Privacy-preserving federated transfer learning system for COVID-19 diagnosis

Helin Yang[1✉], Jun Zhao[1,2✉], Kwok-Yan Lam[1,2], Zehui Xiong[2], Liang Xiao[3] & Qingqing Wu[4]

Abstract

Early identification of COVID-19 patients can effectively assist in patients' treatment and control the spread of the disease. Recently, artificial intelligence (AI) systems have been developed to perform rapid COVID-19 detection. However, the application of AI systems in COVID-19 diagnosis is currently hindered by insufficient medical dataset for model training and validation, due to the restrictions on data sharing governed by medical institutions and patient privacy preservation. To address this issue, we first develop a privacy-preserving federated transfer learning (FTL) system for COVID-19 diagnosis based on chest X-ray and computed tomography (CT). The system collaborates AI-based diagnosis models from different participating medical institutions to train a global AI model through federated learning without leaking private information of patients, and builds a relatively tailored diagnosis model at each medical institution by transfer learning. In a multi-category diagnosis task, the proposed FTL system achieves an area under the receiver-operating characteristics curve of 0.9974 for COVID-19, 0.9752 for normal health, and 0.9613 for common pneumonia on test cohort, respectively. The FTL system outperforms the traditional AI systems in terms of diagnostic performance and privacy preservation. In addition to assisting the world to fight the spread of COVID-19 outbreak, the FTL system has more healthcare applications, such as smart disease identification, lesion detection, tuberculosis diagnosis, lung cancer screening, etc. The code is available at https://www.kaggle.com/yangyang123765/aiforcovid19diagnosis, and we will share all remaining code online in the future.

[1]Strategic Centre for Research in Privacy-Preserving Technologies and Systems (Section of Distributed Artificial Intelligence Systems for Medical Healthcare), Nanyang Technological University, Singapore. [2]School of Computer Science and Engineering, Nanyang Technological University, Singapore. [3]Department of Informatics, Xiamen University, Xiamen, China. [4]Department of Electrical and Computer Engineering, University of Macau, Macau, China. ✉e-mail: helin.yang@ntu.edu.sg; junzhao@ntu.edu.sg

Introduction

The outbreak of global coronavirus disease 2019 (COVID-19) pandemic, caused by a strain of coronavirus named severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), is still spreading the world at a rapid rate[1-3]. A key step in fighting against COVID-19 is to diagnose the patients as early and accurately as possible for controlling the rapid spread of the pandemic and treating patients. Although fast reverse transcription polymerase chain reaction (RT-PCR) testing is available for COVID-19 diagnosis, a number of challenges still remain unsolved, including high false negative rate, time-consuming, laborious, and relatively low sensitivity[4,5].

As a result, using medical imaging, i.e., chest X-ray (CXR) or computed tomography (CT) scan, as a first-line test for the prognosis of COVID-19, can effectively assist in patients' treatment and control the spread of the disease[4, 6-8]. For example, Fang et al.[4] identified that chest CT imaging can provide early lesions in the lung and diagnose COVID-19 patients with high sensitivity, and chest X-ray imaging is widely available and low-cost with rapid triaging[8]. However, the diagnosis of radiography X-ray or CT scans depends on trained radiologists, which leaves some bottlenecks. Firstly, reviewing decades or hundreds of radiography scans by radiologists requires a long time to make a decision. Secondly, in some cases, it is challenging for radiologists to differentiate COVID-19 from common pneumonia (i.e., viral pneumonia and bacterial pneumonia) since they have similar radiographic presentation[9]. Last but not least, as COVID-19 is still rapidly spreading in some countries, such as America and India, it is difficult for the limited number of expert radiologists to quickly complete radiography diagnosis.

To tackle the aforementioned challenges, artificial intelligence (AI)-aided diagnosis systems[10-17] may assist radiologists in more quickly and accurately analyzing radiography scans to diagnose disease. So far, AI has demonstrated its powerful capability in image feature extraction, and has been widely applied for medical imaging-based diagnosis, such as image classification[10,11], lesion identification[12], tuberculosis diagnosis[13,14], and lung cancer detection[15-17]. Thus, the AI diagnosis systems associated with chest X-ray or CT have been recently developed to diagnose COVID-19[18-30]. Zhang et al.[18] proposed an AI-based diagnosis system on chest CT and it can diagnose COVID-19 from other non-COVID-19 scans with a receiver operating characteristic curve (AUC) of 0.980. Similarly, several advanced AI-assisted respiratory detection systems[19-24] based on chest CT were also proposed to assist medical institutions in differentiating COVID-19 from other common pneumonia or normal health in two-category or multi-category classification problems. Besides, there are also some AI-aided COVID-19 diagnosis systems using chest X-ray[25-29], and most of these studies focused mainly on the exploration of deep convolutional neural networks to perform COVID-19 detection. Moreover, Jin et al.[30] developed a deep learning-based COVID-19 diagnosis system on both chest CT and X-ray databases, and compared diagnostic performances of CT and X-ray utilizing paired data.

Even though AI-aided diagnosis systems[18-30] can effectively and readily detect COVID-19 patients, there are some critical challenges for developing a flexible, scalable, and general AI system for COVID-19 diagnosis: 1) AI models are often trained based on sufficient medical data. Unfortunately, due to privacy concerns and restrictions on data sharing, it is not feasible for most medical institutions to share their private raw data in public[31,32], which limits the applications of AI systems. 2) Common AI models can be shared with medical institutions who only have a small amount of medical data by federated learning[33-35], which avoids building a new AI system on insufficient medical data, but different medical institutions have different radiographic features and specific disease characteristics[36-38]. To the best of our knowledge, privacy-preserving AI diagnosis system has not been explored to perform COVID-19 detection while considering the relatively personalized model requirement of each medical institution.

