Published in final edited form as:
Nat Biomed Eng. 2021 June 01; 5(6): 509–521. doi:10.1038/s41551-021-00704-1.

A deep-learning pipeline for the diagnosis and discrimination of viral, non-viral and COVID-19 pneumonia from chest X-ray images

A full list of authors and affiliations appears at the end of the article.

[#] These authors contributed equally to this work.

Abstract

Common lung diseases are first diagnosed via chest X-rays. Here, we show that a fully automated deep-learning pipeline for chest-X-ray-image standardization, lesion visualization and disease diagnosis can identify viral pneumonia caused by Coronavirus disease 2019 (COVID-19), assess its severity, and discriminate it from other types of pneumonia. The deep-learning system was developed by using a heterogeneous multicentre dataset of 145,202 images, and tested retrospectively and prospectively with thousands of additional images across four patient cohorts and multiple countries. The system generalized across settings, discriminating between viral pneumonia, other types of pneumonia and absence of disease with areas under the receiver operating characteristic curve (AUCs) of 0.88–0.99, between severe and non-severe COVID-19 with an AUC of 0.87, and between severe or non-severe COVID-19 pneumonia and other viral and non-viral pneumonia with AUCs of 0.82–0.98. In an independent set of 440 chest X-rays, the system performed comparably to senior radiologists, and improved the performance of junior radiologists. Automated deep-learning systems for the assessment of pneumonia could facilitate early intervention and provide clinical-decision support.

The outbreak of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), and its disease COVID-19, led to a pandemic of the highest concern[1–6]. The genome of the new

[*]Corresponding authors, guangyu.wang24@gmail.com; weimi003@scu.edu.cn; tingchen@mail.tsinghua.edu.cn; kang.zhang@gmail.com, or lintx@mail.sysu.edu.cn.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Author contributions
XH.L., J.S., Z.L., L.Y., C.W., X.W., T.C., K.W., X.Z., Z.Z., J.Y., Y.S., R.D., W.L., T.Y., M.G., J.W., Z.Y., H.C., G.L., L.Z., L.Y., W.X., W.W., A.O., I.Z., C.Z., O.L., W.L., J.L., W.C., W.C., J.S., L.Z., L.Zhang., Z.Y., X.Z., G.L., G.G., L..L., L.M., Y.L., M.R., E.S., C-B.S., M.F., J.Y-N.L., T.X., J.H., W.L., G. W., and K. Z., collected and analyzed the data. J.Y-N.L and T.C. helped to analyze and interpret the data and the revision of the manuscript. K.Z. and GY.W. conceived and supervised the project and wrote the manuscript. All authors discussed the results and reviewed the manuscript.

Competing interests
The authors declare no competing interests.

Peer review information *Nature Biomedical Engineering* thanks Bulat Ibragimov, Eric Oermann and the other, anonymous, reviewer for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

virus, the epidemiological and clinical features of the infection have been reported[1–4]. The viral infection frequently presented as an upper respiratory-tract infection or pneumonia (COVID-19 pneumonia) that can rapidly progress to acute respiratory failure, multi-organ failure, and death. Chest X-ray (CXR) radiography is the mainstay of screening, triaging, and diagnosing varieties of pneumonia, including bacterial, viral, and other types of pneumonia[5–7]. During the flu season, viral pneumonia is prevalent, and CXR plays a critical role in frontline patient care. Radiologists are aware of certain CXR features that may suggest the diagnosis of viral pneumonia; it is multifocal, reflecting the underlying pathogenesis, and may induce more rapid alveolar and potentially endothelial damages.

Recent developments in artificial intelligence (AI) have provided new potential opportunities for the rapid growth of radiological diagnostic applications[8–11]. Previous studies have proposed the concept of radiomics/imageomics, referring to the extraction of quantitative imaging feature information in a high-throughput manner[12]. The AI model also demonstrated general applicability in retinal diseases and childhood diseases with medical images, pretrained with data of conventional approaches based on transfer learning[13]. To diagnose common lung and heart diseases based on CXR, AI models using weakly-supervised classification[14], or attention-based convolution neural network[15] have also been studied.

Although computational methods have been proposed for lung disease detection, there is still a lack of a fully automatic analysis pipeline that is robust toward variable CXR image conditions and meets the standard of actual clinical application[6,16,17]. One of the challenges is anatomical landmark detection, which plays a vital role in medical image analysis. Radiologists routinely align an input image to these landmarks and perform diagnosis and quantification[18–21]. However, most landmark detection methods were developed for facial recognition. Today, it remains a challenge to standardize medical images to facilitate the downstream diagnostic tasks automatically. Other challenges for the translation of AI systems to clinical applications include the lack of the gold standard for clinical evaluation and the generalization of the systems to different populations or new settings. Another critical obstacle for the medical AI system's general use is that deep learning algorithms' inner decision-making processes remain opaque, which hinders the translation into clinical practice. Therefore, under this unprecedented COVID-19 pandemic, it will be of great significance to develop a general AI system for CXR that can give a fast and accurate diagnosis and severity assessment of viral pneumonia even before molecular test results are available. It is of utmost importance to public health as this system can be deployed quickly to healthcare centers to provide the first-line assessment with a quick turn-around time.

Here we aim to develop a comprehensive system to combat the SARS-CoV-2 or any other emerging upper respiratory viral pandemic. The shortcoming of CXR images is evident. A plain CXR image is the summation of the effect of X-ray on all tissues between the X-ray source and the capturing film; tissue structures are less well defined in an X-ray compared to a CT image and lack 3-dimensional information. To overcome these shortcomings, we integrate multiple state-of-the-art computational methods to construct a robust AI system for CXR diagnosis. This CXR diagnostic system detects common thoracic pathologies, performs viral pneumonia diagnosis, and differentiates COVID-19 from other viral pneumonia. Technically, our AI system is a modular analysis pipeline consisting of automated detection of the anatomical landmarks, lung-lesion segmentation, and pneumonia diagnosis prediction, using CXRs as input. In addition, the AI system could assess COVID-19 clinical severity based on the proposed CXR lung-lesion segmentation model (Fig. 1 and Supplementary Fig. 6).

To develop this AI system, we utilized a large-scale hospital-wide dataset (n=120,702) for the detection of common thoracic pathologies and a large multi-center dataset for pneumonia analysis. We also explored the deliverability of the AI system. To assess its real-world clinical performance and generalizability, we applied the system to external datasets collected from different populations from those used for the model training. Furthermore, we compared the performance of the system with that of radiologists in routine clinical practice. The results showed that the AI's performance is accurate and robust across multiple populations and settings. The system might be integrated into the workflow to improve a radiologist's diagnostic performance.

Results

Image characteristics and system overview

We constructed a large CXR dataset based on the China Consortium of Chest X-ray Image Investigation (CC-CXRI) to develop the AI system. The CC-CXRI consisted of two large-scale datasets: the first one, a CXR database for common thoracic diseases containing 145,202 CXR images retrospectively collected from the Memorial Hospital of Sun Yat-sen University (SYSU), and the second, a CXR dataset (CC-CXRI-P) containing 16,196 CXR images for detecting suspicious pneumonia, including COVID-19 pneumonia. In this study, a general AI system was developed for identifying common thoracic diseases and pneumonia diagnoses and triaging patients using CXR images with an application to COVID-19 pneumonia. Our proposed AI system, an automated CXR analysis pipeline, consisting of three modules: (1) a CXR standardization module, (2) a common thoracic disease detection module, and (3) a final pneumonia analysis module.

The CXR standardization module consisted of anatomical landmarks detection and image registration techniques (Fig. 1 and Supplementary Fig. 7). This module was designed to overcome the notorious problem and well-known challenges of data diversity/variations and non-standardization of CXR images. This study used 12 anatomical landmarks labeled on 676 CXR images to train the landmark detection model. We implemented and compared three deep learning models for the landmark detection, including the U-Net[22], fully convolutional networks (FCN)[23], and DeepLabv3[24], using a five-fold cross-validation test (see more details in Supplementary Methods). DeepLabv3 showed the best performance, so we adopted DeepLabv3 for the landmark detection and subsequent analyses (Supplementary Fig. 8 and Supplementary Table 3). Supplementary Fig. 9a showed a visualization example of our AI model compared with the radiologist's annotation, which obtained accurate landmark detection results. Interestingly, we observed that all three models performed better for the right part of landmarks than for the left, probably due to the contrast condition caused by the cardiovascular region (Supplementary Fig. 9b).

The common thoracic disease detection module classified the standardized CXR images into 14 common thoracic pathologies that are frequently observed and diagnosed, including cardiomegaly, consolidation, edema, effusion, emphysema, fibrosis, hernia, infiltration, mass, nodule, pleural thickening, pneumonia, and pneumothorax (Table 1).

The pneumonia analysis module that consists of a lung-lesion segmentation model and a final classification model estimates the subtype of pneumonia (e.g., viral pneumonia) and assesses the severity of COVID-19. We trained the lung-lesion segmentation using 1,016 CXR images that were manually segmented into four anatomical categories and common lesions of opacification (Supplementary Table 4). We implemented and compared the three segmentation models. The results showed that DeepLabv3 outperformed both FCN and U-Net, and its performance was compared to that of manual delineations by radiologists (Supplementary Table 5).

The SCR dataset is a public CXR dataset with annotated landmarks for lung segmentation (https://www.isi.uu.nl/Research/Databases/SCR/)[25,26]. We validated our system on this database and achieved good performance for landmark detection with a mean of 5.568(±6.175) mm on the actual physical distance error. As the SCR database was established to facilitate studies on anatomical segmentation of CXR images, we also validated our lung segmentation model, which showed good accuracy with Dice of 0.954 and 0.961 for segmentation of the left lung field and right lung field, respectively (Supplementary Table 7).

Multi-label classification of common thoracic diseases

Here, a large-scale dataset (SYSU set) from CC-CXRI, which consisted of 120,702 CXR images from 92,327 patients with labels of 14 common thoracic pathologies, was used for model training. All patients were from hospital visits between October 2018 and July 2020. This dataset was randomly partitioned into three subsets with a ratio of 8:1:1 for training, validation, and testing, respectively. The images were first analyzed by automated detection of anatomical landmarks to permit image registration. Then the standardized CXR images were classified into 14 common thoracic pathologies. All 14 labels were common lung pathologies extracted from real-world clinical reports for CXR images. As some pathologies may co-exist or overlap on the same CXR image, we employed a multi-label classification approach instead of a multi-class classification method, where overlaps between labels were allowed, and labels were predicted individually before integrated into a final prediction. The AI system achieved a macro performance with an area under the receiver operating characteristic curve (AUC) of 0.930 on the test set (Supplementary Table 1). Among the 14 pathologies, pneumonia belongs to the category of pulmonary opacity, which represents the pattern of a decrease in the ratio of gas to soft tissue (blood, lung parenchyma, and stroma) in the lung. The opacity can be broadly divided into five levels of atelectasis, mass, edema, pneumonia, and consolidation, which are vital for the differential diagnosis of pneumonia. On the test set, the AI system achieved an AUC of 0.914 for differentiating pneumonia from all other groups and an AUC of 0.935 for the overall classification of lung opacity (Fig. 2a).

