Check for updates

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

Guangyu Wang [1,25] ✉, Xiaohong Liu[2,25], Jun Shen [3,25], Chengdi Wang[4,25], Zhihuan Li[5,25], Linsen Ye[6,25], Xingwang Wu[7,25], 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[6], Zehong Yang[3], Huimin Cai[10], Guangming Lu [12], Lingyan Zhang[13], Lei Yang[14], Wenqin Xu[5], Winston Wang [5], Andrea Olevera[5], Ian Ziyar[5], 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[18], Xiaoguang Zou[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 [5,10] ✉, Weimin Li[4] ✉ and Tianxin Lin [3] ✉

Common lung diseases are first diagnosed using chest X-rays. Here, we show that a fully automated deep-learning pipeline for the standardization of chest X-ray images, for the visualization of lesions and for disease diagnosis can identify viral pneumonia caused by coronavirus disease 2019 (COVID-19) and assess its severity, and can also discriminate between viral pneumonia caused by COVID-19 and other types of pneumonia. The deep-learning system was developed using a heterogeneous multi-centre 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 the absence of disease with areas under the receiver operating characteristic curve (AUCs) of 0.94–0.98; between severe and non-severe COVID-19 with an AUC of 0.87; and between COVID-19 pneumonia and other viral or non-viral pneumonia with AUCs of 0.87–0.97. 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 support for clinical decision-making.

The outbreak of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and its associated disease, COVID-19, led to a pandemic of the highest concern[1–6]. The genome of the new virus and the epidemiological and clinical features of the infection have been reported[1–4]. The viral infection frequently presents as an infection of the upper respiratory tract or pneumonia

[1]School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China. [2]Department of Computer Science and Technology & BNRist, Tsinghua University, Beijing, China. [3]Department of Urology, Department of Radiology, Department of Emergency Medicine, Department of Disciplinary Development and Planning, Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou, China. [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 and University Hospital, Macau, China. [6]Department of Hepatic Surgery and Liver Transplantation Center, The Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, China. [7]Department of Radiology, The First Affiliated Hospital of Anhui Medical University, Hefei, China. [8]The Sun Yat-sen Cancer Center, Sun Yat-sen University, Guangzhou, China. [9]The First College of Clinical Medical Science, China Three Gorges University, Yichang, China. [10]Department of Bioinformatics, Bioland Laboratory (Guangzhou Regenerative Medicine and Health Guangdong Laboratory), Guangzhou, China. [11]Department of Thoracic Surgery and Oncology, The First Affiliated Hospital of Guangzhou Medical University, China State Key Laboratory and National Clinical Research Center for Respiratory Disease, Guangzhou, China. [12]Department of Medical Imaging, Jinling Hospital, Medical School of Nanjing University, Nanjing, China. [13]Department of Medical Imaging, The Third Affiliated Hospital, Southern Medical University, Guangzhou, China. [14]Department of Thoracic Surgery, The First Affiliated Hospital of Sun Yat-sen University, Guangzhou, China. [15]Department of Medical Imaging and Deptartment of Cardiology, Xiangya Hospital, Central South University, Changsha, China. [16]Department of Radiology, Second Xiangya Hospital, Central South University, Changsha, China. [17]Department of Radiology, Taihe Hospital, Hubei University of Medicine, Hubei, China. [18]Department of Radiology, The Second Affiliated Hospital and Yuying Children's Hospital of Wenzhou Medical University, Wenzhou, China. [19]Department of Infectious Disease, Wenzhou Central Hospital, Wenzhou, China. [20]Department of Respiratory and Critical Care Medicine, The First People's Hospital of Kashi Prefecture, Kashi, China. [21]Oncology R&D, AstraZeneca, Cambridge, UK. [22]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK. [23]Department of Radiology and Cancer Research UK Cambridge Centre, University of Cambridge, Cambridge, UK. [24]Department of Applied Biology and Chemical Technology, Hong Kong Polytechnic University, Hong Kong, China. [25]These authors contributed equally: Guangyu Wang, Xiaohong Liu, Jun Shen, Chengdi Wang, Zhihuan Li, Linsen Ye, Xingwang Wu. ✉e-mail: guangyu.wang24@gmail.com; tingchen@mail.tsinghua.edu.cn; kang.zhang@gmail.com; weimi003@scu.edu.cn; lintx@mail.sysu.edu.cn

(COVID-19 pneumonia) that can rapidly progress to acute respiratory failure, multiple 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 radiography has 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[8–10]. Previous studies have proposed the concept of radiomics and imageomics, which refer to the extraction of quantitative imaging feature information in a high-throughput manner[11]. The AI model also demonstrated general applicability in retinal diseases and childhood diseases using medical images, pretrained with data of conventional approaches based on transfer learning[12]. To diagnose common lung and heart diseases on the basis of CXRs, AI models using weakly supervised classification or attention-based convolution neural networks[13] have also been studied.

Although computational methods have been proposed for the detection of lung disease, a fully automatic analysis pipeline that is robust to variable CXR image conditions and that meets the standard of actual clinical application is lacking[6,14]. One of the challenges is anatomical landmark detection, which has a vital role in medical image analysis. Radiologists routinely align an input image to these landmarks and perform diagnosis and quantification[15,16]. However, most landmark-detection methods were developed for facial recognition. To date, 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 a gold standard for clinical evaluation and the generalization of the systems to different populations or new settings. Another critical obstacle for the general use of medical AI systems is that the inner decision-making processes of deep-learning algorithms remain opaque, hindering the translation into clinical practice. Thus, under this unprecedented COVID-19 pandemic, it is particularly relevant to develop a general AI system for CXRs that can provide 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 centres to provide a first-line assessment with a quick turn-around time.

In this Article, we aim to develop a comprehensive system to combat 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 of the tissues between the X-ray source and the capturing film; tissue structures are less-well defined in an X-ray compared with a computed tomography (CT) image and lack three-dimensional information. To overcome these shortcomings, we integrated 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 between COVID-19 and other types of viral pneumonia. Technically, our AI system is a modular analysis pipeline that comprises automated detection of the anatomical landmarks, lung-lesion segmentation and pneumonia diagnosis prediction, using CXRs as input. Furthermore, the AI system could assess the clinical severity of COVID-19 on the basis of the proposed CXR lung-lesion segmentation model (Fig. 1 and Supplementary Fig. 6).

To develop this AI system, we used a large-scale hospital-wide dataset ($n = 120,702$) for the detection of common thoracic pathologies and a large multicentre dataset for pneumonia analysis.

We also examined 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 to train the model. Furthermore, we compared the performance of the system with the performance of radiologists in routine clinical practice. The results show that the performance of the AI is accurate and robust across multiple populations and settings. The system could be integrated into the workflow to improve a radiologist's diagnostic performance.

Results

Image characteristics and system overview. We constructed a large CXR dataset on the basis of the China Consortium of Chest X-ray Image Investigation (CC-CXRI) to develop the AI system. The CC-CXRI consists of two large-scale datasets—the first dataset is a CXR database for common thoracic diseases containing 145,202 CXR images that were retrospectively collected from the Memorial Hospital of Sun Yat-sen University (SYSU), and the second dataset is a CXR dataset (CC-CXRI-P) containing 16,196 CXR images for detecting suspicious pneumonia, including COVID-19 pneumonia. Here, 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, which is an automated CXR analysis pipeline, consists 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 comprises anatomical landmark 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. Here we used 12 anatomical landmarks labelled on 676 CXR images to train the landmark-detection model. We implemented and compared three deep-learning models for landmark detection—including the U-Net[17], fully convolutional networks (FCN)[18] and DeepLabv3 (ref. [19])—using a fivefold cross-validation test (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 Figure 9a shows a visualization example of our AI model compared with the radiologist's annotation, showing that the AI model obtained accurate landmark-detection results. Interestingly, we observed that all three models performed better for the right part of the 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, oedema, effusion, emphysema, fibrosis, hernia, infiltration, mass, nodule, pleural thickening, pneumonia and pneumothorax (Table 1).

