Open Journal of Medical Imaging, 2022, 12, 83-102
https://www.scirp.org/journal/ojmi
ISSN Online: 2164-2796
ISSN Print: 2164-2788

Explainable Analysis of Deep Learning Models for Coronavirus Disease (COVID-19) Classification with Chest X-Ray Images: Towards Practical Applications

Eri Matsuyama[1*], Haruyuki Watanabe[2], Noriyuki Takahashi[3]

[1]Faculty of Informatics, University of Fukuchiyama, Kyoto, Japan
[2]School of Radiological Technology, Gunma Prefectural College of Health Sciences, Gunma, Japan
[3]School of Hearth Sciences, Fukushima Medical University, Fukushima, Japan
Email: *matsuyama-eri@fukuchiyama.ac.jp

How to cite this paper: Matsuyama, E., Watanabe, H. and Takahashi, N. (2022) Explainable Analysis of Deep Learning Models for Coronavirus Disease (COVID-19) Classification with Chest X-Ray Images: Towards Practical Applications. *Open Journal of Medical Imaging*, **12**, 83-102.
https://doi.org/10.4236/ojmi.2022.123009

Received: June 14, 2022
Accepted: July 16, 2022
Published: July 19, 2022

Abstract

In recent years, with numerous developments of convolutional neural network (CNN) classification models for medical diagnosis, the issue of misrecognition/misclassification has become more and more important. Thus, research on misrecognition/misclassification has been progressing. This study focuses on the problem of misrecognition/misclassification of CNN classification models for coronavirus disease (COVID-19) using chest X-ray images. We construct two models for COVID-19 pneumonia classification by fine-tuning ResNet-50 architecture, *i.e.*, a model retrained with full-sized original images and a model retrained with segmented images. The present study demonstrates the uncertainty (misrecognition/misclassification) of model performance caused by the discrepancy in the shapes of images at the phase of model construction and that of clinical applications. To achieve it, we apply three XAI methods to demonstrate and explain the uncertainty of classification results obtained from the two constructed models assuming for clinical applications. Experimental results indicate that the performance of classification models cannot be maintained when the type of constructed model and the geometric shape of input images are not matched, which may bring about misrecognition in clinical applications. We also notice that the effect of adversarial attack might be induced if the method of image segmentation is not performed properly. The results suggest that the best approach to obtaining a highly reliable prediction in the classification of COVID-19 pneumonia is to construct a model using full-sized original images as training data and use full-sized original images as the input when utilized in clinical applications.

Keywords

Explainable AI, CNN, CXR Image, COVID-19 Pneumonia

1. Introduction

With the progress of artificial intelligence (AI) technology, especially deep learning (DL) technology, the range of its utilization has expanded rapidly and has been used in the medical imaging field. Recently, studies on DL-based automated classification of coronavirus disease (COVID-19) pneumonia caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) have been actively conducted [1] [2] [3] [4]. Some of these studies have reported high classification accuracy. However, while the accuracy is high and is expected to be utilized in COVID-19 diagnosis, the evaluation of whether they should be used clinically or not has yet to be determined. Some reasons are due to the fact that the computational process of the model used is complex, its operation is difficult to understand, and it is practically a black box. One of the problems with black boxes in classification is that of misclassification. Ribcage *et al.* reported that in their classification model for wolves and Siberian huskies, the classifier predicts "wolf" if the image has a snow background, and "husky" otherwise, regardless of animal color, position, pose, etc. [5] Zech *et al.* gave an example of a pneumonia detection model for chest X-ray (CXR) images in which L/R markers in the images were learned as strong features for classification [6]. These can be attributed to the quality and bias of the training data set. On the other hand, adversarial perturbations [7] [8] [9] have been demonstrated that they can cause a classification model to misclassify images. Adversarial perturbations, first discovered by Szegedy *et al.* [9], are slight modifications of deep convolutional neural network (CNN) model input that cause misclassification. That is, intentionally giving a small fluctuation (noise) to the input image, which is not easily noticeable to the human eye, can cause the model to make a misclassification and a wrong prediction. Goodfellow *et al.* reported an interesting adversarial example [8]. In this example, a perturbation is applied to an image that the CNN model was originally able to correctly classify as a panda, causing it to be misclassified as a gibbon. However, to a human, the difference from the original image is barely noticeable and the image appears to remain a panda. However, with the DL research progress, there have been studies regarding misrecognition. Therefore, explaining CNN misclassification is a vitally important issue in the development of reliable models.

CNN-based classification models typically learn to use particular features to represent attributes when they appear in the training dataset. Therefore, in order to decrease the misclassification of an image, we believe that it is very important to analyze how the model's performance is affected when an unlearned feature is inputted. It is even more important in the case of medical images. In the lesion classification of medical images, the cropped images of lesioned areas are generally used as the training dataset (data for training and validation). The issue of classification of COVID-19 using CNN is the same as that of other disease classification models. Despite promising results of using CNN to detect COVID-19 on computed tomography (CT) images, the use of CT for this purpose is limited by concerns about time, cost, and radiation exposure. X-ray machines are more affordable and portable, making them an important alternative. Therefore, if the diagnostic performance of CXR is further improved, CXR could be utilized as a tool for the diagnosis of COVID-19.

Studies of disease classification on CXR images have been divided into two main types. Type 1 is a study in which the CXR images themselves (the whole images; hereinafter referred to as the original images) are used as the training dataset [1] [2] [3], and type 2 is a study in which images of the lung region (only the lung regions are extracted from the original images) are used as the training dataset [4] [10] [11]. In other words, in type 1 (the training dataset consists of a set of original images), the model was trained using original images with shapes that contain many structures unrelated to lung lesions. Thus, the model may learn the misplaced regions that are not lesions as features (e.g., when the shoulder joint is in the raised state, the model might learn it as COVID-19 pneumonia). On the other hand, when the constructed model is clinically applied and put into practical use, it is considered that the original images of CXR are inputted. In the case of the model trained on the images of the lung region only (model of type 2), the model will predict an unknown geometric shape image that has not been trained (did not exist at the training phase). In other words, the inputs of a model can be regarded as the cause of the prediction, even if the relationship between the inputs and the outcome to be predicted might not be causal in reality. Therefore, even if the model shows high performance during the construction phase, its performance might not be guaranteed at the stage of practical use. Thus, in order to develop a reliable and high-performance classification model, it is necessary to evaluate the model's accuracy during the construction phase, perform verification assuming practical use, and clarify and analyze the evidence of prediction. For example, in the case of COVID-19 pneumonia, the most frequently observed lesion distribution patterns on CXR and CT images are bilateral involvement, peripheral distribution, and ground-glass opacification (GGO) [12]. Therefore, as a well-designed model, it should reflect these radiological findings and use them as evidence for prediction, when employed in clinical applications.

In this work, we focus on the uncertainty (misrecognition/misclassification) of the constructed, high-performance classification models to be utilized in clinical applications. To this end, we construct two types of classification models for COVID-19 pneumonia (a model retrained with original images and a model retrained with a segmented image having a lung region) and perform verification assuming clinical practical use. We use three explainable AI (XAI) techniques to explain the evidence of prediction of the constructed models. By analyzing the

XAI results and the models' prediction results and prediction scores, we expect that the presence of misrecognition/misclassification and its triggering factor in clinical applications will be clarified.

We used ResNet-50 [13] [14] [15], which has been proven effective in many tasks in the field of medical imaging, as the base for constructing two classification models. ResNet, short for residual network, is a pre-trained model that has been trained on over 1 million images in the ImageNet database [16] and was the winner of the ImageNet Challenge in 2015. ResNet can have very deep networks of up to 152 layers. There are five versions of the ResNet model, each containing 5, 34, 50, 101, and 152 layers, respectively. ResNet-50 corresponds to a 50-layer residual network. However, since ResNet is a network designed for large-scale natural images intrinsically different from CXR images, it is not suitable for performing general transfer learning in our study. Therefore, we use the two types of training datasets described above to respectively perform fine-tuning on all layers of the architecture.

