



OPEN

Advancing COVID-19 diagnosis with privacy-preserving collaboration in artificial intelligence

Xiang Bai^{1,2,24} , Hanchen Wang^{3,24}, Liya Ma^{1,24}, Yongchao Xu^{1,24}, Jiefeng Gan^{1,2,24} , Ziwei Fan², Fan Yang⁴, Ke Ma², Jiehua Yang², Song Bai², Chang Shu², Xinyu Zou², Renhao Huang², Changzheng Zhang⁵, Xiaowu Liu⁵, Dandan Tu⁵, Chuou Xu¹, Wenqing Zhang¹, Xi Wang⁶, Anguo Chen⁷, Yu Zeng⁸, Dehua Yang^{1,9} , Ming-Wei Wang^{1,9} , Nagaraj Holalkere¹⁰, Neil J. Halin¹⁰, Ihab R. Kamel¹¹, Jia Wu¹² , Xuehua Peng¹³, Xiang Wang¹⁴, Jianbo Shao¹³, Pattanasak Mongkolwat¹⁵ , Jianjun Zhang^{16,17} , Weiyang Liu³, Michael Roberts^{18,19} , Zhongzhao Teng²⁰, Lucian Beer²⁰, Lorena E. Sanchez²⁰ , Evis Sala²⁰, Daniel L. Rubin²¹ , Adrian Weller^{3,22}, Joan Lasenby³, Chuansheng Zheng⁴ , Jianming Wang²³ , Zhen Li¹ , Carola Schönlieb^{18,22} and Tian Xia²

Artificial intelligence provides a promising solution for streamlining COVID-19 diagnoses; however, concerns surrounding security and trustworthiness impede the collection of large-scale representative medical data, posing a considerable challenge for training a well-generalized model in clinical practices. To address this, we launch the Unified CT-COVID AI Diagnostic Initiative (UCADI), where the artificial intelligence (AI) model can be distributedly trained and independently executed at each host institution under a federated learning framework without data sharing. Here we show that our federated learning framework model considerably outperformed all of the local models (with a test sensitivity/specificity of 0.973/0.951 in China and 0.730/0.942 in the United Kingdom), achieving comparable performance with a panel of professional radiologists. We further evaluated the model on the hold-out (collected from another two hospitals without the federated learning framework) and heterogeneous (acquired with contrast materials) data, provided visual explanations for decisions made by the model, and analysed the trade-offs between the model performance and the communication costs in the federated training process. Our study is based on 9,573 chest computed tomography scans from 3,336 patients collected from 23 hospitals located in China and the United Kingdom. Collectively, our work advanced the prospects of utilizing federated learning for privacy-preserving AI in digital health.

As the gold standard for identifying coronavirus disease 2019 (COVID-19) carriers, polymerase chain reaction with reverse transcription (RT-PCR) is the primary diagnostic modality to detect viral nucleotide in specimens from cases with suspected infection; however, due to the various disease courses in different patients, the detection sensitivity hovers at around only 0.60–0.71 (refs. ^{1–4}), which results in a considerable number of false

negatives. As such, clinicians and researchers have made tremendous efforts in searching for alternatives^{5–7} and complementary modalities^{2,8–11} to improve the testing scalability and accuracy for COVID-19 and beyond.

It has been reported that coronavirus carriers present certain radiological features in chest computed tomography scans (CTs), including ground-glass opacity, interlobular septal thickening and

¹Department of Radiology, Tongji Hospital and Medical College, Huazhong University of Science and Technology, Wuhan, China. ²School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China. ³Department of Engineering, University of Cambridge, Cambridge, UK. ⁴Department of Radiology, Union Hospital of Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China. ⁵HUST-HW Joint Innovation Lab, Wuhan, China. ⁶CalmCar Inc, Suzhou, China. ⁷Wuhan Blood Centre, Wuhan, China. ⁸MSA Capital, Beijing, China. ⁹The National Centre for Drug Screening, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai, China. ¹⁰CardioVascular and Interventional Radiology, Radiology for Quality and Operations, The Cardiovascular Centre at Tufts Medical Centre, Radiology, Tufts University School of Medicine, Medford, OR, USA. ¹¹Russell H Morgan Department of Radiology & Radiologic Science, Johns Hopkins Hospital & Medicine Institute, Baltimore, MD, USA. ¹²Department of Radiation Oncology, School of Medicine, Stanford University, Palo Alto, CA, USA. ¹³Department of Radiology, Wuhan Central Hospital, Wuhan, China. ¹⁴Department of Radiology, Wuhan Children's Hospital, Wuhan, China. ¹⁵Faculty of Information and Communication Technology, Mahidol University, Salaya, Thailand. ¹⁶Thoracic/Head and Neck Medical Oncology, University of Texas MD Anderson Cancer Centre, Houston, TX, USA. ¹⁷Translational Molecular Pathology, University of Texas MD Anderson Cancer Centre, Houston, TX, USA. ¹⁸Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK. ¹⁹Oncology R&D at AstraZeneca, Cambridge, UK. ²⁰Department of Radiology, University of Cambridge, Cambridge, UK. ²¹Department of Biomedical Data Science, Radiology and Medicine, Stanford University, Palo Alto, USA. ²²Alan Turing Institute, London, UK. ²³Department of Hepatobiliary Pancreatic Surgery, Affiliated Tianyou Hospital, Wuhan University of Science and Technology, Wuhan, China. ²⁴These authors contributed equally: Xiang Bai, Hanchen Wang, Liya Ma, Yongchao Xu, Jiefeng Gan. ✉e-mail: xbai@hust.edu.cn; cszheng@163.com; jmwang@163.com; zhenli@hust.edu.cn; cbs31@cam.ac.uk; tianxia@hust.edu.cn

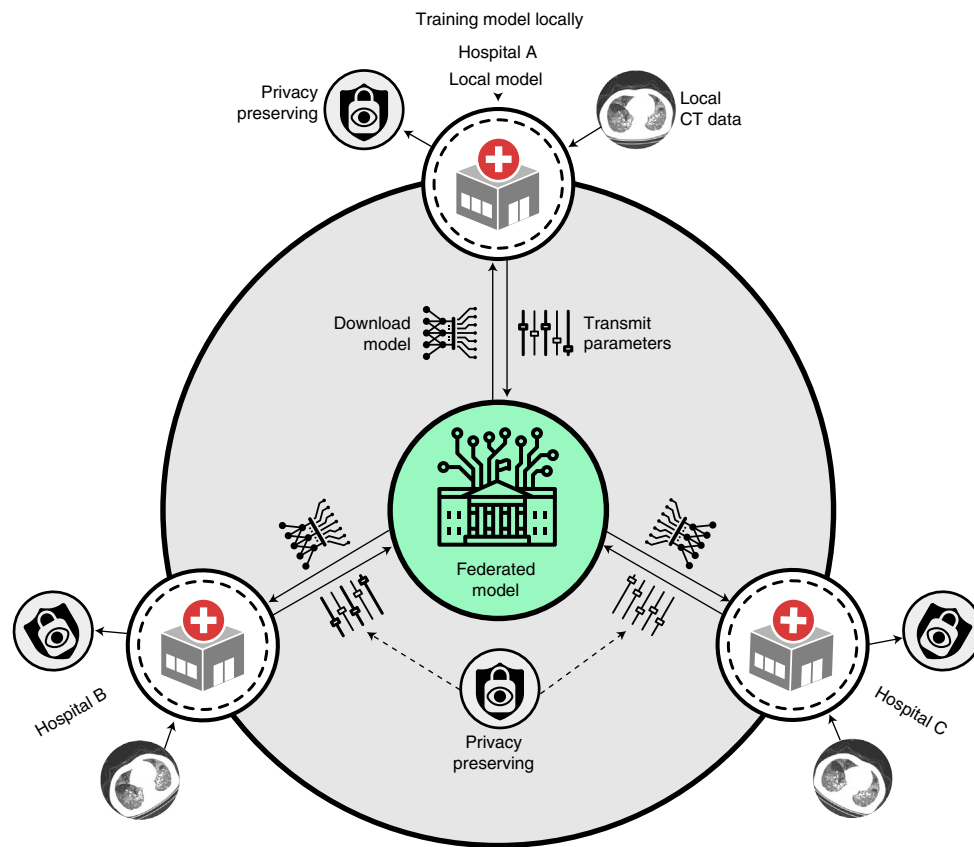


Fig. 1 | Conceptual architecture of UCADI. The participants first download and train the three-dimensional CNN models on the basis of the data of local cohorts. The trained model parameters are then encrypted and transmitted back to the server. Finally, the server produces the federated model via aggregating the contributions from each participant without explicit access to the parameters.