To evaluate the AI system's generalizability across various screening settings, we tested it on a cohort called SYSU-PE, which consisted of additional 24,500 CXR images from 23,585 patients who underwent a routine annual health-check. Compared with the SYSU cohort, there were fewer consolidation or edema cases among the SYSU-PE cohort. The results showed an over-all AUC of 0.916 for multi-label image classification of commonly occurring lung opacity (Fig. 2b). We further applied our AI model to the open public data source RSNA Kaggle competition dataset, and the results also show that our method achieved good performance for lung opacity detection (Supplementary Fig. 1).

Training of the AI system to identify viral pneumonia

To develop a model to differentiate viral pneumonia from other types of pneumonia and absence of pneumonia based on CXR images, we constructed a deep neural network based on the DenseNet-121[27] architecture. The AI system first standardized an input CXR image through anatomical landmark detection and registration before performing lung-lesion segmentation and pneumonia diagnosis (Supplementary Fig. 7).

The diagnosis of pneumonia was verified by a positive polymerase chain reaction (PCR) test or other laboratory test methods, including culture and staining, which served as the ground truth. Medical imaging has been considered part of the diagnostic workup of symptomatic subjects with suspected COVID-19 in settings where laboratory testing information was not available or results are delayed or initially negative in the presence of symptoms attributable to COVID-19[28]. Here, we adopted the terms “gold-standard labels” or “silver-standard labels” to differentiate between the labels obtained from a confirmed laboratory-based ground truth versus clinical and radiographic finding based diagnosis by a consensus of radiologists[29,30]. CXR images were classified into three types, viral pneumonia, other etiologies/types of pneumonia, and absence of pneumonia (normal). The viral pneumonia group consisted of common types of viral pneumonia and COVID-19 patients.

The CXR images in the CC-CXRI-P dataset were all confirmed cases with a definitive “gold-standard label” determined by a gold standard viral RT-PCR or other standard laboratory diagnostic tests. Among the 16,196 images in the CC-CXRI-P dataset, 4,436 were viral pneumonia, which included 1,571 COVID-19 pneumonia, 6,282 were other types of pneumonia, and 5,478 were absence of pneumonia. To train our AI model to be generalizable across different populations and new settings, we purposely included CXRs with “silver-standard labels” from the CheXpert for training. The CheXpert dataset is an open-source retrospective patient cohort containing mixtures of different types of pneumonia and other lung disorders. Our radiologists manually re-graded 13,148 CXR images with a label of “pneumonia” and classified them into 2,840 viral pneumonia, 5,309 other types of pneumonia, and 4,999 absence of pneumonia. This re-annotated pneumonia dataset is named CheXpert-P.

For the model training, we initially trained the AI system with the “golden-standard labels” on the subset of 13,158 images from CC-CXRI, and then tested it on an independent test set with 1,519 CXR images from the CC-CXRI. The CXR images in CC-CXRI were all confirmed with definitive “gold-standard labels” using PCR-based or other standard laboratory diagnostic tests. The three-way classification results showed an overall performance of an AUC with 0.963 (95% CI: 0.955-0.969) (Supplementary Fig. 10a). Next, we added the CheXpert-P dataset with “silver-standard labels” into the training set of CC-

CXRI. Again, we re-trained the AI model and tested it on the same test set from the CC-CXRI. The results showed better performance with AUC of 0.977 (95% CI: 0.971-0.982) for the three-way classification. Thus, we conclude that including the weak labels or “silver-standard labels” for training can potentially lead to improved classification performance. The improvement is due to the AI model being exposed to different types of images. As a result, the AI system differentiated viral pneumonia from the other two groups with 92.94% sensitivity, 87.04% specificity, and an AUC of 0.968 (95% CI: 0.957-0.978) (Fig. 3a and 3b).

To quantify the standardization module’s impact on the diagnostic performance, we evaluated the AI system on the test set with the whole module or part of the module skipped. The AI system performed poorly without the image registration, lesion segmentation, or both (Supplementary Fig. 10b). The whole pipeline demonstrated a statistically significant improvement in absolute specificity from 76.2% to 91.1% (permutation test, $P < 0.001$ for superiority) compared to the baseline model (Supplementary Table 8). The results showed that with a specific decision threshold, the whole pipeline achieved a significantly higher specificity while retaining a sensitivity of 90%. This demonstrated the importance of every component of the pipeline to screen patients for suspicious pneumonia.

External validation in multi-country datasets

To test the AI system’s generalizability to various clinical settings, we conducted four external validations. The first test was performed on a prospective pilot study in a non-epidemic area of China with 1,899 CXR images containing 240 viral pneumonia (including 98 COVID-19 pneumonia), 610 other types of pneumonia, and 1,049 absence of pneumonia (normal). The AI system achieved an average of an AUC of 0.941 (95% CI: 0.931-0.952) in the three-way classification. For differentiating viral pneumonia from other types of pneumonia and normal, it achieved 90.00% sensitivity, 87.40% specificity, and AUC of 0.947 (95% CI: 0.931-0.962) (Fig. 3c and 3d).

The second external validation was performed on another Chinese population screening cohort that included participants in a routine clinical care setting for suspected pneumonia. The external test set contains a total of 1,034 CXR images, including 46 viral pneumonia, 220 other types of pneumonia, and 768 normal (Table 2). The AI model achieved an AUC of 0.938 (95% CI: 0.922-0.955) in the three-way classification, and 89.13% sensitivity, 93.02% specificity, and an AUC of 0.969 (95% CI: 0.943-0.987) (Fig. 3e and 3f), for differentiating viral pneumonia from the other two groups.

The third external validation was performed on an international patient cohort from Ecuador and other open public data sources comprising a total of 650 CXR images (Table 2). Our AI system achieved 0.934 (95% CI: 0.917-0.950) of an AUC for the three-way classification, and an AUC of 0.920 (95% CI: 0.891-0.942) for differentiating viral pneumonia from the other two groups (Supplementary Fig. 2a).

The fourth external validation was performed on an open public Kaggle-pneumonia dataset (https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia). Our AI model achieved an AUC of 0.948 (95% CI: 0.943-0.953) for the three-way classification, and an

AUC of 0.916 (95% CI: 0.907-0.924) for viral pneumonia detection (Supplementary Fig. 2b). Overall, these results firmly demonstrated a high-level consistency of our AI system's performance and proved its generalizability.

Potential for triaging of patients with COVID-19

We attempted to use the AI system to identify COVID-19 pneumonia. A total of 17,883 CXR images, including 1,407 COVID-19 and 5,515 other viral pneumonia, and 10,961 other pneumonia from CC-CXRI, were used to train and validate the AI model (Table 2).

We first evaluated the model on a test set with 164 COVID-19 and 630 other pneumonia, and obtained an AUC of 0.966 (95% CI: 0.955-0.975), 92.07% sensitivity, and 90.12% specificity (Fig. 4a and 4b). A separate, independent dataset containing 164 COVID-19 pneumonia and 190 other types of viral pneumonia was also used to test the model. The results showed an AUC of 0.867 (95% CI: 0.828-0.902), a sensitivity of 82.32%, and a specificity of 72.63% (Fig. 4d and 4e). Both results confirmed that the AI system is sensitive to subtle lesion information CXRs in triaging COVID-19 pneumonia and differentiating it from other pneumonia with reasonable accuracy as a first-line diagnostic tool. We conducted additional experiments to differentiate different subgroups of the COVID-19, severe and non-severe COVID-19 from other types of viral pneumonia. The results showed that the detection of the non-severe COVID-19 had relatively inferior performance than that of the severe COVID-19 (Fig. 4c and 4f).

We then tested the AI system on the public BIMCV dataset from the Valencia region of Spain[31], which includes 663 COVID-19 from BIMCV-COVID19 and 1,277 normal from BIMCV-COVID19-PADCHEST. The results showed an AUC of 0.916 (95% CI: 0.904-0.933) to identify COVID-19 as viral pneumonia and differentiate it from normal. The AI without the image registration and lesion segmentation (the baseline model) obtained inferior performance with an AUC of 0.856 (95% CI: 0.838-0.876) (Supplementary Fig. 3).

Assessing the clinical severity of COVID-19

We next investigated the feasibility of assessing the severity level of COVID-19 pneumonia based on our AI analytic module. We hypothesized that the lung severity could be systematically scored by quantifying a CXR image, which we called the severity index based on the lung-lesion segmentation. Fig. 5e presents an example of viral pneumonia with comparable lung-lesion segmentation by the AI model and human radiologists. Compared to human experts, the AI model produced smoother and clearer lesion segmentation boundaries with higher accuracy. This showed that our AI system could be used as a visualization/ reference tool to highlight the lesion areas for radiologists.

The CXR severity index was determined as follows. Each CXR image was divided into 12 sections defined horizontally by four anatomical categories (lung field, and periphery of the lung field) and vertically by the vertebral column (Supplementary Fig. 4a). Each section was assigned an opacity score from 0 to 4 by a group of trained radiologists, based on the section's percentage with lung lesions. The 1,207 CXR images of the COVID-19 patients were also graded manually with the CXR severity index by radiologists. We evaluated the association between the severity scores by radiologists and by the AI model based on the quantification of the CXR images. The severity index graded by human radiologists and the AI reviewer showed a strong linear relationship, with a Pearson correlation coefficient (PCC) of 0.81 and a mean absolute error (MAE) of 8.64 (Fig. 5a). Bland-Altman plot showed a good agreement between the AI model and the human reviewers with an intraclass correlation coefficient (ICC) of 0.68 (95% CI: 0.60-0.74), while the agreement between radiologists' evaluation achieved an ICC of 0.73 (95% CI: 0.64-0.81) (Fig. 5b).

