and 3 were classified as class 1, while numerous images of class 6 were classified as class 7. This could possibly be attributed to the 8-class dataset having the worst imbalance of data – class 2 and 7 has the least number of images for training.

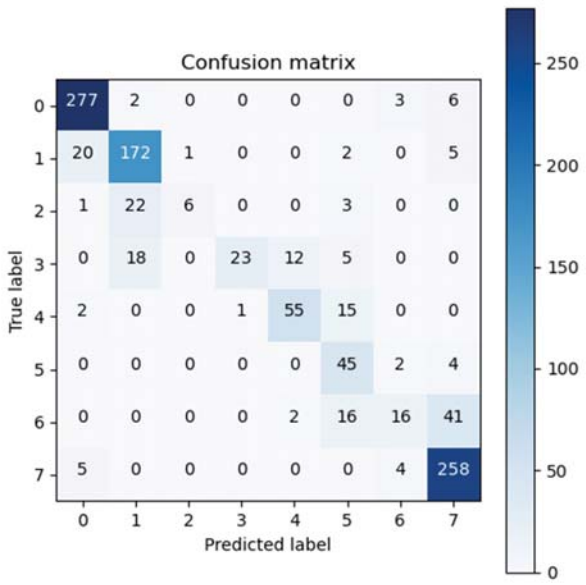

Fig. 4. Confusion Matrix for EfficientNet-B5 network with 8 class dataset

TABLE VII. STATISTIC ANALYSIS OF BEST EFFICIENTNET-B5 NETWORK ON 8 IMAGOLOGICAL FEATURES DATASET

Metric	Clinical Features (Class)								
	0	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*Micro-Average*
f1	0.934	0.831	0.308	0.561	0.775	0.657	0.320	0.888	0.816
Accuracy	0.963	0.933	0.974	0.966	0.969	0.955	0.935	0.938	0.954
Sensitivity	0.962	0.860	0.188	0.397	0.753	0.882	0.213	0.966	0.816
Specificity	0.963	0.950	0.999	0.999	0.986	0.959	0.991	0.928	**0.974**
Precision	0.908	0.804	0.857	0.958	0.797	0.523	0.640	0.822	0.816
False Postive rate	0.037	0.050	0.001	0.001	0.014	0.041	0.009	0.072	**0.026**
False Negative rate	0.038	0.140	0.813	0.603	0.247	0.118	0.787	0.034	0.184
Negative Predictive Value	0.985	0.966	0.975	0.966	0.982	0.994	0.942	0.988	**0.974**

IV. CONCLUSION

Three state of the art CNN-based deep learning models for the detection and classification of pneumonia in LUS images was investigated. The EfficientNet-B5 model achieved the best result and could achieve similar performance with respect to an experience clinician. The model in its current form could potentially aid in easing the workload of clinicians and enable an ultrasound user to rapidly identify patients with pneumonia. This work is a proof of the promising potential of the LUS device in pneumonia diagnosis and proves the viability of deep learning for LUS classification task. In the future, we will extend the models to a larger dataset as more become available, further validating the method's efficacy and accuracy. With further research, we believe LUS device can become the new gold standard for pneumonia diagnosis in the near future.

ACKNOWLEDGMENT

The authors would like to thank The Second Affiliated Hospital of Zhejiang University, School of Medicine, Ultrasound Section for providing the LUS images and label information for the creation of the datasets.

REFERENCES

[1] J. T. Hagaman, G. W. Rouan, R. T. Shipley and R. J. Panos, "Admission chest radiograph lacks sensitivity in the diagnosis of community-acquired pneumonia.," *The American journal of the medical sciences,* vol. 337, no. 4, pp. 236-40, 4 2009.

[2] D. A. Lichtenstein, N. Lascols, G. Meziere and A. Gepner, "Ultrasound diagnosis of alveolar consolidation in the critically ill.," *Intensive care medicine,* vol. 30, no. 2, pp. 276-281, 2 2004.

[3] G. Volpicelli, A. Mussa, G. Garofalo, L. Cardinale, G. Casoli, F. Perotto, C. Fava and M. Frascisco, "Bedside lung ultrasound in the assessment of alveolar-interstitial syndrome.," *The American journal of emergency medicine,* vol. 24, no. 6, pp. 689-96, 10 2006.

[4] S. Parlamento, R. Copetti and S. Di Bartolomeo, "Evaluation of lung ultrasound for the diagnosis of pneumonia in the ED.," *The American journal of emergency medicine,* vol. 27, no. 4, pp. 379-84, 5 2009.

[5] N. Xirouchaki, E. Magkanas, K. Vaporidi, E. Kondili, M. Plataki, A. Patrianakos, E. Akoumianaki and D. Georgopoulos, "Lung ultrasound in critically ill patients: comparison with bedside chest radiography.," *Intensive care medicine,* vol. 37, no. 9, pp. 1488-93, 9 2011.

[6] A. Pagano, F. G. Numis, G. Visone, C. Pirozzi, M. Masarone, M. Olibet, R. Nasti, F. Schiraldi and F. Paladino, "Lung ultrasound for diagnosis of pneumonia in emergency department.," *Internal and emergency medicine,* vol. 10, no. 7, pp. 851-4, 10 2015.

[7] J. Seyedhosseini, G. Bashizadeh-fakhar, S. Farzaneh, M. Momeni and E. Karimialavijeh, "The impact of the BLUE protocol ultrasonography on the time taken to treat acute respiratory distress in the ED," *The American Journal of Emergency Medicine,* vol. 35, pp. 1815-1818, 2017.

[8] R. Copetti and L. Cattarossi, "Ultrasound diagnosis of pneumonia in children.," *La Radiologia medica,* vol. 113, no. 2, pp. 190-8, 3 2008.

[9] T. Liang, "Handbook of COVID-19 prevention and treatment," *The First Affiliated Hospital, Zhejiang University School of Medicine. Compiled According to Clinical Experience,* 2020.

[10] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *arXiv preprint arXiv:1905.11946,* 2019.

[11] Szegedy, Christian, et al. "Rethinking the inception architecture for computer vision." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.