Deep convolution neural networks to differentiate between COVID-19 and other pulmonary abnormalities on chest radiographs: Evaluation using internal and external datasets

Yongwon Cho1 | Sung Ho Hwang1 | Yu-Whan Oh1 | Byung-Joo Ham2 | Min Ju Kim1 | Beom Jin Park1

1Department of Radiology, Korea University Anam Hospital, Seoul, Republic of Korea
2Department of Psychiatry, Korea University Anam Hospital, Seoul, Republic of Korea

Correspondence
Beom Jin Park, Department of Radiology, Korea University Anam Hospital, Korea University College of Medicine, 73, Goryeodae-ro, Seongbuk-gu, Seoul 02841, Republic of Korea.
Email: radiolbj226@gmail.com

Funding information
Korea University Anam Hospital, Grant/Award Numbers: K1809771, O2000221; National Research Foundation of Korea, The Ministry of Education, Grant/Award Numbers: 2018R1D1A1A02085358, 2020R1H1A1A01071600

Abstract
We aimed to evaluate the performance of convolutional neural networks (CNNs) in the classification of coronavirus disease 2019 (COVID-19) disease using normal, pneumonia, and COVID-19 chest radiographs (CXRs). First, we collected 9194 CXRs from open datasets and 58 from the Korea University Anam Hospital (KUAH). The number of normal, pneumonia, and COVID-19 CXRs were 4580, 3884, and 730, respectively. The CXRs obtained from the open dataset were randomly assigned to the training, tuning, and test sets in a 70:10:20 ratio. For external validation, the KUAH (20 normal, 20 pneumonia, and 18 COVID-19) dataset, verified by radiologists using computed tomography, was used. Subsequently, transfer learning was conducted using DenseNet169, InceptionResNetV2, and Xception to identify COVID-19 using open datasets (internal) and the KUAH dataset (external) with histogram matching. Gradient-weighted class activation mapping was used for the visualization of abnormal patterns in CXRs. The average AUC and accuracy of the multiscale and mixed-COVID-19Net using three CNNs over five folds were (0.99 ± 0.01 and 92.94% ± 0.45%), (0.99 ± 0.01 and 93.12% ± 0.23%), and (0.99 ± 0.01 and 93.57% ± 0.29%), respectively, using the open datasets (internal). Furthermore, these values were (0.75 and 74.14%), (0.72 and 68.97%), and (0.77 and 68.97%), respectively, for the best model among the fivefold cross-validation with the KUAH dataset (external) using domain adaptation. The various state-of-the-art models trained on open datasets show satisfactory performance for clinical interpretation. Furthermore, the domain adaptation for external datasets was found to be important for detecting COVID-19 as well as other diseases.

Keywords
chest radiography, computer-aided diagnosis (CAD), COVID-19, deep learning, lung diseases

Abbreviations: AUC, area under the curve; CAD, computer-aided diagnosis; CNN, convolutional neural network; COVID-19, coronavirus disease 2019; CT, computed tomography; CXR, chest radiograph; DL, deep learning; KUAH, Korea University Anam Hospital; ROC, receiver operating characteristic curve.

Cho Yongwon and Sung Ho Hwang contributed equally to this work as co-first authors.
INTRODUCTION

The coronavirus disease 2019 (COVID-19) pandemic, characterized by the severe acute respiratory syndrome and caused by coronavirus 2, has continued almost unabated since the virus was first detected in December 2019.1 There have been around 27 000 000 confirmed cases and 875 000 confirmed deaths worldwide at the time this report was written (September 8, 2020).2 Early detection of COVID-19 is crucial to prevent infection of healthy people owing to the highly contagious nature of the virus. Currently, reverse transcriptase-polymerase chain reaction (RT-PCR), which can detect SARS-CoV-2 RNA in suspected patients, is the primary detection method for COVID-19.3 However, this method is time-consuming (from 3 h to more than 48 h) and involves complicated manual processes.

Chest radiography imaging, including computed tomography (CT) or X-ray imaging is typically used for pneumonia diagnosis. CT screening for the initial diagnosis of COVID-19 has been found to be superior to RT-PCR testing3 and can even confirm COVID-19 infection after a negative or weakly positive diagnosis performed using RT-PCR testing.4 Recently, CT imaging has been employed in various studies for the diagnosis of COVID-19.5,6 However, with the spread of COVID-19, the routine application of CT places a substantial strain on the radiology department. Therefore, the role of chest radiography has become increasingly crucial with regard to the first diagnostic imaging parameters for screening patients with nonspecific thoracic symptoms for pneumonia or COVID-19 in general clinical practice. Generally, chest radiographs (CXRs) reflect thoracic symptoms, including bilateral, peripheral consolidation, and/or ground-glass opacities on CT.5,6 Wong et al.7 investigated the sensitivity of COVID-19 diagnosis using CXRs. Unfortunately, CXR results have been reported to be less sensitive compared with initial RT-PCR tests (69 vs. 91%),7 although 9% of negatively diagnosed cases using RT-PCR were later diagnosed positively using CXRs. In addition, automatic methods to detect subtle abnormalities such as COVID-19 and pneumonia can supplement clinical tools for early diagnosis under conditions such as a large number of potentially infected people and a limited number of trained radiologists. Although CXRs cannot be used to entirely substitute RT-PCR tests, pneumonia is a clinical symptom in high-risk patients who require hospitalization; therefore, CXRs can be used for patient triage, to determine the priority of patient treatment in order to assist overloaded healthcare systems in the midst of a global pandemic. This is especially important because the most frequently known cause of community-acquired pneumonia is bacterial infection.8

Excluding these populations through triage can significantly reduce the amount of medical resources needed.

Accordingly, artificial intelligence, including deep learning (DL), are potential solutions for classification of COVID-19. It is highly challenging to collect a large volume of high-quality curated CXRs to train DL networks in clinical environments. Recently, however, research using small datasets of COVID-19 CXRs has been actively explored.9–17 In particular, Wang and Wong10 presented an open-source network (called the COVID-Net) to detect COVID-19 using CXRs. This network showed good performance, with 80% sensitivity. Minaeea et al.17 analyzed the performance of various open-source (shared online) DL algorithms for COVID-19 detection using CXRs. These algorithms achieved a high sensitivity of 98% (±3%) and a specificity of approximately 90%.

In this study, we aimed to evaluate the COVID-19 diagnosis performance of a customized state-of-the-art (SOTA) DL algorithms. These algorithms were developed using open datasets for training and internal validation. DL architectures, which can used for radiologically interpreting inference results, were trained on a dataset containing a limited number of normal, pneumonia, and COVID-19 images. We provide a statistical analysis of the performance of three models, namely, DenseNet169, InceptionResNetV2, and Xception. Importantly, a Korea University Anam Hospital (KUAH) dataset was used to evaluate the COVID-19 diagnosis performance for external validation using three models trained on multiple open datasets. We also demonstrate that COVID-19 classification performance can be improved by using domain adaptation, such as performing histogram matching on external datasets to reduce the heterogeneities between open (internal) and KUAH (external) datasets.

