A Deep-Learning Diagnostic Support System for the Detection of COVID-19 Using Chest Radiographs

A Multireader Validation Study

Matthias Fontanellaz, MSc, Lukas Ebner, MD,† Adrian Huber, MD, PhD,† Alan Peters, MD,† Laura Löbelenz, MD,† Cynthia Hourscht, MD,† Jeremias Klaus, MD,† Jaro Munz, MD,† Thomas Ruder, MD,† Dionysios Drakopoulos, MD,‡ Dominik Sieron, MD,‡ Elias Primetis, MD,‡ Johannes T. Heverhagen, MD, PhD,† Stavroula Mougiakakou, MD,*† and Andreas Christe, MD†‡*

Objectives: The aim of this study was to compare a diagnosis support system to detect COVID-19 pneumonia on chest radiographs (CXRs) against radiologists of various levels of expertise in chest imaging.

Materials and Methods: Five publicly available databases comprising normal CXR, confirmed COVID-19 pneumonia cases, and other pneumonias were used. After the harmonization of the data, the training set included 7966 normal cases, 5451 with other pneumonia, and 258 CXRs with COVID-19 pneumonia, whereas in the testing data set, each category was represented by 100 cases. Eleven blinded radiologists with various levels of expertise independently read the testing data set. The data were analyzed separately with the newly proposed artificial intelligence–based system and by consultant radiologists and residents, with respect to positive predictive value (PPV), sensitivity, and F-score (harmonic mean for PPV and sensitivity). The χ^2 test was used to compare the sensitivity, specificity, accuracy, PPV, and F-scores of the readers and the system.

Results: The proposed system achieved higher overall diagnostic accuracy (94.3%) than the radiologists (61.4% ± 5.3%). The radiologists reached average sensitivities for normal CXR, other type of pneumonia, and COVID-19 pneumonia of 85.0% ± 12.8%, 60.1% ± 12.2%, and 53.2% ± 11.2%, respectively, which were significantly lower than the results achieved by the algorithm (98.0%, 88.0%, and 97.0%; $P < 0.00032$). The mean PPVs for all 11 radiologists for the 3 categories were 82.4%, 59.0%, and 59.0% for the healthy, other pneumonia, and COVID-19 pneumonia, respectively, resulting in an F-score of 65.5% ± 12.4%, which was significantly lower than the F-score of the algorithm (94.3% ± 2.0%, $P < 0.00001$). When other pneumonia and COVID-19 pneumonia cases were pooled, the proposed system reached an accuracy of 95.7% for any pathology and the radiologists, 88.8%. The overall accuracy of consultants did not vary significantly compared with residents (65.0% ± 5.8% vs 67.4% ± 4.2%); however, consultants detected significantly more COVID-19 pneumonia cases ($P = 0.008$) and less healthy cases ($P < 0.00001$).

Conclusions: The system showed robust accuracy for COVID-19 pneumonia detection on CXR and surpassed radiologists at various training levels.

Key Words: COVID-19, chest radiographs, deep learning, diagnostic support system

(*Invest Radiol* 2020;00: 00–00)

In the current COVID-19 pandemic, radiological imaging studies have rapidly gained significant importance.[1–3] Several reports have demonstrated that computed tomography (CT) might be a highly sensitive diagnostic test to detect COVID-19 pneumonia[4] and assess its severity,[5] and even more sensitive for COVID-19 pneumonia diagnosis than initial polymerase chain reaction testing.[6,7] Therefore, centers in heavily afflicted areas initially used chest CT to screen for COVID-19–positive individuals.[8]

Radiographs are reported to be less sensitive to subtle early imaging findings.[9,10] Two recent studies[11,12] have shown that 2 artificial intelligence (AI)–based systems perform better than human readers. Artificial intelligence has already actively contributed to the fight against the COVID-19 pandemic,[13] particularly in assisting the diagnosis process based on the medical image.[14] Specifically, recent studies have shown that deep learning can accurately detect COVID-19 pneumonia on CT images,[15,16] as well as on chest radiographs (CXRs),[17–19] and differentiate it from other community-acquired pneumonia and lung diseases. Recent studies have even aimed at assessing the severity of COVID-19 pneumonia based on CXRs,[20] as well as the risk of developing critical illness based on other clinical parameters.[21]

We hypothesize that CXR possesses great potential to facilitate management of patients with COVID-19 pneumonia. Recent published recommendations of the Fleischner Society[22] and the European Society for Thoracic Imaging[23] have also emphasized this approach. By appropriate use of CXRs in a predefined patient population, conventional chest imaging can provide critical information on the pulmonary status and facilitate patient management. Moreover, CXR units can be easily deployed to the intensive care units for disease monitoring and could even be applied through glass and/or smart glass doors of isolation rooms.[24] As the COVID-19 pandemic continues to spread and also afflicts developing countries that might not have broad access to CT imaging, CXRs could play a critical role in the diagnosis and management of COVID-19 patients.

However, as outlined previously, CXRs are difficult to read and radiologists require particular expertise.[25,26] To optimize the diagnosis of CXR, we have developed a deep learning diagnostic support system for the detection of COVID-19 pneumonia. The purpose of our study is to assess the system's diagnostic performance, analyze the sensitivity and accuracy for COVID-19 pneumonia and other pneumonias, and compare it to radiologists of different levels of expertise. The proposed system is based on light-weight architecture and does not need any dedicated segmentation module to obtain high diagnostic accuracy. This study is completely based on publicly available data. Furthermore, the working code will be open access. To the best of our knowledge, this is the first published study to compare an open-access system with such a large pool of radiologists with respect to the analysis of publicly available data.

MATERIALS AND METHODS

For this retrospective study, we used open-access databases of CXRs of patients with COVID-19 pneumonia, other pneumonias, or

Received for publication August 18, 2020; and accepted for publication, after revision, October 26, 2020.

From the *ARTORG Center for Biomedical Engineering Research, University of Bern; †Department of Diagnostic, Interventional and Pediatric Radiology, Inselspital, and ‡Department of Radiology, Division City and County Hospitals, Inselgroup, Bern University Hospital, University of Bern, Bern, Switzerland.

M.F., L.E., S.M., and A.C. contributed equally to this study.

Conflicts of interest and sources of funding: none declared.

Correspondence to: Lukas Ebner, MD, Department of Diagnostic, Interventional and Pediatric Radiology, Inselspital, Bern University Hospital, University of Bern, Freiburgstrasse 10, Bern 3010, Switzerland. E-mail: lukas.ebner@insel.ch.