In this paper, we conducted an explainable analysis of deep CNNs for COVID-19 classification. The main contributions of this work include:

- We build two deep learning models based on a well-established ResNet-50 model for the classification of COVID-19 infection, along with evaluating and comparing the performance of the models.
- We employ three widely-used XAI methods, *i.e.*, local Interpretable model-agnostic explanations (LIME), occlusion sensitivity, and gradient-weighted class activation mapping (Grad-CAM), to visually understand and explain the constructed models' predictions. We also qualitatively compare and assess the function of the XAI methods.
- We perform an in-depth analysis of the experimental results and draw some inspiration from explainable COVID-19 disease classification using CXR images in clinical practical applications.

The rest of the paper is organized as follows. Section 2 describes the image data sets, the proposed CNN models, and three XAI methods used in this work. Section 3 presents the experimental results. Section 4 gives the discussion of the results. Section 5 concludes this work comprehensively.

2. Methods

In this study, fine-tuning of the pre-trained ResNet-50 was conducted and re-training with 1200 CXR images was performed with 10-fold cross validation. The classification targets were COVID-19 pneumonia, non-COVID-19 viral pneumonia, bacterial pneumonia, and normal. The following two types of CNN models were constructed. They were chosen based on the assumption that the input image utilized in the COVID-19 disease classification is either an entire image or a segmented image from the entire image. Type 1 is a model retrained on the full-size, original CXR images (hereafter referred to as "model_original"). Type 2 is a model retrained on segmented lung region in CXR images (hereafter referred to as "model_segmented"). It is assumed that the above two types of models will be clinically applied and put into practical use. A total of 20 unused, annotated original images that were not included in the retraining dataset were inputted into the models as test images for analyzing the prediction scores and misrecognition of the models. In addition, in order to evaluate the prediction results obtained from inputting the unused, annotated original (full-sized) images into the two classification models, the prediction results obtained by using the segmented (only lung region) images cropped from the same unused images as inputs to the same two models were also calculated. Three XAI methods, namely, LIME [5] [17] [18], occlusion sensitivity [19], and Grad-CAM [20] were employed for the analysis of misrecognition.

2.1. Image Datasets

The image datasets used in this study are the COVID-19 CXR Datasets publicly published by Unais Sait *et al.* [21] and the image set published by the Radiological Technology Society of Japan [22]. Thus, ethics issues do not arise in this work and the requirement to obtain informed consent was waived.

A total of 1200 images (jpg and DICOM images) were randomly selected from the above described two datasets consisting of 300 images for each of 4 categories, *i.e.*, COVID-19 pneumonia, viral pneumonia, bacterial pneumonia, and normal. We applied 10-fold cross-validation for the network re-training. The re-training dataset used for constructing model_segmented were the images of the lung region obtained by cropping those original images from the dataset which were used for constructing model_original. Therefore, there is no difference in lesion complexity between the training datasets of the two models. As a method for segmentation of lung regions, edge enhancement of the original images was performed using an extended Sobel filter with a mask size of 5 × 5 pixels, and the lung contours were extracted by applying binarization [23]. To account for the overlap between the heart and lungs, the mediastinum was manually adjusted and a segmentation mask was created. **Figure 1** shows some examples of training data for constructing model_original, generated lung masks for segmentation, and training data for constructing model_segmented.

2.2. Overview of ResNet-50-Based Architecture

In this study, we performed re-training based on the pre-trained ResNet-50. The structure of the network is shown in **Figure 2**. ResNet-50 consists of 16 processing blocks and equips with two types of shortcut connections as shown in **Figure 2(a)**. One is a module called convolution block that puts a convolution layer in a shortcut (the input dimension is smaller than the output dimension) (**Figure 2(b)**). The other is a module called identity block (input has the same dimension as output) with no convolution layer in the shortcut (**Figure 2(c)**). Each module has a bottleneck building block structure consisting of 3 layers per block (1 × 1, 3 × 3, and 1 × 1 convolution layers), which allows the number of parameters to be reduced without degrading performance. We retrained all layers of the network with CXR image datasets. In other words, four categories

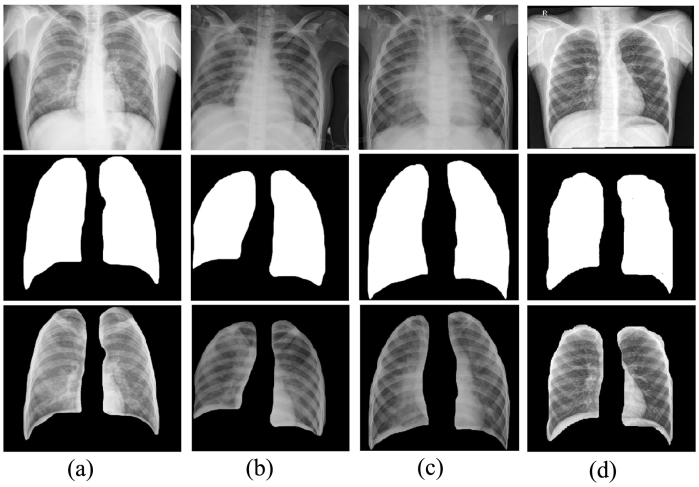


Figure 1. Examples of CXR images from the collected datasets. Upper row: examples of training data used for constructing model_original. Middle row: segmentation masks used for the upper-row images. Lower row: images after performing image processing using the masks shown in the middle low (used as training data of for constructing model_segmented). (a) COVID-19 pneumonia image, (b) Viral pneumonia image, (c) Bacterial pneumonia image, (d) Normal image.

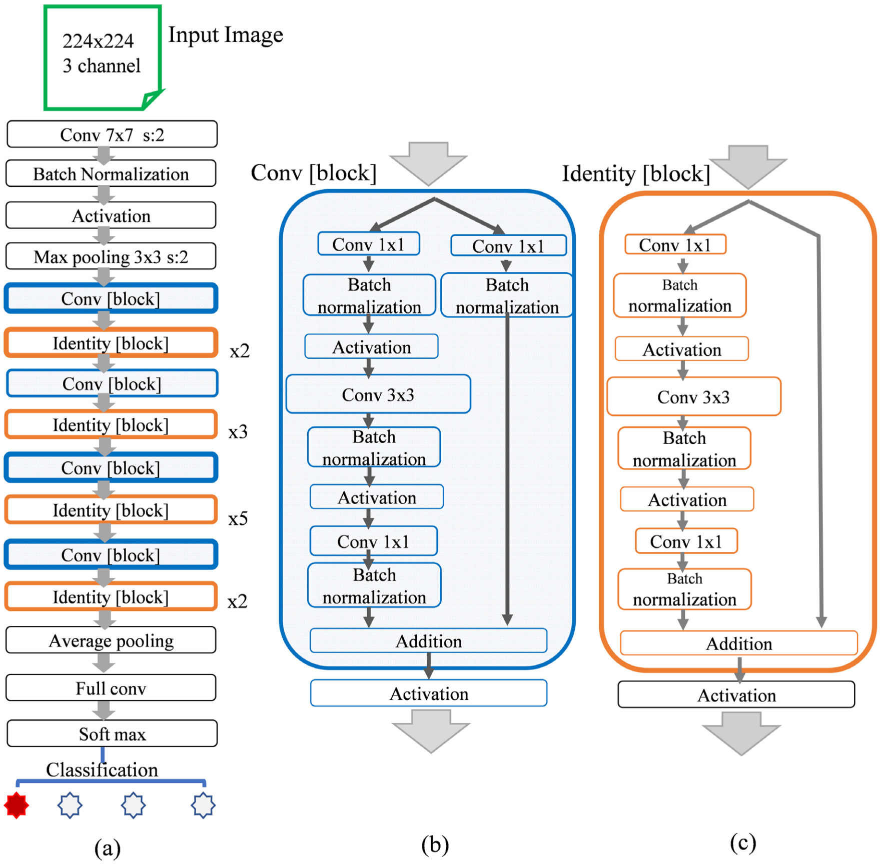


Figure 2. Outline of ResNet-50-based architecture. (a) Overview of the overall structure of the network. The symbols of ×2, ×3, ×5 in the figure are the number of blocks. (b) Structure of a convolution block where input dimension varies. (c) Structure of an identity block where input dimension does not change.

were classified using fine-tuning network without placing frozen layers. The last fully-connected layer and the final classification layer of the network were newly replaced according to the number of categories.