MATERIALS AND METHODS

Our institutional review board approved our retrospective cohort study, and the requirement for informed consent was waived.

Datasets

CXR datasets are of two types: open or local. We employed a CXR dataset that has been used in the development of existing SOTA algorithms, COVID-CXNet9 and COVID-Net.10 The dataset containing COVID-19, normal, and pneumonia images were collected from multiple public sources. The most important feature of this dataset was that it contained a total of 805 COVID-19 images. Because COVID-19 is a new disease, its image
database\textsuperscript{10} is regularly updated on various public sites such as Radiopedia, SIRM, EuroRad, and Hannover Medical School dataset. In addition, a non-COVID-19 dataset, containing a large number of normal and pneumonia images, was collected from the RSNA Pneumonia Detection Challenge 2018. Further, this dataset contains 5000 normal and 4272 pneumonia images. This dataset\textsuperscript{20} includes lung opacity and various conditions such as bleeding, volume loss, pulmonary edema, and lung cancer, which also cause opacity in CXRs. Although this was a segmentation problem to find main lesions, we used classification to classify COVID-19 and other diseases in the present study. A total of 10 077 normal, pneumonia, and COVID-19 images from the open datasets were randomly assigned into training, tuning, and test sets in a 70:10:20 ratio for the final computer-aided diagnosis (CAD) assessment for detecting COVID-19 using open CXRs (Table 1). For external validation, we collected a small dataset from the KUAH. These datasets were selected depending on the availability of the corresponding chest CT images from January 2020 to October 2020, which were confirmed by expert thoracic radiologists. The number of normal, pneumonia, and COVID-19 CXRs were 20, 20, and 18, respectively (Table 1).

The characteristics of the COVID-19 images in the open and KUAH datasets are as follows: In both the cases, COVID-19 infection was confirmed using RT-PCR testing. The COVID-19 patients whose images were included in the open datasets were likely in the early phase of disease progression, whereas those whose images were included in the KUAH dataset were in a critically severe phase of disease progression. Such cases are difficult to distinguish from those of conventional pneumonia, even for expert radiologists, without the aid of CT. In addition, patients with pneumonia and other diseases in the KUAH dataset were at a severe level of disease progression. As the open datasets used for training and internal validation are different from that employed for external validation (KUAH), the interpretation of the results becomes crucial and can be used to determine the way the results should be used in a real medical environment such as at the KUAH. Figure 1(A) shows 15 sample images from multiple open datasets, including five normal and pneumonia images from the RSNA pneumonia challenge (first and second rows), and five COVID-19 images from various public sites (third row). Figure 1(B) shows 15 sample images from the KUAH dataset, including normal (first row), pneumonia (second row), and COVID-19 (third row) images.

### 2.2 Methods

We classified the three classes, including COVID-19 CXRs and others, using a pretrained DenseNet that employed transfer learning,\textsuperscript{21} InceptionResNetV2,\textsuperscript{22} and Xception.\textsuperscript{23}

#### 2.2.1 DenseNet

DenseNet is configured with a dense block that comprises four BN-ReLU-Conv modules, as shown in Figure 2. The colored squares represent feature maps generated at different steps. The convolution layers are most commonly used for image convolution.

#### 2.2.2 InceptionResNetV2

InceptionResNetV2 is formulated based on a combination of the inception structure and residual connections. In the Inception-ResNet block, convolutional filters of various sizes are combined with residual connections. The use of residual connections not only helps us avoid the degradation problem caused by deep structures but

| Classes | The severity of COVID-19 | Number of images in the training set (with tuning set) | Number of images in the test set |
|---------|--------------------------|-------------------------------------------------------|---------------------------------|
| Internal dataset (multiple open datasets) | Normal | 3780 (420) | 800 |
| | Pneumonia | 3484 (388) | 400 |
| | COVID-19 Early phase | 605 (50) | 125 |
| External dataset KUAH | Normal | — | 20 |
| | Pneumonia | — | 20 |
| | COVID-19 Severe phase | — | 18 |

Note: KUAH for external validation.

Abbreviations: CXR, chest radiograph; KUAH, Korea University Anam Hospital.
FIGURE 1  Examples from (A) multiple public datasets for training and internal validation, and (B) Korea University Anam Hospital (KUAH) dataset for external validation. Normal, pneumonia, and COVID-19 images are in the first, second, and third rows, respectively, of (A) and (B).
also reduces the training time. Figure 3 shows the architecture of InceptionResNetV2.

### 2.2.3 | Xception

We proposed a convolutional neural network (CNN) architecture based entirely on depthwise separable convolution layers. In effect, the mapping of cross-channel correlations and spatial correlations in the feature maps of CNNs can be entirely decoupled. Because this hypothesis is a stronger version of the one underlying the inception architecture, we refer to this as the “Xception” architecture, which stands for “extreme inception.” A complete description of the network specifications is presented in Figure 4. The Xception architecture has 36 convolutional layers, forming the feature extraction base of the network.

The network was trained on CXRs based on weak labels and fine-tuned using DenseNet169, InceptionResNetV2, and Xception, all pretrained on the ImageNet dataset to classify normal, pneumonia, and COVID-19 using images from multiple open datasets. The three CNN models were trained on more than 1 million natural images, including 1000 objects in ImageNet. We used transfer learning after the network was trained on a substantially large amount of labeled data (i.e., pretrained on the National Institutes of Health [NIH] dataset). These deep CNN techniques enable the learning of generic image features by employing other domain datasets without the need for training the network from scratch. The pretrained networks serve as feature extractors for generic image features, and the last two layers are fully connected for classification. We trained the models that were pretrained on the NIH dataset with CXRs (Table 1; internal dataset) and only fine-tuned the last layer of each deep CNN model. For evaluation, we used the internal and external datasets corresponding to each model. The image size was 512 x 512 pixels.

