Title: Advancing COVID-19 Diagnosis with Privacy-Preserving Collaboration in Artificial Intelligence

One sentence summary: An efficient and effective privacy-preserving AI framework is proposed for CT-based COVID-19 diagnosis, based on 9,573 CT scans of 3,336 patients, from 23 hospitals in China and the UK.

Author: Xiang Bai[1,2*#], Hanchen Wang[3#], Liya Ma[1#], Yongchao Xu[1#], Jiefeng Gan[2#], Ziwei Fan[2], Fan Yang[4], Ke Ma[2], Jiehua Yang[2], Song Bai[2], Chang Shu[2], Xinyu Zou[2], Renhao Huang[2], Changzheng Zhang[5], Xiaowu Liu[5], Dandan Tu[5], Chuou Xu[1], Wenqing Zhang[1], Xi Wang6, Anguo Chen[7], Yu Zeng[8], Dehua Yang[9], Ming-Wei Wang[9], Nagaraj Holalkere[10], Neil J. Halin[10], Ihab R. Kamel[11], Jia Wu[12], Xuehua Peng[13], Xiang Wang[14], Jianbo Shao[13], Pattanasak Mongkolwat[15], Jianjun Zhang[16,17], Weiyang Liu[3], Michael Robert[18,19], Zhongzhao Teng[20], Lucian Beer[20], Lorena E. Sanchez[20], Evis Sala[20], Daniel L. Rubin[21], Adrian Weller[3,22], Joan Lasenby[3], Chuangsheng Zheng[4*], Jianming Wang[23*], Zhen Li[1*], Carola Schönlieb[18,22*], Tian Xia[2*]

Affiliations: [1]Department of Radiology, Tongji Hospital and Medical College, Huazhong University of Science and Technology, Wuhan, China. [2]School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China. [3]Department of Engineering, University of Cambridge, Cambridge, UK. [4]Department of Radiology, Union Hospital of Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China. [5]HUST-HW Joint Innovation Lab, Wuhan, China. [6]CalmCar Inc, Suzhou, China. [7]Wuhan Blood Centre, Wuhan, China. [8]MSA Capital, Beijing, China. [9]The National Centre for Drug Screening, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai, China. [10]CardioVascular and Interventional Radiology, Radiology for Quality and Operations, The Cardiovascular Centre at Tufts Medical Centre, Radiology, Tufts University School of Medicine, Medford, USA. [11]Russell H Morgan Department of Radiology & Radiologic Science, Johns Hopkins Hospital & Medicine Institute, Baltimore, USA. [12]Department of Radiation Oncology, School of Medicine, Stanford University, Palo Alto, USA. [13]Department of Radiology, Wuhan Central Hospital, Wuhan, China. [14]Department of Radiology, Wuhan Children's Hospital, Wuhan, China. [15]Faculty of Information and Communication Technology, Mahidol University, Thailand. [16]Thoracic/Head and Neck Medical Oncology, [17]Translational Molecular Pathology, University of Texas MD Anderson Cancer Centre, Houston, USA. [18]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK. [19]Oncology R&D at AstraZeneca, Cambridge, Cambridge, UK. [20]Department of Radiology, University of Cambridge, Cambridge, UK. [21]Department of Biomedical Data Science, Radiology and Medicine, Stanford University, Palo Alto, USA. [22]Alan Turing Institute, London, UK [23]Department of Hepatobiliary Pancreatic Surgery, Affiliated Tianyou Hospital, Wuhan University of Science and Technology, Wuhan, China.

# These authors contributed equally to this work. * Correspondence can be addressed to X.B. (xbai@hust.edu.cn), C.Z. (cszheng@163.com), J.W. (jmwang@163.com), Z.L. (zhenli@hust.edu.cn), C.B. (cbs31@cam.ac.uk), and T.X. (tianxia@hust.edu.cn).

Abstract: Artificial intelligence (AI) provides a promising substitution for streamlining COVID-19 diagnoses, owing to its advantages in the efficient evaluation and generalisation power[1–4]. However, concerns surrounding security and trustworthiness[5] impede the collection process of large-scale representative medical data, posing a considerable challenge for training a generalisable model in clinical practices[6–9]. To address this, we launch the Unified CT-COVID AI Diagnostic Initiative (UCADI, in Fig. 1 and 2), where the AI model can be distributedly trained and independently executed at each host institution under a federated learning framework. Our approach thus facilitates training and evaluating a generalised AI model without any data sharing. Moreover, the transmitted information is further homogenously encrypted to preserve privacy and safeguard potential leakage[10–12]. In this study, we conduct the collaborative training of an accurate and efficient AI model for COVID-19 identification nested under UCADI. Our study is based on 9,573 chest computed tomography scans (CTs) from 3,336 patients collected from 23 hospitals located in China and the UK.