The pneumonia analysis module, which consists of a lung-lesion segmentation model and a final classification model, estimates the subtype of pneumonia (for example, 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 with 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/)[20,21]. We validated our system on this database and

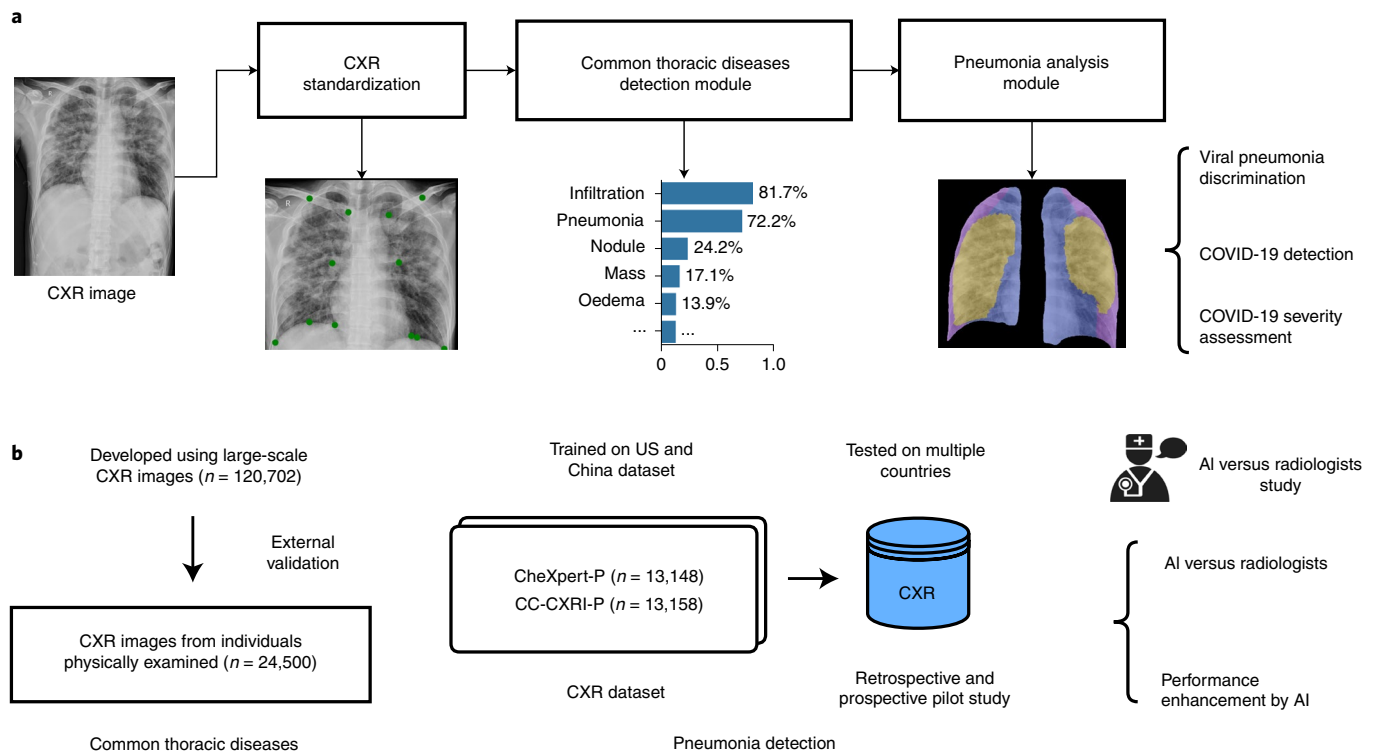**Fig. 1 | The AI system for the detection of viral pneumonia. a**, Model development of the AI system. The system includes a pipeline consisting of a CXR standardization module, a common thoracic diseases detection module and a pneumonia analysis module. The pneumonia analysis module comprises viral pneumonia classification, COVID-19 detection and COVID-19 severity assessment. **b**, Application and evaluation of the AI system. Left: an AI system was trained to identify the presence and absence of 14 common thoracic diseases, and its performance was evaluated in external validation cohorts. Middle: during training with the Chinese cohort (CC-CXRI-P) and the reannotated public US 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 generalizability of the AI system. Right: 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.

achieved a good performance for landmark detection with a mean ($\pm$ s.d.) of 5.568 ± 6.175 mm for the actual physical distance error. As the SCR database was established to facilitate studies on the anatomical segmentation of CXR images, we also validated our lung segmentation model, which showed good accuracy, with a Dice coefficient of 0.954 and 0.961 for segmentation of the left and right lung fields, respectively (Supplementary Table 7).

Multilabel classification of common thoracic diseases. Here, a large-scale dataset (the SYSU set) from CC-CXRI—consisting of 120,702 CXR images from 92,327 patients with labels of 14 common thoracic pathologies—was used to train the model. All of the 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 analysed by automated detection of anatomical landmarks to permit image registration. Next, 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 coexist or overlap on the same CXR image, we used a multilabel classification approach instead of a multiclass classification method, whereby overlaps between labels were allowed, and labels were predicted individually before being integrated into a final prediction. The AI system achieved a macro performance with an 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, oedema, 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 between pneumonia and all of the other groups and an AUC of 0.935 for the overall classification of lung opacity (Fig. 2a).

To evaluate the generalizability of the AI system across various screening settings, we tested it on a cohort called SYSU-PE, which consisted of an 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 oedema cases among the SYSU-PE cohort. The results showed an overall AUC of 0.916 for multilabel 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 detecting lung opacity (Supplementary Fig. 1).

Training the AI system to identify viral pneumonia. To develop a model that could differentiate between viral pneumonia and other types of pneumonia and absence of pneumonia on the basis of CXR images, we constructed a deep neural network based on the DenseNet-121 (ref. [22]) 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).

Table 1 | The CXR datasets for the training, validation and testing of the deep-learning system

Subset	Developmental dataset			External validation (SYSU-PE)
	Training dataset	Tuning dataset	Testing dataset	
Number of images	96,543	12,035	12,124	24,500
Number of individuals	73,917	9,160	9,250	23,585
Inpatients	38,438 (52.0%)	4,761 (52.0%)	4,871 (52.7%)	–
Outpatients	35,479 (48.0%)	4,377 (47.8%)	4,354 (47.1%)	–
Patients for 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%)
Mean age (years) (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%)
Oedema	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%)

IQR, interquartile range. For all the subsets (except 'Number of images'), the numbers given are of individuals.

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 is considered to be part of the diagnostic workup of symptomatic individuals with suspected COVID-19 in settings in which laboratory testing information is not available or results are delayed or initially negative in the presence of symptoms that are attributable to COVID-19 (ref. [23]). Here, we adopted the terms 'gold-standard labels' and 'silver-standard labels' to differentiate between the labels obtained from a confirmed laboratory-based ground truth versus diagnoses based on clinical and radiographic findings by a consensus of radiologists[24,25]. CXR images were classified into three types—viral pneumonia, other aetiologies/types of pneumonia and the absence of pneumonia (normal). The viral pneumonia group consisted of patients with common types of viral pneumonia and with COVID-19.