We further hypothesized that the severity index used in a chest radiograph is correlated with the severity of clinical outcomes. The severe level of a respiratory distress state in the clinical setting was defined by blood oxygen saturation < 92%, respiratory rate < 36, or PO2/FiO2 < 300 mmHg. It usually corresponded to diffuse interstitial pneumonia, which obscured normal lung markings[32]. A total of 1,207 CXRs were manually graded based on clinical diagnoses and classified into 437 severe and 770 non-severe labels. Then we used the severity index scores by the AI model and the radiologist reviewers as an input of a logistic regression model to generate clinical severity prediction (see more details in Methods). The results showed that our AI system could predict the COVID-19 pneumonia severity with an AUC of 0.868 (95% CI: 0.816-0.915), a specificity of 80.65%, and a sensitivity of 82.05% (Fig. 5c), whereas the human radiologists achieved a comparable AUC of 0.832 (95% CI: 0.782-0.885) with a specificity of 74.84% and a sensitivity of 79.49% (Fig. 5c and Supplementary Fig. 4). The results demonstrated that the analytic pipeline could also aid in predicting the severity of COVID-19 pneumonia.

The AI system versus radiologist performance study

An independent test set of 440 CXR images was used to compare the AI system's performance against practicing radiologists in classifying viral pneumonia, other types of pneumonia, and normal. A total of eight radiologists with different levels of clinical experience were enrolled to participate in this study: four junior radiologists with over 10 years of experience and four senior radiologists with over 20 years of experience. The ground truth was determined by positive molecular test results together with the CXR findings verified by another independent group of 3 senior radiologists.

The performance was evaluated by the AUC and the sensitivity/specificity (Fig. 6a and Supplementary Table 6). The AI system achieved comparable performance to the senior radiologists' level with an AUC of 0.981 (95% CI: 0.970-0.990) for the viral pneumonia diagnosis. The operating point, selected from the validation dataset, generated better sensitivity (P < 0.001) and comparable specificity than the average junior radiologists (Supplementary Table 6).

One of our AI system's objectives is to investigate whether it could assist junior radiologists in improving their diagnostic performance. In this experiment, four junior radiologists performed their initial diagnosis, and two weeks later, they were given the diagnosis probability provided by our AI system and asked to repeat the image grading without providing any other prior information. Weighted error, which was calculated based on a penalty score system, was employed as a metric to evaluate and compare the performance of our AI system and the practicing radiologists. The junior radiologists' performance with the

AI assistance yielded an average weighted error of 9.82%, a significant improvement (P < 0.001) compared to that of 27.44% without the AI assistance (Fig. 6b).

We also explored the AI system's potential role in enhancing diagnostic performance by radiologists in the workflow. In this simulated scenario, a specific diagnosis was made by two radiologist readers (see more details in Methods). When there was a disagreement, an "arbitrator" was involved in reaching a decision. The average weighted error was 20.11% when taking a consensus diagnostic decision by the radiologist group. In comparison, when the AI system acted as an "arbitrator", the error was reduced to 16.65%, and when the AI system acted as a second reader, it was further reduced to 7.08% (Fig. 6c). These results demonstrated that the AI system could improve radiologists' performance and reduce imaging reading workload. The details of the ROC curves and confusion matrices of the eight radiologists' performance were given in the Supplementary Fig. 5.

Discussion

This study showed a few crucial points. First, despite the limitation of a plain CXR image, an accurate AI system can assist radiologists in identifying viral pneumonia and COVID-19 accurately, showing that it can be used as a frontline tool in an emergency clinic, remote places, or the developing world. A noteworthy feature of the AI system is that the modular processing pipeline, including anatomical landmark detection, registration, lung-lesion segmentation, and diagnosis prediction, provided robust and explainable results. Second, this AI system can help junior radiologists to perform close to the level of senior radiologists. Finally, this system can differentiate COVID-19 from other types of viral pneumonia with reasonable accuracy. The AI system can also accurately determine the severity of the lesions in patients with established COVID-19. Overall, this diagnostic tool can assist radiologists in managing COVID-19 cases.

A rapid diagnosis of viral pneumonia with high suspicion of COVID-19 is an important first step for clinical management. A positive result should trigger a molecular viral test for SARS-CoV-2, sending the patient to an infectious disease unit with isolation. If confirmed, contact tracing should be initiated quickly. The patient may then receive CT imaging with an AI-based system or CT analysis that is accurate in providing a more detailed description of lesion pathologies[33]. However, the chest CT scan is not a front-line tool, as it takes more time to conduct, is more expensive, and is not readily available in remote places, thereby limiting its application in the general population. In contrast, CXR is a front-line tool with a quick turn-around time and could be used more conveniently in an intensive care setting.

The optimal use of the AI system to improve the clinical workflow remains to be explored. Pneumonia fundamentally is a clinical diagnosis, and in suspected COVID-19, RT-PCR is the reference gold standard. However, due to high rates of false-negative test results for SARS-CoV-2 PCR testing by nasal swab sampling, imaging findings may also be used to make a presumptive diagnosis. Previous studies indicated that CXR images contained specific differences in imaging findings between viral pneumonia and bacterial pneumonia. These differential or subtle features can be detected by the AI system, yet are beyond clinicians' observational ability and comprehension. The specificity advantage exhibited by the AI system suggests that it could help to reduce the false-negative rate of PCR testing. Taken together, CXR has been considered as part of the diagnostic workup of symptomatic subjects with suspected COVID-19 in settings where laboratory testing (RT-PCR) is not available or results are delayed or initially negative in the presence of apparent symptoms attributable to COVID-19[28]. Such workflow could help healthcare/hospital administrators plan and make an informed decision on resource allocation during an epidemic/pandemic.

Although there are published studies that used AI in diagnosing pneumonia, the actual clinical applicability remains unknown since they have not been shown to be free of experimental data bias, and they have not been tested by the peer-reviewed gold standard labels and by external data in different populations and new clinical settings to show generalizability. In this paper, we explored the general applicability of the current AI system. We first trained our AI system using large, heterogenous, multi-center datasets. Then we present evidence of the ability of the AI system to translate between different populations and settings. In particular, we trained a model to detect common thoracic diseases on patients coming for hospital visits (SYSU set), and then measured performance on populations coming for physical examination (SYSU-PE set). Compared with the training set, the external validation set represented populations with less chest pathology. In this context, the system continued to achieve accurate performance. This practice is rare in the current literature.

Notably, the AI system can also assist in the assessment of patient severity. This is particularly important in the intensive care setting or when resources are stretched, as CXR imaging is much easier to perform than a chest CT scan. As a monitoring tool, it will assist the intensive care physicians in assessing patients more comprehensively. In addition, the CXR severity index that is automatically scored by the AI model can be used to assess patients’ risk level of complications and mortality, leading to earlier detection, intervention, and treatment of high-risk patients with COVID-19.

Despite these potential advantages, it is critical to emphasize that this AI system is an assistant to radiologists for diagnosis. A comprehensive analysis of all other clinical and laboratory information is necessary for an accurate diagnosis. Our demonstration that this AI system improved the junior radiologists’ performance proved the benefit of integrating it into radiologists’ current workflow. This integration can be crucial during a pandemic, like the current COVID-19 situation when resources are stretched thinly. Our AI system’s ability to recognize features in the diffuse pattern of lung involvement, which is relatively common among viral diagnostics but difficult to discern by radiologists, may represent an advantage offered by the AI system.

Our study has several limitations, which we hope to address in the future. First, since the AI system was trained in a population where more than 90% are symptomatic patients with abnormal imaging findings, its ability for diagnosing very early COVID-19 cases will need to be validated. Although our AI system achieved good performance with an AUC of 0.901 when evaluated on patients with no apparent findings versus normal x-ray images (using the test set of CC-CXRI), further training with more non-evident COVID-19 cases is necessary to establish its clinical utility in a broad range of populations. Another limitation is its ability to differentiate COVDI-19 from non-focal (diffuse) acute respiratory distress syndrome (ARDS). However, ARDS is a crucial acute condition with associated pulmonary edema; therefore, by additional clinical findings or laboratory testing, it can be differentiated from severe COVID-19.

Finally, this study demonstrated an AI system's value in assisting medical professionals for rapid and accurate diagnoses of pneumonia in a pandemic. Future refinement and improvement will expand its use into diagnostic assessments of other common and routine lung disorders such as tuberculosis and malignancies.

Methods

Images from human subjects

CXR images were extracted from the China Consortium of Chest X-ray Image Investigation (CC-CXRI) data, which were collected from multiple hospitals, including Sun Yat-sen Memorial Hospital and the Third Affiliated Hospital, both affiliated with Sun Yat-sen University, West China Hospital, Guangzhou Medical University First Affiliated Hospital, Nanjing People's Hospital, the First affiliated hospital of Anhui Medical University, and Yichang Central People's Hospital. All CXRs were collected as part of the patients' routine clinical care. For the analysis of CXR images, all radiographs were first de-identified to remove any patient-related information. The CC-CXRI images consisted of both anterior-posterior view and posterior-anterior view of CXR images. There are two sets of data in CC-CXRI: a large-scale dataset for common thoracic disease detection from the Sun Yat-sen University Hospital System (the SYSU set), and a pneumonia assessment survey (CC-CXRI-P). COVID-19 diagnosis was given when a patient had pneumonia with a confirmed viral reverse-transcriptase-PCR test. The other types of pneumonia were diagnosed based on standard clinical, radiological, or culture/molecular assay results (Supplementary Table 9). Institutional Review Board (IRB)/Ethics Committee approvals were obtained from the Sun Yat-sen University Memorial Hospital, West China Hospital, and all patients signed a consent form. The work was conducted in a manner compliant with the United States Health Insurance Portability and Accountability Act (HIPAA). It was adherent to the tenets of the Declaration of Helsinki and in compliance with the Chinese CDC policy on reportable infectious diseases and the Chinese Health and Quarantine Law.

CXR dataset construction of common thoracic diseases

We constructed CXR datasets for the development and evaluation of the AI model for common thoracic diseases. We used an NLP pipeline to extract disease labels from clinical reports for CXR images. The pipeline included disease concept detection and negation classification, similar to CheXpert[34] and NIH Chest X-ray dataset[35] (please see more details in Supplementary Methods).

We selected fourteen common thoracic diseases according to their clinical significance and prevalence, as defined based on the ICD-10 and the NIH Chest X-ray dataset[35]. They were extracted from real-world clinical reports for corresponding CXR images, and each label comes with both the localization of the critical finding and the classification of common thoracic diseases that can be revealed by the CXR image. These disease labels included atelectasis, cardiomegaly, consolidation, edema, effusion, emphysema, fibrosis, hernia, infiltration, nodule, mass, pleural thickening, pneumonia, and pneumothorax. We also defined another label of “No finding” that is positive if and only if all other labels of a CXR image are negative. Thus, each CXR image in the dataset was annotated by the presence or absence of the fifteen labels.