ISSN: 0020-9996/20/0000–0000
DOI: 10.1097/RLI.0000000000000748

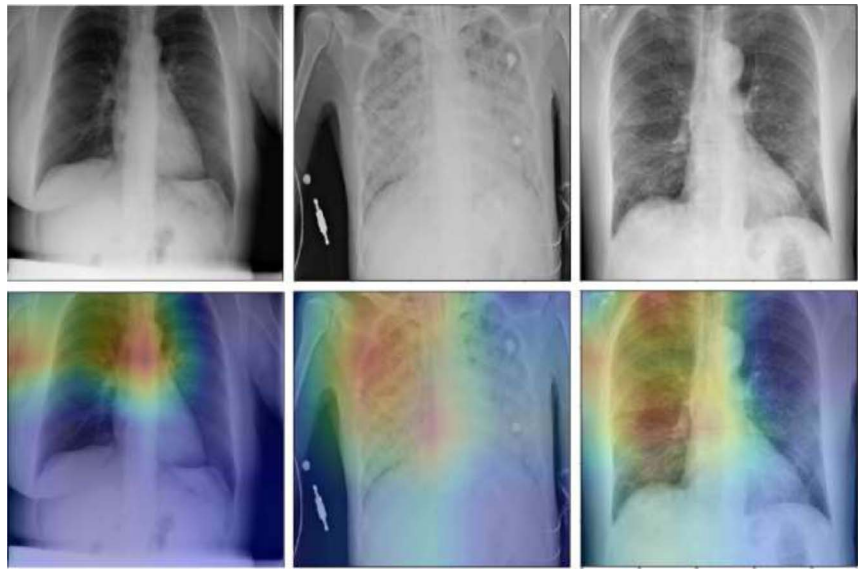

FIGURE 1. Top row, 3 sample chest x-rays (left, sample for healthy class; middle, sample for other types of pneumonia; right, COVID-19 pneumonia); bottom row, overlaid class activation maps.

without lung pathology (Fig. 1). As the study was retrospective and used publicly available, anonymized data, institutional review board approval could be waived. We conformed to the STARD criteria for reporting studies on diagnostic accuracy.[27]

Databases

For the purposes of this study, 5 open-access databases were used: COVID-19 Image Data Collection,[28] Figure 1 COVID-19 Chest X-ray,[29] ActualMed COVID-19 Chest X-ray data set,[30] COVID-19 Radiography Database,[31] and the RSNA Pneumonia Detection Challenge data set.[32] All databases were accessed as of May 27, 2020, and their major characteristics are summarized in Table 1. All COVID-19 pneumonia scans were recorded between December 2019 and May 2020. As shown in Table 1, different data formats are used.

The RSNA data set[32] is mainly used to gather CXR images of healthy individuals and patients with bacterial or viral pneumonia. Furthermore, 19 cases from the COVID-19 Image Data Collection[28] were used to augment the other pneumonia class.[28] However, as shown in Table 1, in some of the used databases, for example, in COVID-19 Image Data Collection[28] and COVID-19 Radiography Database,[31] there is an overlap of COVID-19 pneumonia cases. To avoid duplicate use of a CXR case, we followed the approach proposed by Wang et al.[33] The final merged database contained 13,975 CXR images from 13,870 patients. Of these, 266 patients are confirmed with COVID-19 (358 CXR images), 5538 patients (5551 CXR images) demonstrated bacterial or viral pneumonia with 5551 CXRs, and 8066 were healthy individuals represented by equivalent number of images. All cases in the merged database were annotated either as healthy/normal, pneumonia other than COVID-19, or characteristic findings of COVID-19 pneumonia.

TABLE 1. Major Characteristics of the Databases Used for the Development and Validation of the System

Database	Data Format	Sources	Cases		
			Healthy	Other Pneumonia	COVID-19 Pneumonia
COVID-19 Image Data Collection[28]	.png	• Italian Society of Medical and Interventional Radiology (https://www.sirm.org/category/senza-categoria/covid-19/) • Radiopedia.org (https://radiopaedia.org) • Figure1.com (https://www.figure1.com/covid-19-clinical-cases) • Eurorad.org operated by the European Society of Radiology (https://www.eurorad.org/) • Radiological Society of North America (https://pubs.rsna.org)	—	19	154
Figure 1 COVID-19 Chest X-ray[29]	.jpg	Figure1.com	3	2	35
ActualMed COVID-19 Chest X-ray data set[30]	.png	ActualMed	116	—	50
COVID-19 Radiography Database[31]	.png	• Italian Society of Medical and Interventional Radiology • Radiological Society of North America • Medarxiv.org • New England Journal of Medicine	1341	1345	219
RSNA Pneumonia Detection Challenge data set[32]	DICOM	National Institutes of Health Clinical Center	8851	9555	—

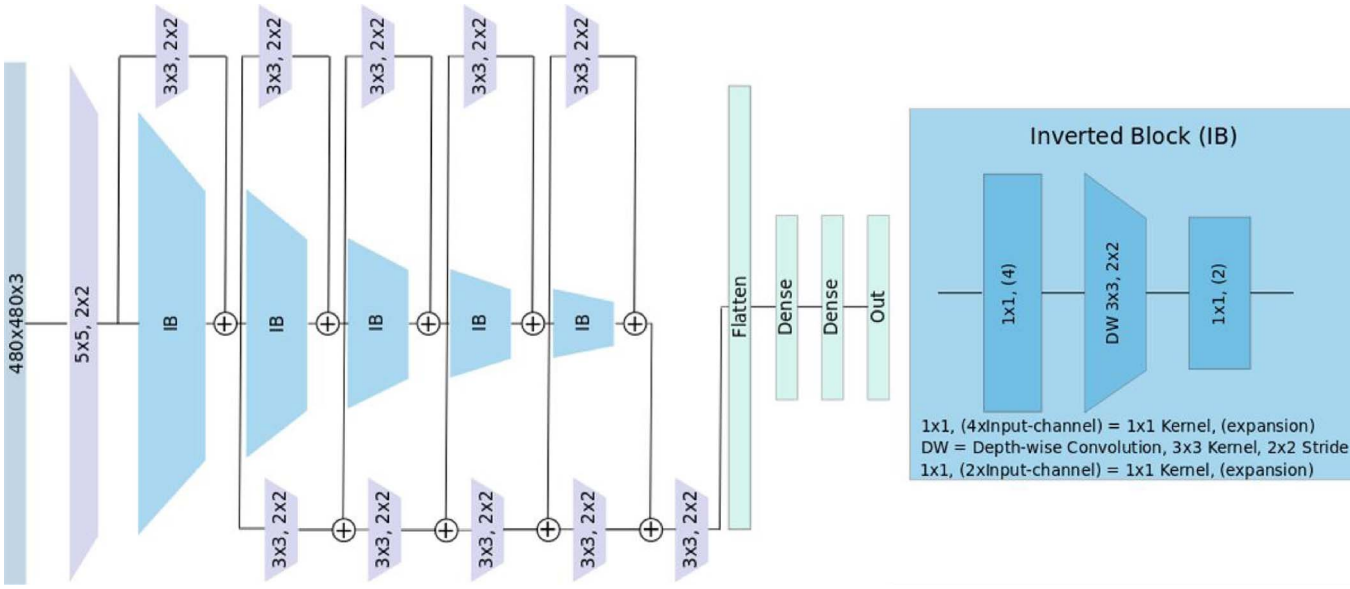

FIGURE 2. The proposed network architecture. Left, overview over the full depth of the network; right, detailed description of the inverted bottleneck block (IB). In the first stage, the feature map channels are expanded by a factor of 4 to compensate the information loss due to rectified linear unit activations. The depth-wise (DW) convolution is used as feature extractor and downsampling stage at once. At the IB's output stage, the feature map channels are collapsed by a factor of 2.

Artificial Intelligence–Based System for Chest Radiograph Analysis

To diagnose COVID-19 pneumonia and distinguish it from other viral/bacterial pneumonias and healthy lungs, a dedicated AI-based system was developed. The system uses a newly introduced deep learning algorithm, which is a modified and enhanced version of the network presented by Wang et al.[33] Either the anteroposterior or the posteroanterior projection of the CXR image was inputted to the system; lateral projections were omitted. The system is composed of an ensemble of 3 individual models, each of them following the architecture shown in Figure 2. Each model uses as input the entire CXR (the pixel values of each image were scaled to a range between 0 and 1), which was downsampled by a learnable strided convolution and further processed by an alteration of the inverted bottleneck blocks (IBs).[34] Inverted bottleneck blocks were used to generate and downsample the feature maps. The combined and downsampled feature maps are fed to a multilayer perceptron, which provides the model's output in the form of a probability for the healthy, other pneumonia, and COVID-19 pneumonia class. Finally, the ensemble is created by using the 3 best performing models during a single training process—checkpoint ensemble.[35] Each of these models predicts a slightly different output probability for the 3 classes that were weighted by the inverse of their entropy for the final prediction.