In this work, we first develop a privacy-preserving federated transfer learning (FTL) system for COVID-19 diagnosis. The FTL system collaborates AI-based diagnosis models from different participating medical institutions (or hospitals) to train a global model without leaking any private medical information of patients to each other and builds a relatively tailored diagnosis model at each hospital side by transfer learning. We also compare the diagnostic performance of the proposed FTL system with that of classical AI systems, and the results demonstrate that the overall performance of the system is higher than that of other systems. To validate the flexibility and practicality of the FTL system, we also evaluate its diagnostic performance on both chest X-ray and CT datasets, and experimental results show that it can effectively differentiate COVID-19 from other common pneumonia or normal health in two-category or multi-category classification problems with high diagnosis accuracy (i.e., AUC of 0.9974 on chest X-ray and 0.9721on chest CT). In addition, we also test the scalability and generalization of our FTL system on three independent COVID-19 databases from different countries. The experimental results confirm the high diagnostic performance, accuracy, and general applicability of our proposed FTL system.

Results

Patient chest X-ray and CT scans for training and testing. We proposed and evaluated a privacy-preserving FTL-based COVID-19 diagnosis system, utilizing multi-category chest X-ray and CT scans. For chest X-ray scans, in total, 5226 chest X-ray scans were used in this study, including 1270 scans with confirmed COVID-19 (COVID-19-infected), 2039 scans with normal health (no infection), and 1917 scans with common pneumonia (bacterial or non-COVID-19 viral infection). The total X-ray scans were divided into three independent cohorts: a training cohort of 2731 scans, a validation cohort of 703 scans, and a test cohort of 1792 scans. For chest CT, in total, 6191 chest CT scans were utilized, including 1781 scans with confirmed COVID-19, 2203 scans

with normal health, and 2207 scans with common pneumonia. The split of chest X-ray and CT scans in training, validation, and testing cohorts are listed in Table 1.

Construction of the piracy-preserving FTL system for COVID-19 diagnosis. We developed an FTL-based system for COVID-19 diagnosis on chest X-ray and CT datasets, as shown in Fig. 1a, where the system in any city or country consists of two components: a central cloud server and multiple participating hospitals (or medical institutions). Each hospital locally trains an AI model based on its COVID-19 chest dataset, rather than uploading a huge amount of raw sensitive dataset (i.e., chest X-ray or CT scans) of patients to the central cloud server. Each hospital acts as a client to communicate with the cloud server at regular intervals to build a global AI model and sends its local trained AI model parameters to the cloud server for global model aggregation. After receiving the local AI model parameters from different participating hospitals, the central cloud server then aggregates the collected model parameters before broadcasting the updated models to the participating hospitals for another round of local model training. After several global aggregation rounds, the target learning accuracy can be achieved. Our proposed FTL system considers a decentralized data sharing scenario and multiple hospitals participate together for COVID-19 detection, where the raw private medical dataset of patients is kept at each participating hospital, which not only preserves private information but also reduces data sharing overhead in computer networks.

If a new hospital joints in the FTL system or an existing hospital wants to build its personalized learning-based diagnosis model, transfer learning can be performed to integrate the cloud server-side AI model with the hospital-side dataset, as shown in Fig. 1b. Note that the AI model sharing does not leak any private medical information of patients during this process. The training and testing structures of deep neural networks (DNN) for COVID-19 diagnosis are illustrated in Fig. 1c.

Performances of diagnosis in FTL system on chest X-ray scans. We adopted receiver operating characteristic curves (AUROC) to analyze the diagnostic accuracy in two popular transfer learning models (i.e., visual geometry group-16 (VGG16) model and residual neural network-18 (ResNet-18) model) on the three-category training-validation dataset, where the operating characteristic curves (ROC) results are shown in Figs. 2a, b with their 95% confidence intervals (CI). On the training-validation cohort, the ROC curve of the VGG16 model (Fig. 2a) illustrated that AUC of three categories were respectively 0.9947 (95% CI: 0.9930–0.9964) for COVID-19, 0.9625 (95% CI: 0.9539–0.9711) for normal, and 0.9436 (95% CI: 0.9303–0.9569) for common pneumonia on training-validation dataset. In addition, as presented in Fig. 2b, the ResNet-18 learning model can also achieve a high AUC value, i.e., the AUC of three categories were respectively 0.9544 (95% CI: 0.9473–0.9615) for COVID-19, 0.9201 (95% CI: 0.9146–0.9257) for normal, and 0.9046 (95%

CI: 0.8981–0.9111) for common pneumonia. The AUROC shown in Figs. 2a,b indicates good discrimination of the proposed FTL system.

After completing the training-validation stage, the trained FTL system was evaluated on the test chest X-ray set. Then the confusion matrices on three diagnostic categories of the two learning models are shown in Figs. 2c, d. Besides, the performance metrics including precision, sensitivity (recall), specificity, F1-score, and AUC values were also calculated, which are reported in Table 2. It is clear from Fig. 2 and Table 2 that the evaluated learning models perform very well in diagnosing COVID-19, normal, and common pneumonia in our proposed FTL system. Specifically, from Fig. 2c,d, we observe that the COVID-19 diagnostic accuracies for the VGG16 model and ResNet-18 model were measured as 0.9795 and 0.9570, respectively, which demonstrates the high quality of our proposed FTL system in solving COVID-19 diagnosis problem. In particular, the VGG16 model correctly identified 528/541 COVID-19-infected scans as COVID-19, and the ResNet-18 model correctly detected 512/541 COVID-19-infected scans. Only 3 normal scans and 4 pneumonia scans were miss-classified as COVID-19 out of 562 normal scans and 556 pneumonia scans, respectively. This indicates that the FTL system is extremely robust in distinguishing COVID-19 scans from normal and common pneumonia on chest X-ray scans.

In addition, it is worth noting that the two learning models are not confusing between COVID-19 and the other two categories, rather than models are more confused between normal and common pneumonia X-ray scans. For example, in VGG16 model (Fig. 2c), the diagnostic accuracies of COVID-19, normal, and common pneumonia are 0.9795, 0.8951, and 0.9497, respectively. However, as illustrated in Table 2, the high precision and F1-score values verify that the two evaluated learning models are performing excellently in diagnosing most of the test X-ray scans reliably in our proposed FTL system. This is essential that the proposed system should not diagnose any COVID-19 patients to normal (or common pneumonia) or vice versa. Although the two learning models shown in Table 2 achieved high diagnostic accuracy, our experiments illustrated that when the system addressed the three-category classification problem, using VGG16 learning model leads to a better COVID-19 diagnosis performance, with a precision of 0.9874 (95% CI: 0.9765-0.9983), a sensitivity of 0. 9495 (95% CI: 0. 9379-0. 9611), a specificity of 0. 9211 (95% CI: 0. 9064-0. 9357), a F1-score of 0.9680 (95% CI: 0.9619-0.9742), and AUC of 0.9974 (95% CI: 0.9963-0.9985). It is evident from Fig. 2 and Table 2 that the two evaluated learning models in the FTL system perform very well in classifying COVID-19, normal, and common pneumonia based on chest X-ray scans in the three-category classification problem.