Two datasets were constructed. The SYSU dataset is composed of 120,702 CXR images from 92,327 patients between October 2018 and July 2020 in both inpatient and outpatient centers. The SYSU-PE dataset is comprised of 24,500 CXR images from 23,585 patients coming for the health check. The SYSU dataset is used for model development and internal validation, and the SYSU-PE dataset is used for external validation. The labels of the validation data were manually reviewed for a reliable evaluation.

Silver-standard labels of pneumonia

Previous works suggested specific differences in CXR imaging findings between viral pneumonia and bacterial pneumonia. Thus, imaging has been considered part of the diagnostic workup of symptomatic subjects with suspected COVID-19 in settings where the laboratory testing (RT-PCR) is not available or results are delayed or initially negative in the presence of symptoms attributable to COVID-19[28].

In this study, we manually curated CheXpert to expand the dataset for training. The CheXpert dataset is a public dataset containing 224,316 CXR images from 65,240 patients. Each image was labeled with the presence or absence of each of 14 common chest radiographic observations. The original CXR images were only given a general diagnosis of pneumonia without a detailed label of viral pneumonia or other types of pneumonia. Here, we considered the manually graded image as a silver standard, contrasting with the ground truth gold-standard labels discussed below. A total of 15 radiologists with over 10 years of clinical experience manually reviewed and graded a subset of CheXpert with pneumonia labels. They labeled them with viral pneumonia, other types of pneumonia (including bacterial pneumonia and mycoplasma pneumonia), and absence of pneumonia (normal). Next, 20% of their results in the dataset were checked and validated by a group of five independent senior radiologists, each with over 20 years of clinical experience. In case of inconsistency, the expert consensus was used to correct labels. A total of 13,148 CXR images from CheXpert were re-labeled into three categories: 2,840 viral pneumonia, 5,309 other types of pneumonia, and 4,999 normal CXRs. We named this re-annotated dataset as CheXpert-P and treated it as a “silver-standard label” dataset for training.

Gold-standard labels and ground truth of pneumonia

All CXR images from the CC-CXRI dataset had definitive diagnosis determined by the gold standard PCR-based / standard laboratory diagnosis; each CXR image was given a specific and definitive diagnosis of COVID-19 pneumonia, other viral, or bacterial pneumonia. The above laboratory test results serve as ground-truth for the data used for validation. More specifically, the CC-CXRI dataset consists of 4,436 viral pneumonia (including 1,571 COVID-19 pneumonia), 6,282 other types of pneumonia, and 5,478 normal CXRs. We used the CC-CXRI for model development and testing. Specifically, patients were randomly assigned for training (80%), validation (10%), or testing (10%) (Table 1).

Quality control of image labels of CXR

For all CXRs for validation/testing, each image went through a tiered grading system consisting of two layers of trained graders of increasing expertise for verification and correction of image labels. Each image imported into the database started with a label matching the diagnosis of the patient. This first tier of graders conducted initial quality control of the image labels to exclude unreadable images, including those missing the whole bilateral lungs or with metal artifacts. The second tier of five senior independent radiologists read and verified the true labels for each image. In case of disagreement, an expert of consensus was used to correct the labels. The resulting labels serve as the ground truth for the evaluation dataset.

Annotation of landmarks and lung-lesion segmentation on CXR

We used 676 manually annotated CXR images from viral pneumonia, other pneumonia, and normal subjects for training the anatomical landmark determination. Twelve anatomical landmarks were labeled on each CXR image: midpoint of clavicle left (MCL) and right (MCR), sternal end of clavicle left (SECL) and right (SECR), hilar angle left (HAL) and right (HAR), costophrenic angle left (CAL) and right (CAR), diaphragmatic dome left (DDL) and right (DDR), cardiac diaphragmatic angle left (CDAL) and right (CDAR).

We manually segmented 1,016 CXR images at the pixel level to train and evaluate our semantic segmentation model. Among these CXR images, 228 were from patients with viral pneumonia (including 121 COVID-19 pneumonia patients), 1,163 from patients with bacterial pneumonia, 187 from patients with other types of pneumonia, and 438 from normal subjects. The annotation was done via polygons. The lung segmentation labels included lung field (left), the periphery of the lung field (left), lung field (right), the periphery of the lung field (right). The lesion segmentation labels consisted of two classes: opacification and interstitial pattern, which were relevant pneumonia lesion features. The segmentations were annotated and reviewed by five senior radiologists. A five-fold cross-validation test was applied for the landmark detection and lung-lesion segmentation.

Performance comparisons with radiologists

To evaluate the performance in classifying the three types of pneumonia, we constructed an independent validation set of 440 CXR images, including 160 viral pneumonia, 160 other types of pneumonia, and 120 normal cases. We used this set to compare the performance of our AI system and the diagnosis of the radiologists. A weighted error scoring was employed to consider that a false negative result (failing to refer to a viral pneumonia case) is more detrimental than a false positive result (making a referral when it was not warranted). Predicted errors based on a weighted penalty table were used to compute a metric to evaluate and compare performance between the AI system and the radiologists. We weighted the misidentification of a “viral pneumonia” as “normal” with an error score of 2, which is larger than the score of 1 for the misidentification of the other two groups (Supplementary

Fig. 5f). This is because if a patient with COVID-19 or other viral pneumonia is mis-triaged to normal, this may cause the spread of the disease.

We conducted a simulation study in which the AI system was deployed firstly as a "second reader" and secondly an "arbitrator" of radiologists' diagnostic decisions. As for the role of a second reader, we used a junior radiologist as the first reader and the AI system as the second reader. Whenever there existed a disagreement, the opinion of the senior reader was introduced. We also simulated the scenario in which the AI system acted as an "arbitrator" by using human radiologists as the first and second readers and the opinion of the AI as a final reader. The weighted error was also calculated. The performance of the AI system was compared with that of the radiologists based on AUC curves, sensitivity, and specificity. The operating point of the AI system was chosen based on the separate validation set. For the statistical significance of the comparison results, we computed confidence intervals and *P*-values using 1,000 random re-samplings (bootstraps).

Transfer learning and deep learning

We trained our AI model using a large number of CXR images from three public datasets, CheXpert dataset[36], MIMIC-CXR dataset[37], and NIH Chest X-ray dataset[14].

Transfer learning was adopted by pretraining a DenseNet-121 model[27] for the CXR image classification. The DenseNet-121 architecture has proven to be effective for CXR classification tasks[36]. The convolutional layers were fine-tuned when transferring to other tasks, while the fully connected layer was trained from scratch. The number of the outputs was also modified in the last fully connected layer to adapt to the appropriate classification task. The softmax operation was used for the classification tasks. For data augmentation, each CXR image is transformed through geometric transformations (such as scale and translation) and changes in contrast and saturation. Four DenseNet-121 models were trained separately to classify common chest diseases, identify pneumonia conditions, differentiate viral pneumonia from other types of pneumonia, triage COVID-19 from other types of viral pneumonia, and predict the severity level of COVID-19 patients. The input CXR images were resized to 512×512 by bilinear interpolation.

We employed the cross-entropy loss function and adopted an Adam optimizer[38] for training, with a learning rate of 0.003, and the batch size was set to 32. All deep learning models were implemented with Pytorch 1.4[39]. A validation set was used for early-stopping with a patience of 10 to avoid overfitting. The model with the best validation loss was finally selected. All training, validation, and testing procedures were conducted on NVIDIA GeForce 1080Ti graphical processing units.

Overview of the AI system

Our proposed AI system applied a modular pipeline approach, which consisted of three main components: a CXR standardization module, a common thoracic disease detection module, and a pneumonia analysis module. A detailed description of the AI system is available in the Supplementary Methods.

The CXR standardization module performs invert-grayscale CXR40 detection, anatomical landmarks detection, and CXR image registration in this study. We first trained an invert-grayscale CXR detection model to detect whether the input of CXR was inverted-grayscale, and if so, the system automatedly converted it into a conventional CXR. After that, the anatomical landmark detection model performs the detection of landmarks of the CXR. Based on detected landmarks, we generated a registered CXR image via the image registration algorithm. These components were specially designed to address the common problems encountered in computer-aided detection with CXR: variations/inconsistency of the radiographs due to orientation, distance, and the difference in imaging pathology area, breathing movement, and spatial alignment. These registered CXR images were used as the input to the model for disease classification or severity prediction.

The common thoracic disease detection module was developed for chest disease detection. As some pathologies may co-exist or overlap on the same CXR image, we employed a multi-label classification approach that could predict multiple categories at the same time and thus is more suitable for clinical settings where combinations or simultaneous occurrences of the categories often exist. Using the standardized CXR images, we trained a multi-label classification model with fifteen binary classifications, including fourteen disease labels and one no-finding label. The number of the output scalars was fifteen with a sigmoid activate function. For the scenario of opacity detection, the case is defined positive if at least one label of atelectasis, mass, edema, pneumonia, and consolidation is present. The predicted probability of opacity was composited by averaging over outputs of atelectasis, mass, edema, pneumonia, and consolidation.

The pneumonia analysis module is a two-stage architecture for pneumonia for identifying the subtype of pneumonia, predicting the presence and absence of COVID-19, and assessing the severity of COVID-19. In the first stage, a lung-lesion segmentation module identifies suspicious regions in the segmented lung region. The networks were trained sequentially: the lung segmentation network was trained using the registered CXR images as inputs, and then extracted lung regions were used to train the lesion segmentation network. Since the raw CXR images may contain irrelevant information for lesion segmentation (e.g. body parts not related to the lungs), a lung segmentation network was trained to discard such information so that the lesion segmentation network can concentrate on the lung area. Next, the diagnostic model, a neural network classifier, made a prediction based on the previous models' outputs, namely the anatomical landmark detection model, the lung segmentation model.

To design the classifier, we conducted experiments to compare the multi-channel model with the single-channel model. The results showed that the multi-channel model had better performance (Supplementary Fig. 11).

Classification of severity levels

We investigated how to score the CXR images to quantify the severity of lung opacity and then investigated whether this CXR severity score is associated with the clinical severity of COVID-19 patients. The clinical severity level is a clinical diagnosis of a respiratory distress state: blood oxygen saturation < 92%; respiratory rate < 36, or PO2/FiO2 < 300 mmHg. It usually corresponds to diffuse interstitial pneumonia, which obscures normal lung markings[32]. For the analysis of the severity of COVID-19 pneumonia patients, a total of 1,207 CXR images were manually graded, resulting in 437 images with “severe” labels and 770 images with “non-severe” labels.