A careful redesign of the workflows with regard to preprocessing, deep CNNs, and computing hardware settings was undertaken. We used geometric augmentation, including zoom, rotation, and shift, on the edges of the images. These augmentations helped alleviate scanner-specific biases and improved the robustness of the CNNs against additional sources of variability, unrelated to the radiological classes in Python 3.6. In addition, we devised multiscale and mixed (MM)-COVID-19Net to train the DL models, as shown in Figure 5. This method was devised based on previous research. It differs from Reference 27 in that the patch images were randomly extracted from the whole images and were rescaled to multiple sizes and various zoom factors. This was intended to reflect the manner in which a radiologist might determine various disease patterns on specific CXRs. Multiscale patch images and whole images were regularly used to train each deep CNN on the CXR data. The input images were resized to 512 x 512 pixels and converted into NumPy arrays. These datasets were loaded...
FIGURE 4  Architecture for Xception: Xception is highly efficient in terms two main points: depthwise shortcuts and separable convolution between convolution blocks, as in ResNet

FIGURE 5  Overall architecture of the multiscale and mixed (MM)-COVID-19Net for the classification of COVID-19 [Color figure can be viewed at wileyonlinelibrary.com]
on a GPU server with Ubuntu 18.04, CUDA 10.2, four 24 GB Titan RTX and Quadro graphics cards, and cuDNN 9.1 (NVIDIA Corporation) with the Keras (TensorFlow) framework. We used the Adam or RAdam optimizer with an initial learning rate of 0.001 for the classification of COVID-19 images. The cross-entropy cost function in binary classification (1.1) is expressed as follows:

\[ L(y, f) = -y \log f - (1 - y) \log (1 - f) \]  

where \( f \) and \( y \) denote the inferred probability and corresponding desired output, respectively.

Tuning errors in the selection of optimized models were minimized by running the backpropagation algorithm over 25 training epochs with a batch size of eight.

After training, we conducted an ablation study on the classification of COVID-19, normal, and pneumonia images using a test dataset comprising images from the open datasets (internal dataset) and the KUAH dataset (external dataset) to confirm the correctness of the results of the feature pyramid and dense connection information. The external validation dataset (KUAH) was preprocessed using histogram matching to enhance the detection rate of COVID-19 by adjusting the external dataset with regard to differences such as texture and intensity from multiple open datasets (internal). This method converts images to a histogram that can be specified arbitrarily. A template image from the internal datasets must be selected for histogram matching, as shown in Figure 6.

Furthermore, this method examines the headers used in all digital imaging and communications in medicine (DICOM) images to determine the intensity range \((L, W)\) because CXRs are not standardized with respect to intensity. The DICOM header includes information regarding the window width (0028, 1051) and window center (0028, 1050). The method uses this information to calculate the highest pixel value (1.2) and the lowest pixel value (1.3) in the image.

\[ P_l = \frac{P_c - P_w}{2} \]  

\[ P_h = \frac{P_c + P_w}{2} \]  

where \( P_l \) is the lowest pixel value, \( P_h \) is the highest pixel value, and \( P_c \) and \( P_w \) are the center and width intensities of the input pixel, respectively.
Because the histograms of the images are distributed in various forms, we investigated the relationship between the histogram and the intensity value of all the CXRs to select a template image for histogram matching. First, we calculated the mean intensity of the selected corresponding images using (1.4).

Template image

\[
\text{Template image} = \text{select} \left( \sum_{i=0}^{c} \text{mean intensity}_i \right) \text{in the number of all images}
\]

(1.4)

The standard template image is a uniformly distributed histogram, as shown in Figure 6.

The performance of CAD for COVID-19 classification was evaluated for each model with and without the new preprocessing method using histogram matching.

In addition, we proposed gradient-weighted class activation mapping (Grad-CAM) to visualize the feature map associated with each class for any convolutional layer of the networks using the gradient generated via backpropagation.

2.3 Statistical analysis

We evaluated the diagnostic performance of the CNN models for the classification of COVID-19, normal, and pneumonia images using fivefold cross-validation analysis. We defined terms that form the confusion matrix as follows: true positive (TP) is the number of labels correctly classified as positive by the algorithms, true negative (TN) is the number of labels correctly classified as negative by the algorithms, false positive (FP) is the number of labels incorrectly classified as positive by the algorithms, and false negative (FN) is the number of labels incorrectly classified as negative by the algorithms. Multiple classifications based on CXRs were assessed in terms of recall, precision, F1-score, and accuracy as follows:

\[
\text{Recall} = \frac{TP}{TP + FN}
\]

(1.5)

\[
\text{Precision} = \frac{TP}{TP + FP}
\]

(1.6)

\[
F1 - \text{score} = \frac{2(\text{Precision} \times \text{Recall})}{\text{Precision} \times \text{Recall}}
\]

(1.7)

\[
\text{Accuracy} = \frac{TN + TP}{TN + TP + FN + FP}
\]

(1.8)

These values were calculated using the scikit-learn Python library. Accuracy is the ratio of the number of correctly classified test samples to the total number of test samples.

For the multiclass case, the area under the curve (AUC) (COVID-19-vs-all) was calculated using the pROC (1.17.0.1) R package.

3 RESULTS

3.1 Comparison of MM-COVID-19Net with three CNNs

To predict COVID-19, MM-COVID-19Net combined with three widely used CNNs, namely, DenseNet169, InceptionResNetV2, and Xception, as the backbone network, as shown in Figure 5, was trained and validated with fivefold cross-validation with internal validation on multiple open datasets containing normal and abnormal classes (COVID-19 and pneumonia, respectively). For additional external validation, we used the KUAH dataset (normal, 20 images; COVID-19, 18 images; and pneumonia, 20 images). The models calculated a probability value for each image to be detected as either a COVID-19 image or an image corresponding to the other classes. We can also use a binary label indicating whether or not an image corresponds to COVID-19 to calculate the recall, precision, F1-score, and accuracy.

The best performing algorithm was chosen from the trained models using fivefold cross-validation. Considering the KUAH dataset (external) is different from the multiple open datasets (internal), as shown in Table 1, an independent external validation is more important than internal validation to accurately evaluate applicability in medical environments; therefore, CAD was conducted on each model, with and without histogram matching for domain adaptation, and the results were evaluated.

The results are presented in Table 2 and Figure 7. A model for COVID-19 detection should have high sensitivity, and thus, the best model was selected based on the recall value. The COVID-19 scores for Xception were as follows: 97.50 ± 0.92 for recall, 98.39 ± 1.28 for precision, 97.50 ± 0.92 for F1-score, 0.99 ± 0.01 for AUC, and 93.57 ± 0.29% for accuracy, averaged over all folds; Xception exhibited its best performance in the fifth-fold with scores of 98.4, 98.4, 98.4, 0.9997, and 93.74% for recall, precision, F1-score, AUC, and accuracy, respectively. The COVID-19 scores for DenseNet169 were 96.00 ± 1.13, 98.53 ± 1.46, 97.25 ± 1.11, 0.99 ± 0.01, and 92.94% ± 0.45% for recall, precision, F1-score, AUC, and accuracy, respectively, averaged over all folds, with
### TABLE 2
Results of the fivefold cross validation for classification (normal, pneumonia, and COVID-19 images) on the internal dataset with (MM)-COVID-19Net-backbone network: DenseNet169, Reference [9], and Reference [10]. Avg = average; SD = standard deviation on multiple open dataset (internal validation)