Comparison of FTL system to other AI systems. To further explore the advantages of the FTL system, we measured the COVID-19 diagnosis performance of FTL and compared it to other AI systems, i.e., the centralized learning system and the local learning system. Note that the centralized learning system requires to collect a huge amount of chest scans from participating

hospitals, which has the risk of leaking the private information of patients. In contrast, the local learning system does not need to share any medical data or model information among different hospitals or the central server, where all AI models are trained locally at each individual hospital based on its dataset.

It is evident from the ROC curves shown in Figs. 3a-c that both the centralized learning system and the FTL system have similar AUC values on the training-validation cohort, and their AUC performance outperforms the local learning system. The ROC curve of the FTL system (Fig. 3b) illustrates that AUC of three categories were respectively 0.9884 (95% CI: 0.9842–0.9925) for COVID-19, 0.9684 (95% CI: 0.9633–0.9735) for normal, and 0.9489 (95% CI: 0.9401–0.9577) for common pneumonia, which revealed good discrimination. Compared to the local learning system (Fig. 3c), the FTL system (Fig. 3b) improves the AUC values of COVID-19 diagnosis, normal diagnosis, and common pneumonia diagnosis by 2.59%, 3.98%, and 7.50%, respectively. The reason lies in the fact that the FTL system utilized federated settings to indirectly excavate more AI model information from distributed datasets to globally optimize a better AI-based diagnosis model, and through the transfer learning, models become more specific to the characteristics of each hospital. In contrast, in the local learning system, as each hospital trains its learning model locally based on its chest scans without sharing any information with other medical institutions, the insufficient dataset may fail to perform model training effectively.

Figs. 3d-f depict the confusion matrices of the three compared AI systems on the test cohort. Similar to the performance shown in Figs. 3a, b, the centralized learning system and the FTL system achieve comparable diagnostic performance in the three-category classification problem. It is shown from Fig. 3b that out of 488 COVID-19 X-ray scans, 455 were correctly identified, 3 scans were miss-classified as normal, and 25 scans were classified as common pneumonia. Besides, only 1 normal scan and 2 pneumonia scans were miss–classified as COVID-19 with quite high diagnostic accuracy. Thus, we achieved an overall diagnostic accuracy of 0.9481 on the test cohort in the FTL system, and obtained high precision rates of 0.9426, 0.9583, and 0.9421 for COVID-19, normal, and pneumonia categories, respectively. However, the diagnostic accuracy of the local learning system is the poorest among the three AI systems as it has an insufficient dataset for global model training. Specifically, among 488 COVID-19 scans, only 315 scans were correctly detected, 72 scans were miss-classified as normal, and 101 scans were labelled as pneumonia. Therefore, in the local learning system, the precision rates of 0.6455, 0.9676, and 0.8216 for COVID-19, normal, and pneumonia categories were obtained, respectively, where the poor precision rate of COVID-19 fails to satisfy the diagnostic accuracy requirement.

In addition to the AUC and confusion matrix analysis of different learning systems, we also measured other metrics in the experiments, including precision, sensitivity, specificity, AUC, and overall diagnosis accuracy (ACC) of the three categories, as shown in Figs. 3g-k. According to Figs. 3g-k, both the centralized learning system and the FTL system still had a comparable performance for classification and their metrics values significantly outperform those of the local

learning system. For instance, the FTL system achieved a very high overall accuracy of 0.9456 (95% CI: 0.9289-0.9623), while the overall accuracy of the local learning system was only 0.7740 (95% CI: 0. 7508-0.7972).

Figure 3*l* compared the data sharing overhead in Mbits when training the learning models using the different learning systems. The centralized learning system has a huge amount of data sharing overhead because the participating hospitals need to upload all the chest X-ray scans to the central server for global AI model training. Although the centralized learning system achieved the best diagnostic accuracy compared with the other two AI systems (Figs. 3a-k), sharing medical data leaks the privacy of patients, which is a challenging task. In contrast, as shown in Fig. 3*l*, the local learning system does not have any data sharing overhead as each hospital only trains its learning model on its local dataset without sharing its information with the central server or other hospitals. However, its diagnostic performances (Figs.3a-k) are the poorest among the three AI systems.

FTL system performance evaluation in international databases. To test the scalability and generalization of our FTL diagnostic system, we evaluated the system performance on three independent COVID-19 databases from different countries. Specifically, the COVID-19 databases were from Germany, China, and Italy without overlapping subjects.

In a retrospective investigation in the country of Germany, as illustrated in Figs. 4a-c, on the test chest X-ray cohort, our FTL diagnostic system achieved AUC of 0.9999 (95% CI: 0.9998–1.0) with a precision of 0.9567 (95% CI: 0.9956–0.9178), a sensitivity of 0.9970 (95% CI: 0.9859–1.0), and a specificity of 0. 9391 (95% CI: 0.9139–0.9643) for COVID-19 diagnosis versus the other two categories. The second prospective investigation was conducted in China (Figs. 4d-f). Our FTL diagnostic system achieved AUC of 0.9974 (95% CI: 0.9973–0.9975) with a precision of 0.9270 (95% CI: 0.8680–0.9860), a sensitivity of 0.9950 (95% CI: 0.9559–1.0), and a specificity of 0.9368 (95% CI: 0.9101–0.9635) for COVID-19 diagnosis versus other two categories. The third prospective study was conducted in Italy (Fig. 4g-i). Our FTL diagnostic system still achieved high AUC of 0.9974 (95% CI: 0.9973–0.9975) with a precision of 0.9576 (95% CI: 0.9185–0.9967), a sensitivity of 0.9575 (95% CI: 0.9184–0.9966), and a specificity of 0.9616 (95% CI: 0.9411–0.9821) for COVID-19 diagnosis versus other two categories.