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 PCR with reverse transcription (RT–PCR) test or other standard laboratory diagnostic tests. Among the 16,196 images in the CC-CXRI-P dataset, 4,436 showed viral pneumonia—including 1,571 COVID-19 pneumonia—6,282 were other types of pneumonia and 5,478 showed an 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 dataset for training. The CheXpert dataset is an open-source retrospective patient cohort that contains mixtures of different types of pneumonia and other lung disorders. Our radiologists manually regraded 13,148 CXR images with the label of pneumonia and classified them into 2,840 viral pneumonia, 5,309 other types of pneumonia and 4,999 absence of pneumonia. This reannotated pneumonia dataset was named CheXpert-P.

To train the model, we initially trained the AI system with the gold-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 tests or other standard laboratory diagnostic tests. The three-way classification results showed an overall performance of an AUC of 0.963 (95% confidence interval (CI) = 0.955–0.969; Supplementary Fig. 10a). We next added the CheXpert-P dataset with silver-standard labels into the training set of CC-CXRI. Again, we retrained the AI model and tested it on the same test set from the CC-CXRI. The results showed better performance, with an 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 an improvement in classification performance. The improvement is due to the AI model being exposed to different types of images. As a result, the AI system differentiated between viral pneumonia and the other two groups with a sensitivity of 92.94%, a specificity of 87.04% and an AUC of 0.968 (95% CI = 0.957–0.978; Fig. 3a,b).

To quantify the impact of the standardization module on the diagnostic performance, we evaluated the AI system on the test set skipping the whole module or part of the module. 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 with the baseline model (Supplementary
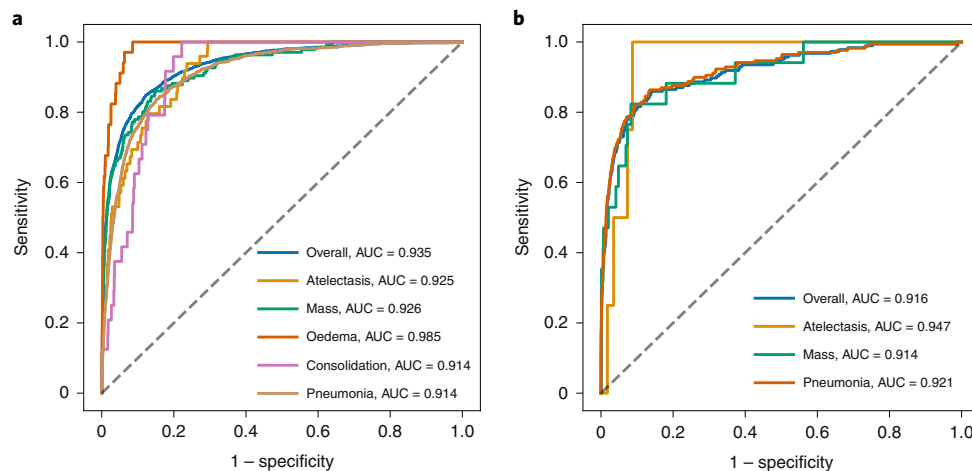
Fig. 2 | The performance of the AI system in the multilabel classification of common chest diseases encompassing opacity. ROC curves of the multilabel classifications. Opacity includes atelectasis, mass, oedema, pneumonia and consolidation. **a**, The performance of the AI system on the held-out test dataset. **b**, The performance of the AI system on the external validation cohort (SYSU-PE) that represents the population for physical examination. Compared with **a**, the external validation cohort shows a different patient distribution with few cases of oedema and consolidation.

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 multicountry datasets. To test the generalizability of the AI system 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 images (including 98 COVID-19 pneumonia), 610 CXRs of other types of pneumonia and 1,049 CXRs of absence of pneumonia (normal images). The AI system achieved an average AUC of 0.941 (95% CI = 0.931–0.952) in the three-way classification. For differentiating between viral pneumonia images and images of other types of pneumonia and normal images, the AI system achieved a sensitivity of 90.00%, a specificity of 87.40% and an AUC of 0.947 (95% CI = 0.931–0.962; Fig. 3c,d).

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 CXRs, 220 CXRs of other types of pneumonia and 768 normal images (Table 2). The AI model achieved an AUC of 0.938 (95% CI = 0.922–0.955) in the three-way classification, and a sensitivity of 89.13%, a specificity of 93.02% and an AUC of 0.969 (95% CI = 0.943–0.987; Fig. 3e,f) for differentiating between viral pneumonia and 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 an AUC of 0.934 (95% CI = 0.917–0.950) for the three-way classification, and an AUC of 0.920 (95% CI = 0.891–0.942) for differentiating between viral pneumonia and the other two groups (Supplementary Fig. 2a).

The fourth external validation was performed on an open public Kaggle-pneumonia dataset (https://www.kaggle.com/paultimothy-mooney/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 detecting viral pneumonia (Supplementary Fig. 2b). Overall, these results firmly

demonstrated that the performance of our AI system showed a high level of consistency, 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 CXRs and 5,515 CXRs of other viral pneumonia, as well as 10,961 CXRs of 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 CXRs and 630 other pneumonia CXRs, and obtained an AUC of 0.966 (95% CI = 0.955–0.975), a sensitivity of 92.07% and a specificity of 90.12% (Fig. 4a,b). A separate, independent dataset containing 164 COVID-19 pneumonia CXRs and 190 CXRs of 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,e). Both results confirmed that the AI system is sensitive to the subtle lesion information of CXRs in triaging COVID-19 pneumonia and differentiating between COVID-19 pneumonia and other pneumonia with a reasonable accuracy as a first-line diagnostic tool. We conducted additional experiments to differentiate between the different subgroups of COVID-19 (severe and non-severe COVID-19) and other types of viral pneumonia. The results showed that the performance for detecting non-severe COVID-19 was relatively inferior performance compared with the performance for detecting severe COVID-19 (Fig. 4c,f).