consolidation, which can be exploited to identify COVID-19 cases. Chest CTs have thus been utilized to diagnose COVID-19 in some countries and regions with reported sensitivity ranging from 0.56 to 0.98 (refs. ^{12–15}); however, these radiological features are not explicitly tied to COVID-19, and the accuracy of CT-based diagnostic tools heavily depends on the radiologists's own knowledge and experience. A recent study¹⁶ has further investigated the substantial discrepancies in differentiating COVID-19 from other viral pneumonia by different radiologists. Such inconsistency is undesirable for any clinical decision system; there is thus an urgent demand to develop an accurate and automatic method to help address the clinical deficiency in current CT-based approaches.

Successful development of an automated method relies on a sufficient amount of data accompanied by precise annotations. We identified three challenges—specifically data-related—for developing a robust and generalized AI model for CT-based COVID-19 identifications. (1) Incompleteness. The high-quality CTs that were used for training were only a small subset of the entire cohort and therefore unlikely to cover the complete set of useful radiological features for identification. (2) Isolation. The CTs acquired across multiple centres were difficult to transfer for training due to security and privacy concerns, whereas a locally trained model may not be generalized to, or improved by, the data collected from other sites. (3) Heterogeneity. Due to the different acquisition protocols (for example, contrast agents and reconstruction kernels), CTs collected from a single hospital are still not yet well standardized; it is therefore challenging to train a precise model on the basis of a simple combination of data¹⁷.

Furthermore, it remains an open question whether the patients with COVID-19 from diverse geographies and varying demographics

show similar or distinct patterns. All of these challenges will impede the development of a well-generalized AI model, and thus, of a global intelligent clinical solution. It is worth noting that these challenges are generally encountered by all of the possible trails in applying AI models in clinical practices, not necessarily COVID-19 related.

We launched the Unified CT-COVID AI Diagnostic Initiative (UCADI; Figs. 1 and 2) to tackle these problems. It was developed on the basis of the concept of federated learning^{18,19}, which enables machine learning engineers and clinical data scientists to collaborate seamlessly without sharing the patient data; thus, in UCADI, every participating institution can benefit from and contribute to the continuously evolving AI model, helping deliver even more precise diagnoses for COVID-19 and beyond.

Results

Developing a local accurate AI diagnostic model. Training an accurate AI model requires comprehensive data collection. We therefore first gathered, screened and anonymized the chest CTs at each institute participating in UCADI (five hospitals in China and 18 hospitals in the United Kingdom), comprising a total of 9,573 CTs from 3,336 patients. We summarized the demographics and diagnoses of the cohort in Supplementary Tables 1 and 2.

Developing an accurate diagnostic model requires a sufficient amount of high-quality data. Consequently, we identified the three branches of Wuhan Tongji Hospital Group (Main Campus, Optical Valley and Sino-French) and the National COVID-19 Chest Imaging Database (NCCID)²⁰ as individual UCADI participants. Each site contains adequate high-quality CTs for the development of the three-dimensional convolutional neural network (CNN) model.

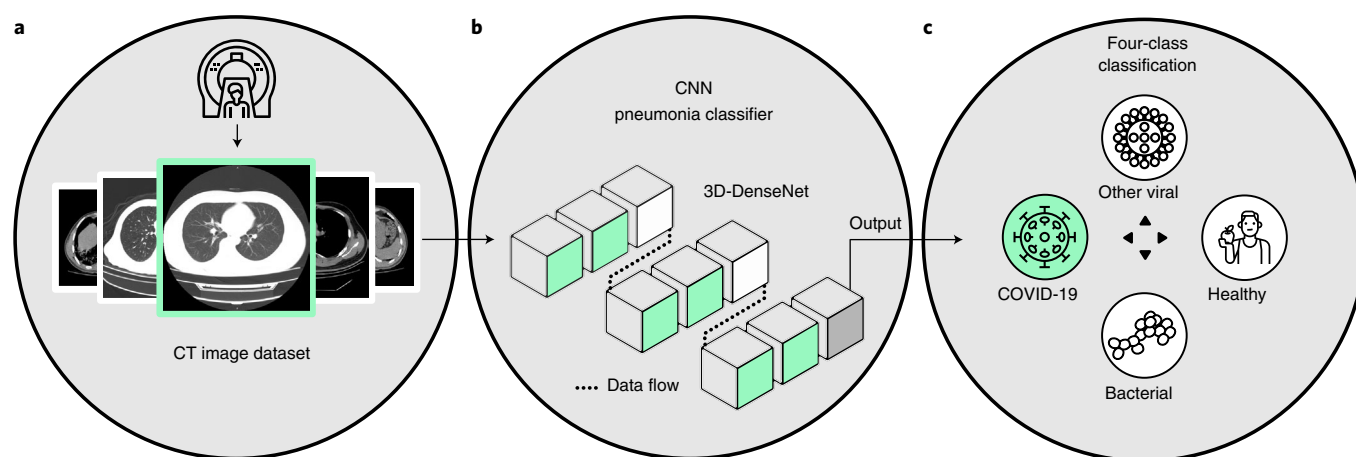


Fig. 2 | Deployment and workflow of UCADI participants. **a**, Data: construct a local dataset based on the high-quality, well-annotated and anonymized CTs. **b**, Flow: the backbone of the 3D-DenseNet model mainly consists of six three-dimensional dense blocks (in green), two three-dimensional transmit blocks (in white) and an output layer (in grey). Computed tomography scans of each case are converted into a (16,128,128) tensor after adaptive sampling, decentralization and trilinear interpolation, and then fed into the three-dimensional CNN model for pneumonia classification. **c**, Process: during training, the model outputs are used to calculate the weighted cross-entropy to update the network parameters. While testing, five independent predictions of each case are incorporated to report the predictive diagnostic results.

We used 80% of the data for training and validation (hereafter referred to as trainval) and the remaining 20% for testing. We also utilize the CTs collected from Tianyou hospital and Wuhan Union hospital as hold-out test sets. We consistently use the same partition in both the local and federated training processes for a fair comparison.