Specifically, we first developed and evaluated a compelling 3D convolutional neural network model based on the locally stored CTs. However, due to the isolation and heterogeneity between different data sources, we found that such trained models showed poor empirical generalisation on the CTs collected at the other sites (test sensitivity/specificity: 0.533/0.540, averaged over four data centres in the UK and China). Therefore, we utilised the FL framework under UCADI to collaboratively train a federated model based on data from the partnership hospitals to resolve such a problem. We showed that the federated model outperformed all the locally trained models by a large margin (test sensitivity/specificity in China: 0.973/0.951, in the UK: 0.730/0.942), and more significantly, it achieved comparable performance with a panel of professional radiologists. We further tested on the hold-out dataset (CTs collected from another two hospitals leave out the federated training), 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.

Introduction

As the gold standard for identifying COVID-19 carriers, reverse transcription-polymerase chain reaction (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[13–16], which results in a considerable number of false negatives. As such, clinicians and researchers have made tremendous efforts searching for alternatives[17–19] and complementary modalities[2,14,20–22] 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 CTs, including ground-glass opacity, interlobular septal thickening, and consolidation, which can be exploited to identify COVID-19 cases. Chest CTs have thus been utilised to diagnose COVID-19 in some countries and regions with reported sensitivity ranging from 0.56 to 0.98[23–26]. However, these radiological features are not explicitly tied to COVID-19, and the accuracy of CT-based diagnostic tools heavily depends on the radiologists' own knowledge and experience. A recent study[27] 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. Therefore, there is 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 generalised AI model for CT-based COVID-19 identifications: **(i) Incompleteness**. High-quality CTs that were used for training was only a small subset of the entire cohort; therefore, they are unlikely to cover the complete set of useful radiological features for identification. **(ii) Isolation**. CTs acquired across multiple centres were difficult to transfer for training due to security and privacy concerns, while a locally trained model may not be generalised to or improved by the data collected from other sites. **(iii) Heterogeneity**. Due to the different acquisition protocols (e.g., contrast agents and reconstruction kernels), CTs collected from a single hospital are still not yet well standardised; therefore, it is challenging to train a precise model based on a simple combination of the data[28].

Furthermore, it remains an open question whether the COVID-19 patients from diverse geographies and varying demographics show similar or distinct patterns. All these challenges will impede the development of a well-generalised AI model, and thus, of a global intelligent clinical solution. It is worth noting that these challenges are generally encountered by all the possible trails in applying AI models in clinical practices, not necessarily COVID-19 related.

To tackle these problems, we launched the Unified CT-COVID AI Diagnostic Initiative (UCADI, in Fig. 1 and 2). It was developed based on the concept of federated learning[29,30], which enables machine learning engineers and clinical data scientists to collaborate seamlessly, yet 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. Therefore, we first gathered, screened, and anonymized the chest CTs at each UCADI participating institute (5 hospitals in China and 18 hospitals in the UK), comprising a total of 9,573 CTs of 3,336 cases. We summarised the demographics and diagnoses of the cohort in the supplementary.

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)[31] as individual UCADI participants. Each site contains adequate high-quality CTs for the development of the 3D CNN model. We used 80% of the data for training and validation (trainval) and the rest 20% for testing. Additionally, we 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.