| Fold | Label     | Recall | Precision | F1-score | AUC for multiclass | Accuracy |
|------|-----------|--------|-----------|----------|--------------------|----------|
| 1    | Normal    | 95.63  | 93.98     | 94.80    | 0.9765             | 93.43    |
|      | Pneumonia | 87.75  | 90.93     | 89.31    | 0.9704             |          |
|      | COVID-19  | 97.6   | 97.6      | 97.6     | 0.9998             |          |
| 2    | Normal    | 95.25  | 93.96     | 94.60    | 0.9746             | 93.13    |
|      | Pneumonia | 88.25  | 89.37     | 88.80    | 0.9675             |          |
|      | COVID-19  | 97.54  | 1.0       | 97.54    | 0.9997             |          |
| 3    | Normal    | 95.63  | 94.56     | 95.09    | 0.9784             | 93.81    |
|      | Pneumonia | 89.92  | 90.61     | 89.92    | 0.9723             |          |
|      | COVID-19  | 97.98  | 99.18     | 97.98    | 0.9997             |          |
| 4    | Normal    | 95.63  | 94.91     | 95.27    | 0.9761             | 93.74    |
|      | Pneumonia | 89.94  | 90.40     | 89.95    | 0.9688             |          |
|      | COVID-19  | 95.97  | 96.75     | 95.97    | 0.9995             |          |
| 5*   | Normal    | 95.25  | 94.66     | 94.95    | 0.9764             | 93.74    |
|      | Pneumonia | 89.25  | 90.38     | 89.81    | 0.9707             |          |
|      | COVID-19  | 98.4   | 98.4      | 98.4     | 0.9997             |          |
| Avg ± SD | Normal | 95.48 ± 0.21 | 94.41 ± 0.42 | 94.94 ± 0.26 | 0.97 ± 0.01            | 93.57 ± 0.29 |
|      | Pneumonia | 89.02 ± 0.99 | 90.34 ± 0.50 | 89.56 ± 0.50 | 0.97 ± 0.01            |          |
|      | COVID-19  | 97.50 ± 0.92 | 98.39 ± 1.28 | 97.50 ± 0.92 | 0.99 ± 0.01            |          |

| Fold | Label     | Recall | Precision | F1-score | AUC for multiclass | Accuracy |
|------|-----------|--------|-----------|----------|--------------------|----------|
| 1*   | Normal    | 94.75  | 94.51     | 94.63    | 0.9794             | 93.36    |
|      | Pneumonia | 89.25  | 89.02     | 89.14    | 0.9717             |          |
|      | COVID-19  | 97.6   | 1.0       | 98.78    | 0.9998             |          |
| 2    | Normal    | 95.50  | 93.63     | 94.55    | 0.9766             | 92.75    |
|      | Pneumonia | 86.50  | 89.64     | 88.04    | 0.9701             |          |
|      | COVID-19  | 95.20  | 96.75     | 95.97    | 0.9996             |          |
| 3    | Normal    | 94.50  | 94.38     | 94.40    | 0.9803             | 92.91    |
|      | Pneumonia | 89.00  | 87.90     | 88.44    | 0.9747             |          |
|      | COVID-19  | 95.20  | 1.0       | 97.54    | 0.9992             |          |
| 4    | Normal    | 95.38  | 94.43     | 94.90    | 0.9806             | 93.36    |
|      | Pneumonia | 88.25  | 89.59     | 88.92    | 0.9748             |          |
|      | COVID-19  | 96.80  | 98.37     | 97.58    | 0.9995             |          |
| 5    | Normal    | 94.25  | 93.90     | 94.07    | 0.9792             | 92.30    |
|      | Pneumonia | 87.50  | 87.50     | 87.50    | 0.9708             |          |
|      | COVID-19  | 95.20  | 97.54     | 96.36    | 0.9994             |          |
| Avg ± SD | Normal | 94.88 ± 0.55 | 94.17 ± 0.38 | 94.51 ± 0.31 | 0.97 ± 0.01            | 92.94 ± 0.45 |
|      | Pneumonia | 88.10 ± 1.13 | 88.73 ± 0.98 | 88.41 ± 0.66 | 0.97 ± 0.01            |          |
|      | COVID-19  | 96.00 ± 1.13 | 98.53 ± 1.46 | 97.25 ± 1.11 | 0.99 ± 0.01            |          |

MM-COVID-19Net-backbone network: InceptionResNetV2

| Fold | Label     | Recall | Precision | F1-score | AUC for multiclass | Accuracy |
|------|-----------|--------|-----------|----------|--------------------|----------|
| 1    | Normal    | 95.63  | 93.75     | 94.68    | 0.9766             | 93.28    |
|      | Pneumonia | 87.50  | 90.90     | 89.17    | 0.9713             |          |
|      | COVID-19  | 96.80  | 97.58     | 97.19    | 0.9997             |          |
| 2    | Normal    | 95.00  | 93.83     | 94.41    | 0.9770             | 92.91    |
|      | Pneumonia | 87.75  | 89.54     | 88.63    | 0.9701             |          |
|      | COVID-19  | 96.00  | 97.56     | 96.77    | 0.9996             |          |
| 3    | Normal    | 95.13  | 94.30     | 94.71    | 0.9794             | 93.21    |

(Continues)
| Fold | Label | Recall | Precision | F1-score | AUC for multiclass Accuracy |
|------|-------|--------|-----------|----------|----------------------------|
| 4    | Pneumonia | 88.25  | 89.82     | 89.03    | 0.9751                     |
|      | COVID-19 | 96.80  | 96.80     | 96.80    | 0.9996                     |
| 5*   | Normal   | 95.75  | 93.19     | 94.45    | 0.9792                     |
|      | Pneumonia| 86.25  | 90.31     | 88.24    | 0.9741                     |
|      | COVID-19 | 95.20  | 98.34     | 96.75    | 0.9997                     |
|      | Normal   | 95.88  | 93.65     | 94.75    | 0.9752                     |
|      | Pneumonia| 87.00  | 91.10     | 89.00    | 0.9693                     |
|      | COVID-19 | 97.60  | 98.39     | 97.99    | 0.9997                     |
|      | Normal   | 95.48 ± 0.39 | 93.74 ± 0.40 | 94.6 ± 0.16 | 0.98 ± 0.01 |
|      | Pneumonia| 87.35 ± 0.76 | 90.33 ± 0.67 | 88.81 ± 0.38 | 0.97 ± 0.02 |
|      | COVID-19 | 96.48 ± 0.91 | 97.73 ± 0.66 | 97.1 ± 0.53 | 0.99 ± 0.01 |