We first calculated the CXR severity index by dividing CXR images into 12 sections. The 12 sections were defined horizontally by the four anatomical parts (including the lung fields and periphery of the lung fields) and vertically by the vertebral column (Supplementary Fig. 4). Each section was assigned a severity index from 0 to 4 to quantify the extent of opacity by radiologists (corresponding to <1, 1-25, 25-50, 50-75, 75-100, respectively), whereas the AI system automatedly segmented the lung-lesion and quantified the CXR image’s severity. Each CXR image of 1,027 COVID-19 patients was given a severity score by a group of radiologists based on the above definition. To evaluate the association between the AI model and radiologists for scoring the CXR severity, we calculated Mean Absolute Error (MAE) and Pearson Correlation Coefficient (PCC). Bland-Altman plot[41] and Intraclass Correlation Coefficient (ICC) were also used for assessing agreement between the AI reviewer and radiologists. We further associated the CXR severity index with clinical outcomes. Instead of directly using the final CXR index, we predicted the clinical severity by using all 12 sections’ scores as input features and adopted the logistic regression as the classification model. An ROC curve and a confusion table were then generated.

Operating point selection

An AI system for pneumonia diagnosis was proposed to produce a probability score for each class. For different clinical applications, the operating point can be set differently to compromise between the true positive rate (TPR) and the false-positive rate (FPR) (Supplementary Table 2).

Statistical analysis

An ROC analysis and AUC were employed to assess model performance for each classification task. For multi-class tasks, the macro-average of ROC and AUC were used as the metrics for each class. The ROC curves were plotted by using the true positive rate (sensitivity) versus the false positive rate (1 – specificity) under different decision thresholds. For a given ROC curve *TPR=f(FPR)*, where *FPR∈[0,1]*, the AUC is defined as: $AUC = \int_0^1 f(x)dx$. Normalized confusion matrices were used to illustrate the classification results. To evaluate the models’ and experts’ performance, the weighted error was calculated by weighting the error of the i-th class being predicted by the j-th class by a defined weight matrix. We evaluated the landmark detection’s performance on our annotated dataset using two evaluation metrics, normalized distance error and successful detection rate. Furthermore, we evaluated landmark detection performance on the external dataset SCR using two additional metrics, pixel distance error and physical distance error. The normalized distance error is defined as the distance between the predicted normalized coordinates and the normalized true coordinates, where original coordinates are normalized with x and y divided by the width and the height of the image, respectively. The successful detection rate (SRD) is defined as the number of accurate detections versus the total number of detections, where an accurate detection is a prediction with a margin error less than or equal to a specified threshold. Physical distance errors were reported when the pixel size was known (e.g. 0.175mm pixel size on the SCR dataset). We evaluated the segmentation model's performance with two evaluation metrics, including Intersection over Union (IOU) and Dice Coefficient (DC). The IOU is the area of the overlap between the predicted segmentation and the ground truth divided by the area of the union. The DC is twice the area of the overlap between the predicted segmentation and the ground truth divided by the sum of the areas of the predicted segmentation and the ground truth.

A bootstrapping strategy (1,000 random re-sampling) was adopted to analyze the confidence intervals (CI) of AUC. The empirical distribution of the test dataset was used to approximate the data distribution and draw n samples from the empirical distribution (n is the size of the test dataset) to calculate an AUC. Repeating such an operation yields the sampling distribution of AUC, from which the CI of AUC was calculated. The shortest two-side 95% CIs of AUC were reported for each experiment. P values for sensitivity, specificity and weighted-error comparisons were generated through a two-sided permutation test of 10,000 random re-samplings. The ROC curves and confusion matrices were generated using the Python scikit-learn library and plotted with the Python matplotlib and seaborn libraries. The measures of sensitivity, specificity, and accuracy were calculated using the Python scikit-learn library.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Authors

Guangyu Wang[#*,1], Xiaohong Liu[#2], Jun Shen[#3], Chengdi Wang[#4], Zhihuan Li[#5], Linsen Ye[#6], Xingwang Wu[#7], Ting Chen[*,2], Kai Wang[2], Xuan Zhang[2], Zhongguo Zhou[8], Jian Yang[9], Ye Sang[9], Ruiyun Deng[10], Wenhua Liang[11], Tao Yu[3], Ming Gao[3], Jin Wang[5], Zehong Yang[3], Huimin Cai[10], Guangming Lu[12], Lingyan Zhang[13], Lei Yang[14], Wenqin Xu[4], Winston Wang[4], Andrea Olevera[5], Ian Ziyar[4], Charlotte Zhang[10], Oulan Li[10], Weihua Liao[15], Jun Liu[16], Wen Chen[17], Wei Chen[18], Jichan Shi[19], Lianghong Zheng[5], Longjiang Zhang[12], Zhihan Yan[19], Xiaoguang Zhou[20], Guiping Lin[3], Guiqun Cao[4], Laurance L. Lau[5], Long Mo[15], Yong Liang[5], Michael Roberts[21,22], Evis Sala[23], Carola-Bibiane Schönlieb[22], Manson Fok[5], Johnson Yiu-Nam Lau[24], Tao Xu[10], Jianxing He[11], Kang Zhang[*,5], Weimin Li[*,4], Tianxin Lin[*,3]

Affiliations

[1]School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China, 100876

[2]Department of Computer Science and Technology & BNRist, Tsinghua University, Beijing, China, 100084

[3]Department of Urology, Department of Radiology, Department of Emergency Medicine, Department of respiratory medicine, Sun Yat-Sen Memorial Hospital, Sun Yat-Sen University, Guangzhou, China, 510235

[4]Department of Respiratory and Critical Care Medicine, Frontiers Science Center for Disease-related Molecular Network, Center for Translational Medicine and Innovations, West China Hospital, West China Medical School, Sichuan University, Chengdu, China

[5]Center for Biomedicine and Innovations, Faculty of Medicine, Macau University of Science and Technology, Macau, China, 999078

[6]Department of Hepatic Surgery and Liver Transplantation Center, the Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, China, 510060

[7]Department of Radiology, The First Affiliated Hospital of Anhui Medical University, Hefei, China. 230022

[8]The Sun Yat-sen Cancer Center, Sun Yat-sen University, Guangzhou, China, 510060

[9]The First College of Clinical Medical Science, China Three Gorges University, Yichang, China, 443002

[10]Bioland Laboratory (Guangzhou Regenerative Medicine and Health Guangdong Laboratory), Guangzhou, China, 510220

[11]Departments of Thoracic Surgery/Oncology, the First Affiliated Hospital of Guangzhou Medical University; China State Key Laboratory and National Clinical Research Center for Respiratory Disease; Guangzhou, China, 510120

[12]Department of Medical Imaging, Jinling Hospital, Medical School of Nanjing University, Nanjing, China, 210093

[13]Department of Medical Imaging, the Third Affiliated Hospital, Southern Medical University, Guangzhou, China, 510630

[14]Department of Thoracic Surgery, the First Affiliated Hospital of Sun Yat-sen University, Guangzhou, China, 510080

[15]Department of Medical Imaging, Dept. of cardiology, Xiangya hospital, Centre-south university, and the National Engineering Laboratory of medical big data application technology, Changsha, China, 410008

[16]Department of Radiology, Second Xiangya Hospital, Central South University, Changsha, China, 430103

[17]Department of radiology, Taihe Hospital, Hubei University of Medicine, Hubei, China, 442000?

[18]Department of Radiology, The Second Affiliated Hospital and Yuying Children's Hospital of Wenzhou Medical University, Wenzhou, China, 325035

[19]Department of Pharmacy, the First People's Hospital of Kashgar Erea, Kashgar, Xiangjiang, China, 844000

[20]Departments of Infectious Disease, Wenzhou Central Hospital, Wenzhou, China, 325000

[21]Oncology R&D, AstraZeneca, Cambridge, UK, CB2 1TN7

[22]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK, CB3 0WA

[23]Department of Radiology and Cancer Research UK Cambridge Center, Cambridge, UK, CB2 0RE

[24]Department of Applied Biology and Chemical Technology, Hong Kong Polytechnic University, Hong Kong, China, 810085

Acknowledgements

This study was funded by the National Key Research and Development Program of China (2019YFB1404804), National Natural Science Foundation of China (grants 61906105, 61872218, 61721003 and 61673241), Innovation and Entrepreneurship Team Program of Guangzhou Regenerative Medicine and Health Guangdong Laboratory (2018GZR0301001), a Macau FDCT grant (0035/2020/A), Recruitment Program of Leading Talents of Guangdong Province (2016LJ06Y375), The Mark Foundation for Cancer Research and Cancer Research UK Cambridge Centre [C9685/A25177]. We thank members of Zhang, Lin and Wang groups for their assistance and helpful discussions and many volunteers and radiologists for curating and grading CXR and medical records.

Data availability

The main data supporting the results in this study are available within the paper and its Supplementary Information. De-identified and anonymised data generated during this study, including source data and the data used to make the figures, were deposited at the China National Center for Bioinformation at the Big Bay Branch at http://miracle.grmh-gdl.cn/chest_xray_ai.

Code availability

The custom codes for the diagnosis and discrimination of viral, non-viral and COVID-19 pneumonia from chest X-ray images is available at the China National Center for Bioinformation at Big Bay Branch at http://miracle.grmh-gdl.cn/chest_xray_ai. The codes are available for download for non-commercial uses.