Besides, we observe that only 1.07% and 3.26% of COVID-19 cases were miss-classified as normal and common pneumonia in Germany (Fig. 4b), respectively, and only 0.06% and 0.16% of normal case and common pneumonia cases were respectively miss-classified as COVID-19 (Fig. 4b) in this prospective pilot study. Similarly, our proposed FTL system can also effectively distinguish COVID-19 from normal or common pneumonia on the test cohorts in China (Fig. 4e) and Italy (Fig. 4h). Overall, the experimental results shown in Figs. 4a-i confirmed the high diagnostic performance, accuracy, ad universal applicability of our proposed FTL system with internationally.

Extension of the FTL system in chest CT dataset. To validate the flexibility and practicability of the FTL system, we further evaluated the FTL system on chest CT scans and showed the experimental results in Fig. 5. Here, we picked two different tasks for the COVID-19 diagnostic study, i.e., a two-category diagnosis task (including COVID-19 and normal categories) and a three-category diagnosis task (including COVID-19, normal, and common pneumonia categories).

For the three-category diagnosis task (Fig. 5a) on the training-validation cohort, the area under the AUC of the COVID-19 diagnosis, normal diagnosis, and common pneumonia diagnosis were 0.9904 (95% CI: 0.9853–0.9955), 0.9631 (95% CI: 0.9508–0.9754), and 0.9347 (95% CI: 0.9190–0.9504), respectively. Besides, the ROC curve of the two-category diagnosis task (Fig. 5b) illustrated that the AUC of the two categories was 0.9834 (95% CI: 0.9671–0.9997) for COVID-19 and 0.9827 (95% CI: 0.9665–0.9989) for normal, respectively. The confusion matrices of three-category and two-category diagnosis tasks are provided in Figs. 5c, d. In the three-category diagnosis task (Fig. 5c), 11 COVID-19 scans were miss-classified into normal out of 495 scans and 14 scans were detected as pneumonia. Besides, only one scan out of 515 normal scans and 6 scans out of 619 pneumonia scans were miss-classified as COVID-19. Fig. 5d shows that 23 COVID-19 CT scans out of 495 scans were miss-classified as normal and 3 scans out of normal 515 scans were labelled as COVID-19, where high precision rates of 0.9535 and 0.9942 for COVID-19 diagnosis and normal diagnosis were obtained, respectively. The results verify that the FTL system is extremely robust in distinguishing COVID-19 scans from normal chest CT scans.

The overall performances of three-category and two-category diagnosis tasks in terms of precision, sensitivity, specificity, F1-score, and AUC on the chest CT test cohort are shown in Table 3. The two diagnosis tasks performed very well in detecting COVID-19 scans in two-category and three-category classification problems. Although the performance difference between the two tasks is very marginal, small overall performance improvement can be achieved in the two-category diagnosis task in comparison to the three-category diagnosis task, because the three-category classification problem is more complex. For example, the two-category diagnosis task achieved a high AUC of 0.9721 (95% CI: 0.9713-0.9729) for COVID-19 diagnosis with a sensitivity of 0.9838 (95% CI: 0.9723-0.9954), a specificity of 0.9528 (95% CI: 0.9422-0.9634), and a F1-score of 0.9662 (95% CI: 0.9587-0.9737), respectively. By contract, the three-category diagnosis task yielded a high AUC of 0.9772 (95% CI: 0.9767-0.9777) for COVID-19 diagnosis with a sensitivity of 0.9468 (95% CI: 0.9280-0.9655), a specificity of 0.9240 (95% CI: 0.9109-0.9369), and a F1-score of 0.9510 (95% CI: 0.9406-0.9613), respectively. Overall, in addition to the great performance achieved in the chest X-ray dataset by using the FTL system (Fig. 2-5, Supplementary Table 2, and Table 3) highlighted the capability of the FTL system to accurately diagnose COVID-19 on chest CT dataset.

Interpreting the FTL system. After comprehensively evaluating the diagnostic performance of the proposed FTL system on large multi-category datasets, we then adopted t-distributed stochastic

neighbor embedding (t-SNE)[39] to assist to analyze the deep features of chest X-ray images on two-dimensional (2D) plane, shown in Fig.6, and learned the characteristic differences between three diagnostic categories in latent space. Similar to the work[30], feature extraction algorithm was adopted to search the discriminative features in identifying COVID-19 from normal healthy and common pneumonia. The selected features were utilized to analyze the imaging characteristics on chest X-ray. Generally, as shown in Fig. 6a, the selected features can significantly distinguish any category from other two categories in the *t-SNE* map. Moreover, we found that the scan features of attentional areas of COVID-19 and other two categories have significant difference, while the results for distinguishing normal and common pneumonia were not obvious. Such feature characteristics can well explain the results (Figs. 2c, d) that only several COVID-19 scans were miss-classified as normal (or common pneumonia) or vice versa, while dozens of normal scans were miss-classified as common pneumonia or vice versa. Thus, we can learn that some scans of one diagnostic category are miss-labelled as another category when their imaging features are located at the attentional areas, which was illustrated at the left of Fig. 6a. The reason lies in the fact that the radiology features of those miss-classified scans from different categories may have small difference, or exist atypical or unclear presentation, thus the diagnosis system had prediction error in our experimental results and it cannot achieve 100% diagnosis accuracy.

We also performed deep feature analysis of COVID-19 chest X-ray scans from three different regions using different datasets, i.e., China, Italy, and Germany, as shown in Fig. 6b. There are no attentional areas of the three countries on COVID-19 chest X-ray, which indicates that the AI models can be shared or transferred with each country to effectively perform COVID-19 diagnosis. Thus, these results shown in Fig. 6b can assist to explain the results (Figs. 3a-i) that the FTL system achieved high diagnostic performance, accuracy, and universal applicability with internationally.