Reference [9]

| COVID-CXNet | Fold | Label | Recall | Precision | F1-score | AUC for multiclass Accuracy |
|-------------|------|-------|--------|-----------|----------|----------------------------|
| 1*          | Normal | 95.13  | 90.38  | 92.69    | 0.9642  | 90.19                     |
|             | Pneumonia | 77.75  | 89.63  | 83.27    | 0.9533  |                         |
|             | COVID-19 | 98.4   | 90.44  | 94.25    | 0.9949  |                         |
| 2           | Normal   | 93.88  | 90.37  | 92.09    | 0.9560  | 89.43                     |
|             | Pneumonia| 81.00  | 84.81  | 82.86    | 0.9438  |                         |
|             | COVID-19 | 88.00  | 98.21  | 92.83    | 0.9928  |                         |
| 3           | Normal   | 97.63  | 88.75  | 92.98    | 0.9653  | 89.81                     |
|             | Pneumonia| 72.75  | 92.97  | 81.63    | 0.9546  |                         |
|             | COVID-19 | 94.4   | 89.39  | 91.83    | 0.9952  |                         |
| 4           | Normal   | 90.38  | 91.87  | 91.15    | 0.9547  | 84.68                     |
|             | Pneumonia| 69.00  | 89.61  | 77.97    | 0.9320  |                         |
|             | COVID-19 | 98.4   | 53.48  | 62.30    | 0.9913  |                         |
| 5           | Normal   | 88.88  | 93.92  | 91.32    | 0.9542  | 86.72                     |
|             | Pneumonia| 89.00  | 73.25  | 80.36    | 0.9439  |                         |
|             | COVID-19 | 65.60  | 1.0    | 79.22    | 0.9882  |                         |
| Avg ± SD    | Normal   | 93.18 ± 3.55 | 91.06 ± 1.94 | 92.05 ± 0.81 | 0.96 ± 0.01 |
|             | Pneumonia| 77.79 ± 7.72 | 86.51 ± 8.04 | 81.22 ± 2.14 | 0.95 ± 0.02 |
|             | COVID-19 | 88.96 ± 13.73 | 86.30 ± 18.93 | 84.09 ± 13.58 | 0.99 ± 0.01 |

Reference [10]

| COVID-Net  | Fold | Label | Recall | Precision | F1-score | AUC for multiclass Accuracy |
|------------|------|-------|--------|-----------|----------|----------------------------|
| 1          | Normal | 88.63  | 94.16  | 91.30    | 0.9595  | 89.13                     |
|            | Pneumonia | 88.50  | 78.84  | 83.39    | 0.9478  |                         |
|            | COVID-19 | 94.4   | 95.93  | 95.16    | 0.9940  |                         |
| 2          | Normal   | 98.50  | 88.04  | 92.97    | 0.9668  | 90.64                     |
|            | Pneumonia| 74.0   | 96.10  | 83.61    | 0.9585  |                         |
|            | COVID-19 | 93.6   | 95.90  | 94.73    | 0.9980  |                         |
| 3          | Normal   | 91.25  | 93.95  | 92.58    | 0.9711  | 90.19                     |
|            | Pneumonia| 88.50  | 81.94  | 85.09    | 0.9650  |                         |
|            | COVID-19 | 88.88  | 95.69  | 92.21    | 0.9959  |                         |
| 4*         | Normal   | 96.63  | 90.19  | 93.30    | 0.9687  | 91.01                     |
|            | Pneumonia| 79.0   | 91.33  | 84.72    | 0.9618  |                         |
|            | COVID-19 | 95.90  | 95.90  | 94.74    | 0.9981  |                         |
| 5          | Normal   | 90.38  | 94.88  | 92.57    | 0.9735  | 90.49                     |
|            | Pneumonia| 90.50  | 80.80  | 85.38    | 0.9661  |                         |
|            | COVID-19 | 91.20  | 99.13  | 99.13    | 0.9974  |                         |
| Avg ± SD   | Normal   | 93.08 ± 4.26 | 92.24 ± 2.98 | 92.54 ± 0.76 | 0.97 ± 0.02 |
|            | Pneumonia| 90.38  | 80.80  | 85.38    | 0.9661  |                         |
|            | COVID-19 | 91.20  | 99.13  | 99.13    | 0.9974  |                         |
its best performance in the first fold, where the scores were 97.6, 100, 98.78, 0.9998, and 93.36% for recall, precision, F1-score, AUC, and accuracy, respectively. The COVID-19 scores for InceptionResNetV2 were 96.48 ± 0.91, 97.73 ± 0.66, 97.10 ± 0.53, 0.99 ± 0.01, and 93.12 ± 0.23% for recall, precision, F1-score, AUC, and accuracy, respectively, averaged over all folds, with its best performance in the fifth-fold, wherein the scores were 96.80, 97.58, 97.19, 0.9997, and 93.28% for recall, precision, F1-score, AUC, and accuracy, respectively.

Although the difference was not significant, the Xception architecture exhibited the best detection performance among the three CNNs in terms of all metrics (*)Xception-fifth-fold: *p*-values: 0.91, 0.83, 0.14, and 0.14 among DenseNet169 and InceptionResNetV2 in MM-COVID-19Net and References [9] and [10] and * denotes the best performance for each model, respectively.

Abbreviations: AUC, area under the curve; MM, multiscale and mixed.

**TABLE 2** (Continued)

| Fold  | Label   | Recall       | Precision | F1-score  | AUC for multiclass | Accuracy   |
|-------|---------|--------------|-----------|-----------|---------------------|------------|
|       | Pneumonia | 84.10 ± 7.20 | 85.80 ± 7.50 | 84.43 ± 0.89 | 0.96 ± 0.03          |            |
| COVID-19 | 92.80 ± 2.77 | 96.51 ± 1.47 | 95.19 ± 2.49 | 0.99 ± 0.01          |            |

Note: In the table, * denotes *p*-values: 0.91, 0.83, 0.14, and 0.14 among DenseNet169 and InceptionResNetV2 in MM-COVID-19Net and References [9] and [10] and * denotes the best performance for each model, respectively.

Abbreviations: AUC, area under the curve; MM, multiscale and mixed.

**FIGURE 7** Confusion matrix on multiple open datasets (internal). The first, middle, and last columns present the results of the multiscale and mixed (MM)-COVID19-Net (Xception, DenseNet169, InceptionResNetV2), COVID-CXNet,[9] and COVID-Net,[10] respectively (0: normal; 1: pneumonia; 2: COVID-19) [Color figure can be viewed at wileyonlinelibrary.com]

3.2 Comparison of the MM-COVID-19Net (the best model), COVID-CXNet, and COVID-Net

We trained COVID-CXNet[9] and COVID-Net[10] on CXRs in the same way that we trained the MM-COVID-19Net. These networks were compared with our network
through statistical analysis on multiple open datasets (internal) and the KUAH dataset (external). Among the trained models with fivefold cross-validation of the backbone network, Xception with MM-COVID-19Net, the best performing network, was selected using the fifth-fold model. The first and fourth folds were selected for COVID-CXNet9 and COVID-Net10 respectively, as presented in Table 2. Using the internal dataset for COVID-19, the scores for our algorithm were as follows: 98.40% for recall, 98.40% for precision, 98.40% for F1-score, 0.9997 for AUC, and 93.74% for accuracy. The corresponding values for COVID-CXNet9 were as follows: 98.40% for recall, 90.44% for precision, 94.25% for F1-score, 0.9949 for AUC, and 90.19% for accuracy; and for COVID-Net10 these values were as follows: 95.90% for recall, 95.90% for precision, 94.74% for F1-score, 0.9981 for AUC, and 91.01% for accuracy.