References

1. Zhou P, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020:1–4.
2. Cohen J. Wuhan seafood market may not be source of novel virus spreading globally. Science. 2020
3. Chan JF, et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. Lancet. 2020
4. Huang C, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet. 2020
5. Qin C, Yao D, Shi Y, Song Z. Computer-aided detection in chest radiography based on artificial intelligence: a survey. Biomedical engineering online. 2018; 17:113. [PubMed: 30134902]

6. Jaiswal AK, et al. Identifying pneumonia in chest X-rays: A deep learning approach. Measurement. 2019; 145:511–518.
7. Pham HH, Le TT, Tran DQ, Ngo DT, Nguyen HQ. Interpreting chest X-rays via CNNs that exploit disease dependencies and uncertainty labels. arXiv preprint. 2019
8. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nature medicine. 2019; 25:44–56.
9. Esteva A, et al. A guide to deep learning in healthcare. Nature medicine. 2019; 25:24–29.
10. Norgeot B, Glicksberg BS, Butte AJ. A call for deep-learning healthcare. Nature medicine. 2019; 25:14–15.
11. Ravizza S, et al. Predicting the early risk of chronic kidney disease in patients with diabetes using real-world data. Nature medicine. 2019; 25:57–59.
12. Lambin P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. European journal of cancer. 2012; 48:441–446. [PubMed: 22257792]
13. Kermany DS, et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell. 2018; 172:1122–1131. [PubMed: 29474911]
14. Wang, X; , et al. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. 2097–2106.
15. Guan Q, et al. Diagnose like a Radiologist: Attention Guided Convolutional Neural Network for Thorax Disease Classification. 2018
16. Franquet T. Imaging of pneumonia: trends and algorithms. The European respiratory journal. 2001; 18:196–208. [PubMed: 11510793]
17. Yao L, Prosky J, Poblenz E, Covington B, Lyman K. Weakly Supervised Medical Diagnosis and Localization from Multiple Resolutions. arXiv: Computer Vision and Pattern Recognition. 2018
18. Ghesu, FC, , et al. An artificial agent for anatomical landmark detection in medical imagesInternational conference on medical image computing and computer-assisted intervention. Springer; 2016. 229–237.
19. Zhang, Z, Luo, P, Loy, CC, Tang, X. Facial landmark detection by deep multi-task learningEuropean conference on computer vision. Springer; 2014. 94–108.
20. Sun, Y; Wang, X; Tang, X. Deep convolutional network cascade for facial point detection. Proceedings of the IEEE conference on computer vision and pattern recognition; 2013. 3476–3483.
21. Yang, S; Luo, P; Loy, C; Tang, X. From Facial Parts Responses to Face Detection: A Deep Learning Approach. 2015 IEEE International Conference on Computer Vision (ICCV); 2015. 3676–3684.
22. Ronneberger, O, Fischer, P, Brox, T. U-net: Convolutional networks for biomedical image segmentationInternational Conference on Medical image computing and computer-assisted intervention. Springer; 2015. 234–241.
23. Long, J; Shelhamer, E; Darrell, T. Fully convolutional networks for semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition; 2015. 3431–3440.
24. Chen L-C, Papandreou G, Schroff F, Adam H. Rethinking atrous convolution for semantic image segmentation. arXiv preprint. 2017
25. Van Ginneken B, Stegmann MB, Loog M. Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database. Medical image analysis. 2006; 10:19–40. [PubMed: 15919232]
26. Shiraishi J, et al. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists’ detection of pulmonary nodules. AJR American journal of roentgenology. 2000; 174:71–74. [PubMed: 10628457]
27. Huang, G; Liu, Z; Van Der Maaten, L; Weinberger, KQ. Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. 4700–4708.

28. Akl EA, Blazic I. Use of Chest Imaging in the Diagnosis and Management of COVID-19: A WHO Rapid Advice Guide. 2020

29. Titano JJ, et al. Automated deep-neural-network surveillance of cranial images for acute neurologic events. 2018; 24:1337–1341.

30. Willemink MJ, Koszek WA. Preparing Medical Imaging Data for Machine Learning. 2020; 295:4–15.

31. Vayá, MdlI; , et al. BIMCV COVID-19+: a large annotated dataset of RX and CT images from COVID-19 patients. 2020

32. Guan WJ, et al. Clinical Characteristics of Coronavirus Disease 2019 in China. The New England journal of medicine. 2020

33. Zhang K, et al. Clinically Applicable AI System for Accurate Diagnosis, Quantitative Measurements and Prognosis of COVID-19 Pneumonia Using Computed Tomography. Cell. 2020

34. Irvin J, et al. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. 2019

35. Wang X, et al. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. arXiv. 2017

36. Irvin, J; , et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. Proceedings of the AAAI Conference on Artificial Intelligence; 2019. 590–597.

37. Johnson AEW, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data. 2019; 6:317. [PubMed: 31831740]

38. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint. 2014

39. Paszke A, et al. PyTorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems. 2019:8024–8035.

40. Musalar E, et al. Conventional vs invert-grayscale X-ray for diagnosis of pneumothorax in the emergency setting. The American journal of emergency medicine. 2017; 35:1217–1221. [PubMed: 28343817]

41. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986; 1:307–310. [PubMed: 2868172]

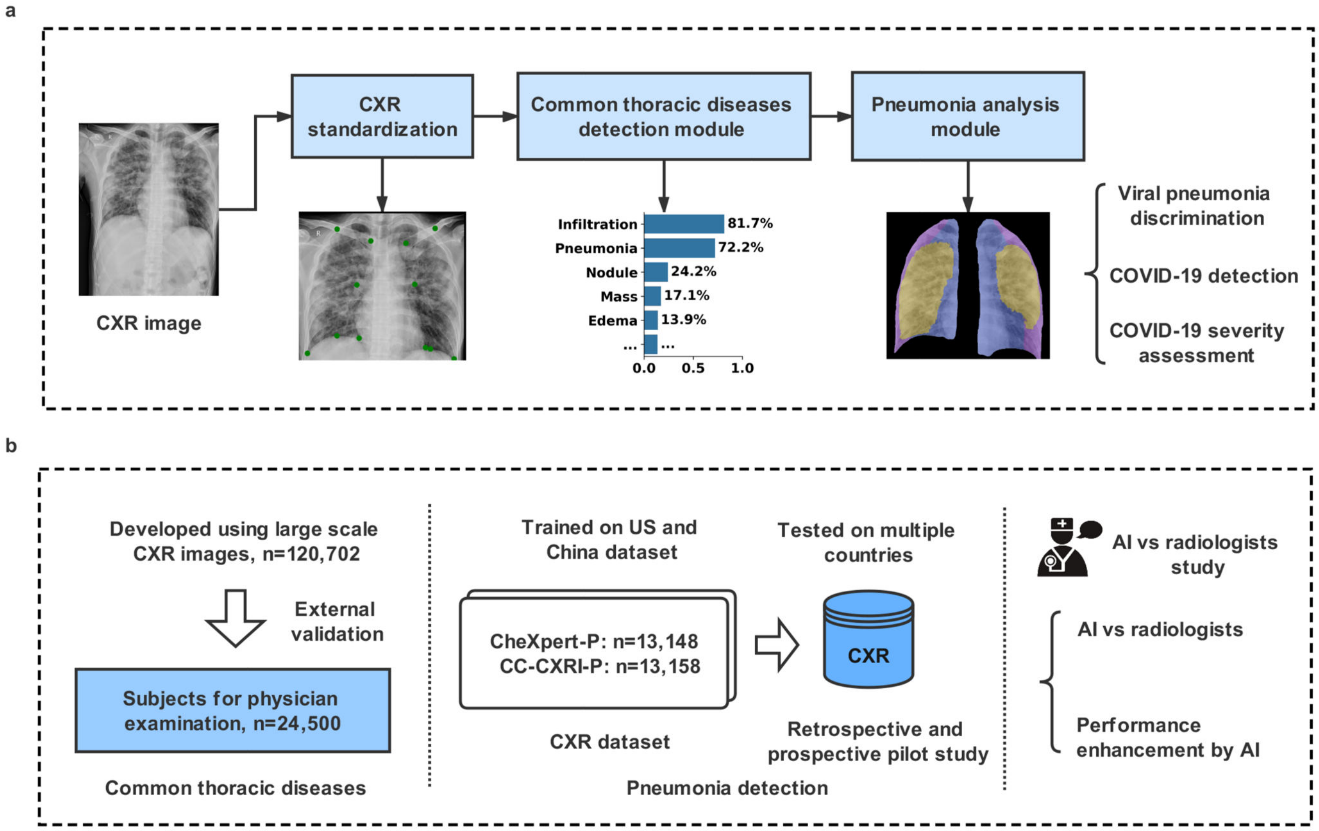
a
CXR image
CXR standardization
Common thoracic diseases detection module
Pneumonia analysis module
Infiltration 81.7%
Pneumonia 72.2%
Nodule 24.2%
Mass 17.1%
Edema 13.9%
... ...
0.0 0.5 1.0
Viral pneumonia discrimination
COVID-19 detection
COVID-19 severity assessment
b
Developed using large scale CXR images, n=120,702
External validation
Subjects for physician examination, n=24,500
Common thoracic diseases
Trained on US and China dataset
Tested on multiple countries
CheXpert-P: n=13,148
CC-CXRI-P: n=13,158
CXR dataset
CXR
Retrospective and prospective pilot study
Pneumonia detection
AI vs radiologists study
AI vs radiologists
Performance enhancement by AI

Fig. 1. The AI system for the detection of viral pneumonia.

a, Model development of the AI system. The system included a pipeline consisting of a CXR standardization module, a common chest thoracic disease detection module, and a pneumonia analysis module. The pneumonia analysis module consisted of viral pneumonia classification, COVID-19 detection, and COVID-19 severity assessment. **b,** Application and evaluation of the AI system. Left panel: An AI system was trained to identify the presence and absence of 14 common thoracic pathologies, and its performance was evaluated in external validation cohorts. Middle panel: In training with the Chinese cohort (CC-CXRI-P) and the re-annotated public dataset (CheXpert-P), the AI system made a diagnosis of viral pneumonia (including COVID-19 pneumonia). The model was then tested on external cohorts to assess the AI system's generalizability. Right panel: the performance of the AI system was compared with the performances of radiologists and with the performance of the combination of human and machine intelligence.

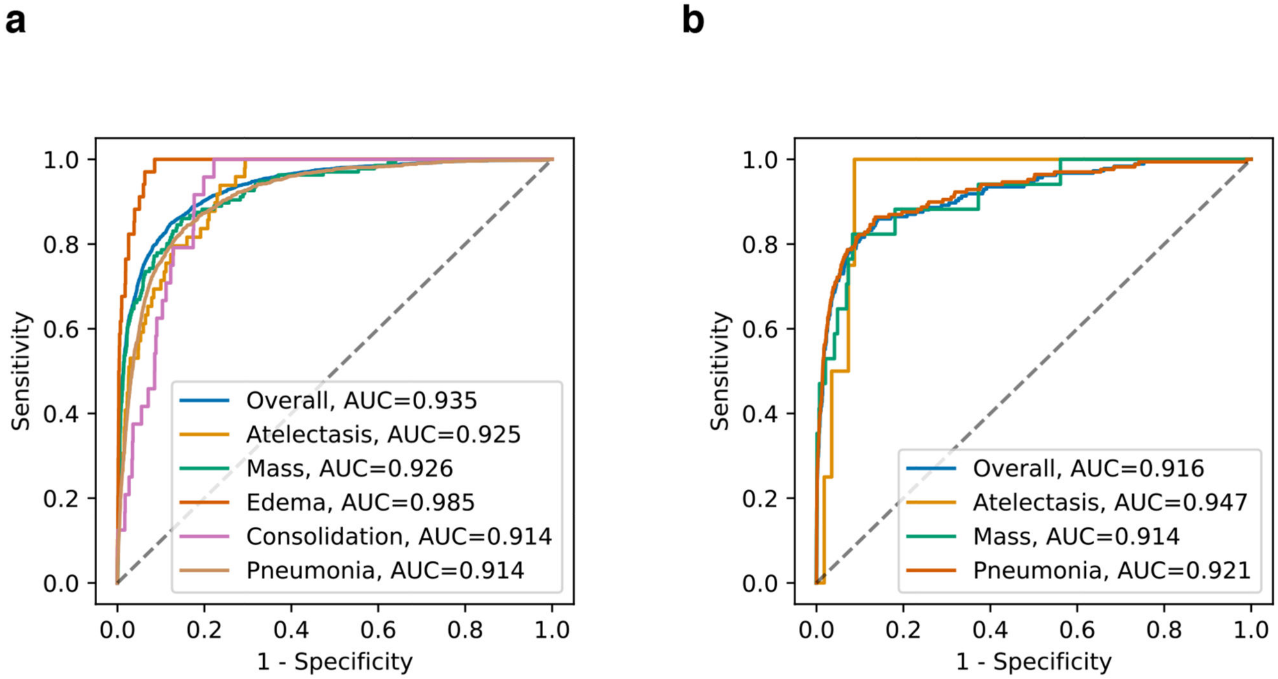

Fig. 2. Performance of the AI system in the multi-label classification of common chest diseases encompassing opacity.