Discussion

In this study, we have developed an FTL-based diagnosis system, the first privacy-preserving federated transfer learning framework for the diagnosis of COVID-19, normal health (no infection), and common pneumonia (non-COVID-19 infection) based on chest X-ray scans and CT scans. The FTL system collaborates the scattered AI models from different participating hospitals to train a global COVID-19 diagnosis model, without leaking any private information of each hospital, and achieves a tailored model at the hospital-side through AI model transferring. Missing medical data or insufficient medical data frequently occur, especially in some small or poorly equipped medical institutions, which degrades the COVID-19 diagnosis model training efficiency. Our proposed FTL system can address this challenge by transferring a high-quality learning model from the medical institutions (having sufficient medical data) to other institutions. However, to take full advantage of this system, we recommend that medical institutions or hospitals in the world can collaboratively to share their AI models[31]. The use of chest X-ray or CT scans for the objective of diagnosing COVID-19 patients has been investigated[19-29], where AI techniques were adopted to

tackle COVID-19 diagnosis and as high as 95% detection accuracy was reported in these studies. However, sufficient chest X-ray or CT data need to be collected for model training in these work, which may leak the private information of patients, and medical institutions may not be allowed to share the patients' medical data for the public to a certain degree[31,32]. The proposed privacy-preserving FTL system not only protects the privacy security of medical institutions, but also archives a quite high diagnostic accuracy on both chest X-ray and CT dataset.

Two popular transfer learning models, i.e., VGG16 and ResNet-18[35,40,41], were adopted in the FTL system to evaluate the diagnostic accuracy in the three-category classification problem. Here, on the training-validation dataset, we obtained 0.9947 AUC of 703 chest X-ray scans resulting in 0.9495 (95% CI: 0.9379-0.9611) accuracy for diagnosis of COVID-19 in VGG16, and achieved 0.9544 AUC resulting in 0.9496 (95% CI: 0.9286-0.9707) accuracy for diagnosis of COVID-19 in ResNet-18 model, as shown in Fig. 2. Besides, the performance of a very high metric (i.e., precision, sensitivity, specificity, F1-score, and AUC, shown in Table 2) of COVID-19 diagnosis, normal diagnosis, and pneumonia diagnosis have clearly revealed the good discrimination of the proposed FTL system. We also observed that the overall performance of VGG16 is better than that of ResNet-18, but the former model is slightly more complex than the latter one.

To further understand the advantages of the FTL system, we evaluated the COVID-19 diagnosis performance of FTL and compared it to other AI systems, i.e., the centralized learning system and the local learning system, and the performance comparisons were shown in Fig. 3. The FTL system can achieve comparable performance to the centralized learning system (Figs. 3a-k), but the centralized learning system needs to collect a huge volume of chest scans (huge shared data size, shown in Fig. 3*l*) from participating hospitals, which leaks the private information of patients. Compared with the centralized system, the proposed FTL system utilized very little of data sharing overhead, because the participating hospitals only upload AI model parameters rather than the entire chest X-ray scans to the central server. Consequently, the FTL system does not have the overhead of collecting the huge volume of raw chest scan dataset, and it still achieves comparable diagnostic performances (i.e., the overall diagnostic accuracy of 0.9456 (95% CI: 0.9289-0.9623)) compared to the centralized learning system (i.e., the overall diagnostic accuracy of 0.9506 (95% CI: 0.9402-0.9611)), as shown in Fig. 3k. It is worth noting that in order to protect patients' privacy, the medical institutions do not allow private information leakage or data sharing across medical institutions. Thus, the centralized learning system may not be practical for COVID-19 diagnosis. However, the proposed FTL system can effectively protect medical data privacy by enabling the participating hospitals to train learning models locally without sharing the huge amount of sensitive medical data, and only the weights of the AI model need to be shared for global model optimization.

In addition, compared with the local learning system, the FTL system obtained the values of a quite higher metric (i.e., precision, sensitivity, specificity, F1-score, and overall accuracy, shown in Figs. 3a-k) of COVID-19 diagnosis, normal diagnosis, and common pneumonia diagnosis,

respectively. As shown in Fig. 3j, the COVID-19, normal, and common pneumonia categories with a precision of 0.9434±0.0266, 0.9490±0.0158, and 0.9405±0.0118 at 95% CI were respectively obtained in the FTL system, while the local learning system yielded the precision with 0.6782±0.0097, 0.9170±0.0086, and 0.8298±0.0053 for COVID-19, normal, and pneumonia categories at 95% CI, respectively. This is most likely because the insufficient model information or chest dataset in the local learning system makes it more difficult for the model to be optimized.

We also evaluated the generalization and applicability of the proposed FTL system, by testing the system diagnostic performance on three independent COVID-19 datasets from different centuries (i.e., Germany, China, and Italy). The experimental results shown in Fig. 4 demonstrated that the FTL system achieved high diagnostic accuracy for COVID-19 diagnosis in the three different countries. For example, the system obtained a high precision of 0.9567, 0.9270, and 0.9576 on the test cohort from Germany, China, and Italy, respectively. Thus, the FTL system can effectively address the regional variations and general applicability issue, and it can be applied for COVID-19 diagnosis within internationally. Besides, we also provided the visualization feature analysis of chest X-ray scans by using *t-SNE* **(**Fig. 6a**).** *T-SNE* clearly illustrated that deep features provided by the proposed FTL system can effectively distinguish different categories of chest X-ray scans, especially COVID-19 cases were clearly identified from the other two categories (normal and common pneumonia) with a significant attentional area. Thus, the FTL system achieved high diagnostic accuracy for COVID-19 diagnosis. Further, a visual interpretation of COVID-19 scans from three different countries was presented in this study. There are no attentional areas of the three countries on COVID-19 scans (Fig. 6b), thus the FTL system achieved the favorable diagnostic performances (Fig. 4**)** on the test cohorts from these countries as their COVID-19 databases have significant similar features and their trained AI models can be effectively shared or transformed with each other for model optimization.

We investigated the flexibility and practicability of the FTL system by extending it to chest CT scans, as CT is also regarded as an available way to diagnose COVID-19. Our experimental results further demonstrated that the FTL system still achieves excellent diagnostic performance for three-category diagnosis task and two-category diagnosis task on chest CT scans. As shown in Fig. 5c, among 495 COVID-19 scans, only 11 scans and 22 scans are respectively miss-classified as normal and common pneumonia in the three-category diagnosis task, achieving a high precision of 0.9495. Besides, only 23 COVID-19 scans were miss-labelled as normal in the two-category diagnosis task, achieving a high precision of 0.9535. According to Fig. 2-5, Table 2, and Table 3, both the chest X-ray and CT had very good diagnostic accuracy. To the best of our knowledge, we are the first to provide comprehensive evaluations of both CT and CXR performances for COVID-19 diagnosis, which can make our experimental results more convincing. Moreover, the overall performance of CT-based COVID-19 diagnosis (Table 3a) slightly outperforms that of X-ray-based COVID-19 diagnosis (Table 2a) in VGG16 model. These results coincide with the results of the existing AI-based COVID-19 diagnosis system[30].