Next, we tested the AI system on the public BIMCV dataset from the Valencia region of Spain[26], including 663 COVID-19 images from BIMCV-COVID19 and 1,277 normal images from BIMCV-COVID19-PADCHEST. The results showed an AUC of 0.916 (95% CI = 0.904–0.933) for identifying COVID-19 as viral pneumonia and differentiating 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 on the basis of 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. Figure 5e presents an example of viral
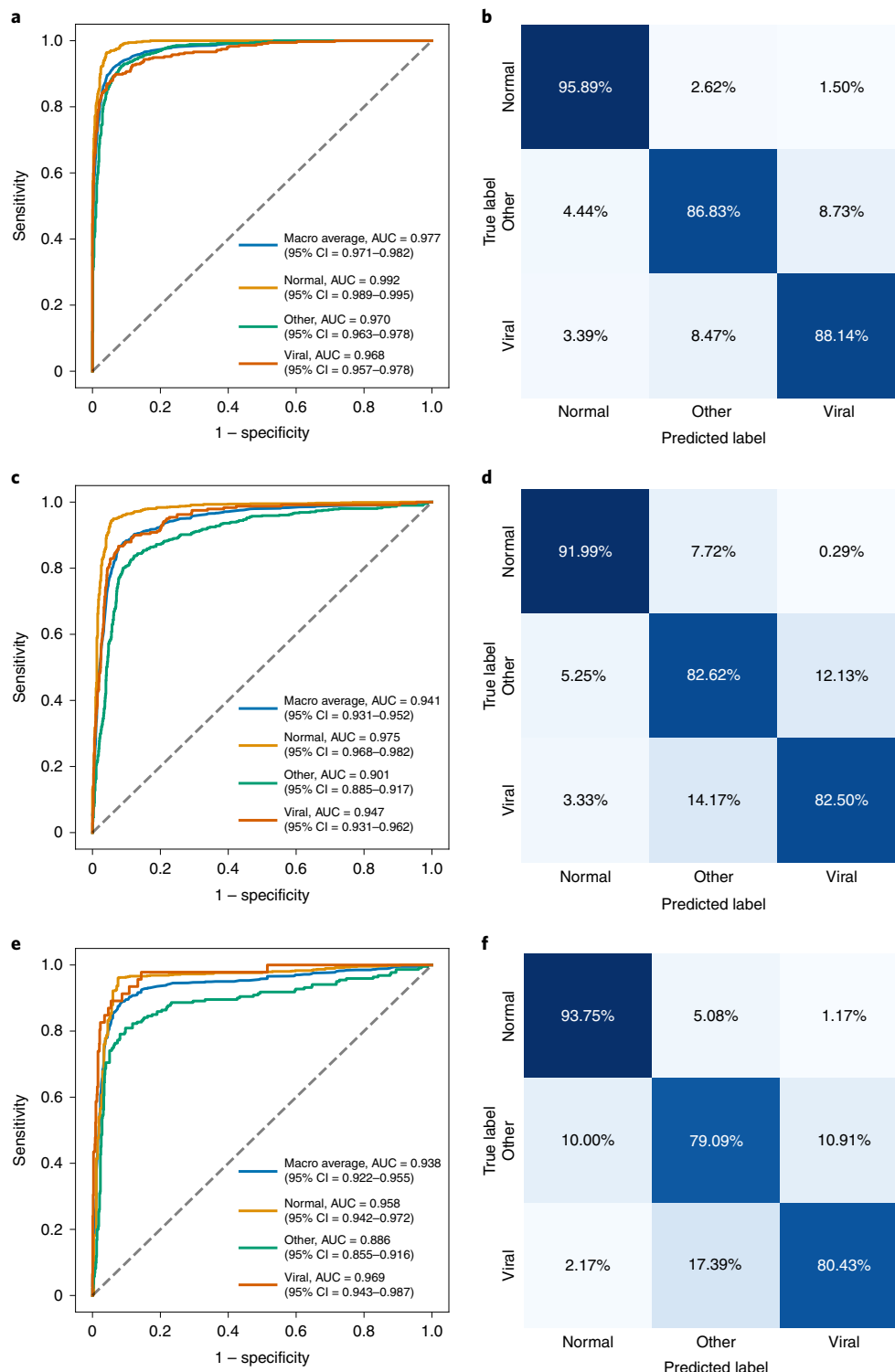
Fig. 3 | The performance of the AI system in discriminating between viral pneumonia, other types of pneumonia and the absence of pneumonia from CXR images. a–f, ROC curves (**a,c,e**) and normalized confusion matrices (**b,d,f**) of multiclass classifications. **a,b**, The performance of the AI system on the held-out test dataset. **c,d**, The performance of the AI system on the first independent external validation set in the China cohort. **e,f**, The performance of the AI system on the second external validation set of individuals screening for suspicious pneumonia.

pneumonia with comparable lung-lesion segmentation by the AI model and human radiologists. Compared with the 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 of 0 to 4 by a group of trained radiologists on the basis the percentage of lung lesions in

Table 2 | The number of CXR images for training, validation and testing in differentiating between viral pneumonia, other types of pneumonia and the absence of pneumonia (normal)

Cohorts	Viral pneumonia		Other types of pneumonia	Normal	Total
	Other types of viral pneumonia	COVID-19 pneumonia			
Training					
Gold-standard labels China (CC-CXRI)	2,506	1,248	5,015	4,389	13,158
Silver-standard labels United States (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
First external validation					
Gold-standard labels (China cohort)	142	98	610	1,049	1,899
Second external validation					
Gold-standard labels (China screening cohort)	46	0	220	768	1,034
Third external validation					
Gold-standard labels on COVID-19 (international cohort)	63	132	226	229	650

the section. The 1,207 CXR images of the patients with COVID-19 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 radiologists and the AI reviewer showed a strong linear relationship, with a Pearson correlation coefficient of 0.81 and a mean absolute error of 8.64 (Fig. 5a). A Bland–Altman plot showed a good agreement between the AI model and the radiologists, with an intraclass correlation coefficient (ICC) of 0.68 (95% CI = 0.60–0.74), whereas the agreement between evaluations of the radiologists 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 $PO_2/FiO_2 < 300$ mmHg. It usually corresponded to diffuse interstitial pneumonia, which obscured normal lung markings[27]. A total of 1,207 CXRs were manually graded on the basis of clinical diagnoses and classified into 437 severe and 770 non-severe labels. We next used the severity index scores by the AI model and the radiologist reviewers as an input for a logistic regression model to generate a clinical severity prediction (Methods). The results showed that our AI system could predict the severity of COVID-19 pneumonia 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 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 help to predict 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 performance of the AI system against practicing radiologists in classifying the viral pneumonia, other types of pneumonia and normal groups. A total of eight radiologists with different levels of clinical experience were enrolled to participate in this study: four junior radiologists with longer than 10 years of experience and four senior radiologists with longer than 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 three senior radiologists.

The performance was evaluated on the basis of AUC and the sensitivity and specificity (Fig. 6a and Supplementary Table 6). The AI system achieved comparable performance to the level of the senior radiologists, 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 compared with the average junior radiologists (Supplementary Table 6).

One of our objectives was to investigate whether the AI system 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 on the basis of a penalty score system, was used as a metric to evaluate and compare the performance of our AI system and the practicing radiologists. The performance of the junior radiologists with the AI assistance yielded an average weighted error of 9.82%, a significant improvement ($P < 0.001$) compared with that of 27.44% the radiologists without the AI assistance (Fig. 6b).