NCCID is an initiative established by NHSX, providing massive CT and CXR modalities of COVID-19 and non-COVID-19 patients from over 18 partnership hospitals in the UK. Since each hospital's data quantity and categorial distribution are quite uneven, we pooled all 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 with contrast materials (e.g., iodine). These contrast materials are usually utilized to block X-rays and appeared with higher attenuation on CTs, which could help emphasise tissues such as blood vessels and intestines (in Supplementary). However, in practice, we found that a simple combination of the contrast and the non-contrast CTs did not back the training of a well-generalized model since their intrinsic differences induced in the acquisition procedures[32]. Therefore, to overcome the data heterogeneity between the contrast and non-contrast CTs in the NCCID, we applied an unpaired image-to-image translation method called CycleGAN[33] to transform the contrast CTs into non-contrast variants as augmentations during the local model training. In Supplementary, we have compared CycleGAN with two other recent image translation methods (CouncilGAN[34] and ACL-GAN[33]). 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 3D convolutional neural network (CNN) model based on the massive cohort collection towards delivering precise diagnoses with AI approaches. We term it 3D-DenseNet and report its architectural designs and training optimisations in the Methods and Supplementary. 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 at distinguishing COVID-19 (Fig. 3a and Supplementary) from healthy cases and two other pneumonia types, namely non-COVID-19 viral and bacterial pneumonia (Fig. 3b). We preferred a four-class taxonomy since further distinguishment of COVID-19 with community-acquired pneumonia (CAP)[35,36] can help deliver more commendatory clinical treatments, where the bacterial and the viral are two primary pathogens of CAP[37] (Fig. 2c). However, given different institutions accompanied by various annotating protocols, it is more feasible for the model to learn to discriminate COVID-19 from all non-COVID-19 cases. Therefore, we 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' cohort in Supplementary.

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), with the help of CycleGAN to mitigate the heterogeneity between contrast and non-contrast CTs, 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. In Supplementary, we further compared 3D-DenseNet with two other 3D CNN baseline models: 3D-ResNet[38] and 3D-Xception[39]. 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)[40] analysis on the CTs from the test set. We visualised the featured regions that lead to identification decisions. It has been found that the generated heatmaps (Fig. 3c) primarily characterised local lesions that are highly overlapped with the radiologist's annotations, suggesting the model is capable of learning robust radiologic features rather than simply overfitting[41]. This heatmap can help the radiologists localise the lesions quicker for delivering diagnoses in an actual clinical environment. Moreover, localising the lesions will also provide a guide for further CT acquisition and clinical test. A similar idea has been described as "region-of-interest (ROI) detection" in a previous similar study[4].

To examine the cross-domain generalisation ability of the locally trained models, we tested China's locally trained model on Britain's test set and vice versa. We reported the numerical results in Fig. 4. However, due to incompleteness, isolation, and heterogeneity in the various data resources, we found that all 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. Next, we describe how to incorporate federated learning for the cross-continent privacy-preservation collaboration on training a generalised 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[4,28], 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 practises; hospitals are usually reluctant (and often not permitted) for data disclosure due to privacy concerns and legislation[42]. On the other side, the federated learning technique proposed by Google[43], in contrast, is an architecture where the AI model is distributed to and executed at each host institution without data centralisation. 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. Recently, Swarm Learning[44] is proposed towards the model decentralisation via edge computation. However, we conjecture it is immature for the privacy-preserving machine learning[45] applications based on massive data collection and participants due to the exponential increase in computation.

In UCADI, we have provided: (i) An online diagnostic interface allowing people to query the diagnostic results on identifying COVID-19 by uploading their chest CTs; (ii) 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 customised 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) [46] to encrypt the transmitted model parameters. By so doing, 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 utilising LWE encryption, enabling the continuation of the model optimisation at the local level. Our framework is designed to be highly flexible, allowing dynamic participation and breakpoint resumption (detailed in 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 Methods). We compared the test sensitivity and specificity of the federated model to the local variations (Fig. 4). We plotted the ROC curves and calculated the corresponding AUC scores, along with 95% confidence intervals (CI) and p-values, to validate the model's performance (Fig. 4 and Supplementary). As confirmed by the curves and numbers, the federated model outperformed all the locally trained ones on the same test splits collected from China and the UK. 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. In addition, 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 study[4].

We illustrated that the federated framework is an effective solution to mitigate against the issue that we cannot centralise 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 9 years 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 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. 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 models' performance on the hold-out test sets (645 cases from Wuhan Tianyou Hospital and 506 cases from Wuhan Union Hospital) in Supplementary. We proved that the federatively trained model also performed better on these two hold-out datasets.