Since ResNet structurally requires input data size to be 224 × 224, the input images were all resized with the bi-cubic interpolation. We performed a 10-fold cross-validation for the network re-training. Of the total 1200 images, 1080 images, 90% of the total images, were used for re-training and the remaining 120 images for validation. The average of these classification accuracies is taken as the accuracy of the network. The mini-batch size was 36 and the optimizer used was Adam. Adam is an optimization algorithm that can be used as an alternative to the classical stochastic gradient descent method to iteratively update network weights based on training data [24]. In the re-training, in order to improve the accuracy, the parameters were adjusted so that the learning speed advanced faster in the newly replaced fully connected layer. On the contrary, the learning speed decreased in the transfer layer. Also, parameters were adjusted so that the learning rate decreased every 5 epochs. Furthermore, an L2 norm regularization was applied to the cost function (also referred to as loss function) to prevent overfitting. The epoch setting was determined by performing accuracy verification at each iteration cycle, and re-training stops after 5 consecutive iterations when the accuracy has stopped improving.

2.3. Explainable Artificial Intelligence

The outlines of three widely used XAI methods, LIME [5], occlusion sensitivity [19] and Grad-CAM [20] adopted in this study for visual analysis are briefly described in the following sub-sections, respectively.

2.3.1. Local Interpretable Model-Agnostic Explanations (LIME)

LIME is an algorithm proposed by Ribeiro *et al.* [5] that explains the individual predictions of a black box model. The key idea of the method is to use a simpler glass box model that is easier to interpret to locally approximate the black box model. LIME creates perturbations by turning on and off a portion of the super pixels in the image. To create a human-readable representation, LIME attempts to determine the importance of contiguous super pixels in a source image with respect to the output class. It is a procedure that allows one to understand how the input features of a deep learning model affect its predictions. The procedure is to divide the input image into super pixels. A super pixel is an interconnected pixel of location and color (similar colors); the number of super pixels makes the region segmentation finer and more complex. **Figure 3** shows an example of region segmentation by super pixels. In this study, both lungs of each image were segmented into at least upper, middle, and lower lung fields, respectively, and the shoulder and diaphragm were distinguished. In addition, the number of super pixels was set to 40 to avoid complexity. Here, it is possible to add variations to the image (to change image features) by randomly turning on (active) or off (inactive) the individual regions divided by the super pixels. Therefore, multiple

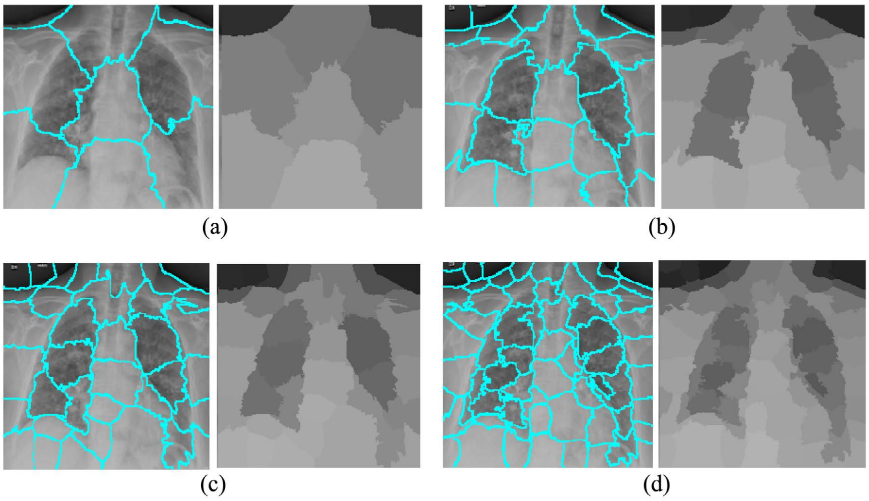


Figure 3. An example of region segmentation using super pixels. The features and resolution obtained differ depending on the number of super pixels. (a) 10 super pixels, (b) 30 super pixels, (c) Segmentation result using 40 super pixels adopted in this study, (d) 60 super pixels.

sample images that are similar to the input image can be generated by performing on (active) and off (inactive) settings. There are 10,000 types of sample images generated in this study. LIME predicts each generated sample image with a respective black box model and then trains the simple classifier (regression model) based on the sample image set. The cosine distance between each sample image and the original image was calculated, and the higher the similarity between the sample image and the original image, the greater its weight and importance. After being weighting, it was fitted to a linear regression model and the approximation coefficients for each feature were acquired. Features with large coefficients play a major role in the prediction of the black box model. In this work, we used a linear regression model based on Lasso regression. More details about Lasso regression can be found in [25].

2.3.2. Occlusion Sensitivity

Occlusion sensitivity is an approach for understanding which parts of an image are most important for classification. It helps us understand the learning behavior of the underlying task by determining whether the network is actually categorized based on task-specific features [19]. Specifically, different small regions of the input image are sequentially blocked with an occlusion mask (a rectangular mask) and the change in probability score for a predetermined class is measured as a function of mask position. The procedure of the approach is as follows.

Step 1: Classify the target image with a black box model and confirm the probability score of the classification class.

Step 2: Replace a small region of the input image with an occlusion mask to give the input image a small variation.

Step 3: Input the variation image obtained from step 2 to the black box model and calculate the probability score of the classification class.

Step 4: Slide the occlusion mask region in the column direction to calculate the probability score of the classification class.

Step 5: Repeat the steps until the occlusion mask moves across the entire image.

The procedure can highlight which regions of the image are most important in the classification. In this study, since we want to target the scattered and relatively wide lesion regions such as multiple patchy shadows and consolidation, the mask size is set to the integer value, which is nearest to 30% of the input image dimension. The width of shift is an integer value that is nearest to 20% of the input image size.

2.3.3. Gradient-Weighted Class Activation Mapping (Grad-CAM)

The CNN model consists of a feature extraction module and a classification module. Generally, the classification module contains a fully connected neural network model, and the extracted features are converted into a probability score for each category in the softmax layer. The final prediction classification result of the network is the category with the highest probability score. Grad-CAM [20] is a class-discriminative localization method that can generate visual explanations without requiring architectural changes or retraining. It localizes the relevant image area, and uses the gradient (derivative factor) of the feature map of the final convolutional layer of the network to emphasize which part of the image has the greatest effect on the probability score of the final prediction. The areas where the gradient is large are the areas that have a great effect on the prediction results. **Figure 4** shows the flowchart of how to implement Grad-CAM. More details about Grad-CAM can be found in [20].

2.3.4. Brief Comparison of XAI Methods Used for Explanation of Pneumonia in CXR Images

The XAI is used to provide local explanations. The local explanation is an explanation given to each prediction result. This requires two steps: 1) calculating the impact of each input on the output, and 2) expressing it in a human-understandable way. All of the above-mentioned three XAI methods include these two steps, but the methods are different from each other, and the results obtained are also different. For example, because occlusion sensitivity measures important features by stride in rectangular regions, the importance of the combination of features between regions is unknown. In LIME, the predictions of a simple classifier used for approximation are not always correct. Grad-CAM shows which regions of the image affected the probability score of the final prediction and it does not always provide evidence for COVID-19 pneumonia. Therefore, we believe that the prediction explanations using a combination of the three XAI methods together would be useful. **Table 1** shows a summary comparison of the three methods.