We also examined the potential role of the AI system in enhancing the diagnostic performance of radiologists in the workflow. In this simulated scenario, a specific diagnosis was made by two radiologist readers (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. By contrast, when the AI system acted as an 'arbitrator', the error was reduced to 16.65%; furthermore, when the AI system acted as a second reader, the error was further reduced to 7.08% (Fig. 6c). These results demonstrated that the AI system could improve the performance of radiologists and reduce image-reading workload. The details of the receiver
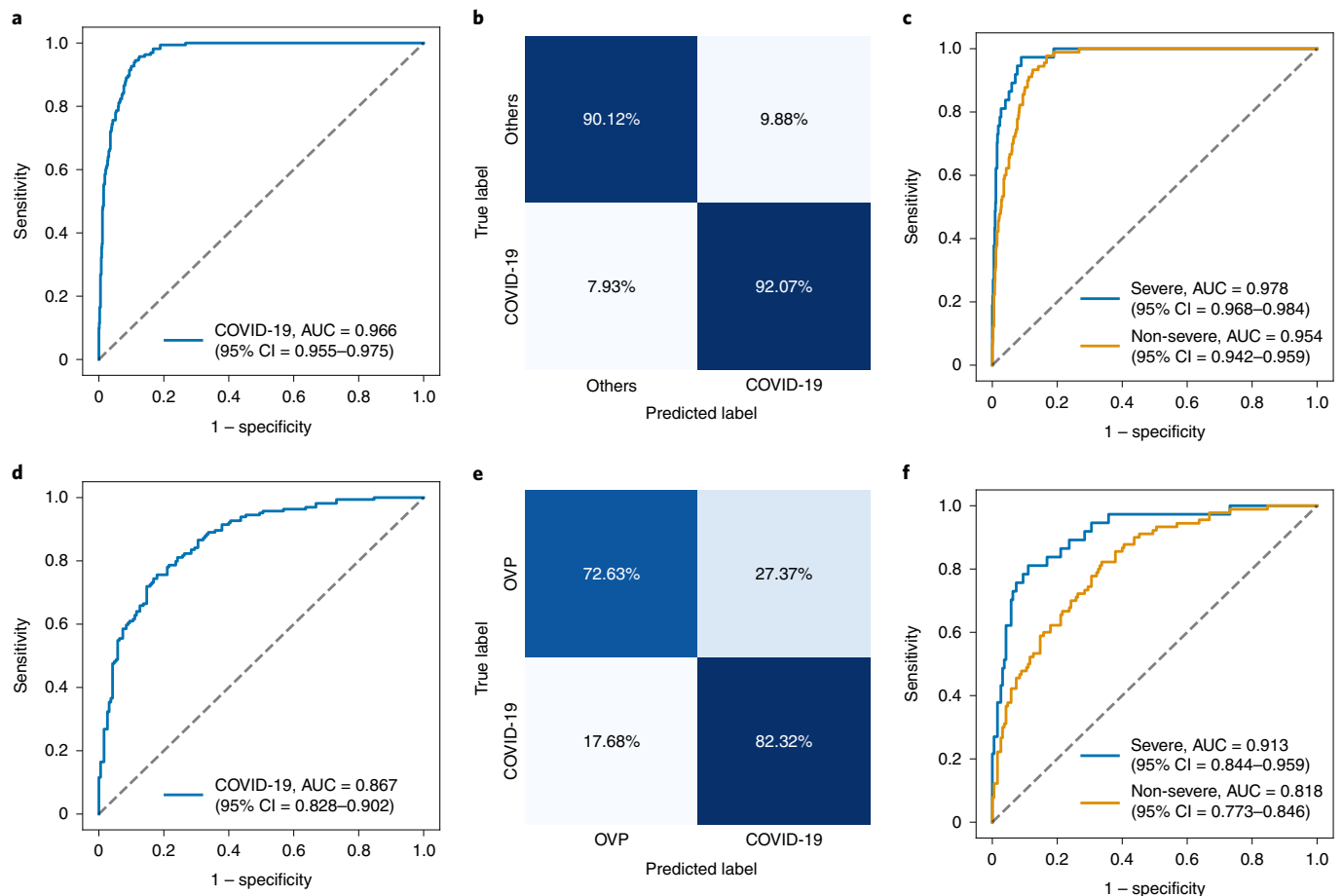
a

b

c

d

e

f

Fig. 4 | The performance of the AI system in identifying COVID-19 pneumonia from CXR images. a,b,d,e, ROC curves (**a,d**) and normalized confusion matrices (**b,e**) for binary classification. **a,b**, The performance of the AI system in differentiating between COVID-19 pneumonia and other pneumonia ('Others', for example, bacterial pneumonia) on the test dataset: AUC = 0.966 (95% CI = 0.955–0.975), sensitivity = 92.07%, specificity = 90.12%. **d,e**, The performance of the AI system in differentiating between COVID-19 pneumonia and other viral pneumonia (OVP) on the test dataset: AUC = 0.867 (95% CI = 0.828–0.902), sensitivity = 82.32%, specificity = 72.63%. **c,f**, ROC curves showing the performance of the AI system in identifying severe or non-severe COVID-19 from other pneumonia (**c**) and other types of viral pneumonia (**f**).

operating characteristic (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 between COVID-19 and 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[28]. 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. By contrast, CXR is a front-line tool with a quick turnaround 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 is fundamentally a clinical diagnosis and, in suspected COVID-19, RT–PCR is the reference gold standard. However, owing 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 the observational ability and comprehension of the clinicians. 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 imaging has been considered to be part of the diagnostic workup of symptomatic individuals with suspected COVID-19 in settings in which laboratory testing
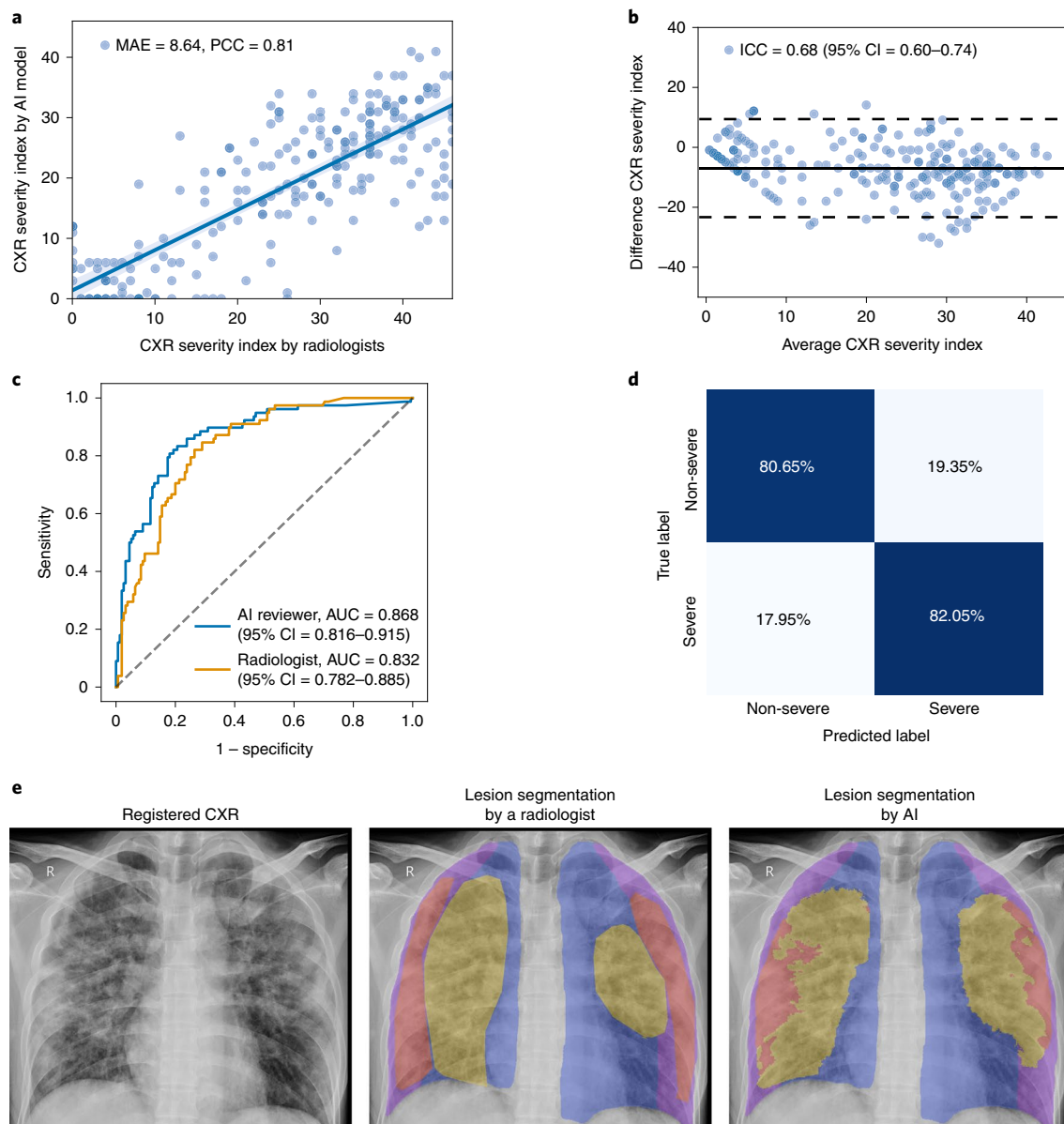
Fig. 5 | Severity analysis of patients with COVID-19 pneumonia from CXR images. a, The correlation of the CXR severity index generated by the AI model versus the CXR severity index based on the radiologist's assessment. MAE, mean absolute error; PCC, Pearson correlation coefficient. **b**, Bland–Altman plot of the agreement between the AI-predicted severity index and the radiologist-assessed severity index. The x axis represents the mean of the two measurements, and the y axis represents the difference between the two measurements. The black solid line represents the mean of the difference, and the two dashed lines represent 95% limits of agreement (mean ± 1.96× s.d.). ICC, intraclass correlation coefficient. **c**, ROC curves for the binary classification of the clinical severity. The blue curve represents the severity prediction using the AI-predicted severity index as the input: AUC = 0.868 (95% CI = 0.816–0.915). The orange curve represents the severity prediction using the radiologist-assessed severity index as the 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.