During the federated training process, each participant is required to synchronise the model weights with the server every few training epochs using web sockets. Intuitively, more frequent communication should lead to better performance. However, each synchronisation 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 synchronisations. We report the models' 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 180 million people have been infected worldwide, with hundreds of thousands hospitalized and mentally affected[47,48], and as of June 2021, above four million are reported to have died. 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 multination collaborative AI framework, UCADI, to assist radiologists in streamlining and accelerating CT-based COVID-19 diagnoses. Firstly, we developed a new CNN model that achieved performance comparable to expert radiologists in identifying COVID-19. The predictive diagnoses can be utilised as references while the generated heatmap helps with faster lesion localisation and further CT acquisition. Then, we 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. Because of the limited medical understanding of the vast majority of diseases, including pathogenesis, pathological process, treatment, etc., 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. Thus, radiologists are 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 3D CNN model still has space for improvements (in Supplementary). There are always engineering advancements that can be incorporated to refine the framework.

Methods

We first described how we constructed the dataset, then we discussed 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 1.25mm and 1mm 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 the Chinese-derived CTs are taken without the intravenous injection of iodine contrast agent (i.e., non-contrast CTs). Regarding the acquisition date, 2,723 CTs of the 432 COVID-19 patients were enrolled, selected, and annotated from January 7, 2020; 3,017 CTs from other three categories were then retrieved from the databases of these three hospitals, with an event horizon going back to 2016.

As detailed in Supplementary, 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.

Additionally, we 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 generalisation of the locally trained models as well as the federated model. Since the data source only contained COVID-19 cases, we did not utilize it during the training process. We also summarised and reported the demographic information (i.e., gender and age) of the cohort in Supplementary.

- **UK dataset development (NCCID)**

For the total 2,682 CTs that were acquired from the 18 partner hospitals located in the United Kingdom (See Supplementary Table. 3), the acquisition devices and protocols varied from hospital to hospital. There are over 14 types of utilised 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 (see Fig. 4c). Regarding the material differences, 2,145 out of 2,682 CTs were taken after the injection of an iodine contrast agent (i.e., contrast CTs). As pointed out by previous study[32], 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, we used CycleGAN[33] to incorporate both contrast and non-contrast CTs, and shall elaborate upon such settings in the following section.

As detailed in Supplementary, 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 (i.e., gender and age) of the cohort in Supplementary. 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 UK were with larger averaged ages and had more male patients. These demographical differences might also explain why the UK locally trained model failed to perform well when applied to the CTs acquired from China.

- **Data pre-processing, model architecture and training setting**

We pre-processed the raw acquired CTs for standardisation 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 standardised 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 utilised trilinear interpolation to resize each slice from 512 to 128 pixels along each axis and rescaled the lung windows to a range between -1200 and 600 Hounsfield units before feeding into the network model.

We named our developed model 3D-DenseNet (Supplementary). It was developed based on DenseNet[49], a densely connected convolutional network model that performed remarkably well in classifying 2D images. To incorporate such design with the 3D CT representations, we adaptively customized the model architecture into fourteen 3D-convolution layers distributed in six dense blocks and two transmit blocks (insets of Supplementary Fig. 1). Each dense block consists of two 3D convolution layers and an inter-residual connection, whereas the transmit blocks are composed of a 3D convolution layer and an average pooling layer. We placed a 3D DropBlock[50] instead of simple dropout[51] before and after the six dense blocks, which proved to be more effective in regularising the training of convolution neural networks. We set the momentum of batch normalisation[52] 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 utilised SGD optimiser 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, since using a large learning rate at the very beginning of the training may result in numerical instability[53]. We then used cosine annealing[54] 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 five-fold cross-validation on trainval split, and then selected the best model and reported their test performance (in Fig. 4 and Supplementary).

- **Federated learning and privacy preservation**

At the central server, we adapted the FedAvg[43] algorithm to aggregate the updated model parameters from all clients (i.e., UCADI participants), that is, to combine the weights with respect to clients' dataset sizes and the number of local training epochs between consecutive communications. To ensure secure transmissions between the server and the clients, we used an encryption method called "Learning with Errors" (LWE)[46] to further protect all the transmitted information (i.e., model parameters and metadata). LWE 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 privacy (DP)[55], Moving Horizon Estimations (MHE)[56] and Model Predictive Control