3. Results

The pneumonia classification models constructed in this study are a model

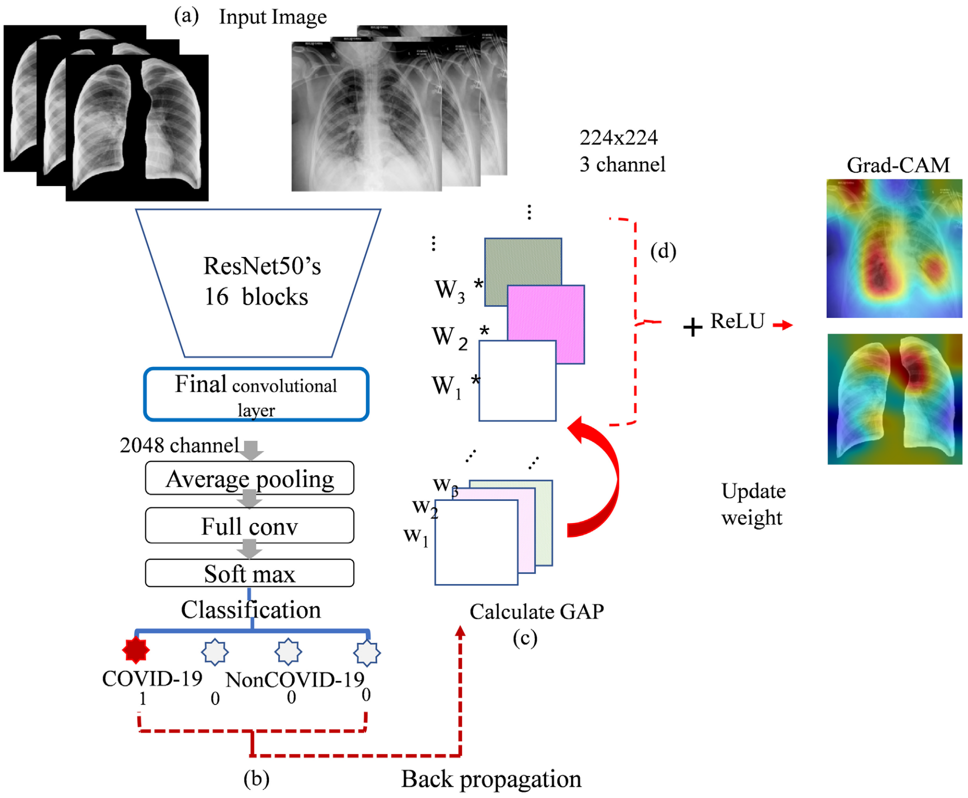


Figure 4. Schematic of the Grad-CAM. (a) Input test images to the network and obtain prediction results. (b) Back propagate with COVID-19 class as 1 and other classes as 0. (c) Calculate the global average pooling (GAP) of the gradient for each channel and use it as the weight for the network. (d) Perform a weighted combination of forward activation maps, and follow it by a ReLU to obtain Grad-CAM.

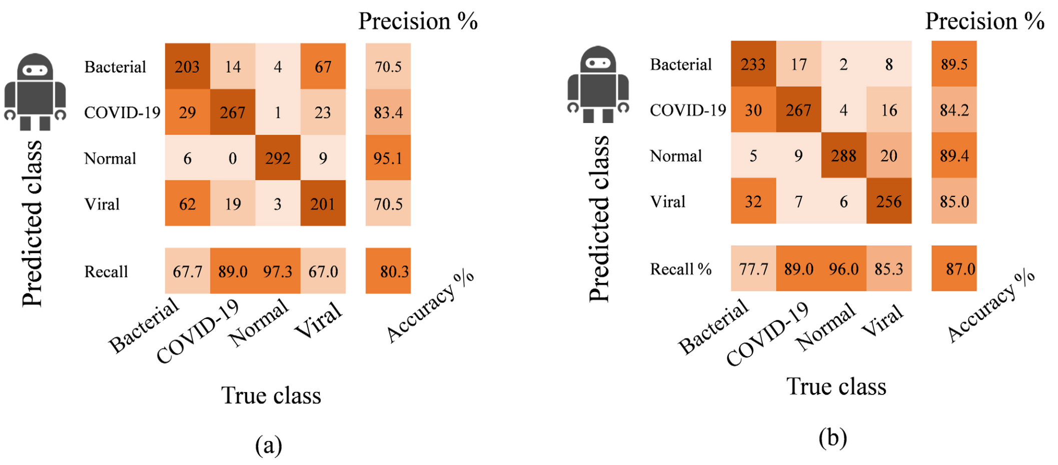
Table 1. Summary comparison of the characteristics of the three XAI methods.

XAI Method	Characteristics	Description of Pneumonia CXR Image	Spatial Resolution
LIME	*A method of approximating the prediction result of a black box model with another model. *Approximate classifiers are not always correct.	*Multiple lesions (GGO, consolidation, etc.) in the image can be effectively visualized. *Highlight which superpixel region of the image is most important for classification.	*The spatial resolution is adjusted according to the number of superpixels, allowing for a high degree of flexibility in feature investigation.
Occlusion Sensitivity	*A method of making data partially obstructive and measuring the importance of a region using a black box model. *The importance of combination features cannot be expressed.	*Multiple lesions (GGO, consolidation, etc.) in the image can be visualized. *Highlight which region of the image is most important for classification.	*The spatial resolution is adjusted according to the mask size and the size of the stride, allowing for a high degree of flexibility in feature investigation.
Grad-CAM	*A method to make the black box model itself have an evidence for judgment. *Describes the regions that affects the final score.	*Focuses on a wide range of a major lesion and therefore cannot identify (visualize) multiple lesions in the image [26]. *Highlight the pixels that contribute to change the final decision.	*The spatial resolution of the feature map in the final convolution layer is low because it is 7 × 7 (in the case of ResNet-50).

trained on whole CXR images (model_original and a model trained on lung region images (model_segmented). The overall accuracy at the phase of the construction of both models were 0.803 and 0.870, respectively, and the area under the receiver operating characteristic (ROC) curve (AUC) values were 0.969 and 0.898, respectively. Figure 5 shows the confusion matrices of the two constructed models.

The input images to both constructed models assumed for clinical application and practical use were 20 unused COVID-19 pneumonia CXR images with a variety of different findings. In this study, we used Grad-CAM to identify the feature regions that was utilized by the models to make the final decision during clinical application. On the other hand, the important feature regions that influence the class scores were identified by occlusion sensitivity. Furthermore, LIME was used to approximate the model's prediction results with a linear regression model to identify important features that have a strong influence on the class scores.