(RT–PCR) is not available or results are delayed or initially negative in the presence of apparent symptoms that are attributable to COVID-19 (ref. [23]). Such a workflow could help healthcare/hospital administrators to 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 as 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. Here we examined the general

applicability of the present AI system. We first trained our AI system using large, heterogenous multicentre datasets. We next presented 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 in patients coming for hospital visits (SYSU set), and then measured the performance in 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.
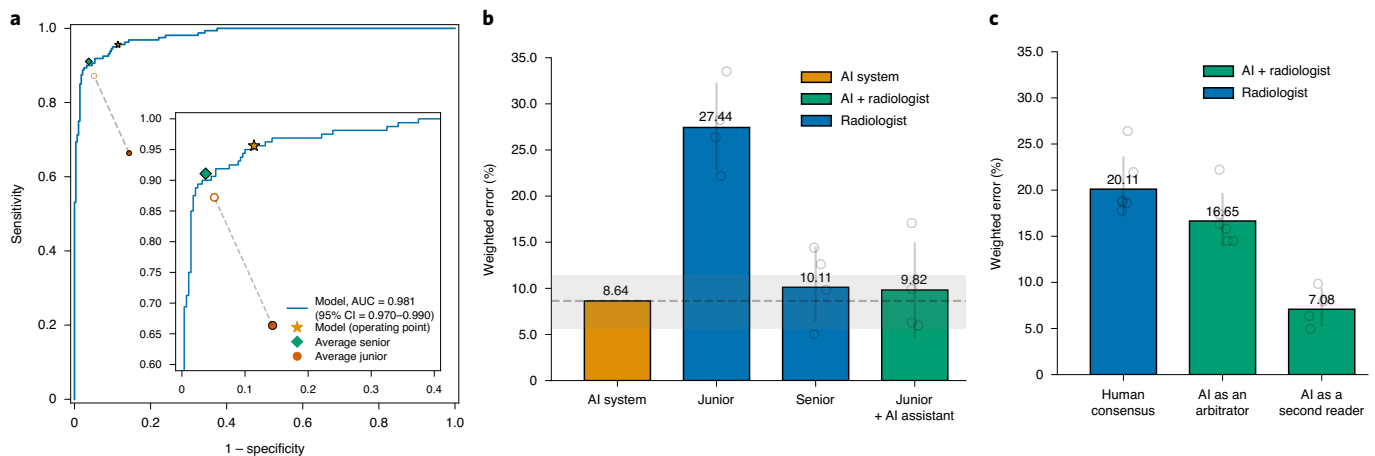
Fig. 6 | The performance of the AI system and of radiologists in identifying pneumonia conditions from CXR images. Performance comparison of four groups: the AI system; an average of a group of four junior radiologists; an average of a group of four senior radiologists; and an average of the group of four junior radiologists with AI assistance. **a**, The ROC curves for diagnosing viral pneumonia from other types of pneumonia and from the absence of pneumonia. The star denotes the operating point of the AI system. The filled dots denote the performance of the junior and senior radiologists, and the hollow dots denote the performance of the junior group with assistance from the AI system. The dashed lines link the paired performance values of the junior group. Inset: magnification of the plot. **b**, Weighted errors of the four groups on the basis of a penalty metric. The grey dashed line represents the performance of the AI system, and the grey shaded region represents the 95% confidence interval. $P < 0.001$ computed using a two-sided permutation test of 10,000 random resamplings. **c**, An evaluation experiment on diagnostic performance when the AI system acted as a 'second reader' or an 'arbitrator'.

Notably, the AI system can also assist in assessing the severity of a patient's condition. 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. Furthermore, 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' present workflow. This integration can be crucial during a pandemic, such as the current COVID-19 situation, when resources are stretched thinly. The ability of our AI system 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, as the AI system was trained in a population in which more than 90% were symptomatic patients with abnormal imaging findings, its ability to diagnose 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 in 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 will be necessary to establish its clinical use in a broad range of populations. Another limitation is its ability to differentiate between COVID-19 and non-focal (diffuse) acute respiratory distress syndrome. However, acute respiratory distress syndrome is a crucial acute condition with associated pulmonary oedema; therefore, through 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 during 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 patients. CXR images were extracted from the CC-CXRI data, which were collected from multiple hospitals, including the Sun Yat-sen Memorial Hospital and the Third Affiliated Hospital, which are 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 the 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 an 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 RT–PCR test. The other types of pneumonia were diagnosed on the basis of 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 of the patients signed a consent form. The research 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 a natural language processing (NLP) pipeline to extract disease labels from clinical reports for CXR images. The pipeline included disease concept detection and negation classification, similar to CheXpert[29] and the NIH Chest X-ray dataset[13] (Supplementary Methods).

We selected 14 common thoracic diseases according to their clinical importance and prevalence, as defined on the basis of the ICD-10 and the NIH Chest X-ray dataset. 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, oedema, effusion, emphysema, fibrosis, hernia, infiltration, nodule, mass, pleural thickening, pneumonia and pneumothorax. We also defined another label of 'no finding' that was positive if and only if all of the other labels of a CXR image were negative. Thus, each CXR image in the dataset was annotated by the presence or absence of the 15 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 centres. The SYSU-PE dataset is comprised of 24,500 CXR images from 23,585 patients coming for the health check. The SYSU dataset was used for model development and internal validation, and the SYSU-PE dataset was used for external validation. The labels of the validation data were manually reviewed for a reliable evaluation.

Silver-standard labels of pneumonia. Previous research suggested specific differences in CXR imaging findings between viral pneumonia and bacterial pneumonia. Thus, imaging has been considered to be part of the diagnostic workup of symptomatic individuals with suspected COVID-19 in settings in which the laboratory testing (RT–PCR) is not available or results are delayed or initially negative in the presence of symptoms attributable to COVID-19 (ref. [23]).