As shown in Table 2, the merged database is used to formulate the training and testing sets. A total of 258 COVID-19 pneumonia CXRs, including follow-up scans or multiview examinations (anteroposterior and posteroanterior projection), were used for training, whereas the remaining 100 were used for testing. Using multiple views as well as follow-up scans might be considered as a special case of data augmentation. The other pneumonia class is composed of 5440 and 98 unique patient scans[32] for the training and testing sets, respectively.[28,32] Of the 8066 healthy individuals, 7966 unique patient scans were extracted for training and 100 for testing. Table 2 provides statistics for both training and testing sets.

As depicted in Table 2, the number of cases was unbalanced among the classes considered. To mitigate the negative effects of such a skewed class distribution, oversampling of COVID-19 pneumonia cases was applied. Further, conventional data augmentation techniques, such as slight translation (10%), rotation (±10 degrees), horizontal flip, zoom (10%), and intensity shift (10%), were randomly applied on all images. This kind of data augmentation aims at solving the risk of overfitting due to multiple appearances of oversampled COVID-19 pneumonia cases. To emphasize the importance of COVID-19 detection, the corresponding error was weighted by a factor of 4. An additional performance boost had been achieved by pretraining our network on a subset of the ImageNet[36] collection.

To implement our system, we used the Keras[37] with the Tensoflow framework in the 2.2.4 and 1.12.0 versions. The model was trained on an Nvidia Titan X (12 GB) graphics processing unit for 15 hours. Further,

TABLE 2. Data Allocation Into Training and Testing Sets

	Training			Testing		
	Healthy	**Other Pneumonia**	**COVID-19 Pneumonia**	**Healthy**	**Other Pneumonia**	**COVID-19 Pneumonia**
Patients	7966	5440	192	100	98	74
Multiple views (AP, PA)	0	0	15	0	0	13
Follow-up CXR	0	11	51	0	2	13
Total CXR	7966 (58.3%)	5451 (39.8%)	258 (1.9%)	100 (33.3%)	100 (33.3%)	100 (33.3%)

AP, anteroposterior projection; PA, posteroanterior projection; CXR, chest x-ray examination.

our machine used an Intel Core i7-5960X central processing unit at 3.00 GHz and 128 GB random-access memory.

Experimental Reader Setup

Eleven readers, blinded to the diagnosis and system's output, independently scored the same testing set of CXR as applied to the AI-based system (100 normal, 100 COVID-19 pneumonias, and 100 other pneumonias). For this purpose, the readers received individual access to the image database. All images consisted of either the anteroposterior or the posteroanterior projection, no lateral projections were used. All radiologists performed the CXR image analysis on a picture archiving and communication system workstation using Barco professional medical display monitors (MDCC-6430; BARCO, Kortrijk, Belgium). The readers were asked to rate the CXR images as either "normal," "pneumonia other than COVID-19," or "pneumonia with findings characteristic for COVID-19." The criteria used are shown in Table 3.

Five readers were residents with 2 to 4 years of training. Six readers were board-certified, consulting radiologists, including 2 dedicated chest radiologists (with 6 and 20 years of experience), 1 subspecialized emergency radiologist (with 10 years of experience), and 3 general radiologists (with 6, 13, and 19 years of experience).

Every radiologist rated the 300 cases individually; results were calculated for each radiologist. There was no consensus reading. To calculate sensitivity, specificity, and accuracy, the mean of each group of interest was calculated. In addition, the F-score was analyzed (F-score [harmonic mean] = [2 × PPV × SENS]/[PPV + SENS]; PPV being positive predictive value and SENS standing for sensitivity).

All cases with COVID-19 pneumonia and other pneumonia patterns were pooled together to the pathology group (all cases displaying any pneumonia, without the healthy cases). Sensitivity, specificity, and accuracy for detecting pathology were calculated for the system, the consultants, residents, and all human readers separately. Moreover, precision (PPV) and recall (sensitivity) for other types of pneumonia versus COVID-19 pneumonia was analyzed separately for each group. The χ^2 test was used to compare the sensitivity, specificity, accuracy, PPV, and F-scores of the readers and the proposed system. Interrater agreement was calculated among the readers with the Fleiss κ test: poor ($\kappa < 0$), slight ($\kappa < 0.2$), fair ($\kappa < 0.4$), moderate ($\kappa < 0.6$), substantial ($\kappa < 0.8$), and almost perfect agreement ($\kappa \leq 1$). The agreement between the consultants and the radiologists was compared with the Z-score. The significance level was set to 0.05, and MedCalc Statistical Software version 19.3 (MedCalc Software Ltd, Ostend, Belgium) was used for computation.

TABLE 3. CXR Findings for Pneumonias With Different Causes

Type of Pneumonia		Findings
Typical COVID-19 pneumonia findings on CXR[2]		Consolidation is the most common finding (47%) followed by ground-glass opacities (33%). Pathologies on CXR have a peripheral distribution (41%) and lower zone predominance (50%) with bilateral involvement (50%). Pleural effusion is uncommon (3%).
Non–COVID-19 pneumonias	Bacterial pneumonia[38,39]	Consolidation confined to a lung segment or 1 lobe is typically seen in bacterial pneumonias. Bulging fissure signs are attributed to typical bacterial pneumonias. Bronchopneumonias (focal segmental distribution) after aspiration may be found on both sides in the lower lobes.
	Atypical pneumonia (PCP, fungal, viral, pneumonia other than COVID) patterns on CXR[40,41]	Most common finding is reticulation with or without ground-glass opacities (90%) followed by consolidation (50%) and nodules (10%). The abnormalities are usually diffuse and symmetric on both sides.

CXR, chest x-ray examination; PCP, pneumocystis pneumonia.

TABLE 4. Performance of the Proposed Approach on the Testing Set

	Proposed System			
	Sensitivity (%)	Specificity (%)	PPV (%)	F-Score (%)
Healthy	98.0	94.5	89.9	93.8
Other pneumonia	88.0	98.5	96.7	92.1
COVID-19 pneumonia	97.0	98.5	97.0	97.0
Mean ± SD	94.3 ± 4.5	97.2 ± 1.9	94.5 ± 3.3	94.3 ± 2.0

For our evaluation, we present means and standard deviations.
PPV, positive predictive value.

RESULTS

System Performance

The proposed system achieved an average accuracy of 94.3%. Table 4 and Figure 3 summarize the results on the testing set. The introduced system primarily misclassified other pneumonia cases as healthy, thus reducing the sensitivity for this class. Furthermore, the 3 false-positive COVID-19 pneumonia cases reduced the PPV for the class of interest. Figure 1 shows examples of detected pathological lung tissue.

Comparison Between Human Reader and System

The 11 radiologists reached an overall accuracy of 61.4% ± 5.3%. The readout results for all radiologists are summarized in Table 5 and Figures 3 and 4. The combined pattern sensitivity (66.1% ± 13.7%) as well as the combined F-score (65.5% ± 12.4%) are significantly lower than for the proposed system (94.3% ± 4.5% [$P < 0.00032$] and 94.3% ± 2.0% [$P < 0.00001$], respectively).