The NCCID is an initiative established by NHSX (a joint unit of the National Health Service (NHS) England and the Department of Health and Social Care (DHSC)), providing massive CT and chest X-ray modalities of COVID-19 and non-COVID-19 patients from over 18 partnership hospitals in the United Kingdom. As each hospital's data quantity and categorial distribution are quite uneven, we pooled all of the CTs and identified the entire NCCID cohort as a single participant. Unlike the CTs procured from China, which are all non-contrast, around 80% of CTs from NCCID are acquired using contrast materials (for example, iodine). Such materials are usually utilized to block X-rays and appeared with higher attenuation on CTs, which could help emphasize tissues such as blood vessels and intestines (in Supplementary Fig. 1 and Table 3); however, in practice, we found that a simple combination of the contrast and non-contrast CTs did not back the training of a well-generalized model as their intrinsic differences induced in the acquisition procedures²¹. To overcome the data heterogeneity between the contrast and non-contrast CTs in the NCCID, we therefore applied an unpaired image-to-image translation method called CycleGAN²² to transform the contrast CTs into non-contrast variants as augmentations during the local model training. In Supplementary Table 4, we have compared CycleGAN with two other recent image translation methods (CouncilGAN²³ and ACL-GAN²²). We showed that the model trained on CycleGAN transformed contrast CTs has the best performance (test on the non-contrast CTs); however, this modality transformation is not always helpful, as the performance degenerated when training on the raw plus translated contrast CTs.

We developed a densely connected three-dimensional convolutional neural network model—3D-DenseNet—on the basis of the massive cohort collection towards delivering precise diagnoses with AI approaches; we report its architectural designs and training optimizations in the Methods and Supplementary Fig. 2. We examined the predictive power of 3D-DenseNet on a four-class pneumonia classification task as well as COVID-19 identification. In the first task we aimed to distinguish COVID-19 (Fig. 3a, Supplementary Fig. 3 and Table 5) from healthy cases and two other pneumonia

types, namely non-COVID-19 viral and bacterial pneumonia (Fig. 3b). We preferred a four-class taxonomy, as further distinguishing COVID-19 from community-acquired pneumonia^{24,25} can help deliver more commendatory clinical treatments where the bacterial and the viral are two primary pathogens of community-acquired pneumonia²⁶ (Fig. 2c); however, given that different institutions are accompanied by various annotating protocols, it is more feasible for the model to learn to discriminate COVID-19 from all non-COVID-19 cases. We therefore base the experimental results on this two-category classification in the main text. We report the four-class experiments based on the Wuhan Tongji Hospital Group's cohort in Supplementary Fig. 3 and Table 5.

For the three UCADI data centres in China (Main Campus, Optical Valley and Sino-French branches of Wuhan Tongji Hospital Group), the locally trained 3D-DenseNet achieved an average test sensitivity/specificity of 0.804/0.708 for identifying COVID-19. As for the collection from Britain (NCCID), the test sensitivity/specificity (on non-contrast CTs) of the local model can be improved from 0.703/0.961 to 0.784/0.961 with the help of CycleGAN to mitigate the heterogeneity between contrast and non-contrast CTs. We further compared 3D-DenseNet with two other 3D CNN baseline models: 3D-ResNet²⁷ and 3D-Xception²⁸ (Supplementary Tables 6 and 7). As a result, we demonstrated that 3D-DenseNet had better performance and smaller size, presenting it as highly suitable for federated learning.

To interpret the learned features of the model, we performed gradient-weighted class activation mapping (GradCAM)²⁹ analysis on the CTs from the test set. We visualized the featured regions that lead to identification decisions and found that the generated heatmaps (Fig. 3c) primarily characterized local lesions that highly overlap with the radiologists's annotations, suggesting that the model is capable of learning robust radiologic features rather than simply overfitting³⁰. This heatmap can help the radiologists localize the lesions quicker for delivering diagnoses in an actual clinical environment. Moreover, localizing the lesions will also provide a guide for further CT acquisition and clinical testing. A similar idea has been described as region-of-interest detection in a similar study³¹.

To examine the cross-domain generalization ability of the locally trained models, we tested China's locally trained model on Britain's test set and vice versa, reporting the numerical results in Fig. 4; however, due to incompleteness, isolation and heterogeneity in the

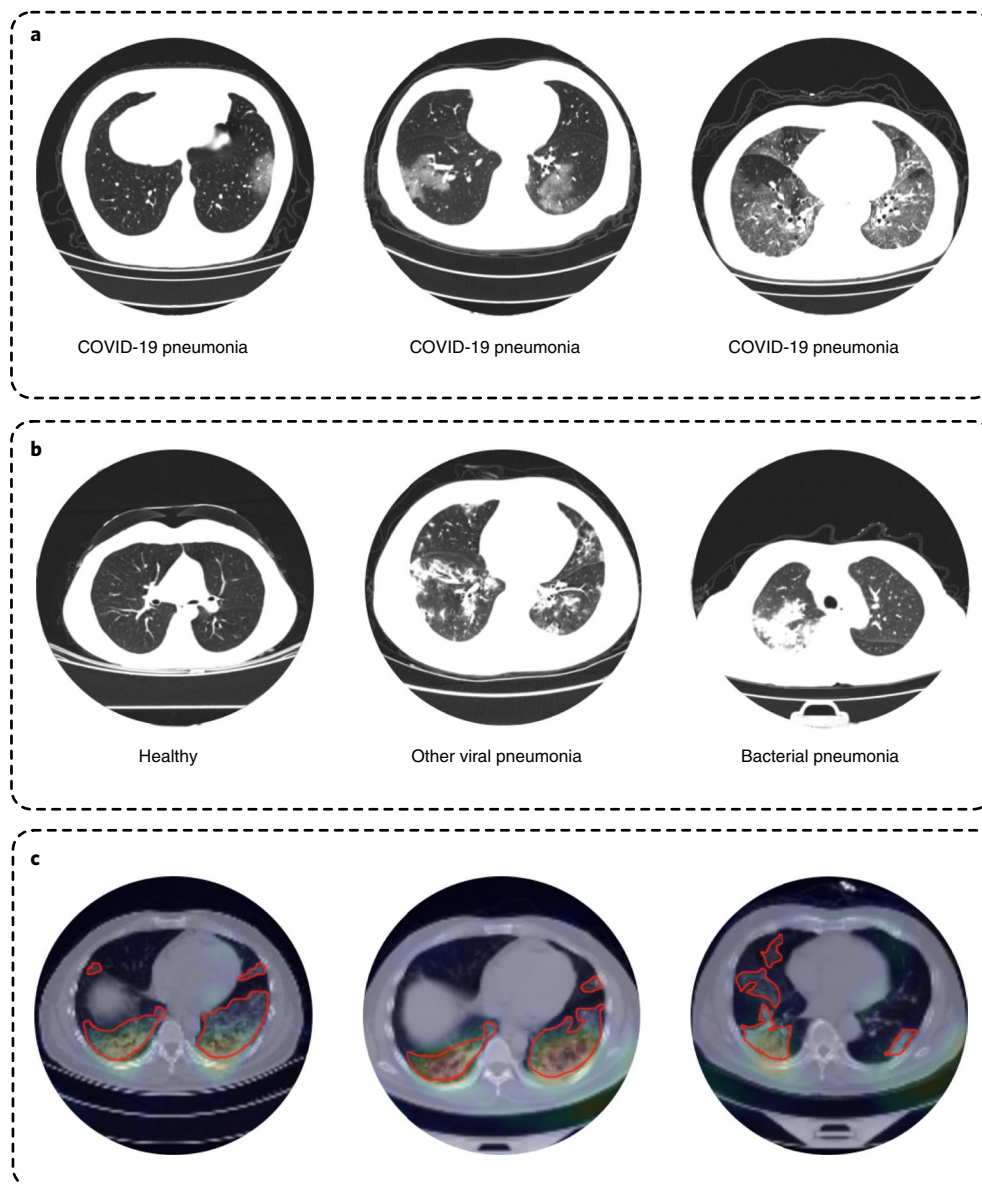


Fig. 3 | Overview of CTs. **a**, Radiological features correlated with COVID-19 pneumonia cases: ground glass opacity, interlobular septal thickening and consolidation are shown from left to right. **b**, Other non-COVID-19 cases, including healthy, other viral and bacterial pneumonia. **c**, Localized class-discriminative regions generated by GradCAM (in the heatmap) and annotated by professional radiologists (circled in red), for COVID-19 cases.

various data resources, we found that all of the locally trained models exhibited less-than-ideal test performances on other sources. Specifically, the model trained on NCCID non-contrast CTs had a sensitivity/specificity/AUC of 0.313/0.907/0.745 in identifying COVID-19 on the test set of China, which is lower than locally trained ones, and vice versa. We next describe how to incorporate federated learning for the cross-continent privacy-preservation collaboration on training a generalized AI diagnostic model, mitigating the domain gaps and data heterogeneity.