Here 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 labelled with the presence or absence of each of 14 common chest radiographic observations. The original CXR images were given only a general diagnosis of pneumonia without a detailed label of viral pneumonia or other types of pneumonia. Here, we considered the manually graded image to be the silver standard, in contrast to the ground truth gold-standard labels discussed below. A total of 15 radiologists with longer than 10 years of clinical experience manually reviewed and graded a subset of CheXpert with pneumonia labels. They labelled 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 longer than 20 years of clinical experience. In the case of inconsistency, the expert consensus was used to correct labels. A total of 13,148 CXR images from CheXpert were relabelled into three categories: 2,840 viral pneumonia, 5,309 other types of pneumonia and 4,999 normal CXRs. We named this reannotated dataset CheXpert-P and treated it as the silver-standard label dataset for training.

Gold-standard labels and ground truth of pneumonia. All of the CXR images from the CC-CXRI dataset had a 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 pneumonia or bacterial pneumonia. The above laboratory test results served 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 that were 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 the case of disagreement, an expert of consensus was used to correct the labels. The resulting labels served as the ground truth for the evaluation dataset.

Annotation of landmarks and lung-lesion segmentation on CXRs. We used 676 manually annotated CXR images from individuals with viral pneumonia and other pneumonia, as well as healthy individuals to train the anatomical landmark determination. Twelve anatomical landmarks were labelled on each CXR image: midpoint of clavicle left and right, sternal end of clavicle left and right, hilar angle left and right, costophrenic angle left and right, diaphragmatic dome left and right, cardiac diaphragmatic angle left and right.

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 patients with COVID-19 pneumonia), 1,163 from patients with bacterial pneumonia, 187 from patients with other types of pneumonia and 438 from healthy individuals. The annotation was performed using polygons. The lung segmentation labels included lung field (left), the periphery of the lung field (left), lung field (right) and 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 fivefold cross-validation test was applied for the landmark detection and lung-lesion segmentation.

Performance comparisons with radiologists. To evaluate the performance of our AI system in classifying the three types of pneumonia, we constructed an independent validation set of 440 CXR images, including 160 CXRs of viral pneumonia, 160 CXRs of other types of pneumonia and 120 CXRs of 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 used 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 mistriaged to normal, this may cause the spread of the disease.

We conducted a simulation study in which the AI system was deployed first as a 'second reader' and second as an 'arbitrator' of the diagnostic decisions of radiologists. 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 on the basis of the separate validation set. For the statistical significance of the comparison results, we computed CIs and P values using 1,000 random resamplings (bootstraps).

Transfer learning and deep learning. We trained our AI model using a large number of CXR images from three public datasets—CheXpert dataset[29], MIMIC-CXR dataset[30] and NIH Chest X-ray dataset[13].

Transfer learning was adopted by pretraining a DenseNet-121 model[22] for the CXR image classification. The DenseNet-121 architecture has been proven to be effective for CXR classification tasks[29]. The convolutional layers were fine-tuned when transferring to other tasks, whereas 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 was 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 patients with COVID-19. The input CXR images were resized to 512 × 512 by bilinear interpolation.

We used the cross-entropy loss function and adopted an Adam optimizer[31] for training, with a learning rate of 0.003, and the batch size was set to 32. All of the deep-learning models were implemented with Pytorch v.1.4 (ref. [32]). 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. For our proposed AI system, we 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 provided in the Supplementary Methods.

The CXR standardization module performed invert-greyscale CXR[33] detection, anatomical landmark detection and CXR image registration in this study. We first trained an invert-greyscale CXR detection model to detect whether the input of CXR was inverted-greyscale and, if so, the system automatically converted it into a conventional CXR. The anatomical landmark-detection model next performs the detection of landmarks of the CXR. On the basis of the detected landmarks, we generated a registered CXR image using the image registration algorithm. These components were specially designed to address the common problems that are encountered in computer-assisted detection with CXRs, including variations/inconsistency in 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 coexist or overlap on the same CXR image, we used a multilabel classification approach that could predict multiple categories at the same time and is therefore more suitable for clinical settings in which combinations or simultaneous occurrences of the categories often exist. Using the standardized CXR images, we trained a multilabel classification model with 15 binary classifications, including 14 disease labels and 1 no-finding label. The number of the output scalars was 15 with a sigmoid activate function. For the scenario of opacity detection, the case was defined as positive if at least one label of atelectasis, mass, oedema, pneumonia and consolidation was present. The predicted probability of opacity was composited by averaging over outputs of atelectasis, mass, oedema, pneumonia and consolidation.

The pneumonia analysis module is a two-stage architecture 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 extracted lung regions were then used to train the lesion segmentation network. As the raw CXR images may contain irrelevant information for lesion segmentation (for example, body parts that are not related to the lungs), a lung segmentation network was trained to discard such information such that the lesion segmentation network can concentrate on the lung area. Next, the diagnostic model—a neural network classifier—made a prediction on the basis of the outputs of the previous models, namely, the anatomical landmark-detection model and 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 patients with COVID-19. The clinical severity level is a clinical diagnosis of a respiratory distress state: blood oxygen saturation < 92%; respiratory rate < 36 or $PO_2/FiO_2 < 300$ mmHg. It usually corresponds to diffuse interstitial pneumonia, which obscures normal lung markings[27]. For the analysis of the severity of patients with COVID-19 pneumonia, a total of 1,207 CXR images was 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 and 75–100, respectively), whereas the AI system automatedly segmented the lung lesion and quantified the severity of the CXR image. Each CXR image of 1,027 patients with COVID-19 was given a severity score by a group of radiologists on the basis of the above definition. To evaluate the association between the AI model and radiologists for scoring the CXR severity, we calculated the mean absolute error and Pearson correlation coefficient. Bland–Altman plots[34] and ICC were also used to assess the 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 the scores of all 12 sections as input features and adopted the logistic regression as the classification model. A 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[35], 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. A ROC analysis and AUC were used to assess model performance for each classification task. For multiclass tasks, the macro averages 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 model f and a given ROC curve TPR = f(FPR), where FPR $\in$ [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 performance of the models and experts, the weighted error was calculated by weighting the error of the ith class being predicted by the jth class by a defined weight matrix. We evaluated the performance of the landmark-detection model 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 the original coordinates are normalized with x and y divided by the width and the height of the image, respectively. The successful detection rate is defined as the number of accurate detections versus the total number of detections, where an accurate detection is a prediction with a margin of error of less than or equal to a specified threshold. Physical distance errors were reported when the pixel size was known (for example 0.175 mm pixel size on the SCR dataset). We evaluated the performance of the segmentation model with two evaluation metrics, including intersection over union and Dice coefficient. The intersection over union is the area of the overlap between the predicted segmentation and the ground truth divided by the area of the union. The Dice coefficient 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 resamplings) was adopted to analyse the CI of the AUC[36]. 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 the

AUC was calculated. The shortest two-sided 95% CIs of the AUC were reported for each experiment. P values for sensitivity, specificity and weighted-error comparisons were generated using two-sided permutation tests of 10,000 random resamplings. 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.