(MPC)[57], LWE differentiates itself by essentially enabling the clients to achieve identical performance with the variants trained without decryption. However, the LWE method would add additional 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 2.7s (average over 100 trials under a test environment consisting of a single CPU (Intel Xeon E5-2630 v3 @ 2.40GHz) and the encrypted model size arises from 2.8MB to 62 MB, which increases the transmission time from 3.1s to 68.9s, in a typical international bandwidth environment[58] of 900KB/s (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 (average of 9 years of clinical experience, range from 4 to 18 years) from the Tongji Hospital Group, to make the diagnoses based on 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 vs non-COVID-19 CTs in Fig. 4 (detailed comparisons are presented in Supplementary). 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 work[32], we first extracted and converted the slices from contrast and non-contrast CTs of NCCID into JPEG format images with a resolution of 512px by 512px. 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 ResNet[59] backbone as the feature encoder and set the remaining parts in concordance with the original literature[32]. 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 2e-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 utilising CycleGAN for augmentation, we first trained the 3D DenseNet on trainval set of: (i) only non-contrast; (ii) non-contrast and CycleGAN synthesized non-contrast; (iii) only contrast and (iv) contrast and CycleGAN synthesized contrast CTs. In Supplementary, 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.

References

1. Ferretti, L. *et al.* Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science (80-.).* **368**, (2020).
2. Mei, X. *et al.* Artificial intelligence–enabled rapid diagnosis of patients with COVID-19. *Nat. Med.* **26**, (2020).
3. Yan, L. *et al.* An interpretable mortality prediction model for COVID-19 patients. *Nat. Mach. Intell.* **2**, (2020).
4. Ardila, D. *et al.* End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat. Med.* **25**, (2019).
5. Cohen, I. G., Evgeniou, T., Gerke, S. & Minssen, T. The European artificial intelligence strategy: implications and challenges for digital health. *The Lancet Digital Health* vol. 2 (2020).
6. Esteva, A. *et al.* A guide to deep learning in healthcare. *Nature Medicine* vol. 25 (2019).
7. Hinton, G. Deep learning-a technology with the potential to transform health care. *JAMA - Journal of the American Medical Association* vol. 320 (2018).
8. Jiang, F. *et al.* Artificial intelligence in healthcare: Past, present and future. *Stroke and Vascular Neurology* vol. 2 (2017).
9. Liu, X. *et al.* A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit. Heal.* **1**, (2019).
10. Hitaj, B., Ateniese, G. & Perez-Cruz, F. Deep Models under the GAN: Information leakage from collaborative deep learning. in *Proceedings of the ACM Conference on Computer and Communications Security* (2017).
11. Zhu, L., Liu, Z. & Han, S. Deep leakage from gradients. in *Advances in Neural Information Processing Systems, NeurIPS* (2019).
12. Melis, L., Song, C., De Cristofaro, E. & Shmatikov, V. Exploiting unintended feature leakage in collaborative learning. in *IEEE Symposium on Security and Privacy* vols 2019- (2019).
13. 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**, (2020).
14. Fang, Y. *et al.* Sensitivity of chest CT for COVID-19: Comparison to RT-PCR. *Radiology* vol. 296 (2020).
15. Kanne, J. P., Little, B. P., Chung, J. H., Elicker, B. M. & Ketai, L. H. Essentials for radiologists on COVID-19: An update-radiology scientific expert panel. *Radiology* vol. 296 (2020).
16. 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. *Annals of internal medicine* vol. 173 (2020).
17. Ackerman, C. M. *et al.* Massively multiplexed nucleic acid detection with Cas13. *Nature* **582**, (2020).
18. Chen, J. S. *et al.* CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity. *Science (80-.).* **360**, (2018).
19. Broughton, J. P. *et al.* CRISPR–Cas12-based detection of SARS-CoV-2. *Nat. Biotechnol.* **38**, (2020).
20. 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**, (2020).

21. 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**, (2020).
22. Cleverley, J., Piper, J. & Jones, M. M. The role of chest radiography in confirming covid-19 pneumonia. *The BMJ* vol. 370 (2020).
23. Vaseghi, G. *et al.* Clinical characterization and chest CT findings in laboratory-confirmed COVID-19: a systematic review and meta-analysis. *medRxiv* (2020).
24. Shi, H. *et al.* Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: a descriptive study. *Lancet Infect. Dis.* **20**, (2020).
25. Kanne, J. P. Chest CT findings in 2019 novel coronavirus (2019-NCoV) infections from Wuhan, China: Key points for the radiologist. *Radiology* vol. 295 (2020).
26. Chung, M. *et al.* CT imaging features of 2019 novel coronavirus (2019-NCoV). *Radiology* **295**, (2020).
27. Bai, H. X. *et al.* Performance of Radiologists in Differentiating COVID-19 from Non-COVID-19 Viral Pneumonia at Chest CT. *Radiology* **296**, (2020).
28. Harmon, S. A. *et al.* Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets. *Nat. Commun.* **11**, (2020).
29. Konecný, J. *et al.* Federated Learning: Strategies for Improving Communication Efficiency. in *NeurIPS Workshop on Private Multi-Party Machine Learning* (2016).
30. Bonawitz, K. *et al.* Towards Federated Learning at Scale: System Design. in *Proceedings of Machine Learning and Systems, MLSys* (2019).
31. National COVID-19 Chest Image Database (NCCID). https://nhsx.github.io/covid-chest-imaging-database/.
32. 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**, (2019).
33. Zhu, J. Y., Park, T., Isola, P. & Efros, A. A. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. in *Proceedings of the IEEE International Conference on Computer Vision, ICCV* (2017).
34. Nizan, O. & Tal, A. Breaking the cycle-colleagues are all you need. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 7860–7869 (2020).
35. 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).
36. 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).
37. 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).