Although the diagnostic performance of COVID-19 diagnosis is encouraging, there are still some limitations and future works of this study. Firstly, although this study utilized a publicly available dataset of 11417 chest scans (including 5226 chest X-ray scans and 6191 CT scans), due to the diversity and heterogeneity of COVID-19, the scalability and robustness of the proposed FTL system require to be further validated and tested under other huge and diverse databases. Secondly, in the FTL system, during the training process, each participating medical institution may be undedicated to the model aggregation/transferring task at hand and most institutions are not active on any given training iteration. In this case, the synchronous and asynchronous schemes need to be analyzed via distributed optimization. Thirdly, combining chest X-ray and CT together to diagnose COVID-19 patients may be an effective way to improve the diagnostic accuracy, which will enable additional evaluation of our system and the development of more functionality. This work is not available in this study yet, because there is currently no enough dataset with paired CT and CXR scans captured. Last but not least, we will intend to share our proposed privacy-preserving FTL system into a free online platform for COVID-19 diagnosis. In this way, hospitals and medical institutions around the world can detect patients on chest X-ray or CT scans without the requirement for building their diagnosis platforms. However, CT scans are not used widely for COVID-19 diagnosis[42]. For example, the American College of Radiology suggests that medical institutions should not use CT as a first-line test for COVID-19 diagnosis, which may limit the popularization of our proposed FTL system[42].

In conclusion, a privacy-preserving FTL system was proposed to perform COVID-19 diagnosis in this study. Unlike the centralized learning system under the requirements of a huge number of sensitive medical data, the FTL system aggregates the AI model parameters from different participating medical institutions to prevent the private information leakage of patients and obtains a tailored learning model through knowledge transfer. The comprehensive experimental results demonstrated that the system obtained a favorable diagnostic accuracy in the chest X-ray and CT datasets, and it achieved a significant higher diagnostic performance than the traditional AI system in diagnosing COVID-19. In addition, we also evaluated the diagnostic performance of our FTL system on three independent COVID-19 databases from different countries. The experimental results confirm the high diagnostic performance, accuracy, and general applicability of our proposed FTL system with internationally. Radiologists around the world can use the FTL system to perform a personalized AI model of COVID-19 diagnosis, providing a new driving force for stopping the fast spread of COVID-19 outbreak.

Methods

The setting of experimental data. This research is a retrospective study based on radiologic assessments, including chest X-ray or CT. We used 11417 chest scans (including 5226 chest X-ray scans and 6191 chest CT scans) from different databases shown in data availability. All chest

X-ray or CT scans were randomly divided into training, validation, and testing cohorts, which have been listed in Table 1.

1. Training cohort: For the chest X-ray dataset, 2731 scans were randomly assigned to a training set which included 609 confirmed COVID-19, 1098 normal, and 1024 common pneumonia scans. For the chest CT dataset, 3529 scans were randomly assigned into the training cohort which had 991 confirmed COVID-19, 1379 normal, and 1219 common pneumonia scans.
2. Validation cohort: The chest scan volumes of the validation cohort are smaller than those of training cohort. For the chest X-ray dataset, 703 scans were assigned to the validation cohort, which was comprised of 173 confirmed COVID-19, 293 normal, and 237 common pneumonia scans. For the chest CT dataset, it had 973 scans in the validation cohort, and the cohort contained 295 confirmed COVID-19, 309 normal, and 369 common pneumonia scans.
3. Testing cohort: In order to evaluate the diagnostic performance of the developed FTL system and make our experimental results more convincing, the chest scan volumes of the test cohort should be sufficient. Thus, 1792 scans and 1629 scans were randomly assigned to the X-ray testing cohort and CT testing cohort, respectively. The former cohort included 488 confirmed COVID-19, 648 normal, and 656 common pneumonia X-ray scans, while the latter cohort included 495 confirmed COVID-19, 515 normal, and 619 common pneumonia CT scans.

Note that the pneumonia category contains viral pneumonia and bacterial pneumonia, which together are the common causes of pneumonia and come from clinical, radiological, and hospital results. Moreover, this study considers that there exist four participating hospitals in the FTL systems, where the training data set are divided into four parts for four hospitals. For example, at the first hospital, it has 98 confirmed COVID-19, 217 normal, and 203 common pneumonia scans on chest X-ray.

Federated learning model. Federated learning has become an important paradigm aiming to train a collaborative AI model while keeping all the training data localized[43] and confidential. Thus, federated learning is recently applied to hold great promises on healthcare data analytics[33-35], which enables hospitals to train their AI models locally without sharing sensitive medical data of patients to the central server. In our considered privacy-preserving FTL system, as shown in Fig. 1a, the chest X-ray or CT scan dataset of patients stored locally at different participating hospitals is employed to train the local models, where the learning-based diagnosis models are communicated the central cloud server for model aggregation without uploading the private medical data. The cloud server then collects these AI model parameters to optimize a better global model which is further broadcasted to different participating hospitals for COVID-19 diagnosis.