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

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

Code availability
The custom codes for the diagnosis and discrimination between viral, non-viral and COVID-19 pneumonia from chest X-ray images are available at the China National Center for Bioinformation at the Big Bay Branch (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* **579**, 270–273 (2020).
2. Cohen, J. Wuhan seafood market may not be source of novel virus spreading globally. *Science* https://www.sciencemag.org/news/2020/01/wuhan-seafood-market-may-not-be-source-novel-virus-spreading-globally (26 January 2020).
3. Chan, J. F. 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* **395**, 514–523 (2020).
4. Huang, C. et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **395**, 497–506 (2020).
5. Qin, C., Yao, D., Shi, Y. & Song, Z. Computer-aided detection in chest radiography based on artificial intelligence: a survey. *Biomed. Eng. Online* **17**, 113 (2018).
6. Jaiswal, A. K. et al. Identifying pneumonia in chest X-rays: a deep learning approach. *Measurement* **145**, 511–518 (2019).
7. Pham, H. H., Le, T. T., Tran, D. Q., Ngo, D. T. & Nguyen, H. Q. Interpreting chest X-rays via CNNs that exploit disease dependencies and uncertainty labels. *Neurocomputing* **437**, 186–194 (2021).
8. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).
9. Esteva, A. et al. A guide to deep learning in healthcare. *Nat. Med.* **25**, 24–29 (2019).
10. Norgeot, B., Glicksberg, B. S. & Butte, A. J. A call for deep-learning healthcare. *Nat. Med.* **25**, 14–15 (2019).
11. Lambin, P. et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur. J. Cancer* **48**, 441–446 (2012).
12. Kermany, D. S. et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **172**, 1122–1131 (2018).
13. Wang, X. et al. Chestx-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Deep Learning and Convolutional Neural Networks for Medical Imaging and Clinical Informatics* (eds Lu, L. et al.) 369–392 (Springer International Publishing, 2019).
14. Franquet, T. Imaging of pneumonia: trends and algorithms. *Eur. Respir. J.* **18**, 196–208 (2001).
15. Ghesu, F. C. et al. An artificial agent for anatomical landmark detection in medical images. In *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention* (eds Ourselin, S. et al.) 229–237 (Springer, 2016).
16. Zhang, J., Liu, M. & Shen, D. Detecting anatomical landmarks from limited medical imaging data using two-stage task-oriented deep neural networks. *IEEE Trans. Image Processing* **26**, 4753–4764 (2017).
17. Ronneberger, O., Fischer, P. & Brox, T. U-net: convolutional networks for biomedical image segmentation. In *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention* (ed. Navab, N.) 234–241 (Springer, 2015).
18. Long, J., Shelhamer, E. & Darrell, T. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Analysis Mach. Intel.* **39**, 640–651 (2015).

19. Chen, L.-C., Papandreou, G., Schroff, F. & Adam, H. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans. Pattern Analysis Mach. Intel.* **40**, 834–848 (2018).
20. Van Ginneken, B., Stegmann, M. B. & Loog, M. Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database. *Med. Image Anal.* **10**, 19–40 (2006).
21. & 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 Am. J. Roentgenol.* **174**, 71–74 (2000).
22. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 4700–4708 (IEEE, 2017).
23. Akl, E. A. et al. Use of chest imaging in the diagnosis and management of COVID-19: a WHO rapid advice guide. *Radiology* **298**, E63–E69 (2021).
24. Titano, J. J. et al. Automated deep-neural-network surveillance of cranial images for acute neurologic events. *Nat. Med.* **24**, 1337–1341 (2018).
25. Willemink, M. J. & Koszek, W. A. Preparing medical imaging data for machine learning. *Radiology* **295**, 4–15 (2020).
26. Vayá, M. d. l. I. et al. BIMCV COVID-19+: a large annotated dataset of RX and CT images from COVID-19 patients. Preprint at https://arxiv.org/abs/2006.01174 (2020).
27. Guan, W. J. et al. Clinical characteristics of coronavirus disease 2019 in China. *New Eng. J. Med.* **382**, 1708–1720 (2020).
28. Zhang, K. et al. Clinically applicable AI system for accurate diagnosis, quantitative measurements and prognosis of COVID-19 pneumonia using computed tomography. *Cell* **181**, 1423–1433 (2020).
29. Irvin, J. et al. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In *Proc. 33rd AAAI Conference on Artificial Intelligence* 590–597 (AAAI Press, 2019).
30. & Johnson, A. E. W. et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci. Data* **6**, 317 (2019).
31. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. Preprint at https://arxiv.org/abs/1412.6980 (2014).
32. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. In *Proc. Advances in Neural Information Processing Systems* (eds Wallach, H. et al.) 8024–8035 (Curran Associates, 2019).
33. Musalar, E. et al. Conventional vs invert-grayscale X-ray for diagnosis of pneumothorax in the emergency setting. *Am. J. Emerg. Med.* **35**, 1217–1221 (2017).
34. Bland, J. M. & Altman, D. G. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **1**, 307–310 (1986).
35. McKinney, S. M. et al. International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89–94 (2020).
36. Efron, B. Bootstrap methods: another look at the jackknife. *Ann. Stat.* **7**, 1–26 (1979).

Author contributions

X.L., J. Shen, Z.L., L. Ye, C.W., X.W., T.C., K.W., X. Zhang, Z.Z., J.Y., Y.S., R.D., W. Liang, T.Y., M.G., J.W., Z. Yang, H.C., G. Lu, Lingyan Zhang, L. Yang, W.X., W.W., A.O., I.Z., C.Z., O.L., W. Liao, J.L., Wen Chen, Wei Chen, J. Shi, L. Zheng, Longjiang Zhang, Z. Yan, X. Zou, G. Lin, G.C., L.L.L., L.M., Y.L., M.R., E.S., C.-B.S., M.F., J.Y.-N.L., T.X., J.H., W. Li, G.W. and K.Z. collected and analysed the data. J.Y.-N.L. and T.C. helped to analyse and interpret the data and revise the manuscript. K.Z. and G.W. conceived and supervised the project and wrote the manuscript. All of the authors discussed the results and reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41551-021-00704-1.

Correspondence and requests for materials should be addressed to G.W., T.C., K.Z., W.L. or T.L.

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

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

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s):   Kang Zhang

Last updated by author(s):   Feb 19, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No special software or code was used to collect the data.
Data analysis	Pytorch and python libraries.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on 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.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	A total of 145,202 chest x-ray images were used to train and validate our AI models for diagnosis and discrimination of viral, non-viral and COVID-19 pneumonia.
Data exclusions	No data were excluded after passing the initial image-quality-control step.
Replication	Replication was not relevant. We used independent validation cohorts.
Randomization	Samples were randomly allocated to the training, tuning and testing sets.
Blinding	All images were de-identified during image processing to remove any patient-related information.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	Antibodies
☒	Eukaryotic cell lines
☒	Palaeontology
☒	Animals and other organisms
☐	☒ Human research participants
☐	☒ Clinical data

Methods

n/a	Involved in the study
☒	ChIP-seq
☒	Flow cytometry
☒	MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	Chest x-ray photographs were obtained as a part of routine clinical care.
Recruitment	Participants were recruited from multiple hospitals.
Ethics oversight	The China Consortium of Chest X-ray Image Investigation (CC-CXRI) Ethics Committee.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	The Institutional Review Boards of Sun Yat-sen University Second and Third Affiliated Hospitals and West China Hospital approved the study protocols. All participants were informed about the study objectives and signed a written informed consent.
Study protocol	The training included studying tutorials of each disease image atlas and a quiz on 1,000 images of pneumonia with a passing grade of over 95% accuracy. Phase-II graders consisted of general radiology specialists who individually reviewed every image classified by phase-I graders to ensure diagnostic accuracy. To improve consistency among phase-II graders, 20% of images were randomly selected and reviewed by senior experts. After establishing the consensus diagnoses, images were transferred to the AI team to undergo a final stage of data preprocessing.

Data collection	Chest x-ray photographs were obtained as a part of routine clinical care.
Outcomes	The AUC, sensitivity and specificity of the automated deep-learning pipeline for diagnosis and discrimination of viral, non-viral and COVID-19 pneumonia.