Figure 6(a) shows an example of 10 input images (both the original and segmented images of COVID-19 pneumonia CXR) assuming for clinical application purposes. Figure 6(b) and Figure 6(c) are the probability scores when these two types of images were respectively inputted to both the model_original and the model_segmented. Figure 7(a) is the original image of img.8 shown in Figure 6(a), and the lesions on the image are marked in Figure 7(b). Figure 7(c) is the segmented image of img.8, and the lesions are noted in the same way as shown in Figure 7(b). The locations marked with three asterisks "⁂" in the figure indicate the ground-glass shadows, the area marked with orange circles show hilar lymphadenopathy, and the areas marked with yellow ellipses indicate breast shadows. Figure 8 and Figure 9 illustrate the respective results when the original image (entire image) and the segmented image were inputted to both the model_original and the model_segmented. The top row of Figure 8 and Figure 9



Predicted class	Bacterial	COVID-19	Normal	Viral	Precision %
Bacterial	203	14	4	67	70.5
COVID-19	29	267	1	23	83.4
Normal	6	0	292	9	95.1
Viral	62	19	3	201	70.5
Recall	67.7	89.0	97.3	67.0	80.3

True class / Accuracy %

(a)

Predicted class	Bacterial	COVID-19	Normal	Viral	Precision %
Bacterial	233	17	2	8	89.5
COVID-19	30	267	4	16	84.2
Normal	5	9	288	20	89.4
Viral	32	7	6	256	85.0
Recall %	77.7	89.0	96.0	85.3	87.0

True class / Accuracy %

(b)

Figure 5. The confusion matrices of the two constructed models. The values in the bottom row of the matrices are recalls, those in the rightmost column are precision. The value in the cell located at the lower right corner is overall accuracy. (a) Result of model_original. (b) Result of model_segmented.

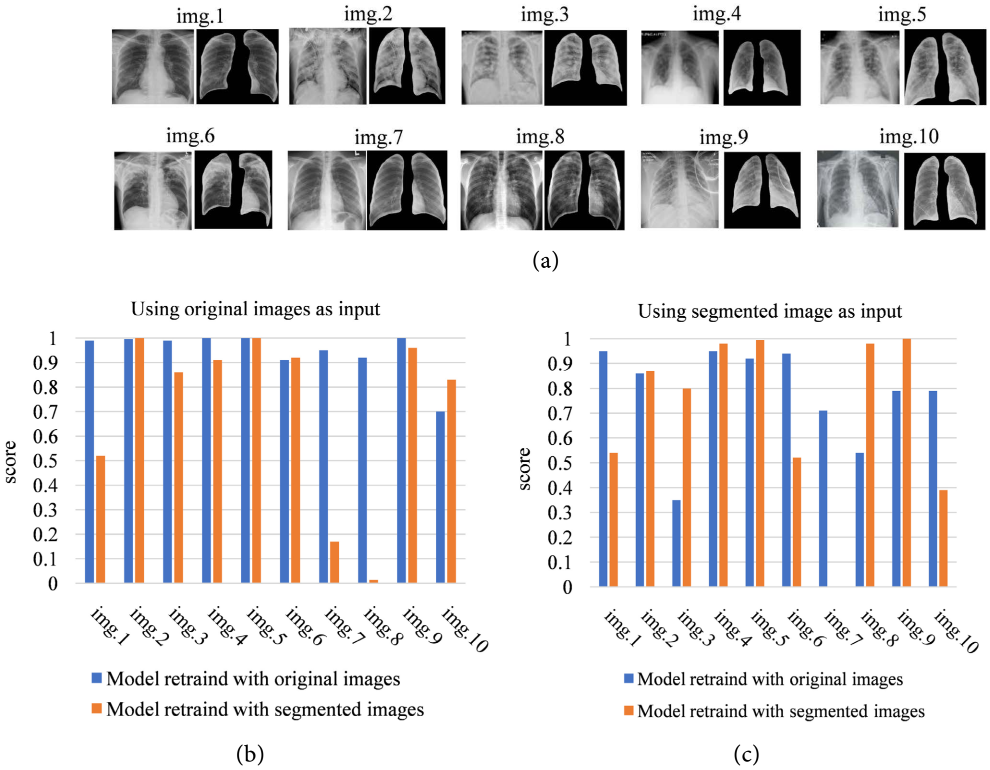


Figure 6. An example of the results of inputting COVID-19 pneumonia CXR images to each of the two constructed models. (a) Input images: original image on the left and segmented image on the right of the paired images. (b) Prediction scores of the two models when the original images were inputted, (c) Prediction scores of the two models when the segmented images were inputted. Here, it is notable that the prediction score of img.7 using the model_segmented is 0.

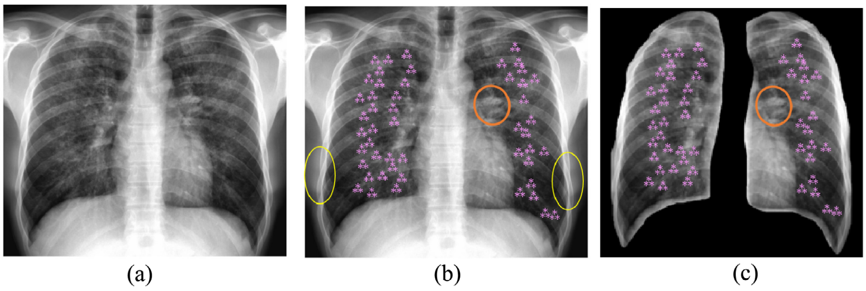


Figure 7. (a) The original image of img.8 shown in **Figure 6(a)**. (b) The lesions on the image are marked. (c) The segmented image of img.8. The locations marked with three asterisks "⁂" in the figures indicate the ground-glass shadows, the area marked with orange circles show hilar lymphadenopathy, and the areas marked with yellow ellipses indicate breast shadows.

shows the input image (COVID-19 CXR image), the numerical values are the scores for each class. Below the numerical values are the LIME results, and in the bottom row is the result of occlusion sensitivity maps. In both cases, the red regions of the heatmaps are the feature regions that have a great influence on the class score, and the influence decreases as it gets closer to blue. On the right side of the input image is the result of Grad-CAM. The regions that contributed most

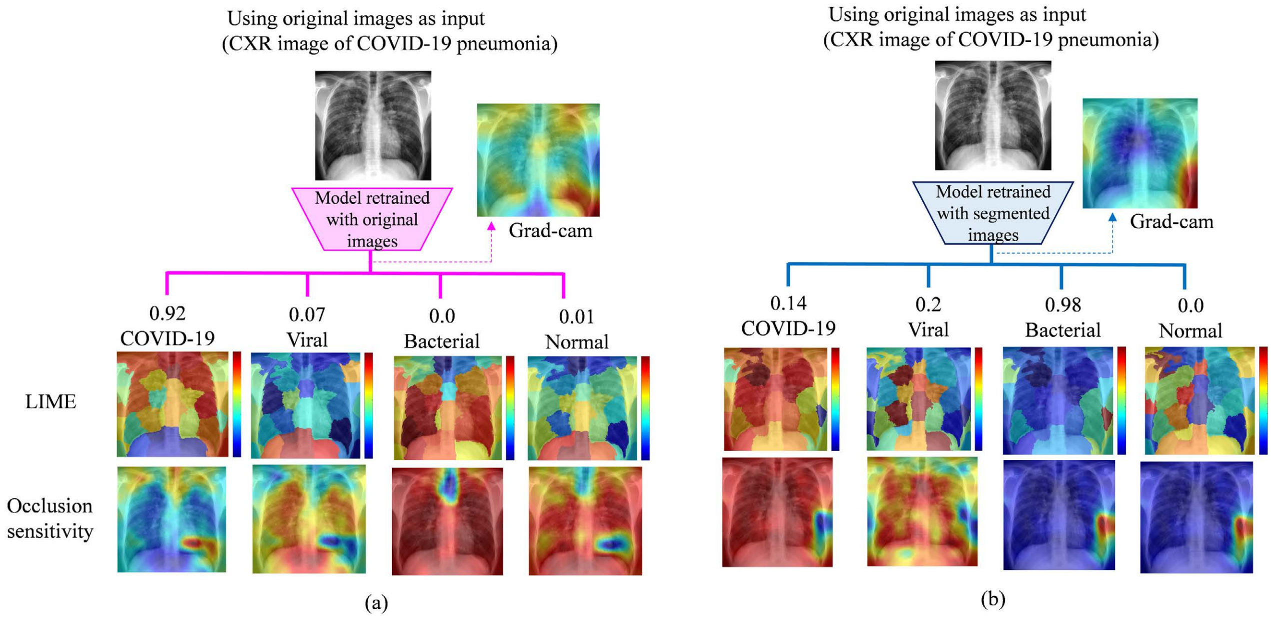

Using original images as input
(CXR image of COVID-19 pneumonia)
Model retrained with original images
Grad-cam
0.92 COVID-19
0.07 Viral
0.0 Bacterial
0.01 Normal
LIME
Occlusion sensitivity
(a)
Using original images as input
(CXR image of COVID-19 pneumonia)
Model retrained with segmented images
Grad-cam
0.14 COVID-19
0.2 Viral
0.98 Bacterial
0.0 Normal
(b)

Figure 8. An example of the results of inputting the original image to the two constructed models (in the case of img.8 shown in **Figure 6(a)**). (a) In the case of model_original. (b) In the case of the model_segmented.