Enable multination privacy-preserving collaboration with federated learning

We developed a federated learning framework to facilitate the collaboration nested under UCADI and NCCID, integrating diverse cohorts as part of a global joint effort on developing a precise and robust AI diagnostic tool. In traditional data science approaches^{17,31}, sensitive and private data from different sources are directly gathered and transported to a central hub where the models are deployed;

however, such procedures are infeasible in real clinical practices as hospitals are usually reluctant (and often not permitted) to disclose data due to privacy concerns and legislation³². On the other side, the federated learning technique proposed by Google³³, by contrast, is an architecture in which the AI model is distributed to and executed at each host institution without data centralization. Furthermore, transmitting the model parameters effectively reduced the latency and the cost associated with sending large amounts of data during internet connections. More importantly, the strategy to preserve privacy by design enables medical centres to collaborate on developing models without sharing sensitive clinical data with other institutions. Swarm Learning³⁴ was recently proposed towards model decentralization via edge computation; however, we conjecture that it is immature for the privacy-preserving machine learning³⁵ applications based on massive data collection and participants due to the exponential increase in computation.

With UCADI, we have provided: (1) an online diagnostic interface allowing people to query the diagnostic results on identifying

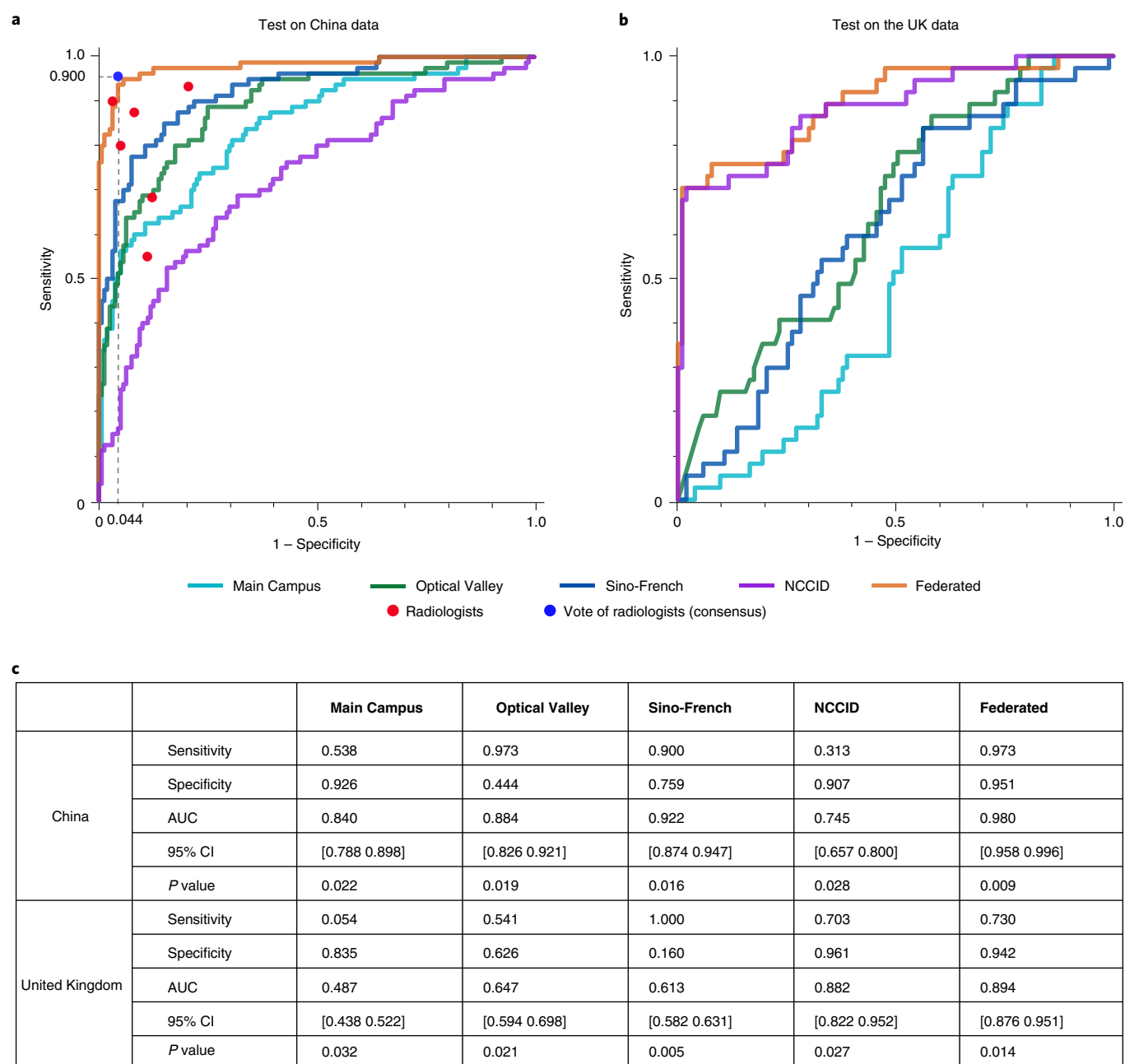


Fig. 4 | COVID-19 pneumonia identification performance of three-dimensional CNN models trained on four different data resources (Main Campus, Optical Valley, Sino-French and NCCID) individually and federatively. **a**, Receiver operating characteristic curves when the models are tested on the data from China, in comparison with six professional radiologists, **b**, Receiver operating characteristic curves of the CNN models tested on the data from the United Kingdom. **c**, Numeric results of the test sensitivity, specificity and area under the curve (AUC, with 95% confidence intervals and *P*-values).

COVID-19 by uploading their chest CTs; and (2) a federated learning framework that enables UCADI participants to collaboratively contribute to improving the AI model for COVID-19 identification. Each UCADI participant will send the model weights back to the server via a customized protocol during the collaborative training process every few iterations. To further mitigate the potential for data leaks during such a transmission process, we applied an additive homomorphic encryption method called Learning with Errors (LWE)³⁶ to encrypt the transmitted model parameters. By doing so, participants will keep within their data and infrastructure, with the central server having no access whatsoever. After receiving the transmitted packages from the UCADI participants, the central

server then aggregates the global model without comprehending the model parameters of each participant. The updated global model would then be distributed to all participants, again utilizing LWE encryption, enabling the continuation of the model optimization at the local level. Our framework is designed to be highly flexible, allowing dynamic participation and breakpoint resumption (detailed in the Methods).