Receiver operating characteristic curves (ROC) and normalized confusion matrices of the classification model. Opacity included atelectasis, mass, edema, pneumonia, and consolidation. **a,** The AI system’s performance on the hold-out test dataset. **b**, The AI system’s performance on the external validation cohorts that represent the population for physical examination. Compared with the patient distribution from **a,** there existed merely edema, and consolidation.

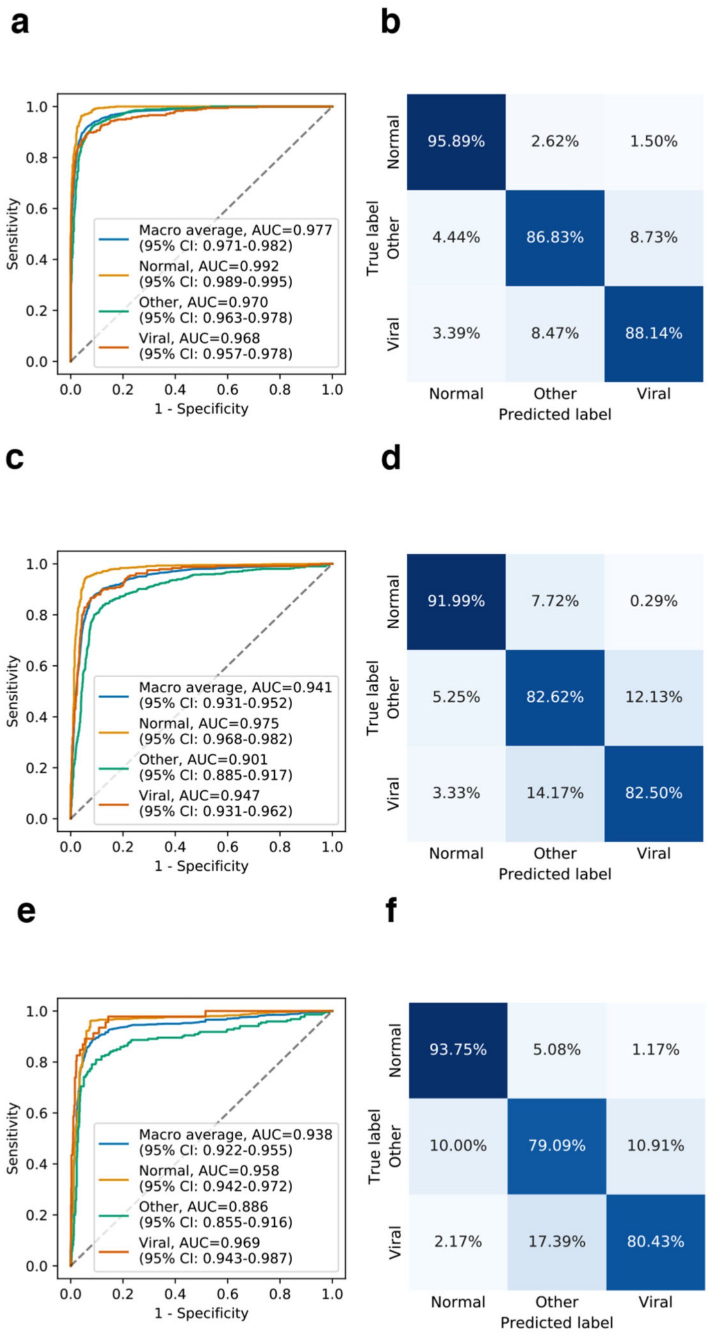

Fig. 3. Performance of the AI system in the discrimination of viral pneumonia, other types of pneumonia, and absence of pneumonia, from CXR images.

Receiver operating characteristic curves (ROC) and normalized confusion matrices of the classification model. **a** and **b,** AI system's performance on the hold-out test dataset. **c** and **d**, The AI system's performance on the independent external validation data in the China cohort. For the three-way classification. **e** and **f**, The AI system's performance on the external validation set for subjects screening for suspicious pneumonia. CI, confidence interval.

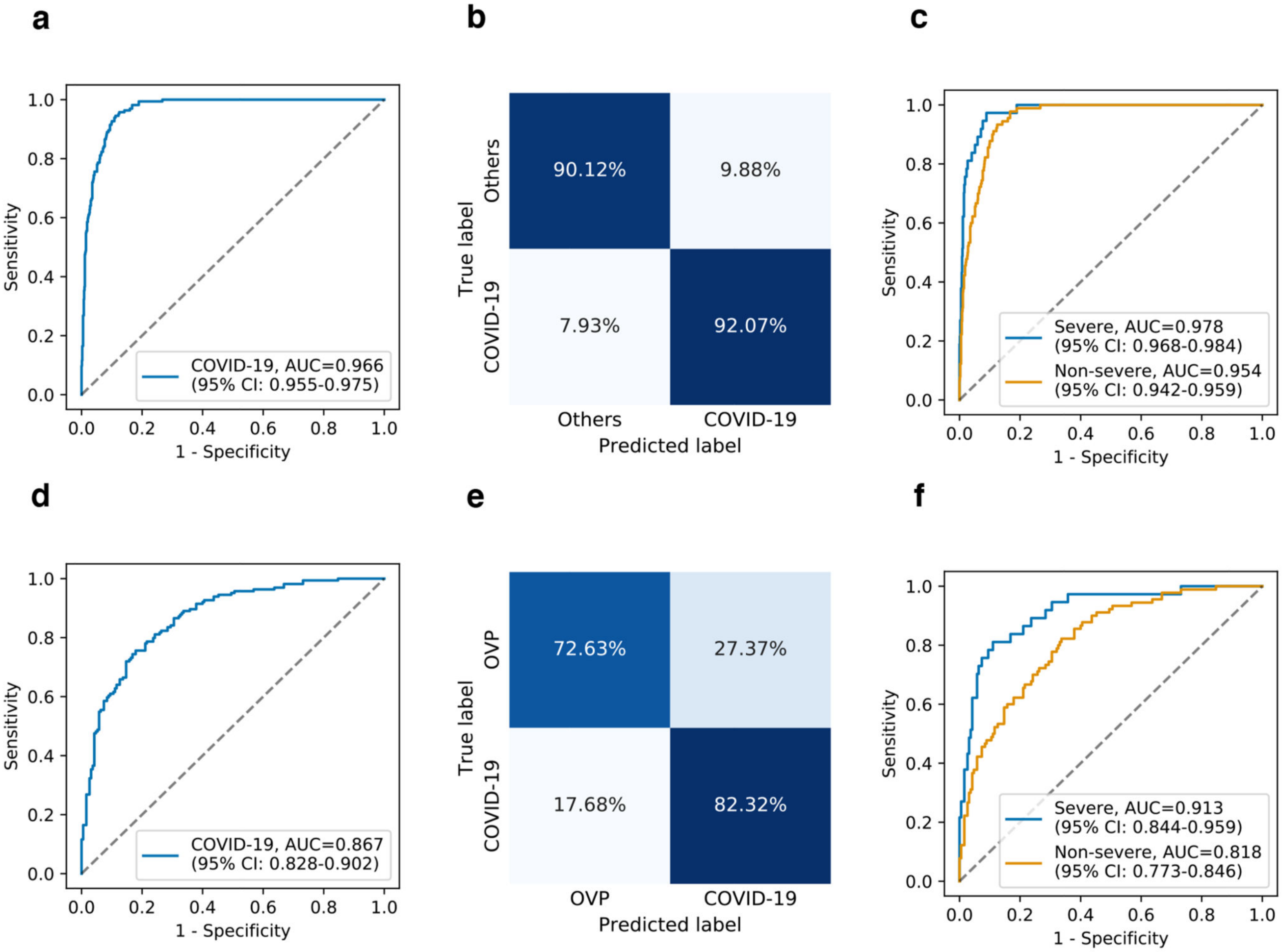

Fig. 4. Performance of the AI system in the identification of COVID-19 pneumonia from CXR images.

ROC curves and normalized confusion matrices for binary classification. **a** and **b**, The AI system’s performance on differentiating COVID-19 pneumonia from others (e.g., bacterial pneumonia) on test dataset: AUC = 0.966 (95% CI: 0.955-0.975), sensitivity = 92.07%, specificity = 90.12%. **d** and **e**, The AI system’s performance on differentiating COVID-19 pneumonia from other viral pneumonia (OVP) on the test dataset: AUC = 0.867 (95% CI: 0.828-0.902), sensitivity = 82.32%, specificity = 72.63%. **c** and **f**, ROC curves showing the AI system’s performance on identifying severe or non-severe COVID-19 from others pneumonia (**c**) (e.g., bacterial pneumonia) and other types of viral pneumonia (**f**).