Using the external dataset, among the trained models with fivefold cross-validation of the backbone network, the scores of the MM-COVID-19Net using DenseNet169 for our algorithm were as follows: 55.56% for recall, 66.67% for precision, 60.60% for F1-score, 0.7472 for AUC, and 74.14% for accuracy. The corresponding values for COVID-CXNet9 were as follows: 0% for recall, zero for precision, zero for F1-score, 0.3444 for AUC, and 44.83% for accuracy (p-value = 0.004); whereas the scores for COVID-Net10 were as follows: 33.33% for recall, 85.71% for precision, 48.00% for F1-score, 0.5918 for AUC, and 77.59% for accuracy (p-value = 0.82).

4 | DISCUSSION AND CONCLUSION

We analyzed the feasibility of classifying COVID-19 images using SOTA DL algorithms trained on a dataset containing a limited number of normal, pneumonia, and COVID-19 images obtained from open datasets to radiologically interpret the inference results. For external validation, we used the KUAH dataset, which was confirmed by expert radiologists using CT. Furthermore, we developed a new training architecture based on SOTA to reflect the manner in which a radiologist might determine various disease patterns on specific CXRs.
Previous studies\textsuperscript{10,17} used statistical analyses to assess the AUC, accuracy, sensitivity, and specificity. These analyses were conducted using open datasets; however, we aimed to determine whether CNN architectures trained using multiple open datasets (internal) performed sufficiently well for use in real clinical environments. In addition, previous studies have been conducted on the application of CAD in response to nodules\textsuperscript{29–30} using high-quality CXRs and without an open dataset. We herein investigated the detection and classification of abnormalities, including thoracic disease patterns, and provided insights for the detection of COVID-19.

Our results indicate that all the models have similar COVID-19 detection performance. Among the three CNNs, the accuracy of Xception was 93.57% ± 0.29% with 97.50% recall and 0.99 ± 0.01 AUC for COVID-19 on the internal datasets. A comparison between our algorithm and Reference 9, which achieved an AUC of 0.99 ± 0.01 and accuracy of 88.17 ± 2.38%, and Reference 10, which achieved an AUC of 0.99 ± 0.01 and accuracy of 90.29 ± 0.71%, showed no significant difference between our algorithm and other algorithms. However, it is important to indirectly compare the characteristics of open datasets with those of datasets obtained in a specific medical environment. Although the number of images in the KUAH dataset (external) was limited, as presented in Table 1, and the results presented in Table 3 are inferior to those obtained using multiple open datasets (internal), as presented in Table 2, COVID-19 images could be detected by the CNNs trained using the open datasets. DenseNet169 proved to be the best model, with an accuracy of 68.97% and recall of 38.89%. These results suggest that COVID-19 patients whose images were included in the multiple open datasets (internal) are likely in the early stages of disease progression and those whose images are included in the KUAH dataset (external) are in highly severe stages of disease progression, which are difficult to distinguish from pneumonia without the aid of CT, even for expert radiologists. Therefore, we conducted histogram matching on the KUAH dataset, that is, domain adaptation, to improve the results using the external dataset.


table

| With histogram matching | Label       | Recall | Precision | F1-score | AUC for multiclass | Accuracy |
|-------------------------|-------------|--------|-----------|----------|-------------------|----------|
| MM-COVID-19Net-backbone network: Xception | Normal     | 90.00  | 85.71     | 87.80    | 0.9540            | 68.97    |
|                         | Pneumonia  | 65.00  | 65.00     | 65.00    | 0.7763            |          |
|                         | COVID-19   | 50.00  | 52.94     | 51.40    | 0.7153            |          |
| No                      | Normal     | 95.00  | 70.37     | 80.85    | 0.9197            | 62.07    |
|                         | Pneumonia  | 55.00  | 61.11     | 57.89    | 0.7855            |          |
|                         | COVID-19   | 33.33  | 46.15     | 38.70    | 0.6431            |          |
| MM-COVID-19Net-backbone network: DenseNet169 | Normal     | 95.00  | 79.11     | 86.36    | 0.9303            | 74.14    |
|                         | Pneumonia  | 70.00  | 73.68     | 71.95    | 0.8105            |          |
|                         | COVID-19   | 55.56  | 66.67     | 60.60    | 0.7472            |          |
| No                      | Normal     | 95.00  | 76.00     | 84.44    | 0.9105            | 68.97    |
|                         | Pneumonia  | 70.00  | 66.67     | 68.29    | 0.8569            |          |
|                         | COVID-19   | 38.89  | 58.33     | 46.67    | 0.6819            |          |
| MM-COVID-19Net-backbone network: InceptionResNetV2 | Normal     | 95.00  | 73.08     | 82.61    | 0.9500            | 68.97    |
|                         | Pneumonia  | 60.00  | 70.59     | 64.86    | 0.8319            |          |
|                         | COVID-19   | 50.00  | 60.00     | 54.54    | 0.7653            |          |
| No                      | Normal     | 90.00  | 64.29     | 75.00    | 0.9290            | 60.34    |
|                         | Pneumonia  | 50.00  | 71.42     | 58.82    | 0.8658            |          |
|                         | COVID-19   | 38.89  | 43.75     | 41.72    | 0.6556            |          |

Note: The performance on the external dataset was evaluated with the best model for each architecture (with histogram matching). *p-value: 0.0123 without histogram matching.

Abbreviations: AUC, area under the curve; KUAH, Korea University Anam Hospital; MM, multiscale and mixed.
performance of the algorithm after domain adaptation, such as performing histogram matching (accuracy: 74.14%, recall of COVID-19: 55.56%, and AUC of COVID-19: 0.7472), was better than that without histogram matching on the KUAH dataset (Figure 9 and Table 3). In addition, we compared our algorithm with others on the KUAH dataset (external). The results for Reference 9 (accuracy: 44.83%, recall of COVID-19: 0%, and AUC of COVID-19: 0.3444) and Reference 10 (accuracy: 77.59%, recall of COVID-19: 33.33%, and AUC of COVID-19: 0.5918) were lower than those of our algorithm, as presented in Table 4. Therefore, our algorithm can be useful when training (internal—open) and testing (external—KUAH) with different datasets.

Although the COVID-19 classification performance was insufficient on the external dataset, it is vital for distinguishing other viral pneumonia from COVID-19 pneumonia. Even though highly accurate detection was not achieved on the KUAH dataset (external), distinguishing the risk groups for the purpose of screening is still important. In addition, if internal or external datasets are well refined and various preprocessing techniques are used, the performance can be improved for distinguishing traditional pneumonia from COVID-19-induced pneumonia.