With this framework, we deployed the same experimental configurations to validate the federated learning concept for developing a generalized CT-based COVID-19 diagnostic model (detailed in the Methods). We compared the test sensitivity and specificity of the federated model to the local variations (Fig. 4). We plotted

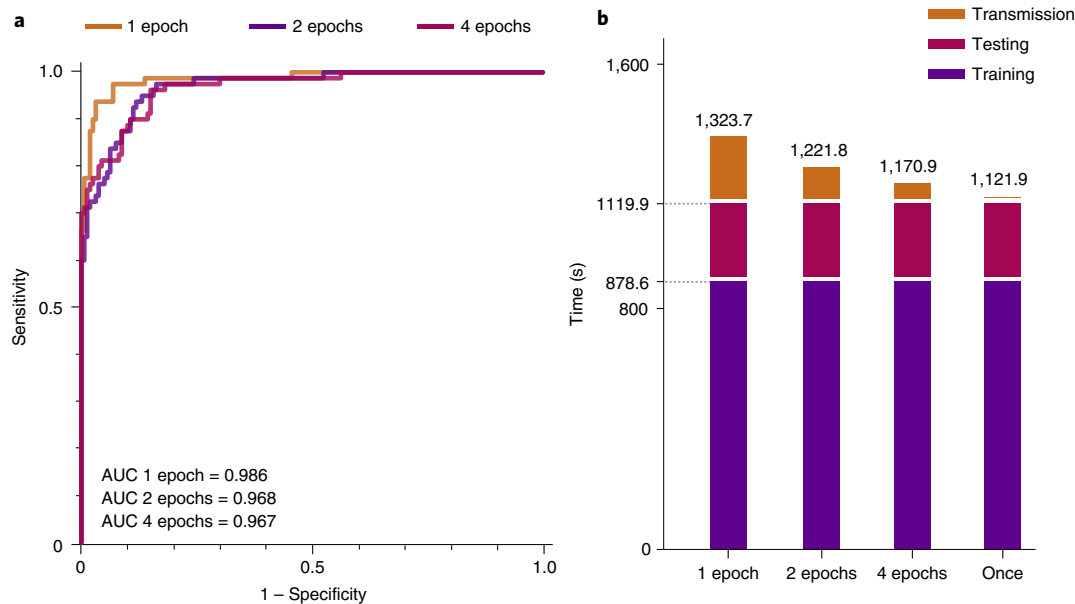


Fig. 5 | Trade-off on the performance and communication cost in federated training. **a**, Relationships between transmission expense and model generalization. **b**, Estimated time spent at different communication/synchronization intervals. The statistics is measured based on a joint FL training of two clients. Each client has 200 CTs and 100 CTs for training and testing, respectively. The client's software infrastructure is a single-core of GPU (NVIDIA GTX 1080Ti) and a CPU (Xeon(R) CPU E5-2660 v4 @ 2.00 GHz). The bandwidth for transmission is around 7.2 Mb s^{-1} (900 KB s^{-1}), which is the average broadband speed.

the receiver operating characteristic curves and calculated the corresponding AUC scores—along with 95% confidence intervals and *P*-values—to validate the model's performance (Fig. 4). As confirmed by the curves and numbers, the federated model outperformed all of the locally trained ones on the same test splits collected from China and the United Kingdom. Specifically, for the test performance on the 1,076 CTs of 254 cases in China (all from the three branches of Wuhan Tongji Hospital Group), the federated model achieved a sensitivity/specificity/AUC of 0.973/0.951/0.980, respectively, outperforming the models trained locally at Main Campus, Optical Valley, Sino-French and NCCID. Furthermore, the federated model achieves a sensitivity/specificity/AUC of 0.730/0.942/0.894 for COVID-19 classification when applied to the test set of the NCCID (from 18 UK hospitals), vastly outperforming all the locally trained models. We based the performance measure on the CT level instead of the patient level, coherent with the prior study³¹.

We illustrated that the federated framework is an effective solution to mitigate against the issue that we cannot centralize medical data from hospitals worldwide due to privacy and legal legislation. We further conducted a comparative study on the same task with a panel of expert radiologists. With an average of nine years's experience, six qualified radiologists from the Department of Radiology, Wuhan Tongji Hospital (Main Campus) were asked to make diagnoses on each CT from China, as one of the four classes. The six experts were first asked to provide diagnoses individually, then to address integrated diagnostic opinions via majority votes (consensus) in a plenary meeting. We presented the radiologists and AI models with the same data partition for a fair comparison. In differentiating COVID-19 from the non-COVID-19 cases, the six radiological experts obtained an average 0.79 in sensitivity (0.88, 0.90, 0.55, 0.80, 0.68, 0.93, respectively), and 0.90 in specificity (0.92, 0.97, 0.89, 0.95, 0.88, 0.79, respectively). In reality, the consideration of a clinical decision is usually made by consensus decision among the experts. Here we use the majority votes among the six expert radiologists to represent such a decision-making process. We provide the

detailed diagnostic decisions of each radiologist in Supplementary Table 5. We found that the majority vote helps reduce the potential bias and risk: the aggregated diagnoses are with the best performance among individual radiologists. In Fig. 4a, we plotted the majority votes in blue markers (sensitivity/specificity: 0.900/0.956) and remarked that the federatively trained 3D-DenseNet had shown comparable performance (sensitivity/specificity: 0.973/0.951) with the expert panel. We have further presented and discussed the model's performance on the hold-out test sets (645 cases from Wuhan Tianyou Hospital and 506 cases from Wuhan Union Hospital) in Supplementary Table 8. We proved that the federatively trained model also performed better on these two hold-out datasets, yet the confidence sometimes is not well calibrated.

During the federated training process, each participant is required to synchronize the model weights with the server every few training epochs using web sockets. Intuitively, more frequent communication should lead to better performance. However, each synchronization accumulates extra time. To investigate the trade-off between the model performance and the communication cost during the federated training, we conduct parallel experiments with the same settings but different training epochs between the consecutive synchronizations. We report the models's subsequent test performance in Fig. 5a and time usage in Fig. 5b. We observe that, as expected, more frequent communication leads to better performance. Compared with the least frequently communication scenario, to download the model from the beginning and train locally without intermediate communications, synchronizing at every epoch will achieve the best test performance with less than 20% increment in time usage.

Discussion

COVID-19 is a global pandemic. Over 200 million people have been infected worldwide, with hundreds of thousands hospitalized and mentally affected^{37,38}, and above four million are reported to have died as of October 2021. There are borders between countries, yet the only barrier is the boundary between humankind and the

virus. We urgently demand a global joint effort to confront this illness effectively. In this study, we introduced a multinational collaborative AI framework, UCADI, to assist radiologists in streamlining and accelerating CT-based COVID-19 diagnoses. First, we developed a new CNN model that achieved performance comparable with expert radiologists in identifying COVID-19. The predictive diagnoses can be utilized as references while the generated heatmap helps with faster lesion localization and further CT acquisition. We then formed a federated learning framework to enable the global training of a CT-based model for precise and robust diagnosis. With CT data from 22 hospitals, we have herein confirmed the effectiveness of the federated learning approach. We have shared the trained model and open-sourced the federated learning framework. It is worth mentioning that our proposed framework is with continual evolution, is not confined to the diagnosis of COVID-19 but also provides infrastructures for future use. The uncertainty and heterogeneity are the characteristics of clinical work. Due to the limited medical understanding of the vast majority of diseases, including pathogenesis, pathological process, treatment and so on the medical characteristics of diseases can be studied by the means of AI. Along with this venue, research can be more instructive and convenient in dealing with large (sometimes isolated) samples, especially suitable for transferring knowledge in studying emerging diseases.

However, certain limitations are not well addressed in this study. First is the potential bias in the comparison between experts and models. Due to legal legislation, it is infeasible and impossible to disclose the UK medical data with radiologists and researchers in China or vice versa. Radiologists are thus from nearby institutions. Though their diagnostic decisions are quite different, it is not unrealistic to conclude that our setting and evaluation process eliminate biases. The second is engineering efforts. Although we have developed mechanisms such as dynamic participation and breakpoint resumption, the participants still happened to drop from the federated training process for the unstable internet connection. Also, the computation efficiency of the three-dimensional CNN model still has space for improvements (in Supplementary Table 7). There are always engineering advancements that can be incorporated to refine the framework.