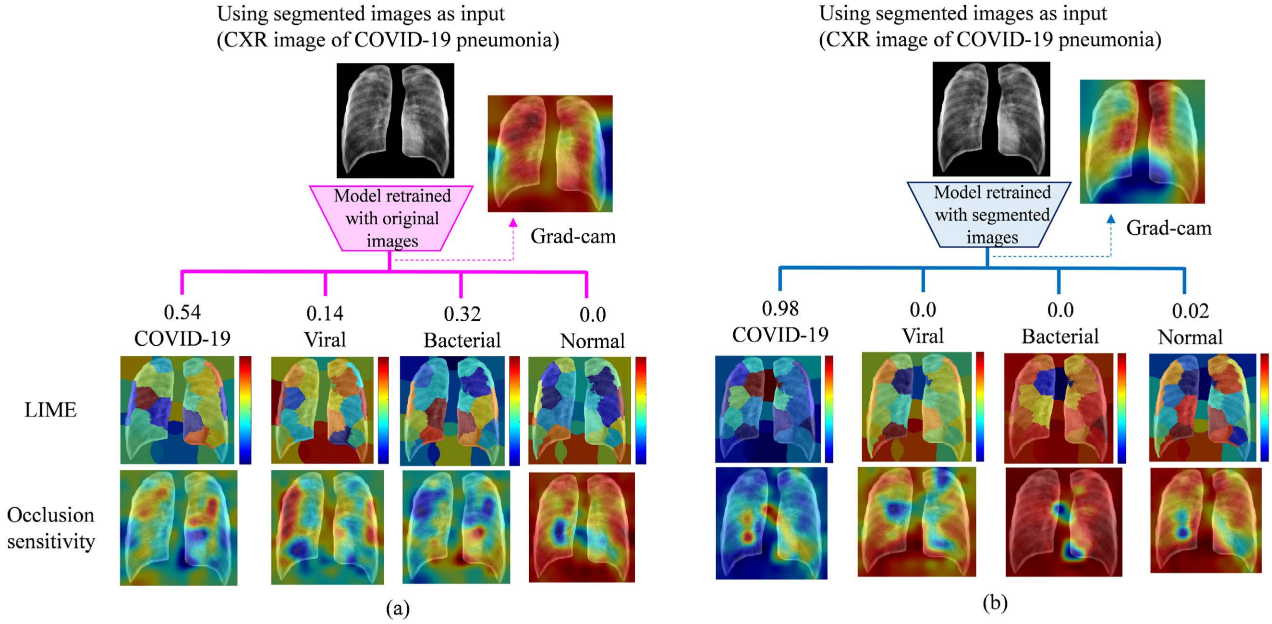

Using segmented images as input
(CXR image of COVID-19 pneumonia)
Model retrained with original images
Grad-cam
0.54 COVID-19
0.14 Viral
0.32 Bacterial
0.0 Normal
LIME
Occlusion sensitivity
(a)
Using segmented images as input
(CXR image of COVID-19 pneumonia)
Model retrained with segmented images
Grad-cam
0.98 COVID-19
0.0 Viral
0.0 Bacterial
0.02 Normal
(b)

Figure 9. An example of the results of inputting the segmented image to the two constructed models (in the case of img.8 shown in **Figure 6(a)**). (a) In the case of model_original. (b) In the case of the model_segmented.

to the final prediction are shown in red. When the color is closer to blue, it indicates a weaker contribution to the prediction.

Figure 10 shows the prediction results when the original image of number img.10 shown in **Figure 6** was inputted to the model_segmented (probability of COVID-19 is 83%). Also, in this figure, the important features that contributed significantly to the prediction are illustrated by Grad-CAM, LIME, and occlusion

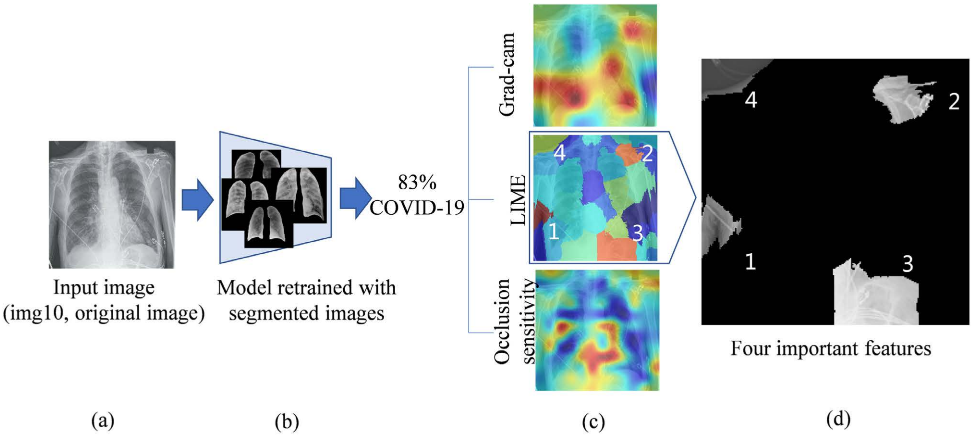


Figure 10. An example of inputting the original image to the model_segmented. (a) Input image. (b) Model_segmented and the prediction score. (c) Important features that provide the evidence for the predictions shown by the three XAI methods. (d) Magnified view of the 4 important features (super-pixel regions) shown by LIME. Numbers 1 to 3 in the figure are electrocardiogram cables outside the lung field, and number 4 is an oxygen mask tube in the background.

sensitivity. The feature regions that have a large effect on the class score are the red regions of the heatmaps, and the effect decreases as it turns to blue. Figure 10(d) is the magnified view of four feature regions (super pixels) that have a great influence on the class score as indicated by LIME.

4. Discussion

We constructed two models for COVID-19 pneumonia classification based on ResNet-50, and demonstrated their performance for clinical applications. As shown in Figure 5, the overall accuracy at the phase of construction is 80.3% in model_original and 87.0% in model_segmented, respectively, while the recall for COVID-19 pneumonia was both 89%, and the precision for that were 83.4% and 84.2%, respectively. In other words, the model_segmented is more accurate compared with the model_original. This result shows the superiority of using segmented images without unnecessary structures as training data to build a model. On the other hand, the false positive (FP) rates are 6.8% and 5.6%, respectively. The model_segmented has a slightly higher FP rate. However, in the classification of pneumonia, including COVID-19 pneumonia, it is important that the false negative (FN) rate (the percentage of COVID19 pneumonia misclassified as other pneumonia) should be low. The FN rates for both models are 10.3% and 10.4%, respectively. There is no statistically significant difference between the two models. This suggests that there might be no difference between the two models for detecting COVID-19 pneumonia. However, for example, when the original image of img.1 shown in Figure 6(a) was inputted to the model_segmented, the classification score was 52% , and when it was inputted to the model_original, the classification score was 99% (see Figure 6(b)). The results indicate that the model_segmented provides a low classification score and is not a reliable prediction. For other COVID-19 pneumonia original images, the model_segmented tended to classify them with lower scores. Furthermore, in the cases of the original images of img.7 and img.8, the model_segmented made misclassification (see **Figure 6(b)**). More specifically, in the case of img.7, the prediction scores for the 4 categories are: COVID-19 (17%), viral (34%), bacterial (49%), and normal (0%). Similarly, in the case of img.8, the prediction results are: COVID-19 (1.4%)) viral (0.2%), bacterial (98.4%), and normal (0%). In both cases, the model_segmented misclassified COVID-19 as "bacterial". These results suggest the risks of using the model_segmented. On the other hand, in the original image of img.10, the model_segmented classified COVID-19 pneumonia with a higher score, and the accuracy of the model_segmented at the phase of construction appeared to be maintained (see **Figure 6(b)**).