TABLE 5. Performance (Mean ± SD) of the 11 Radiologists

	Sensitivity (%)	Specificity (%)	PPV (%)	F-Score (%)
Healthy	85.0 ± 12.8	90.7 ± 2.7	82.4 ± 3.1	83.0 ± 6.4
COVID-19 pneumonia	53.2 ± 11.2	81.0 ± 6.0	59.0 ± 6.8	55.2 ± 7.2
Mean ± SD	66.1 ± 13.7	83.0 ± 5.6	66.8 ± 11.0	65.5 ± 12.4

PPV, positive predictive value.

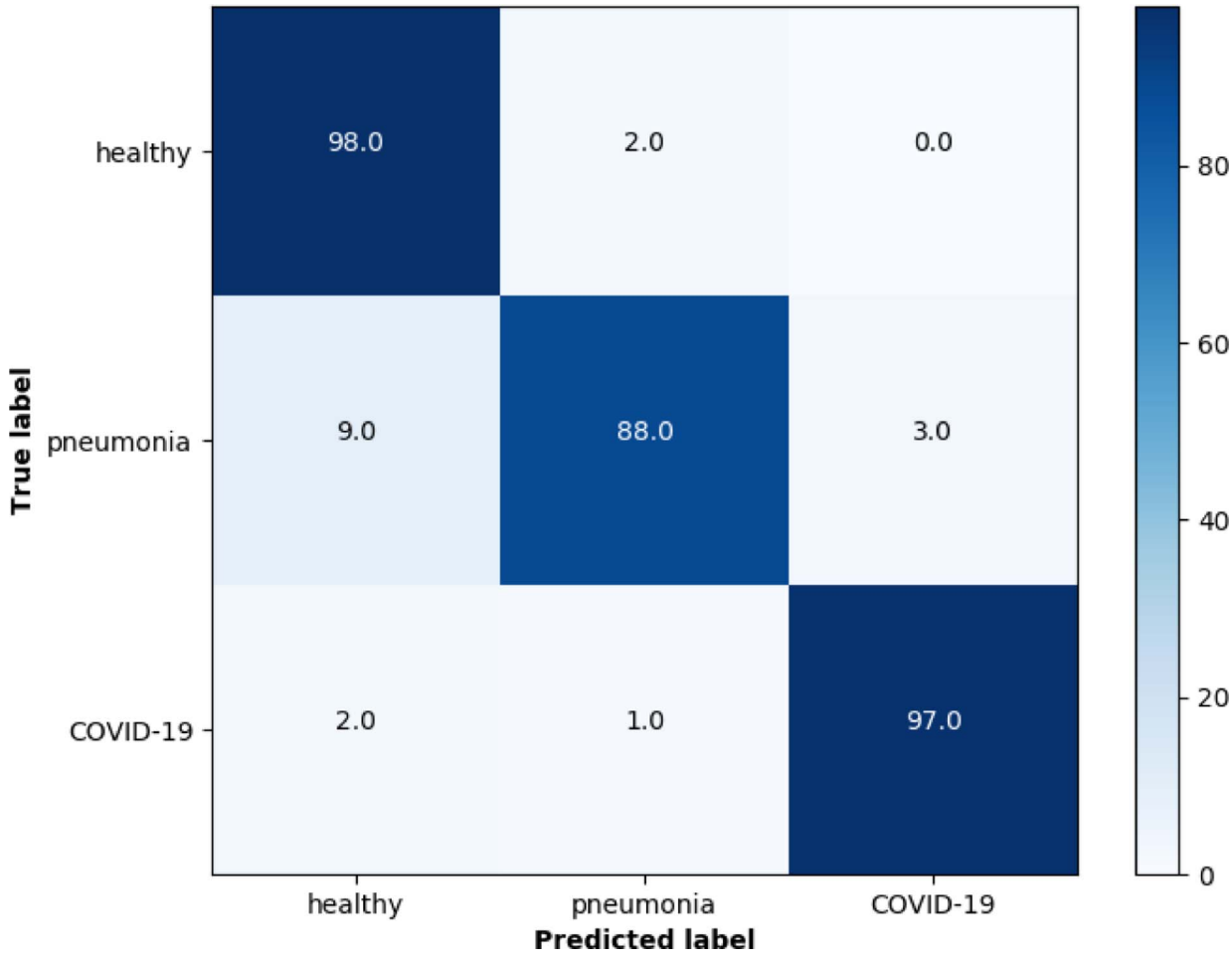

FIGURE 3. Confusion matrix for our ensemble system with entropy voting.

Comparison Between Consultant and Resident Radiologists

Consultant radiologists reached an overall accuracy of 65.0% ± 5.8%. The readout results for the consultant radiologists are summarized in Table 6 and Figure 5. On the other hand, the resident radiologists demonstrated a similar overall accuracy of 67.4% ± 4.2%. Consultants detected significantly more COVID-19 pneumonia cases ($P = 0.008$) and fewer healthy cases ($P < 0.00001$) than the residents did (Table 7, Fig. 6).

Comparison Between Pathological and Healthy Chest Radiographs

When other types of pneumonia and COVID-19 pneumonia cases were pooled, the proposed system detected pathology in 94.6% of patients (sensitivity) with a specificity of 97.4%, leading to an accuracy of 95.7%. The human readers reached lower numbers of 90.7%, 85.0%, and 88.8% ($P = 0.074$, $P = 0.00032$, $P = 0.00023$, respectively). The diagnostic accuracy of the residents (90.2%) was significantly higher than the accuracy of the consultants for detecting pathology (87.7%, $P = 0.021$).

Comparison Between COVID-19 Pneumonia and Other Pneumonias

Human readers reached similar F-sores for other types of pneumonia and COVID-19 pneumonia of 58.4% and 55.2% ($P = 0.111$). The proposed system was superior in detecting COVID-19 pneumonia than other types pneumonia (F-score 97.0% vs 92.1%; $P = 0.038$). The system demonstrated significantly better results than the human readers ($P < 0.00001$ for pneumonia and COVID-19).

Interreader Agreement

The interreader agreement was moderate: the average weighted κ of the consultant radiologists was 0.515 ± 0.07 and did not differ significantly from the weighted κ of the residents (0.535 ± 0.08, $P = 0.43$).

DISCUSSION

In this study, we aimed to develop a deep learning algorithm for the diagnosis of COVID-19 pneumonia on CXRs. The results from this study demonstrated that the proposed system can reliably detect pneumonia on CXRs. In addition, the developed algorithm is capable of separating COVID-19 pneumonia from other forms of pneumonia with high diagnostic accuracy when compared with human readers.

The system's development takes advantage of the availability of open-access databases. The open-access databases allow the head-to-head comparison of algorithms, support transparency and repeatability in research, and boost the AI research in the field of diagnosis and treatment of COVID-19. However, the used databases introduce several limitations. Specifically, the provided metadata (eg, case-specific history, disease severity, course of the disease, acquisition times, and acquisition machine) are not standardized and thus scarcely provided, whereas different image data formats are supported. In general, a certain inhomogeneity of data has to be taken into account when using public data sets.