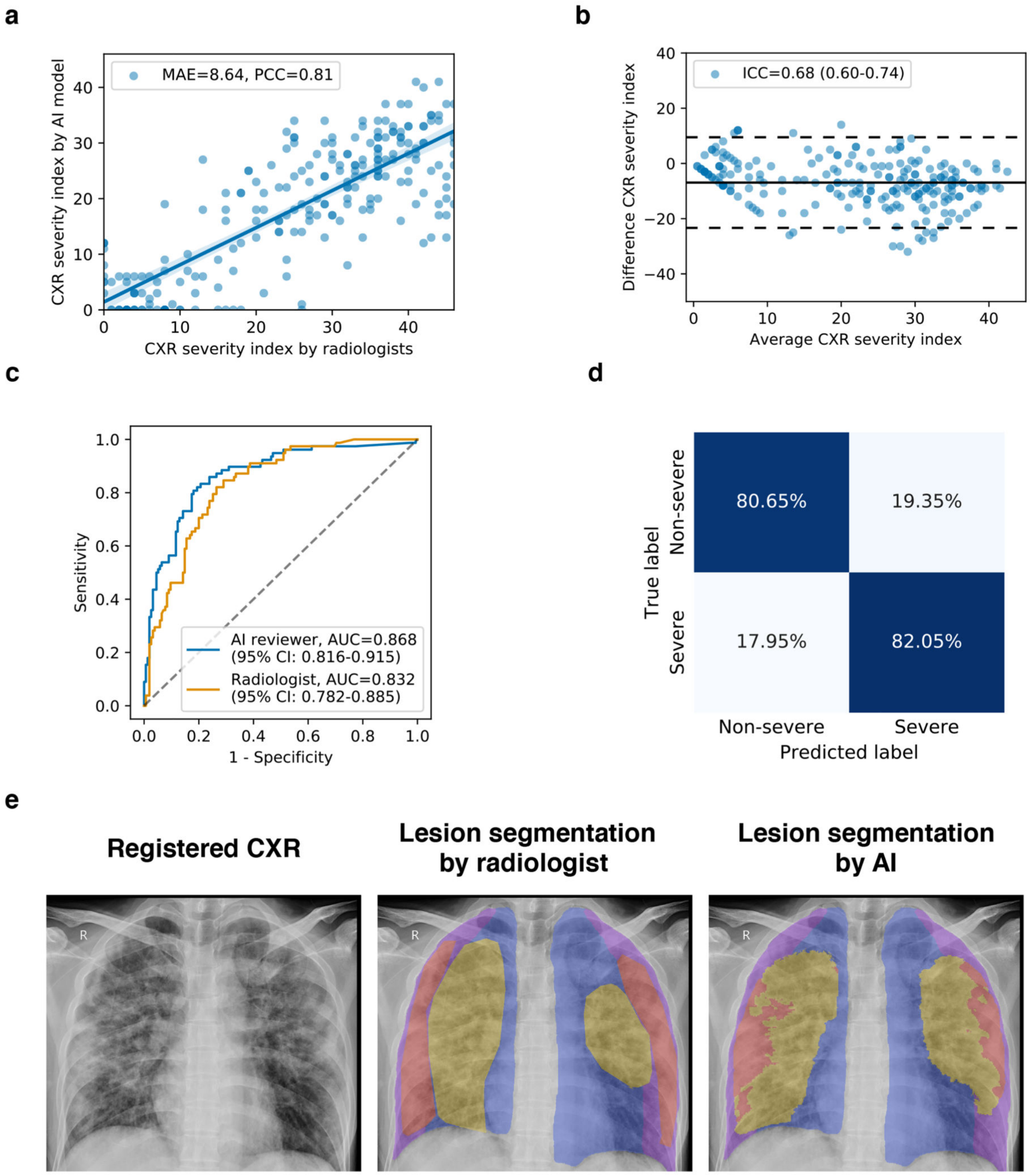

Fig. 5. Severity analysis of COVID-19 pneumonia patients from CXR images.

a, Scatter plot showing the correlation of the CXR severity index by the AI model versus the CXR severity index by the radiologist's assessment. **b**, Bland-Altmann plot showing the agreement between the AI predicted severity index and the radiologist assessed severity index. X-axis represents the mean of the two measurements, and the Y-axis represented the difference between the two measurements. **c**, ROC curves for the binary classification of the clinical severity. The blue curve represented the severity prediction by using the AI predicted severity index as input: AUC = 0.868 (95% CI: 0.816-0.915). The orange curve represented the severity prediction by using the radiologist assessed severity index as input: AUC = 0.832 (95% CI: 0.782-0.885). **d,** Confusion matrix for the binary classification of the clinical severity. The performance of the AI reviewer: accuracy = 81.12%, sensitivity = 82.05%, specificity = 80.65%. **e,** An example of lung-lesion segmentation of viral pneumonia of a CXR image. PCC, Pearson correlation coefficient; MAE, mean absolute error; ICC, Intraclass correlation coefficient.

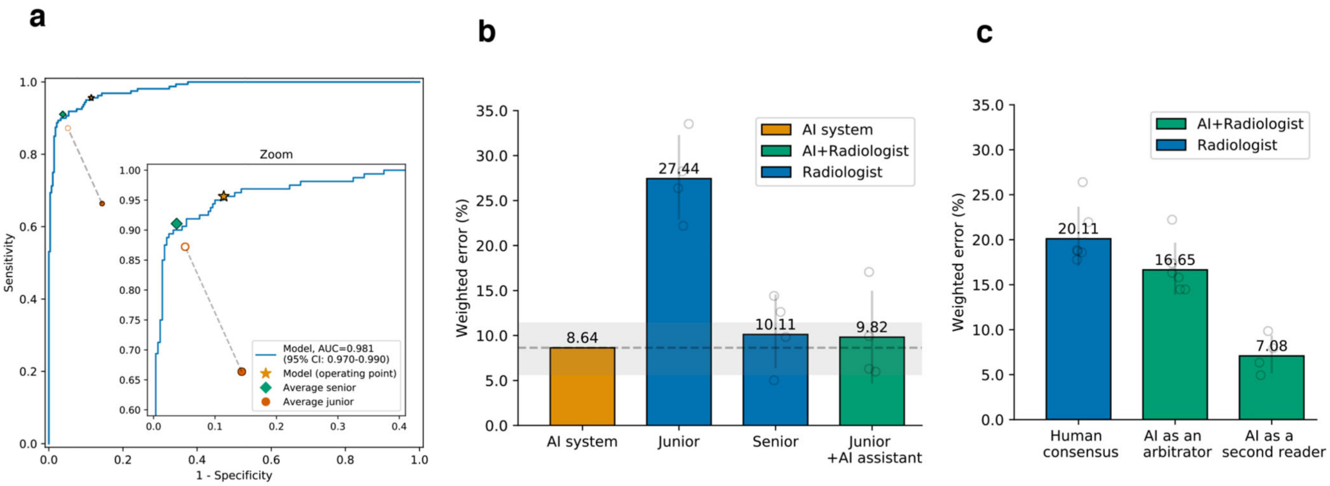

Fig. 6. Performance of the AI system and of radiologists in identifying pneumonia conditions from CXR images.

The performance comparison of four groups: the AI system, an average of a group of four junior radiologists, an average of a group of four senior radiologists, and an average of the group of four junior radiologists with AI assistance. **a,** The ROC curves for diagnosing viral pneumonia from the rest (other types of pneumonia and normal). The star denoted the operating point of the AI system. Filled dots denoted the junior and senior radiologists' performance, while the hollow dots denoted the performance of the junior group with the AI's assistance. Dashed lines linked the paired performance values of the junior group. **b,** Weighted errors of the four groups based on a penalty metric. P < 0.001 computed using a two-sided permutation test of 10,000 random re-samplings. **c,** An evaluation experiment on diagnostic performance when the AI system acted as a "second reader" or an "arbitrator".

Table 1

The CXR datasets for the training, validation and testing of the deep-learning system.

Cohorts	**Developmental Dataset**			**External validation**
	Training dataset	**Tuning dataset**	**Testing dataset**	**(SYSU-PE)**
Number of images	96,543	12,035	12,124	24,500
Number of subjects	73,917	9,160	9,250	23,585
Inpatient	38,438 (52.0%)	4,761 (52.0%)	4,871 (52.7%)	--
Outpatient	35,479 (48.0%)	4,377 (47.8%)	4,354 (47.1%)	--
Physical Examination	--	22 (0.2%)	25 (0.2%)	23,585 (100.0%)
Male (%)	31,019 (42.0%)	3,840 (41.9%)	3,850 (41.6%)	11,868 (50.3%)
Age (years), mean (IQR)	44.9 (32-59)	45.1 (32-60)	44.9 (32-59)	37.8 (28-46)
Atelectasis	167(0.23%)	26(0.28%)	22(0.24%)	4(0.02%)
Cardiomegaly	1,828(2.47%)	242(2.64%)	239(2.58%)	46(0.20%)
Fibrosis	4,405(5.96%)	523(5.71%)	560(6.05%)	431(1.83%)
Infiltration	7,085(9.59%)	914(9.98%)	886(9.58%)	88(0.37%)
Mass	708(0.96%)	86(0.94%)	82(0.89%)	17(0.07%)
Nodule	4,187(5.66%)	550(6.00%)	554(5.99%)	463(1.96%)
Pleural thickening	4,192(5.67%)	545(5.95%)	544(5.88%)	412(1.75%)
Pneumonia	8,099(10.96%)	1,015(11.08%)	1,042(11.26%)	164(0.70%)
Pneumothorax	552(0.75%)	67(0.73%)	61(0.66%)	0(0.00%)
Consolidation	118(0.16%)	12(0.13%)	12(0.13%)	0(0.00%)
Edema	133(0.18%)	12(0.13%)	21(0.23%)	0(0.00%)
Effusion	3,903(5.28%)	485(5.29%)	462(4.99%)	43(0.18%)
Hernia	23(0.03%)	3(0.03%)	1(0.01%)	1(0.01%)
Emphysema	715(0.97%)	84(0.92%)	84(0.91%)	29(0.12%)
No finding	55,320(74.84%)	6,823(74.49%)	6,882(74.40%)	22,319(94.63%)

Table 2

Number of CXR images for training, validation and testing in differentiating among viral pneumonia, other types of pneumonia, and absence of pneumonia (normal).

Cohorts		Viral pneumonia				
		Other types of viral pneumonia	**COVID-19 pneumonia**	**Other types of pneumonia**	**Normal**	**Total**
Training	“Gold-standard labels” China (CC-CXRI)	2,506	1,248	5,015	4,389	13,158
	“Silver-standard labels” US (CheXpert-P)	2,840	--	5,309	4,999	13,148
Validation “Gold-standard labels” (CC-CXRI)		169	159	637	554	1,519
Testing “Gold-standard labels” (CC-CXRI)		190	164	630	535	1,519
External validation “Gold-standard labels”		142	98	610	1,049	1,899
Population Study “Gold-standard labels”		46	0	220	768	1,034
International cohort “Gold-standard labels” on COVID-19		63	132	226	229	650