Methods

We first describe how we constructed the dataset and then discuss the details of our implementations for collaboratively training the AI model, we provided further analysis of our methods at the end of this section.

CN dataset development (UCADI). A total of 5,740 chest CT images that are acquired from the three branches (Main Campus, Optical Valley and Sino-French) of Tongji Hospital Group located in Wuhan, China, using similar acquisition protocols. Three scanners are used to obtain the CTs: GE Medical System/LightSpeed16, GE Medical Systems/Discovery 750 HD and Siemens SOMATOM Definition AS+. The scanning slice thickness is set as 1.25 mm and 1 mm for the GE and the Siemens scanners, respectively. The reconstruction protocols include a statistical iteration (60%) and sinogram affirmed iteration for the GE and the Siemens devices, respectively. All of the Chinese-derived CTs are taken without the intravenous injection of iodine contrast agent (that is, non-contrast CTs). Regarding the acquisition date, 2,723 CTs of the 432 patients with COVID-19 were enrolled, selected and annotated from 7 January 2020; 3,017 CTs from other three categories were then retrieved from the databases of these three hospitals, with an event horizon dating back to 2016.

As detailed in the Supplementary Information, the chest CTs were then divided into a training/validation (hereafter: trainval) split of 1,095 cases, and a testing split of 254 cases. The trainval split consists of 342 cases (1,136 CTs) for healthy individuals, 405 cases (2,200 CTs) for those COVID-19 positive, 56 cases (250 CTs) for other viral pneumonia and 292 cases (1,078 CTs) for bacterial pneumonia. For the test split, we considered a balanced distribution over the four classes, consisting of 80 cases (262 CTs) for healthy individuals, 94 cases (523 CTs) for the COVID-19-positive instances, 20 cases (84 CTs) for other viral pneumonia and 60 cases (207 CTs) for bacterial pneumonia. Specifically, the virus types that are regarded as other viral pneumonia include respiratory syncytial, Epstein-Barr, cytomegalovirus, influenza A and parainfluenza.

We also collected independent cohorts including 507 COVID-19 cases from Wuhan Union Hospital and 645 COVID-19 cases from Wuhan Tianyou Hospital.

These hold-out test sets were used for testing the generalization of the locally trained models as well as the federated model. As the data source only contained COVID-19 cases, we did not utilize it during the training process. We also summarized and reported the demographic information (that is, gender and age) of the cohort in Supplementary Table 1.

UK dataset development (NCCID). For the total 2,682 CTs that were acquired from the 18 partner hospitals located in the United Kingdom (Supplementary Table 3), the acquisition devices and protocols varied from hospital to hospital. There are over 14 types of utilized CT scanners: Siemens Sensation 64; Siemens SOMATOM Drive; Siemens SOMATOM Definition AS/AS+/Edge/Flash; GE Medical Systems Optima CT660; GE Medical Systems Revolution CT/EVO; GE Medical Systems LightSpeed VCT; Canon Medical Systems Aquilion ONE; Philips Ingenuity Core 128 and Toshiba Aquilion ONE/PRIME. Settings such as filter sizes, slice thickness and reconstruction protocols are also quite diverse among these CTs. This might explain the reason why the NCCID locally trained model failed to perform as well as the Chinese locally trained variant (Fig. 4c). Regarding the material differences, 2,145 out of 2,682 CTs were taken after the injection of an iodine contrast agent (that is, contrast CTs). As pointed out by previous study²¹, contrast and non-contrast CTs have different feature distributions in terms of attenuation and brightness; it is therefore infeasible to simply mix all the CTs together for local or federated training. The reported numbers in Fig. 3 are based on the non-contrast CTs, while in Supplementary Table 3, we used CycleGAN²² to incorporate both contrast and non-contrast CTs, and shall elaborate upon such settings in the following section.

As detailed in Supplementary Information, CTs from NCCID were first partitioned into two types: contrast and non-contrast. Such division is based on the metadata provided in the CTs as well as validated from the professional radiologists. For the contrast CTs, the trainval produces a split of 421 cases, and a testing split of 243 cases. The trainval split consists of 276 cases (1,097 CTs) for non-COVID-19 and 145 cases (491 CTs) for the COVID-19 positive cases. The test split contains 160 cases (259 CTs) for non-COVID-19 and 83 cases (138 CTs) for the COVID-19 positives. The non-contrast CTs is fewer in quantity compared with the contrast ones. It has 116 cases (394 CTs) for non-COVID-19 and 54 cases (163 CTs) for the COVID-19 positive cases. Moreover, there are 75 cases (103 CTs) for non-COVID-19 and 27 cases (37 CTs) for the COVID-19 positive cases for the test split.

We also noticed that a small subset of the CTs only contained partial lung regions, we removed these insufficient CTs whose number of slices are less than 40. As for our selection criteria in this regard, although the partial lung scans might be infeasible for training segmentation or detection models, we believe that a sufficient number of slices is enough to ensure the model effectively captures the requisite features and thereby help with the precise classification in medical diagnosis.

We reported patient demographical information (that is, gender and age) of the cohort in Supplementary Table 2. However, the reported demographics is not inclusive since the demographical attributes of non-COVID-19 cases are not recorded. In comparison to the demographical information of the COVID-19 cases acquired from China, COVID-19 cases in the United Kingdom were with larger averaged ages and had more male patients. These demographical differences might also explain why the United Kingdom locally trained model failed to perform well when applied to the CTs acquired from China.

Data preprocessing, model architecture and training setting. We pre-processed the raw acquired CTs for standardization as well as to reduce the burden on computing resource. We utilized an adaptive sampling method to select 16 slices from all sequential images of a single CT case using random starting positions and scalable transversal intervals. During the training and validation process, we sampled once for each CT study, while in testing we repeated the sampling five independent times to obtain five different subsets. We then standardized the sampled slices by removing the channel-wise offsets and rescaling the variation to uniform units. During testing, the five independent subsets of each case were fed to the trained CNN classifier to obtain the prediction probabilities of the four classes. We then averaged the predictive probabilities over these five runs to make the final diagnostic prediction for that case. By so doing, we can effectively include impacts from different levels of lung regions as well as to retain scalable computations. To further improve the computing efficiency, we utilized trilinear interpolation to resize each slice from 512 to 128 pixels along each axis and rescaled the lung windows to a range between -1,200 and 600 Hounsfield units before feeding into the network model.

We named our developed model 3D-DenseNet (Supplementary Fig. 2). It was developed based on DenseNet³⁹, a densely connected convolutional network model that performed remarkably well in classifying two-dimensional images. To incorporate such design with the three-dimensional CT representations, we adaptively customized the model architecture into fourteen three-dimensional convolution layers distributed in six dense blocks and two transmit blocks (insets of Supplementary Fig. 2). Each dense block consists of two three-dimensional convolution layers and an inter-residual connection, whereas the transmit blocks are composed of a three-dimensional convolution layer and an average pooling

layer. We placed a 3D DropBlock⁴⁰ instead of simple dropout⁴¹ before and after the six dense blocks, which proved to be more effective in regularizing the training of convolution neural networks. We set the momentum of batch normalization⁴² to be 0.9, and the negative slope of LeakyReLU activation as 0.2.