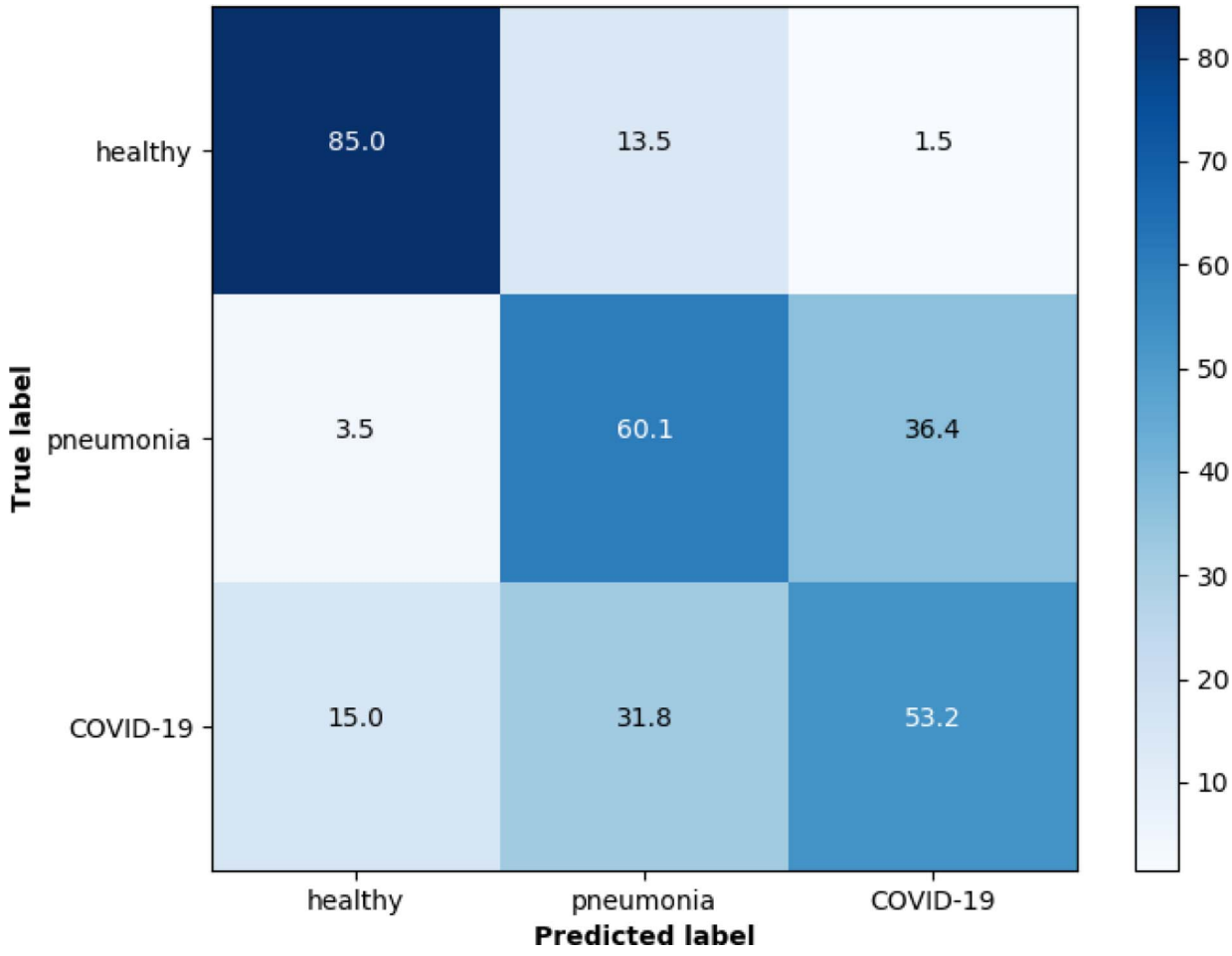

FIGURE 4. Confusion matrix for the 11 radiologists.

This also applies for indication, disease status, examination type, and nonstandardized examination parameters and apparatures. Furthermore, the number of available lateral views, although they are very informative, is limited and thus not included.

The introduced deep learning system is based on an ensemble of 3 individual models, whereas its major innovation characteristic is its ability to be installed and used in portable devices, while keeping high performance. Specifically, by introducing the IBs in the architecture, the number of used layers and thus the network complexity is significantly reduced. In contrast to previous approaches,[33,34] we used stride convolutions in our IBs to combine downsampling and feature extraction. The information shortcuts between the downsampling stages preserve the high accuracy, which is further supported by using checkpoint ensemble.

As regards the system performance in the clinical application, the algorithm showed high sensitivity for the detection of healthy patients and COVID-19 pneumonia cases (98.0% and 97.0%, respectively). However, the sensitivity for the detection of pneumonia other than COVID-19 pneumonia was significantly lower (88.0%). The strength of our system compared with that in the study by Wang et al[33] was the 6% increase in COVID-19 pneumonia sensitivity. This performance gain, however, came at the cost of pneumonia sensitivity, leading to only a marginal overall accuracy increase of 0.3% from 94%[33] to 94.3%.

Despite the sacrifice in sensitivity for non–COVID-19 pneumonia, the proposed system was able to surpass radiologists of any training level in the detection of COVID-19 pneumonia. The radiologists in particular struggled with the discrimination between COVID-19 pneumonia and other types of pneumonia. We observed that many non–COVID-19 pneumonia CXRs were misinterpreted as COVID-19 pneumonia. Conversely, a considerable number of COVID-19 pneumonia cases were regarded as non–COVID-19 pneumonia. The obvious explanation for the first observation is that, in the current pandemic, readers tend to overcall consolidations on radiographs as COVID-19 pneumonia related. This effect was more prevalent with consultant radiologists, who detected more cases of COVID-19 pneumonia while conversely overcalling the disease. As compared with the residents, the diagnostic threshold of board-certified radiologists was lower for COVID-19 pneumonia.

Especially when used as a screening tool for disease, imaging studies bear the risk of overdiagnosis bias.[42,43] The latter finding that COVID-19 pneumonia was misclassified as other pneumonia is probably associated with more advanced COVID-19 disease, which

TABLE 6. Performance (Mean ± SD) of the 6 Consultant Radiologists

	Sensitivity (%)	Specificity (%)	PPV (%)	F-Score (%)
Healthy	79.0 ± 14.3	92.0 ± 2.5	83.5 ± 3.1	80.4 ± 7.3
Other pneumonia	59.2 ± 14.0	75.5 ± 11.9	56.7 ± 11.0	56.6 ± 8.4
COVID-19 pneumonia	56.8 ± 7.4	80.0 ± 7.2	60.1 ± 7.5	57.7 ± 3.5
Mean ± SD	65.0 ± 9.9	82.5 ± 7.0	66.8 ± 11.9	64.9 ± 11.0

PPV, positive predictive value.