Salehi et al. described COVID-19 patterns in CXRs, including peripheral distribution, ground-glass opacification, and bilateral involvement. The results of various DL algorithms should reflect such radiological findings and patterns. Therefore, Grad-CAM was used to visualize the interpretation of the DL models, making them more transparent. This approach showed that the
**FIGURE 10** Confusion matrix for the Korea University Anam Hospital (KUAH) dataset (external) in Table 1. The first, middle, and last columns show the results of multiscale and mixed (MM)-COVID19-Net (backbone: DenseNet169) COVID-CXNet,9 and COVID-Net,10 respectively. (0: normal; 1: pneumonia; 2: COVID-19) [Color figure can be viewed at wileyonlinelibrary.com]

**FIGURE 11** Gradient-weighted class activation mapping (Grad-CAM) for COVID-19 (ground truth) on chest radiographs. (A) Negative results of each model corresponding to normal or pneumonia images without histogram matching. (B) Positive results of each model indicating COVID-19 with histogram matching [Color figure can be viewed at wileyonlinelibrary.com]
models derive information that indicates not only the presence of the disease but also, indirectly, the disease location. The visualization results presented in the first and second rows depicted in Figure 11 coincide with the opinions of radiologists, although some radiologists might question whether the last column actually represents the location of the lesion. Nevertheless, Grad-CAM shows that the model with histogram matching is more accurate than that without histogram matching, as demonstrated in Figure 11.

Our study has several limitations. First, there is a lack of well-curated CXRs (open datasets with COVID-19 images) that could be obtained from multiple public datasets\(^\text{10}\) (see Table 1). Moreover, the open COVID-19 datasets for training were largely collected websites and online publications; thus, strict standards were not applied when collecting them. The characteristics of COVID-19 are not diverse, this could have an impact when training the models using open datasets and testing them for external validation on real-world clinical datasets, such as the KUAH dataset. In addition, no comparisons were made between the results of the human observers and those of the models in the detection of COVID-19 and other diseases. Furthermore, owing to a lack of GPU memory, all the CXRs were down sampled to \(512 \times 512\) pixels, which could decrease the clinical classification and detection validity. Finally, although our algorithm can help improve the detection of COVID-19, we must train our algorithm using various real-world clinical datasets.

In the future, we intend to train the algorithm using various CXRs including COVID-19 and pneumonia to improve the detection performance, particularly for COVID-19, in clinical settings. After collecting more CXRs that include various disease patterns independently confirmed by expert radiologists in our institutions and additional centers, we intend to develop algorithms that can diagnose COVID-19 and other diseases based on multiple CXRs. Finally, histogram matching will be used to detect lesions using datasets different from those of the training datasets (multiple open datasets). We will therefore develop a superior domain adaptation method to improve the detection of COVID-19 based on CXRs. In addition, we will train our models on real clinical datasets (the KUAH dataset and various multicenter datasets) to develop our algorithms on CXRs and evaluate these models using multiple open datasets or others. Most importantly, the diagnosis results obtained by human radiologists should be compared with those obtained by CNN algorithms through reader tests.\(^\text{31}\)

In conclusion, we evaluated three widely used CNNs for the classification of COVID-19 and other diseases using a limited number of CXRs for internal (open datasets) and external validation (KUAH). These models, trained using open datasets, were validated on an external dataset (KUAH), which was representative of actual clinical environments. Although the diagnostic performances of the models using the internal dataset (open datasets) were desirable (Table 2), the performance of the same models on the external dataset (KUAH) was insufficient for routine clinical use. To overcome this problem, we proposed domain adaption, such as histogram matching, to detect or classify COVID-19 and other diseases more accurately on external datasets. This method produced better results for dataset obtained from special hospitals. Although COVID-19 detection using CXRs (internal or external datasets) has the limitation in that it cannot enable a direct confirmation of COVID-19 in the

| Table 4 | Comparison of the best models for classification (normal, pneumonia, and COVID-19) on the KUAH dataset (external with domain adaptation for histogram matching) in MM-COVID-19Net-backbone network: DenseNet169, COVID-CXNet,\(^\text{9}\) and COVID-Net\(^\text{10}\) |
|---------|---------------|-------------|-----------------|-------------|-----------------|
|         | Label         | Recall      | Precision       | F1-score     | AUC for        | Accuracy       |
|         |               |             |                 |              | multiclass     |                |
| MM-COVID-19Net-backbone network: DenseNet169* | Normal      | 95.00        | 79.11           | 86.36        | 0.9303         | 74.14          |
|         | Pneumonia     | 70.00        | 73.68           | 71.95        | 0.8105         |                |
|         | COVID-19      | **55.56**    | **66.67**       | **60.60**    | **0.7472**     |                |
| Reference [9] COVID-CXNet | Normal      | 95.00        | 54.29           | 69.09        | 0.8750         | 44.83          |
|         | Pneumonia     | 35.00        | 41.17           | 37.88        | 0.5171         |                |
|         | **COVID-19**  | **0**        | **0**           | **0**        | **0.3444**     |                |
| Reference [10] COVID-Net | Normal      | 95.00        | 70.37           | 80.85        | 0.9658         | 77.59          |
|         | Pneumonia     | 1.00         | 83.33           | 90.09        | 0.8948         |                |
|         | **COVID-19**  | **33.33**    | **85.71**       | **48.00**    | **0.5918**     |                |

Note: Performance on the external dataset was evaluated with the best model for MM-COVID-19Net backbone network: DenseNet169 (with histogram matching). *p-value: 0.004 and 0.82 among and References [9] and [10].

Abbreviations: AUC, area under the curve; MM, multiscale and mixed.
same way as a diagnostic kit using an RT-PCR test, it can still assist in identifying lung diseases caused by respiratory infections in clinical environments. Our empirical evaluation of the diagnosis of COVID-19 can be extended to the development of CAD algorithms for COVID-19 based on DL for use in real-world clinical environments.

CONFLICT OF INTEREST

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as potential conflicts of interest.

AUTHOR CONTRIBUTIONS

Yongwon Cho, Beom Jin Park, and Sung Ho Hwang: Wrote the manuscript. Yongwon Cho: Performed the experiments and prepared the Figures. Yu-Whan Oh, Beom Jin Park, Byung-Joo Ham, and Sung Ho Hwang: Prepared the dataset and confirmed the datasets. Min Ju Kim: Confirmed the datasets. All authors reviewed the manuscript. All authors were involved in writing the paper and approved the final submitted and published versions.

DATA AVAILABILITY STATEMENT

Even though the limitations of our dataset (KUAH) for public use are regulated by the Personal Information Protection Act in South Korea, we expect to share our datasets and source code as requests.