38. Kataoka, H., Wakamiya, T., Hara, K. & Satoh, Y. Would mega-scale datasets further enhance spatiotemporal 3d cnns? *arXiv Prepr. arXiv2004.04968* (2020).
39. Chollet, F. Xception: Deep learning with depthwise separable convolutions. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 1251–1258 (2017).
40. Selvaraju, R. R. *et al.* Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. in *IEEE International Conference on Computer Vision, ICCV* (2017).
41. DeGrave, A. J., Janizek, J. D. & Lee, S.-I. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nat. Mach. Intell.* 1–10 (2021).
42. Commission, E. *reform of EU data protection rules*. (2018).
43. 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, AISTATS* (2016).
44. Warnat-Herresthal, S. *et al.* Swarm Learning for decentralized and confidential clinical machine learning. *Nature* **594**, 265–270 (2021).
45. Kaissis, G. *et al.* End-to-end privacy preserving deep learning on multi-institutional medical imaging. *Nat. Mach. Intell.* **3**, 473–484 (2021).
46. 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**, (2018).
47. Li, L. Z. & Wang, S. Prevalence and predictors of general psychiatric disorders and loneliness during COVID-19 in the United Kingdom. *Psychiatry Res.* **291**, (2020).
48. 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**, (2020).
49. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR* (2017).
50. Ghiasi, G., Lin, T. Y. & Le, Q. V. Dropblock: A regularization method for convolutional networks. in *Advances in Neural Information Processing Systems, NeurIPS* (2018).
51. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res. JMLR* **15**, (2014).
52. Ioffe, S. & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. in *nternational Conference on Machine Learning, ICML* (2015).
53. He, T. *et al.* Bag of tricks for image classification with convolutional neural networks. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR* (2019).
54. Loshchilov, I. & Hutter, F. SGDR: Stochastic gradient descent with warm restarts. in *International Conference on Learning Representations, ICLR* (2017).
55. Dwork, C. & Roth, A. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* **9**, (2013).
56. Michalska, H. & Mayne, D. Q. Moving Horizon Observers and Observer-Based Control. *IEEE Trans. Automat. Contr.* **40**, (1995).

57. García, C. E., Prett, D. M. & Morari, M. Model predictive control: Theory and practice-A survey. *Automatica* **25**, (1989).
58. International Telecommunication Union. *Yearbook of Statistics, Telecommunication/ICT Indicators 2007-2018*. (2019).
59. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR* (2016).

Ethics Approval: The UK data used in this study is under approval by Control of patient information (COPI) 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 CTs used or generated in this study is available by enquiries.

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

Funding: This study was supported by HUST COVID-19 Rapid Response Call (No. 2020kfyXGYJ031, No. 2020kfyXGYJ093, No. 2020kfyXGYJ094) and the National Natural Science Foundation of China (61703171 and 81771801). This work was also supported in part by a grant from the National Cancer Institute, National Institutes of Health, U01CA242879, and Thammasat University Research fund under the NRCT, Contract No. 25/2561, for the project of "Digital platform for sustainable digital economy development", based on the RUN Digital Cluster collaboration scheme. The funders did not participate in the research or review any details of this study.

Author contributions: TX, XB, ZL, CZ, and CS conceived the work; HW, JG and ZF contributed to the design & development of the methods and models, all the authors help interpret and analyse the experimental results; TX, HW, JG, ZF, XB, YX, WL, JL, AW and CS wrote the paper.

Competing interests: The authors declare no competing interests.