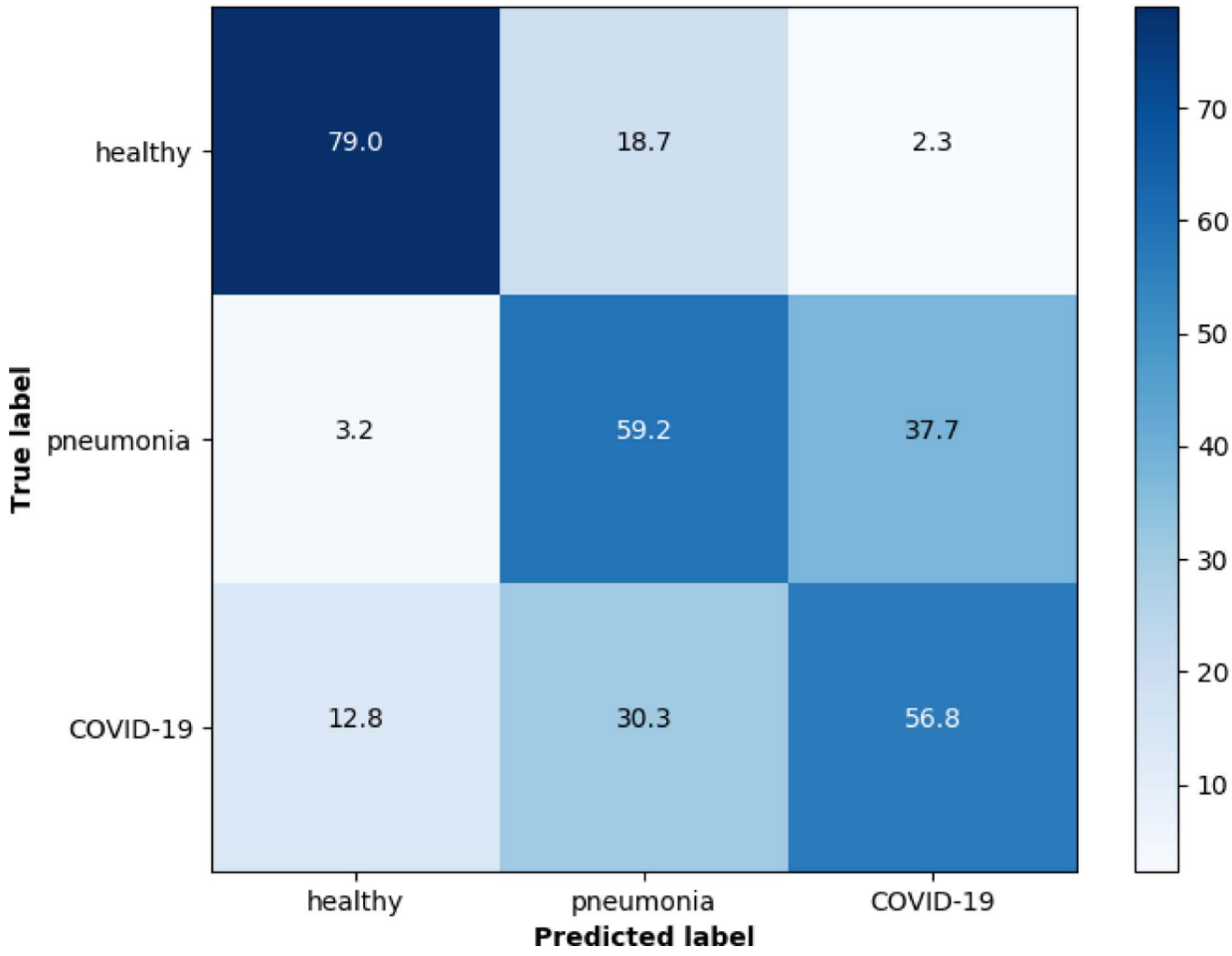

FIGURE 5. Readout for 6 consultants.

typically presents with extensive consolidations. It is well documented that advanced COVID-19 infection can be complicated by bacterial superinfection and adult respiratory distress syndrome, characterized by consolidations and lobar or diffuse distribution in the lungs.[44–46] Visual discrimination between viral and bacterial pneumonia is not feasible at this point. Although the algorithm showed better diagnostic accuracy, both the proposed system as well as the human readers exhibited acceptable accuracy in determining pathology on radiographs. Discriminating pneumonia/COVID-19 can be considered a secondary task and might also be accomplished by polymerase chain reaction testing.

A critical limitation of CXR is its inferior sensitivity to detecting ground-glass opacifications. Moreover, ground-glass opacifications—mainly in a basal and peripheral, subpleural locations—have been reported as being one of the hallmark findings in early COVID-19 pneumonia.[46] However, the deployed CXR-based system gave comparable values for sensitivity, PPV, and F-score for COVID-19 pneumonia diagnosis to those found from CT images (Table 8), although we used data from a different pool of patients, which makes a direct comparison more difficult. As shown in Table 8, a CXR-based system holds the potential to support radiologists in the initial diagnosis, whereas CT-based systems could be additionally used for risk assessment and treatment optimization during the course of COVID-19 pneumonia.

In the utilization for initial diagnosis and risk estimation, imaging in general plays a vital role in the follow-up of patients. Numerous studies using chest CT have already shown that the extent and morphology of pulmonary infiltrates at presentation permit risk estimation and could indicate whether the course of the disease will be moderate or severe.[50,51] Recent reports suggest that the same quantification and risk prediction might be feasible using CXRs.[52–54] Deep learning–assisted detection systems for CXR in conjunction with integrated disease quantification could potentially serve as a valid surrogate for chest CT in the current pandemic.

The presented study has several limitations. First, the sample size, in particular for the development of a deep learning system, is relatively small. However, we have been spurred on by these promising results, and prospective trials are in place to further develop the proposed system. Another critical aspect is the heterogeneous database comprising mild, moderate, and severe cases. However, this could also be seen as another challenge that was overcome by the system and is supported by the positive test performance. Finally, this first iteration of the algorithm was solely dedicated to deploy an AI-assisted diagnostic tool; disease quantification has not been addressed in this version. Upcoming versions are

TABLE 7. Performance (Mean ± SD) of the 5 Residents

	Sensitivity (%)	Specificity (%)	PPV (%)	F-Score (%)
Healthy	92.2 ± 4.8	89.2 ± 2.0	81.1 ± 2.6	86.2 ± 2.6
Other pneumonia	61.2 ± 9.4	79.6 ± 8.6	61.8 ± 8.8	60.5 ± 4.4
COVID-19 pneumonia	48.8 ± 13.3	82.3 ± 4.3	57.7 ± 5.6	52.1 ± 9.1
Mean ± SD	67.4 ± 18.3	83.7 ± 4.0	66.9 ± 10.2	66.3 ± 14.5

PPV, positive predictive value.