To perform the privacy-preserving model aggregation, in the system, a set of hospitals $\mathcal{N}=\{1,\cdots,N\}$ participate in a global AI model training (i.e., training federation for COVID-19-infected patients detection) with a central could server communicating with these hospitals. Each hospital *n* adopts its chest X-ray or CT scan dataset $\mathcal{D}_n$ to train its local AI model parameters $\boldsymbol{\theta}_n$ without sharing the raw medical data with the cloud server. In the federated learning scenario, the loss function $f(\boldsymbol{\theta}_n;\mathbf{x}_{n,i},\mathbf{y}_{n,i})$ is introduced to quantify the federated performance error over the input data sample vector $\mathbf{x}_{n,i}$ on the training model $\boldsymbol{\theta}_n$ and the desired output scale vector $y_{n,i}$ for each input sample *i* at the *n*-th hospital. Accordingly, the local loss function on the training set $\mathcal{D}_n$ at the *n*-th hospital can be expressed as

$$F(\boldsymbol{\theta}_n)\triangleq\frac{1}{|\mathcal{D}_n|}\sum_{i\in\mathcal{D}_n}f(\boldsymbol{\theta}_n;\mathbf{x}_{n,i},\mathbf{y}_{n,i}), \tag{1}$$

where $|\mathcal{D}_n|$ denotes the cardinality of the set $\mathcal{D}_n$. At the central cloud server-side, the global loss function with the local datasets of participating hospitals can be expressed as

$$F(\boldsymbol{\theta})\triangleq\sum_{n\in\mathcal{N}}\frac{|\mathcal{D}_n|}{|\mathcal{D}|}F(\boldsymbol{\theta}_n)=\frac{1}{|\mathcal{D}_n|}\sum_{n\in\mathcal{N}}\sum_{i\in\mathcal{D}_n}f(\boldsymbol{\theta}_n;\mathbf{x}_{n,i},\mathbf{y}_{n,i}), \tag{2}$$

where $\boldsymbol{\theta}$ denotes the global model parameters at the central cloud server and $\mathcal{D}$ is the sum data samples from all participating hospitals. The objective of the federated task is to find an optimal model parameter $\boldsymbol{\theta}^*$ by minimizing the global loss function as

$$\boldsymbol{\theta}^*=\arg\min F(\boldsymbol{\theta}). \tag{3}$$

It is worth noting that the global loss function $F(\boldsymbol{\theta})$ cannot be directly calculated by using the raw datasets from participating hospitals, as the chest X-ray or CT scans of patients are sensitive and the volumes are quite large. The central cloud server iteratively updates the aggregated model through the local training procedure across different participating hospitals until the model converges to a certain learning accuracy target.

To improve the leaning efficiency and accuracy, the central cloud server can select a part of participating hospitals with high computation capacity and training accuracy to perform federated model aggregation. The reason lies in the fact that the hospitals with low-quality learning models on insufficient medical data will reduce the aggregated learning accuracy and the hospitals with low computational capacities increase global learning delay.

Transfer learning model. Transfer learning is a powerful AI tool for improving learning efficiency by transferring the learned knowledge or trained model from one related task to another new task. Recently, transfer learning has widely been applying to analyze medical images[27,36-38], such as image classification, lesion identification, and pattern recognition.

In the proposed FTL system, two major ways of transfer learning are applied to complete COVID-19 detection, i.e., ConvNet as fixed feature extractor and fine-tuning ConvNet. In the former, as depicted in Fig. 7, the pre-trained model of COVID-19 detection can be treated as a fixed feature extractor for the new chest X-ray or CT scan dataset, while the last two fully-connected layers are modified based on the classification objective of a new task. In a later way, the whole DNN or a subset of the network is fine-tuned on a new task. In this context, the top-most layers of ConvNet are replaced and retained on new chest X-ray images, and the corresponding weights are fine-tuned for a new COVID-19 detection task.

Here, we take the fixed feature extractor as an example to illustrate the transfer learning process of the proposed federated CNN-based COVID-19 diagnosis model. As studied in[27,40,41], the chest X-ray or CT image features in the DNN transition from the lower layers with low-level features to the higher layers with specific high-level features. In this context, once the model parameters are achieved, we can perform transfer learning at hospitals to learn their personalized COVID-19 diagnosis model for the following conditions: a new hospital joins the federated leaning system who wants to build a new learning model; an existing hospital needs to build its personalized learning model; a hospital from another city (City B) wants to share the learning model from the current city (City A) due to the insufficient chest scans.

Figure 5 also illustrates the transfer learning process, where the network includes two convolution layers, two pooling layers, two fully connected layers, and one soft-max output layer for COVID-19 diagnosis. As the objective of both the convolution and pooling layers is to extract general features (low-level) of chest X-ray or CT images, we keep these layers frozen and do not need to update the corresponding parameters in backpropagation. By contrast, the fully connected layers at the higher-level focus on analyzing specific features; thus we update the parameters when training on a new task.

In the FTL system, for privacy preservation restrictions, we only have the cloud server-side learning model and the local dataset of the hospital, which is different from existing transfer learning frameworks where both the source and target datasets are available[41]. Therefore, similar to[40], the correlation alignment layer is adopted before the soft-max layer to adapt the second-order statistics of the source and target domains. Accordingly, the loss of the correlation alignment is calculated by

$$Loss = \frac{1}{4d^2}\left\| \boldsymbol{\Psi}_S - \boldsymbol{\Psi}_T \right\|_F^2 \tag{4}$$

where d denotes the dimension of the embedding features and $\|\cdot\|_F^2$ stands for the Frobenius norm. $\boldsymbol{\Psi}_S$ and $\boldsymbol{\Psi}_T$ are the covariance matrices of the source and target domains, respectively, which can be computed by the study[44]. The loss function on the n-th hospital-side can be expressed as

$$\bar{F}(\boldsymbol{\theta}_n) = \frac{1}{|\mathcal{D}_n|}\sum_{i\in\mathcal{D}_n} f(\boldsymbol{\theta}_n; \mathbf{x}_{n,i}, \mathbf{y}_{n,i}) + \mu Loss_n \tag{5}$$

where $\mu > 0$ denotes the transfer learning rate which is also the tradeoff parameter. Note that the second part ($\mu Loss_n$) in (5) will be further reduced after each training step to gradually remove the effect of the server-side transferred model on the hospital-side learning model.