For verification, we used all unused, annotated images as input images and analyzed the evidence of prediction using the three XAI methods together. **Figure 10** illustrates the evidence for prediction for the original image of img.10 shown in **Figure 6(a)**. From **Figure 10(d)**, it can be seen that the important features explained with LIME are the electrocardiogram cables outside the lung field and the tube of the oxygen mask. Although the important features (regions shown by the heat maps) shown by the three XAI methods do not completely match, the Grad-CAM shown in **Figure 10(c)** highlights the electrocardiogram cables inside and outside the lung field. It is clear from **Figure 10(c)** that the occlusion sensitivity also highlights the electrocardiogram cables. That is, with respect to the original image of img.10, even though the model_segmented highlighted wrong features, it predicted a high score. These results suggest that when the original image is inputted to the model_segmented, unknown features may induce misrecognition and make prediction unreliable.

In this study, the segmented images were also inputted into both models for investigation in consideration of the possibility that segmented images might be used in clinical applications. As shown in **Figure 6(c)**, it is difficult to determine which model is superior when using segmented images as input. On the other hand, looking at the image of img.8, when the original image was inputted to the model_segmented, misclassification occurred (see **Figure 6(b)**). However, in the case that when the segmented image was inputted to the model_segmented, the model made the correct classification with a dramatically higher score (98%) (see **Figure 6(c)**). While these experimental results reflect the accuracy of the model_segmented at the phase of construction, they also suggest that the prediction result may vary significantly depending on the shape of the input image, for example, with or without segmentation or the image shape after segmentation. In other words, the reason why it is difficult to determine which model is more useful in **Figure 6(c)** is that depending on the method of segmentation, an effect such as Adversarial Attack [9] (giving a small perturbation to the image causes the model to misrecognize) might be brought about.

Figure 8 and **Figure 9** are the results of inputting the original and segmented images of img.8 into each of the two models. The right side of each input image is the result of Grad-CAM. The red regions are the areas that contributed most to the final classification, and the closer it is to the blue color, the lower the contribution. The numerical values in the figure are the prediction probabilities for each class. The middle and bottom rows below the numerical values are the results for LIME and occlusion sensitivity, respectively. In both cases, the regions that are important for classification are clearly shown on the heat maps. **Figure 8(a)** is the result of inputting the original image into the model_original. In this case, the probability of having COVID-19 pneumonia is 92%. The LIME results indicate that the important feature regions (red regions) that affect the prediction are the left lung field and the right upper middle lung field areas, while the subdiaphragmatic region (blue region) is less important. In occlusion sensitivity, the ground-glass shadows (red area) in the lower left lung field are indicated as important features. All of these XAI methods capture the lesion regions, indicating that the feature recognition is correctly performed. From the results of Grad-CAM, it can be seen that the model_original recognized the ground-glass shadows of the lower left lung field as the strongest evidence for prediction (92% as COVID-19). In this case, the probability of viral is 7%. The important feature of the viral (red region) as indicated by LIME in the subdiaphragmatic region. In other words, the condition of the subdiaphragmatic region (e.g., the lung region overlapping with the diaphragm) is similar to viral. The features of viral as indicated by occlusion sensitivity are ground glass shadows in the right lung, hilar and bronchi, ground glass shadows in the left lung, and subdiaphragmatic. These states suggest a similarity to viral. The probability of "Normal" obtained from the experiments is 1%. The result of LIME is a decision that the subdiaphragmatic region slightly resembles "Normal" case. The occlusion sensitivity indicates that the probability of "Normal" is 1%, excluding both the ground glass shadow of the left lower lung field and the area of the upper trachea (blue region). Based on these results, we believe that a highly reliable prediction can be obtained by inputting the original images to the model_original as a decision support tool utilized in clinical practice.

Figure 8(b) shows the results of inputting the original COVID-19 pneumonia image into the model_segmented. In this case, the prediction score for "bacterial" is 98%. LIME and occlusion sensitivity indicates that the important feature of "bacterial" is the left breast shadow outside the lung field. This is not the result of the model's mislearning as "bacterial", when breast shadows are present. The model used here is the model_segmented, and the breast shadow itself did not exist at the phase of retraining. In other words, it can be said that the inputted, unknown feature (the breast shadow outside the lung field) triggered the model to make mis-recognition. As a result, the model_segmented misclassified COVID-19 pneumonia with a predictive score of 14%. Although the important region for COVID-19 pneumonia, as indicated by LIME and occlusion sensitivity contains sufficient lesion areas in both lungs, it is not recognized as COCID-19 pneumonia. In Grad-CAM, the left breast shadow outside the lung field was used as evidence for the final prediction. Furthermore, both lung fields are indicated in blue, and it is clear that they were rarely used in the final prediction. In other words, when the model_segmented inputted the images with an unknown shape that did not exist at the phase of training, the model might make an uncertain prediction like the under-learned model. Figure 9 illustrates the results of inputting the segmented images of img.8 into the two constructed models. As shown in Figure 9(a), though the model_original correctly classified COVID-19 pneumonia, the prediction is disturbed and the confidence score is low (54%). In the LIME and occlusion sensitivity, the small lesion regions inside the lung field are regarded as an important feature. On the other hand, the heatmap colors in the background region and in the lung field are very similar. This suggests that the model_original also featured the region (body shape outside the lung field which was removed by the segmentation) that existed at the phase of retraining as a feature for judgment. From the results of Grad-CAM, it is also clear that the evidence of model_original judging the category as COVID-19 pneumonia, is based on the background area without structures in addition to the ground-glass shadows present in both lungs. In other words, it is considered the shape region that is present only at the training phase becomes a trigger factor of misrecognition, and as a result, a prediction with low certainty similar to lack of learning is made.

Figure 9(b) is the result of using the model_segmented. In this case, the score for COVID-19 pneumonia is 98% and is considered highly confident. However, both LIME and occlusion sensitivity regard the space between the left and right lungs (the area that has been removed by segmentation) as one of the important features. This suggests that the shape obtained after performing a certain segmentation processing may be a factor that induces misrecognition. In addition, LIME and occlusion sensitivity respectively indicate the important features of different areas of the lung regions. However, they both regard intrapulmonary lesions as important features. When these are combined, they are almost the same as the regions indicated by Grad-CAM. In other words, they represent the evidence for the final prediction. These results suggest that a more detailed and definitive predictive explanation can be obtained by using all three types of XAI methods together.

Our study had some limitations. First, we only constructed two types of constructed models using ResNet-50 as the backbone network. The development of a highly reliable classification model by integrating AI and machine-learning technologies is our next task. Second, the accuracy of the models at the phase of construction was 0.80 for the model retrained with full-size original images and 0.87 for the model retrained with the segmented images, respectively. While the achieved performance is very encouraging, further analysis is required on a larger set of COVID-19 images, to have a more reliable estimation of accuracy rates. Also, a more detailed validation is required to improve the accuracy of the classification models for widespread use. Further adjustment of the parameters of the network may be required. Third, The segmentation method used did not take quantitative considerations into account, such as to what extent the mediastinum should be removed. Our future work will include a detailed examination of whether the method of segmentation works on a negative effect of adversarial attack and to what extent it affects the prediction. Finally, we did not quantitatively compare the performance of the three XAI employed in this study. Our future tasks include a quantitative assessment of XAI methods using metrics such as Max-sensitivity metric, file size, and computation time. Additionally, the field of XAI is still in its infancy and XAI methods should be developed and selected with care.