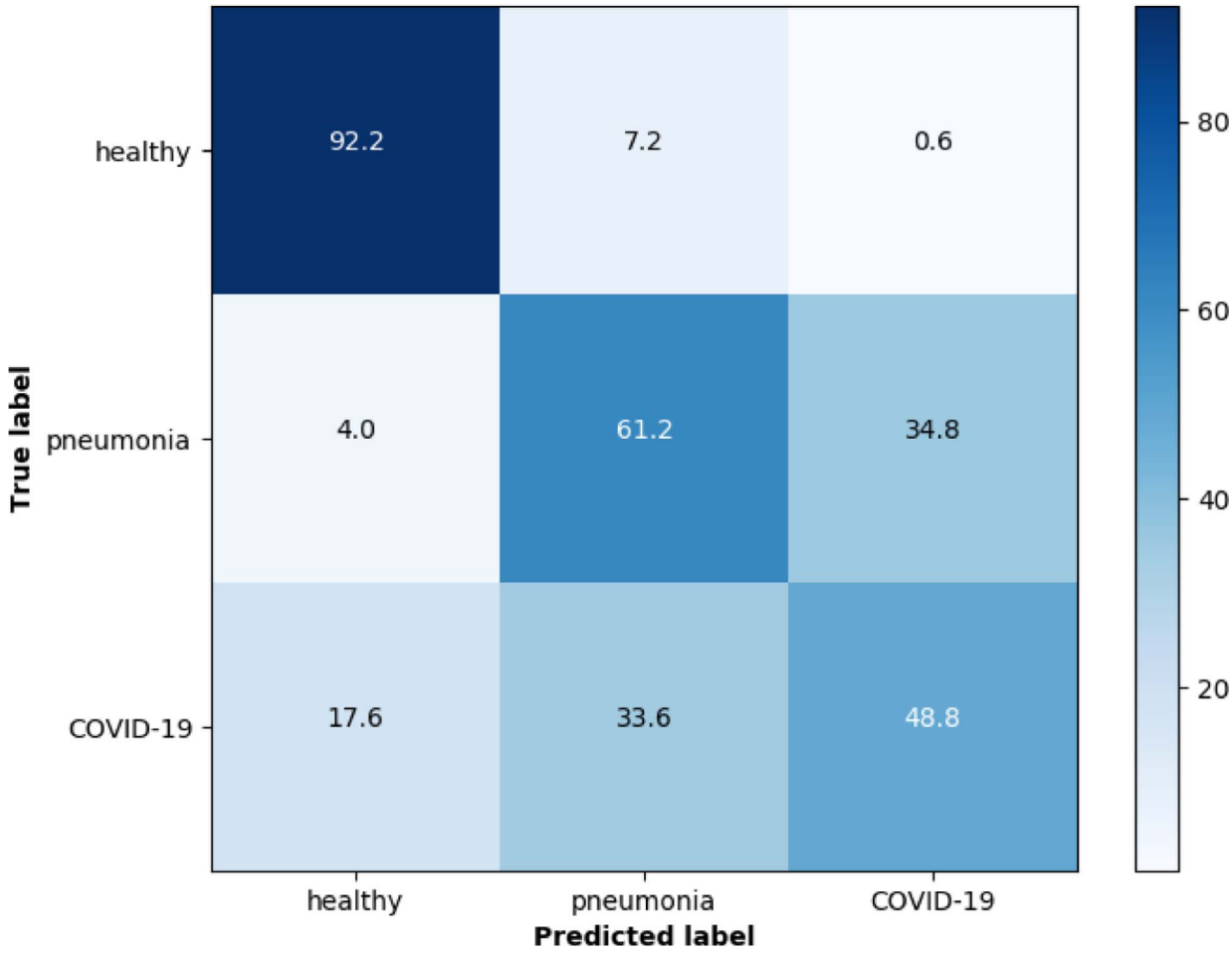

FIGURE 6. Readout for 5 residents.

planned to incorporate a multivariate system, including extent of disease, as well as correlation with clinical parameters and outcome.

In conclusion, we proposed a deep learning algorithm for the diagnosis of COVID-19 pneumonia on CXRs. Our data suggests that this approach is robust and capable to diagnose and differentiate COVID-19 pneumonia from other types of pneumonia. In this pilot study, the deployed system surpassed radiologists at all levels of training. The presented results emphasize the potential of CXRs in combination with AI-supported detection systems for patient triage and to ultimately mitigate the impact on radiology departments and health care providers globally during the COVID-19 pandemic.

TABLE 8. High Level Comparison Between Artificial Intelligence Driven Diagnostic Systems Using CT and CXR

Method	Imaging Modality	Sensitivity (%)	Specificity (%)	PPV (%)	F-Score (%)
Xu et al[47]	CT	86.7	—	81.3	83.9
Gozes et al[48]	CT	98.2	0.922		—
Chen et al (retrospective/ prospective)[49]	CT	100/100	99.2/81.8	88.4/88.9	93.8/94.1

CT, computed tomography; CXR, chest x-ray examination; and PPV, positive predictive value.

REFERENCES

1. Wen Z, Chi Y, Zhang L, et al. Coronavirus disease 2019: initial detection on chest CT in a retrospective multicenter study of 103 Chinese subjects. *Radiol Cardiothorac Imaging*. 2020;2:e200092.
2. Wong HYF, Lam HYS, Fong AH-T, et al. Frequency and distribution of chest radiographic findings in patients positive for COVID-19. *Radiology*. 2020;296:E72–E78.
3. Li K, Wu J, Wu F, et al. The clinical and chest CT features associated with severe and critical COVID-19 pneumonia. *Invest Radiol*. 2020;55:327–331.
4. Salehi S, Abedi A, Balakrishnan S, et al. Coronavirus disease 2019 (COVID-19): a systematic review of imaging findings in 919 patients. *AJR Am J Roentgenol*. 2020;215:87–93.
5. Lyu P, Liu X, Zhang R, et al. The performance of chest CT in evaluating the clinical severity of COVID-19 pneumonia: identifying critical cases based on CT characteristics. *Invest Radiol*. 2020;55:412–421.
6. Ai T, Yang Z, Hou H, et al. Correlation of chest CT and RT-PCR testing for coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases. *Radiology*. 2020;296:E32–E40.
7. Xie X, Zhong Z, Zhao W, et al. Chest CT for typical coronavirus disease 2019 (COVID-19) pneumonia: relationship to negative RT-PCR testing. *Radiology*. 2020;296:E41–E45.
8. Caruso D, Zerunian M, Polici M, et al. Chest CT features of COVID-19 in Rome, Italy. *Radiology*. 2020;296:E79–E85.
9. Choi H, Qi X, Yoon SH, et al. Extension of coronavirus disease 2019 (COVID-19) on chest CT and implications for chest radiograph interpretation. *Radiol Cardiothorac Imaging*. 2020;2:e200107.