ORCID

Yongwon Cho https://orcid.org/0000-0001-8092-5799

REFERENCES

1. Helmy YA, Fawzy M, Elaswad A, Sobieh A, Kenney SP, Shehata AA. The COVID-19 pandemic: a comprehensive review of taxonomy, genetics, epidemiology, diagnosis, treatment, and control. J Clin Med. 2020;9(4):1225. https://doi.org/10.3390/jcm9041225.
2. World Health Organization. WHO Coronavirus Disease (COVID-19) Dashboard. https://covid19.who.int.
3. Wang W, Xu Y, Gao R, et al. Detection of SARS-CoV-2 in different types of clinical specimens. JAMA. 2020;323(18):1843-1844. https://doi.org/10.1001/jama.2020.3786.
4. Xie X, Zhong Z, Zhao W, Zheng C, Wang F, Liu J. Chest CT for typical 2019-ncov pneumonia: relationship to negative RT-PCR testing. Radiology. 2020;296(2):200343. https://doi.org/10.1148/radiol.2020200343.
5. Fang Y, Zhang H, Xie J, et al. Sensitivity of chest CT for COVID-19: comparison to RT-PCR. Radiology. 2020;296(2):200432. https://doi.org/10.1148/radiol.2020200432.
6. Ai T, Yang Z, Hou H, et al. Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases. Radiology. 2020;296(2):200642. https://doi.org/10.1148/radiol.2020200642.
7. Wong HYF, Lam HYS, Fong AH-T, et al. Frequency and distribution of chest radiographic findings in COVID-19 positive patients. Radiology. 2020;296(2):201160.
8. Brown PD, Lerner SA. Community-acquired pneumonia. The Lancet. 1998;352(9136):1295-1302. https://doi.org/10.1016/S0140-6736(98)02239-9.
9. A. Haghanifar, M. M. Majdabadi, Y. Choi, S. Deivalakshmi, S. Ko, COVID-CXNet: Detecting COVID-19 in Frontal Chest X-Ray Images using Deep Learning. arXiv preprint, arXiv:2006.13807.
10. Wang L, Wong A. COVID-net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest radiography images. arXiv preprint arXiv:2003.09871. 2020.
11. Narin A, Kaya C, Pamuk Z. Automatic detection of coronavirus disease (COVID-19) using X-ray images and deep convolutional neural networks. arXiv preprint arXiv:2003.10849. 2020.
12. Apostolopoulos ID, Mpesiana TA. COVID-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks. Phys Eng Sci Med. 2020;43(2):635-640. https://doi.org/10.1007/s11246-020-00865-4.
13. Hemdan EE-D, Shouman MA, Karar ME. COVIDX-Net: a framework of deep learning classifiers to diagnose COVID-19 in X-ray images. arXiv preprint arXiv:2003.11055. 2020.
14. Apostolopoulos I, Aznaouridis S, Tzani M. Extracting possibly representative COVID-19 biomarkers from X-ray images with deep learning approach and image data related to pulmonary diseases. arXiv preprint arXiv:2004.00338. 2020.
15. Farooq M, Hafeez A. COVID-ResNet: a deep learning framework for screening of COVID19 from radiographs. arXiv preprint arXiv:2003.14395. 2020.
16. Afshar P, Heidarian S, Naderkhani F, Oikonomou A, Plataniotis KN, Mohammadi A. COVID-CAPS: a capsule network-based framework for identification of COVID-19 cases from X-ray Images. arXiv preprint arXiv:2004.02696. 2020.
17. Minaeaa S, Kafiehb R, Sonkac M, Yazdanid S, Soufie GJ. DeepCOVID: predicting COVID-19 from chest X-ray images using deep transfer learning. Med Image Anal. 2020;65:101794.
18. Salehi S, Abedi A, Balakrishnan S, Gholamrezaezhad A. Coronavirus disease 2019 (COVID-19): a systematic review of imaging findings in 919 patients. Am J Roentgenol. 2020;215(1):1-7. https://doi.org/10.2214/AJR.20.23034.
19. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization. in Proceedings of the IEEE international conference on computer vision 2017, pp. 618–626.
20. Radiological Society of North America. RSNA Pneumonia Detection Challenge. https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data
21. Huang G, Liu Z, Maaten LVD, Weinberger KQ. Densely Connected Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

22. Szegedy C, Ioffe S, Vanhoucke V. Inception-v4, inception-ResNet and the impact of residual connections on learning. arXiv:1602.07261v2. 2016.

23. Chollet F. Xception: deep learning with depthwise separable convolutions. arXiv:1610.02357v3. 2017.

24. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. Imagenet: A Large-Scale Hierarchical Image Database. Paper presented at Computer Vision and Pattern Recognition. CVPR 2009. IEEE Conference on 2009. 2009.

25. Kumar A, Kim J, Lyndon D, Fulham M, Feng D. An ensemble of fine-tuned convolutional neural networks for medical image classification. IEEE J Biomed Health Inform. 2017;31-40. https://doi.org/10.1109/JBHI.2016.2635663.

26. National Institute of Health. NIH clinical center provides one of the largest publicly available chest X-ray datasets to scientific community. https://www.nih.gov/news-events/news-releases/nih-clinical-center-provides-one-largest-publicly-available, 2017.

27. Park B, Cho Y, Lee G, et al. A curriculum learning strategy to enhance the accuracy of classification of various lesions in chest-PA X-ray screening for pulmonary abnormalities. Sci Rep. 2019;9:15352. https://doi.org/10.1038/s41598-019-51832-3.

28. Cho Y, Lee SM, Cho Y-H, et al. Deep chest X-ray: detection and classification of lesions based on deep convolutional neural networks. Int J Imaging Syst Technol. 2020;31:1-10. https://doi.org/10.1002/ima.22508.

29. Cho Y, Kim Y, Lee SM, et al. Reproducibility of abnormality detection on chest radiographs using convolutional neural network in paired radiographs obtained within a short-term interval. Sci Rep. 2020;10:17417. https://doi.org/10.1038/s41598-020-74626-429.

30. Kim Y-G, Lee SM, Lee KH, Jang R, Seo JB, Kim N. Optimal matrix size of chest radiographs for computer-aided detection on lung nodule or mass with deep learning. Eur Radiol. 2020;30:4943-4951. https://doi.org/10.1007/s00330-020-06892-9.

31. Murphy K, Smits H, Knoops AJG, et al. COVID-19 on chest radiographs: a multi reader evaluation of an artificial intelligence system. Radiology. 2020;296(2):E166-E172. https://doi.org/10.1148/radiol.2020201874.

How to cite this article: Cho Y, Hwang SH, Oh Y-W, Ham B-J, Kim MJ, Park BJ. Deep convolutional neural networks to differentiate between COVID-19 and other pulmonary abnormalities on chest radiographs: Evaluation using internal and external datasets. Int J Imaging Syst Technol. 2021;1–18. https://doi.org/10.1002/ima.22595