During training, the 3D-DenseNet took the pre-processed CT slice sequences as the input, then output a prediction score over the four possible outcomes (pneumonia types). Due to the data imbalance, we defined the loss function as the weighted cross-entropy between predicted probabilities and the true categorical labels. The weights were set as 0.2, 0.2, 0.4, 0.2 for healthy, COVID-19, other viral pneumonia and bacterial pneumonia cases, respectively. We utilized SGD optimizer with a momentum of 0.9 to update parameters of the network via backpropagation. We trained the networks using a batch size of 16. At the first five training epochs, we linearly increased the learning rate to the initial set value of 0.01 from zero. This learning rate warm-up heuristic proved to be helpful, as using a large learning rate at the very beginning of the training may result in numerical instability⁴³. We then used cosine annealing⁴⁴ to decrease the learning rate to zero over the remaining 95 epochs (100 epochs in total).

During both local and federated training processes, we utilized a fivefold cross-validation on trainval split, and then selected the best model and reported their test performance (in Fig. 4 and Supplementary Fig. 2).

Federated learning and privacy preservation. At the central server, we adapted the FedAvg³³ algorithm to aggregate the updated model parameters from all clients (UCADI participants) to combine the weights with respect to clients's dataset sizes and the number of local training epochs between consecutive communications. To ensure secure transmissions between the server and the clients, we used LWE³⁶ to further protect all the transmitted information (that is, model parameters and metadata). The LWE method is an additively homomorphic variant of the public key encryption scheme, therefore the participant information cannot even leak to the server, which is to say, that the server has no access to the explicit weights of the model. Compared with other encryption methods, such as differential privacy⁴⁵, Moving Horizon Estimations⁴⁶ and Model Predictive Control⁴⁷, LWE differentiates itself by essentially enabling the clients to achieve identical performance with the variants trained without decryption; however, the LWE method would add further costs to the federated learning framework in terms of the extra encryption/decryption process and the increased size of the encrypted parameters during transmission. The typical time usage of a single encryption-decryption round is 2.7 s (average over 100 trials under a test environment consisting of a single CPU (Intel Xeon E5-2630 v3 @ 2.40 GHz) and the encrypted model size arises from 2.8 MB to 62 MB, which increases the transmission time from 3.1 s to 68.9 s, in a typical international bandwidth environment⁴⁸ of 900 KB s⁻¹ (Fig. 5).

Comparing with professional radiologists. We further conducted a comparative study on this four-type classification between the CNN model and expert radiologists. We asked six qualified radiologists (with an average of nine years's clinical experience, ranging from four to eighteen years) from the Tongji Hospital Group to make the diagnoses on the basis of the CTs. We provided the radiologists with the CTs and their labels from the China-derived trainval split. We then asked them to diagnose each CT from the test split into one of the four classes. We reported the performance of each single radiologist and the majority votes on the COVID-19 versus non-COVID-19 CTs in Fig. 4 (detailed comparisons are presented in Supplementary Table 5 and 9). If there are multiple majority votes for different classes, the radiologist panel will make further discussions until reaching a consensus.

Augmented contrast/non-contrast CTs with CycleGAN. Following similar procedures as previous work²¹, we first extracted and converted the slices from contrast and non-contrast CTs of NCCID into JPEG format images with a resolution of 512 px × 512 px. The trainval and test splits of the contrast CTs contain 932 images (23 cases) and 139 images (22 cases), respectively. For the non-contrast CTs, there are 1,233 images (26 cases) and 166 images (26 cases) for the trainval and test splits, respectively. For the architecture of the CycleGAN, we use ResNet⁴⁹ backbone as the feature encoder and set the remaining parts in concordance with the original literature²¹. For the training settings of CycleGAN, we used a batch size of 12 for the total number of 200 epochs. We used the same settings on the trade-off coefficients in the adversarial loss. We started with a learning rate of 2×10^{-4} , kept it constant for the first 100 epochs, then decayed it to zero linearly over the next 100 epochs.

To evaluate the effectiveness of utilizing CycleGAN for augmentation, we first trained the 3D-DenseNet on trainval set of: (1) only non-contrast; (2) non-contrast and CycleGAN synthesized non-contrast; (3) only contrast; and (4) contrast and CycleGAN synthesized contrast CTs. In Supplementary Table 3, we reported the test performance of these trained models on the non-contrast and contrast CTs respectively. We observed that augmenting the non-contrast CTs with CycleGAN would result in a better identification ability of the model while this was not held when converting the non-contrast ones into contrast.

Ethics approval. The UK data used in this study is under approval by Control of Patient Information (COP) notice issued by The Secretary of State for Health

and Social Care. The CN data usage is approved by the Ethics Committee Tongji Hospital, Tongji Medical College of Huazhong University of Science and Technology.

Data availability

The clinical data collected from the 23 hospitals that utilized in this study remains under their custody. Part of the data are available via applications from qualified teams. Please refer to the NCCID website (<https://www.nhs.uk/covid-19-response/data-and-covid-19/national-covid-19-chest-imaging-database-nccid/>) for more details.

Code availability

The online application to join UCADI is provided at <http://www.covid-ct-ai.team>. Codes are publicly available at <https://github.com/HUST-EIC-AI-LAB/UCADI> (ref. 50), which is released under a Creative Commons Attribution-NonCommercial 3.0 Unported License (CC BY-NC 3.0).

Received: 2 March 2021; Accepted: 27 October 2021;
Published online: 15 December 2021

References

1. Ai, T. et al. Correlation of chest CT and RT-PCR testing for coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases. *Radiology* **296**, E32–E40 (2020).
2. Fang, Y. et al. Sensitivity of chest CT for COVID-19: comparison to RT-PCR. *Radiology* **296**, E115–E117 (2020).
3. Kanne, J. P., Little, B. P., Chung, J. H., Ellicker, B. M. & Ketani, L. H. Essentials for radiologists on COVID-19: an update-radiology scientific expert panel. *Radiology* **296**, E113–E114 (2020).
4. Kucirka, L. M., Lauer, S. A., Laeyendecker, O., Boon, D. & Lessler, J. Variation in false-negative rate of reverse transcriptase polymerase chain reaction-based SARS-CoV-2 tests by time since exposure. *Ann. Intern. Med.* **173**, 262–267 (2020).
5. Ackerman, C. M. et al. Massively multiplexed nucleic acid detection with Cas13. *Nature* **582**, 277–282 (2020).
6. Chen, J. S. et al. CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity. *Science* **360**, 436–439 (2018).
7. Broughton, J. P. et al. CRISPR-Cas12-based detection of SARS-CoV-2. *Nat. Biotechnol.* **38**, 870–874 (2020).
8. He, J. L. et al. Diagnostic performance between CT and initial real-time RT-PCR for clinically suspected 2019 coronavirus disease (COVID-19) patients outside Wuhan, China. *Respir. Med.* **168**, 105980 (2020).
9. Ravi, N., Cortade, D. L., Ng, E. & Wang, S. X. Diagnostics for SARS-CoV-2 detection: a comprehensive review of the FDA-EUA COVID-19 testing landscape. *Biosens. Bioelectron.* **165**, 112454 (2020).
10. Mei, X. et al. Artificial intelligence-enabled rapid diagnosis of patients with COVID-19. *Nat. Med.* **26**, 1224–1228 (2020).
11. Cleverley, J., Piper, J. & Jones, M. M. The role of chest radiography in confirming COVID-19 pneumonia. *BMJ* **370**, m2426 (2020).
12. Vaseghi, G. et al. Clinical characterization and chest CT findings in laboratory-confirmed COVID-19: a systematic review and meta-analysis. Preprint at *medRxiv* <https://doi.org/10.1101/2020.03.05.20031518> (2020).
13. Shi, H. et al. Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: a descriptive study. *Lancet Infect. Dis.* **20**, 425–434 (2020).
14. Kanne, J. P. & Chest, C. T. findings in 2019 novel coronavirus (2019-nCoV) infections from Wuhan, China: key points for the radiologist. *Radiology* **295**, 16–17 (2020).
15. Chung, M. et al. CT imaging features of 2019 novel coronavirus (2019-nCoV). *Radiology* **295**, 202–207 (2020).
16. Bai, H. X. et al. Performance of radiologists in differentiating COVID-19 from non-COVID-19 viral pneumonia at chest CT. *Radiology* **296**, E46–E54 (2020).
17. Harmon, S. A. et al. Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets. *Nat. Commun.* **11**, 4080 (2020).
18. Konečný, J. et al. Federated learning: strategies for improving communication efficiency. Preprint at <https://arxiv.org/abs/1610.05492> (2016).
19. Bonawitz, K. et al. Towards federated learning at scale: system design. Preprint at <https://arxiv.org/abs/1902.01046> (2019).
20. National COVID-19 Chest Image Database (NCCID) (NHSX, 2020); <https://nhsx.github.io/covid-chest-imaging-database/>
21. Sandfort, V., Yan, K., Pickhardt, P. J. & Summers, R. M. Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks. *Sci. Rep.* **9**, 16884 (2019).
22. Zhu, J. Y., Park, T., Isola, P. & Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. IEEE International Conference on Computer Vision* 2242–2251 (IEEE, 2017).