10. Weinstock MB, Echenique A, Joshua W, et al. Chest x-ray findings in 636 ambulatory patients with COVID-19 presenting to an urgent care center: a normal chest x-ray is no guarantee. *J Urgent Care Med*. 2020;14:13–18.
11. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. *Nature*. 2020;577:89–94.
12. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542:115–118.
13. Bullock J, Luccioni A, Pham KH, et al. Mapping the landscape of artificial intelligence applications against COVID-19. *ArXiv* e-prints 2003.11336. April 23, 2020. Available at: https://arxiv.org/abs/2003.11336v1. Accessed August 5, 2020.
14. Shi F, Wang J, Shi J, et al. Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for COVID-19. *IEEE Rev Biomed Eng*. 2020:PP.
15. Bai HX, Wang R, Xiong Z, et al. Artificial intelligence augmentation of radiologist performance in distinguishing COVID-19 from pneumonia of other origin at chest CT. *Radiology*. 2020;296:E156–E165.
16. Narin A, Kaya C, Pamuk Z. Automatic detection of coronavirus disease (COVID-19) using x-ray images and deep convolutional neural networks. *ArXiv*. 2020;2003:10849. Available at: https://arxiv.org/abs/2003.10849. Accessed June, 2020.
17. Asif S, Wenhui Y, Jin H, et al. Classification of COVID-19 from chest x-ray images using deep convolutional neural networks. *medRxiv* preprint: 2020.05.01.20088211. Available at: https://doi.org/10.1101/2020.05.01.20088211. Accessed August 2020.
18. Khan AI, Shah JL, Bhat MM. CoroNet: a deep neural network for detection and diagnosis of COVID-19 from chest x-ray images. *Comput Methods Programs Biomed*. 2020;196:105581. Accessed June 2020.
19. Ozturk T, Talo M, Yildirim EA, et al. Automated detection of COVID-19 cases using deep neural networks with x-ray images. *Comput Biol Med*. 2020;121:103792.
20. Zhu J, Shen B, Abbasi A, et al. Deep transfer learning artificial intelligence accurately stages COVID-19 lung disease severity on portable chest radiographs. *PLoS One*. 2020;15:e0236621.
21. Liang W, Yao J, Chen A, et al. Early triage of critically ill COVID-19 patients using deep learning. *Nat Commun*. 2020;11:3543.
22. Rubin GD, Ryerson CJ, Haramati LB, et al. The role of chest imaging in patient management during the COVID-19 pandemic: a multinational consensus statement from the Fleischner society. *Radiology*. 2020;296:172–180.
23. Revel MP, Parker AP, Prosch H, et al. COVID-19 patients and the radiology department—advice from the European Society of Radiology (ESR) and the European Society of Thoracic Imaging (ESTI). *Eur Radiol*. 2020;30:4903–4909.
24. Rai A, MacGregor K, Hunt B, et al. Proof of concept: phantom study to ensure quality and safety of portable chest radiography through glass during the COVID-19 pandemic. *Invest Radiol*. 2020. doi:10.1097/RLI.0000000000000716.
25. Padley SPG, Hansell DM, Flower CDR, et al. Comparative accuracy of high resolution computed tomography and chest radiography in the diagnosis of chronic diffuse infiltrative lung disease. *Clin Radiol*. 1991;44:222–226.
26. Ebner L, Bütikofer Y, Ott D, et al. Lung nodule detection by microdose CT versus chest radiography (standard and dual-energy subtracted). *Am J Roentgenol*. 2015;204:727–735.
27. Cohen JF, Korevaar DA, Altmann DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open*. 2016;6:e012799.
28. Cohen JP, Morrison P, Dao L, et al. COVID-19 Image Data Collection. Available at: https://github.com/ieee8023/covid-chestxray-dataset. 2020. Updated July 31, 2020. Accessed May 2020.
29. Wang L, Lin ZQ, Wong A, et al. DarwinAI Corp, Canada and Vision and Image Processing Research Group, University of Waterloo, Canada. Figure 1 COVID-19 Chest X-ray data initiative. Available at: https://github.com/agchung/Figure1-COVID-chestxray-dataset. Updated May 4, 2020. Accessed May 2020.
30. Wang L, Lin ZQ, Wong A, et al. DarwinAI Corp, Canada and Vision and Image Processing Research Group, University of Waterloo, Canada. Actualmed COVID-19 Chest X-ray data initiative. Available at: https://github.com/agchung/Actualmed-COVID-chestxray-dataset. Updated May 1, 2020. Accessed May 2020.
31. Chowdhury M, Rahman T, Khandakar A. Qatar University, Doha, Qatar and University of Dhaka, Bangladesh. Can AI help in screening Viral and COVID-19 pneumonia? *IEEE Access*. 2020;8:132665–132676. COVID-19 Radiography Database. Available at: https://www.kaggle.com/tawsifurrahman/covid19-radiography-database. Updated March 28, 2020. Accessed May 2020.
32. Radiological Society of North America. RSNA Pneumonia Detection Challenge. Available at: https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data. Updated November 1, 2018. Accessed May 2020.
33. Wang L, Lin ZQ, Wong A, et al. COVID-net: a tailored deep convolutional neural network Design for Detection of COVID-19 cases from chest X-ray images. *ArXiv e-prints*. 2003.09871. May 11, 2020. Available at: https://arxiv.org/abs/2003.09871. Accessed June 13, 2020.
34. Sandler M, Howard A, Zhu M, et al. Mobilenetv2: inverted residuals and linear bottlenecks. *Proc IEEE Conf Comp Vis Pattern Recog*. 2018;4510–4520.
35. Chen H, Lundberg S, Lee S-I. Checkpoint ensembles: ensemble Methods from a single training process. ArXiv:1710.03282. October 9, 2017. Available at: https://arxiv.org/abs/1710.03282. Accessed October 5, 2020.
36. Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database. *Proc IEEE Conf Comp Vis Pattern Recog*. 2009;248–255.
37. Keras. Version 2.2.4. Francois Chollet. GitHub. 2015. Available at: https://keras.io. Accessed June 2020.
38. Collins J, Stern EJ. *Chest Radiology: The Essentials*. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2014.
39. Vilar J, Domingo ML, Soto C, et al. Radiology of bacterial pneumonia. *Eur J Radiol*. 2004;51:102–113.
40. Christe A, Walti L, Charimo J, et al. Imaging patterns of pneumocystis jirovecii pneumonia in HIV-positive and renal transplant patients—a multicentre study. *Swiss Med Wkly*. 2019;149:w20130.
41. Franquet T, Giménez A, Hidalgo A. Imaging of opportunistic fungal infections in immunocompromised patient. *Eur J Radiol*. 2004;51:130–138.
42. Davies L, Petitti DB, Martin L, et al. Defining, estimating, and communicating overdiagnosis in cancer screening. *Ann Intern Med*. 2018;169:36–43.
43. Welch HG, Prorok PC, O'Malley JA, et al. Breast-cancer tumor size, overdiagnosis, and mammography screening effectiveness. *N Engl J Med*. 2016;375:1438–1447.
44. Jajodia A, Ebner L, Heidinger B, et al. Imaging in corona virus disease 2019 (COVID-19)—a scoping review. *Eur J Radiol Open*. 2020;100237:7.
45. Bernheim A, Mei X, Huang M, et al. Chest CT findings in coronavirus disease-19 (COVID-19): relationship to duration of infection. *Radiology*. 2020;295:200463.
46. Xiong Y, Sun D, Liu Y, et al. Clinical and high-resolution CT features of the COVID-19 infection: comparison of the initial and follow-up changes. *Invest Radiol*. 2020;55:332–339.
47. Xu X, Jiang X, Ma C, et al. A deep learning system to screen novel coronavirus disease 2019 pneumonia. *Engineering (Beijing)*. 2020. February 21, 2020. Available at: https://doi.org/10.1016/j.eng.2020.04.010. Accessed August 2020.
48. Gozes O, Frid-Adar M, Greenspan H, et al. Rapid AI development cycle for the coronavirus (COVID-19) pandemic: initial results for automated detection & patient monitoring using deep learning CT image analysis. *ArXiv*. e-prints: 2003.05037. March 24, 2020. Available at: https://arxiv.org/abs/2003.05037. Accessed August 2020.
49. Chen J, Wu L, Zhang J, et al. Deep learning–based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography: a prospective study. *medRxiv*. preprint: 2020.02.25.20021568. Available at: https://doi.org/10.1101/2020.02.25.20021568. Accessed August 2020.
50. Leonardi A, Scipione R, Alfieri G, et al. Role of computed tomography in predicting critical disease in patients with covid-19 pneumonia: a retrospective study using a semiautomatic quantitative method. *Eur J Radiol*. 2020;130:109202.
51. Grégory J, Raynaud L, Galy A, et al. Extension of COVID-19 pulmonary parenchyma lesions based on real-life visual assessment on initial chest CT is an independent predictor of poor patient outcome. *Infect Dis (Lond)*. 2020;52:838–840.
52. Murphy K, Smits H, Knoops AJG, et al. COVID-19 on chest radiographs: a multireader evaluation of an artificial intelligence system. *Radiology*. 2020;296:E166–E172.
53. Li MD, Arun NT, Gidwani M, et al. Automated assessment and tracking of COVID-19 pulmonary disease severity on chest radiographs using convolutional Siamese neural networks. *Radiol Artif Intell*. 2020;2:4.
54. Toussie D, Voutsinas N, Finkelstein M, et al. Clinical and chest radiography features determine patient outcomes in young and middle-aged adults with COVID-19. *Radiology*. 2020;297:E197–E206.