5. Conclusions

In this paper, we constructed two models for COVID-19 pneumonia classification by fine-tuning ResNet-50, *i.e.*, a model retrained with full-sized original images and a model retrained with segmented images. We applied three XAI techniques to demonstrate and explain the uncertainty of classification results obtained from the two models assuming clinical applications. In the construction of the pneumonia classification models, as a training dataset, the use of segmented images which are the images of only lung regions with removed unnecessary structures obtained higher accuracy as compared to that of full-sized original images. On the other hand, in clinical applications of model_segmented, when the full-sized original image was used as the input for prediction, the prediction might be erroneous resulting from being confused by the structures and shapes of the input image because those structures and shapes are unknown to the model_segmented. Moreover, the results showed that the model could predict with a high score, despite emphasizing incorrect features. We also found that the way of segmentation used may bring about adversarial-like attack effects. In the clinical application of the model_original, high-performance classification can be achieved, if the whole-sized original images are used as input. However, when segmented images are inputted, the model_original may focus on the background region without structures and this may cause a disturbance in the prediction, in turn, resulting in less confident predictions.

In conclusion, we believe that the best approach to obtaining a highly reliable prediction in the classification of COVID-19 pneumonia is to construct a model using full-sized original images as training data and use full-sized original images as the input when utilized in clinical applications.

Acknowledgements

This work was supported in part by JSPS KAKENHI (Grant-in-Aid for Scientific Research) Grant Number 18K15641. The authors would like to thank Professor Du-Yih Tsai for the extremely useful discussions and for reviewing the paper prior to submission.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Wang, L., Lin, Z.Q. and Wong, A. (2020) COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images. *Scientific Reports*, **10**, Article No.19549. https://doi.org/10.1038/s41598-020-76550-z

[2] Guefrechi, S., Jabra, M.B., Ammar, A., Koubaa, A. and Hamam, H. (2021) Deep Learning Based-Detection of COVID-19 from Chest X-Ray Images. *Multimedia Tools and Applications*, **80**, 31803-31820. https://doi.org/10.1007/s11042-021-11192-5

[3] Wu, C., Khishe, M., Mohammadi, M., Karim, S.H.T. and Rashid, T.A. (2021) Evolving Deep Convolutional Neutral Network by Hybrid Sine-Cosine and Extreme Learning Machine for Real-Time COVID19 Diagnosis from X-Ray Images. *Soft Computing*. https://doi.org/10.1007/s00500-021-05839-6

[4] Karthik, R., Menaka, R. and Hariharan, M. (2020) Learning Distinctive Filters for COVID-19 Detection from Chest X-Ray Using Shuffled Residual CNN. *Applied Soft Computing*, **99**, Article ID: 106744. https://doi.org/10.1016/j.asoc.2020.106744

[5] Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 1135-1144. https://doi.org/10.1145/2939672.2939778

[6] Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J.J. and Oermann, E.K. (2018) Variable Generalization Performance of a Deep Learning Model to Detect Pneumonia in Chest Radiographs: A Cross-Sectional Study. *PLOS Medicine*, **6**, e1002683. https://doi.org/10.1371/journal.pmed.1002683

[7] Narodytska, N. and Kasiviswanathan, S.P. (2016) Simple Black-Box Adversarial Perturbations for Deep Networks. 2017 *IEEE Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*), Honolulu, 21-26 July 2017, 1310-1318. https://doi.org/10.1109/CVPRW.2017.172

[8] Goodfellow, I.J., Shlens, J. and Szegedy, C. (2015) Explaining and Harnessing Adversarial Examples. arXiv:1412.6572v3.

[9] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. and Fergus, R. (2014) Intriguing Properties of Neural Networks. arXiv:1312.6199.

[10] Kholiavchenko, M., Sirazitdinov, I., Kubrak, K., Badrutdinova, R., Kuleev, R., *et al.* (2020) Contour-Aware Multi-Label Chest X-ray Organ Segmentation. *International Journal of Computer Assisted Radiology and Surgery*, **15**, 425-436.

[11] Yu, P., Xu, H., Zhu, Y., Yang, C., Sun, X. and Zhao, J. (2011) An Automatic Computer-Aided Detection Scheme for Pneumoconiosis on Digital Chest Radiographs. *Journal of Digital Imaging*, **24**, 382-393.

[12] Salehi, S., Abedi, A., Balakrishnan, S. and Gholamrezanezhad, A. (2020) Coronavirus Disease 2019 (COVID-19): A Systematic Review of Imaging Findings in 919 Patients. *American Journal of Roentgenology*, **215**, 87-93.

[13] Narayanan, B.N., Silva, M.S.D., Hardie, R.C., Nathan K. Kueterman, N.K. and Ali, R. (2019) Understanding Deep Neural Network Predictions for Medical Imaging Applications. arXiv:1912.09621v1.

[14] Narayanan, B.N., Davuluru, V.S.P. and Hardie, R.C. (2020) Two-Stage Deep Learning Architecture for Pneumonia Detection and Its Diagnosis in Chest Radiographs. *Proceedings of SPIE Medical Imaging* 2020, Houston, 2 March 2020, 113180G. https://doi.org/10.1117/12.2547635

[15] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/CVPR.2016.90

[16] ImageNet. http://www.image-net.org

[17] Samek, W., Wiegand, T. and Müller, K.R. (2017) Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. arXiv:1708.08296. https://arxiv.org/abs/1708.08296

[18] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Pedreschi, D. and Giannotti, F. (2018) A Survey of Methods For Explaining Black Box Models. *ACM Computing Surveys*, **51**, Article No. 93. https://doi.org/10.1145/3236009

[19] Rajaraman, S., Silamut, K., Hossain, A., Ersoy, I., Maude, R.J., Jaeger, S., *et al.* (2018) Understanding the Learned Behavior of Customized Convolutional Neural Networks toward Malaria Parasite Detection in Thin Blood Smear Images. *Journal of Medical Imaging*, **5**, Article ID: 034501. https://doi.org/10.1117/1.JMI.5.3.034501

[20] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D. (2017) Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. 2017 *IEEE International Conference on Computer Vision*, Venice, 22-29 October 2017, 618-626. https://doi.org/10.1109/ICCV.2017.74

[21] Sait, U., Gokul, L., Sunny, P., Rahul, B, Tarun, K., Sanjana S. and Kriti, B. (2021) Curated Dataset for COVID-19 Posterior-Anterior Chest Radiography Images (X-Rays). Mendeley Data.

[22] http://imgcom.jsrt.or.jp/download/

[23] Hiura, M., Kido, S. and Shouno, H. (2005) Development of Pulmonary Nodule Detection Method on Chest Radiographs. *Medical Imaging Technology*, **23**, 250-258.

[24] Brownlee, J. (2021) Gentle Introduction to the Adam Optimization Algorithm for Deep Learning. Machine Learning Mastery. https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning/

[25] Musoro, J.Z., Zwinderman, A.H., Puhan, M.A., Riet, G.T. and Geskus, R.B. (2014) Validation of Prediction Models Based on Lasso Regression with Multiply Imputed Data. *BMC Medical Research Methodology*, **14**, Article No. 116. https://doi.org/10.1186/1471-2288-14-116

[26] Oh, Y., Park, S. and Ye, J.C. (2020) Deep Learning COVID-19 Features on CXR Using Limited Training Data Sets. *IEEE Transactions on Medical Imaging*, **39**, 2688-2700. https://doi.org/10.1109/TMI.2020.2993291