23. Nizan, O. & Tal, A. Breaking the cycle-colleagues are all you need. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 7860–7869 (IEEE, 2020).
24. Baek, M. S. et al. Clinical and radiological findings of adult hospitalized patients with community-acquired pneumonia from SARS-CoV-2 and endemic human coronaviruses. *PLoS ONE* **16**, e0245547 (2021).
25. Li, L. et al. Using artificial intelligence to detect COVID-19 and community-acquired pneumonia based on pulmonary CT: evaluation of the diagnostic accuracy. *Radiology* **296**, E65–E71 (2020).
26. Cao, B. et al. Diagnosis and treatment of community-acquired pneumonia in adults: 2016 clinical practice guidelines by the Chinese Thoracic Society, Chinese Medical Association. *Clin. Respir. J.* **12**, 1320–1360 (2018).
27. Kataoka, H., Wakamiya, T., Hara, K. & Satoh, Y. Would mega-scale datasets further enhance spatiotemporal 3D CNNs? Preprint at <https://arxiv.org/abs/2004.04968> (2020).
28. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 1251–1258 (IEEE, 2017).
29. Selvaraju, R. R. et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision* 618–626 (IEEE, 2017).
30. DeGrave, A. J., Janizek, J. D. & Lee, S.-I. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nat. Mach. Intell.* **3**, 610–619 (2021).
31. Ardila, D. et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat. Med.* **25**, 1319–1319 (2019).
32. *Reform of EU Data Protection Rules* (European Commission, 2018).
33. McMahan, H. B., Moore, E., Ramage, D., Hampson, S. & Arcas, B. A. Y. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics* 1273–1282 (PMLR, 2017).
34. Warnat-Herresthal, S. et al. Swarm learning for decentralized and confidential clinical machine learning. *Nature* **594**, 265–270 (2021).
35. Kaissis, G. et al. End-to-end privacy preserving deep learning on multi-institutional medical imaging. *Nat. Mach. Intell.* **3**, 473–484 (2021).
36. Phong, L. T., Aono, Y., Hayashi, T., Wang, L. & Moriai, S. Privacy-preserving deep learning via additively homomorphic encryption. *IEEE Trans. Inf. Forensics Secur.* **13**, 1333–1345 (2018).
37. Li, L. Z. & Wang, S. Prevalence and predictors of general psychiatric disorders and loneliness during COVID-19 in the United Kingdom. *Psychiatry Res.* **291**, 113267 (2020).
38. Holman, E. A., Thompson, R. R., Garfin, D. R. & Silver, R. C. The unfolding COVID-19 pandemic: a probability-based, nationally representative study of mental health in the United States. *Sci. Adv.* **6**, eabd5390 (2020).
39. 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* 2261–2269 (IEEE, 2017).
40. Ghiasi, G., Lin, T. Y. & Le, Q. V. Dropblock: a regularization method for convolutional networks. *Adv. Neural Inf. Process. Syst.* **31**, 10727–10737 (2018).
41. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
42. Ioffe, S. & Szegedy, C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning* 448–456 (JMLR, 2015).
43. He, T. et al. Bag of tricks for image classification with convolutional neural networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 558–567 (IEEE, 2019).
44. Loshchilov, I. & Hutter, F. SGDR: stochastic gradient descent with warm restarts. Preprint at <https://arxiv.org/abs/1608.03983> (2016).
45. Dwork, C. & Roth, A. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* **9**, 211–407 (2013).
46. Michalska, H. & Mayne, D. Q. Moving horizon observers and observer-based control. *IEEE Trans. Automat. Contr.* **40**, 995–1006 (1995).
47. García, C. E., Prett, D. M. & Morari, M. Model predictive control: theory and practice—a survey. *Automatica* **25**, 335–348 (1989).
48. *Yearbook of Statistics: Telecommunication/ICT Indicators 2007–2018* (International Telecommunication Union, 2019).
49. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* Vol. 37, 770–778 (IEEE, 2016).
50. Wang, H. et al. HUST-EIC-AI-LAB/UCADI: COVID-19 diagnosis with federated learning. *Zenodo* <https://doi.org/10.5281/ZENODO.5560136> (2021).

Acknowledgements

This study was supported by the HUST COVID-19 Rapid Response Call (grant nos. 2020kfyXGYJ021, 2020kfyXGYJ031, 2020kfyXGYJ093 and 2020kfyXGYJ094), the National Natural Science Foundation of China (grant nos. 61703171 and 81771801), National Cancer Institute, National Institutes of Health (grant no. U01CA242879) and Thammasat University Research fund under the NRCT (contract no. 25/2561) for the Digital platform for sustainable digital economy development project based on the RUN Digital Cluster collaboration scheme. H.W. acknowledges support from Cambridge Trust, Kathy Xu Fellowship, Centre for Advanced Photonics and Electronics and the Cambridge Philosophical Society. M.R. acknowledges support from AstraZeneca, Intel and the DRAGON consortium. A.W. acknowledges support from a Turing AI Fellowship under grant no. EP/V025379/1, The Alan Turing Institute and the Leverhulme Trust via CFI. C.B.S. acknowledges support from DRAGON, Intel, the Philip Leverhulme Prize, the Royal Society Wolfson Fellowship, the EPSRC grants EP/S026045/1 and EP/T003553/1, EP/N014588/1, EP/T017961/1, the Wellcome Innovator Award RG98755, the Leverhulme Trust project Unveiling the invisible, the European Union Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 777826 NoMADS, the Cantab Capital Institute for the Mathematics of Information and the Alan Turing Institute. The funders did not participate in the research or review any details of this study.

Author contributions

T.X. and X.B. conceived the work. H.W., J.G., Z.F., F.Y. and K.M. contributed to the design and development of the models, software and the experiments. T.X., X.B., H.W., J.G., Z.F., K.M., J.L., M.R. and C.S. interpreted, analysed and presented the experimental results. T.X., H.W., J.G., Z.F., K.M., X.B., Y.X., W.L., A.W., J.L., M.R. and C.S. contributed to drafting and revising the manuscript. All of the authors contributed to the data acquisition and resource allocation.

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/s42256-021-00421-z>.

Correspondence and requests for materials should be addressed to Xiang Bai, Chuangsheng Zheng, Jianming Wang, Zhen Li, Carola Schönlieb or Tian Xia.

Peer review information *Nature Machine Intelligence* thanks Jimeng Sun 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.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021, corrected publication 2022