CNN-based COVID19 diagnosis model. The workflow of deep learning-based COVID-19 diagnosis has been shown in Fig. 1c, which can be composed of three main steps, i.e., augmentation of raw chest X-ray or CT scans, training of diagnosis model, and testing of the trained optimization model. The deep learning model is implemented to classify COVID-19-infected patients from the chest dataset, where there have three-category of X-ray or CT scans including, COVID-19, normal health, and common pneumonia. Firstly, the three-category raw chest scans are processed by using an augmentation scheme to address the problem of images' uneven sample distributions. After augmentation, we divide the augmented dataset into three subsets, i.e., training cohort, validation cohort, and testing cohort. The training and validation cohorts act as the input of the training and model optimization in the deep learning-based diagnosis model, while the testing dataset is used to be the testing input. To detect the COVID-19-infected patients, the features of chest X-ray or CT scans are utilized to accurately obtain each patient classification whether he or she belongs to COVID-19 category or not. Finally, the best-trained model is achieved and employed for the decision-making process in the testing cohort. The obtained best deep learning-based diagnosis model can identify the infection categories and classification performances can be determined.

In this research, two transfer learning models are adopted to evaluate the COVID-19 diagnosis performance, called VGG16 and ResNet-18, which are the two most popular CNN architectures to provide easier gradient flow for more efficient imaging training. Both these two models accept color images (i.e., X-ray or CT scans) as an input with the size of 224 x 224 and three channels, i.e. red, green, and blue. VGG-16 follows the arrangement of convolution as well as max pool layers consistently throughout the whole CNN network. At the end of the architecture, there exist two fully connected layers which are followed by a Soft-max for output. The term "16" in VGG16 means that it has 16 layers with network weights. Generally, VGG16 is a large-scale network that has almost 130 million parameters in the whole architecture. ResNet-18 builds a deeper neural network architecture by using skip or shortcut connections to skip one or more layers. This process can assist the learning network in achieving a direct forward to the very early layers, thus making the parameters update for these layers much quicker. Similarly, the term "18" in ResNet-18 represents that it is 18 layers deep.

Model evaluation metrics. Diagnostic performance was evaluated by the following metrics: AUC, overall accuracy (ACC), true positive rate (TPR), false positive rate (FPR), precision, sensitivity (which is also equal to recall or TPR), specificity, and F1-score for correctly

distinguishing between COVID-19 and other cases (i.e., normal health and common pneumonia), where the metrics can be respectively defined as

$$ACC = \frac{True\ \ positive\ (TP) + True\ nagative\ (TN)}{Total\ number\ of\ \ tested\ \ scans}, \tag{6}$$

$$Precision = \frac{TP}{TP + False\ \ positive(FP)}, \tag{7}$$

$$Sensitivity = \frac{TP}{TP + False\ negative(FN)}, \tag{8}$$

$$Specificity = \frac{TN}{FP + TN}, \tag{9}$$

$$FPR = \frac{FP}{FP + TN}, \tag{10}$$

$$F1 - score = \frac{2 \times Precision \times Sensitivity}{Precision + Sensitivity}, \tag{11}$$

where true positive (TP) denotes the number of correctly identified COVID-19 scans, true negative (TN) is the number of currently identified Non-COVID-19 scans, false negative (FN) is the number of COVID-19 scans miss-classified as Non-COVID-19, and false positive (FP) is the number of Non-COVID-19 miss-classified as COVID-19.

FTL system versus centralized and local learning systems. In this study, we have compared the diagnostic accuracy of the proposed FTL system to classical centralized and local learning systems. The centralized learning system has all chest X-ray and CT scans collected from all the participating hospitals. Thus, the chest scans of patients at each hospital need to be uploaded to the central server for global diagnosis model optimization, and a huge amount of chest scans is required to be shared with the central cloud server. In contrast, in the local learning system, each hospital acts as an independent unit to perform model training based on its observed medical data instead of sharing any information with other medical institutions or the central server. Therefore, we could observe that the centralized learning system achieved the best diagnostic performance (Fig.3 and Table 3) based on a sufficient training dataset, while the local learning system obtained the poorest performance due to the insufficient training dataset or knowledge. However, the private information of patients will be leaked in the centralized learning system; thus medical institutions, patients, and governments may not agree to share their raw medical data in public. The proposed FTL system only collaborates the AI model parameters from participating hospitals to train a global model instated of uploading the sensitive medical data of patients, which does not leak any private information to each other.

Data availability

The databases used in this study are mostly publicly available. For the chest X-ray dataset, the corresponding databases can be accessed at https://www.kaggle.com/tawsifurrahman/covid19-radiography-database, https://github.com/ieee8023/covid-chestxray-dataset, https://www.kaggle.com/prashant268/chest-xray-covid19-pneumonia, and https://github.com/rgbnihal2/COVID-19-X-ray-Dataset. For the chest CT dataset, a portion of this research used data from publicly available datasets at https://www.kaggle.com/anaselmasry/ct-images-for-covid-normal-pneumonia-mendeley, https://www.kaggle.com/plameneduardo/a-covid-multiclass-dataset-of-ct-scans, and https://github.com/mohammad2682/Covid19-Dataset. Interested readers can contact the corresponding author about the datasets used in this study.

Code availability

All models presented in the research were developed using the python3 platform. To assist researchers in reproducing the COVID-19 learning-based diagnosis models and experimental results, some corresponding code and models have been provided at https://www.kaggle.com/yangyang123765/aiforcovid19diagnosis, and more codes will be provided in this website when the paper is accepted or published online. Besides, a part of the analytic code has been submitted in the journal system for reviewers' checking.

References

1. Li, Q. et al. Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. *N. Engl. J. Med.* **382**, 1199–1207 (2020).
2. Mahase, E. Coronavirus covid-19 has killed more people than SARS and MERS combined, despite lower case fatality rate. *BMJ* **368**, m641 (2020).
3. World Health Organization. *Coronavirus 2019 (COVID-19)* (World Health Organization, 2020). https://covid19.who.int/.
4. Fang, Y. et al. Sensitivity of chest CT for covid-19: comparison to RT-PCR. *Radiology* **296**, E115–E117 (2020).
5. Ai, T. et al. Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases. *Radiology* **296**, E32–E40 (2020).
6. Huang, C. et al. Clinical features of patients infected with 2019 Novel Coronavirus in Wuhan China. *Te Lancet* **395**, 497–506 (2020).
7. Feng, Z. et al. Early prediction of disease progression in COVID-19 pneumonia patients with chest CT and clinical characteristics. *Nat. Commun.* **11**, 4968 (2020).
8. Rubin, G. D. et al. The role of chest imaging in patient management during the COVID-19 pandemic: a multinational consensus statement from the Fleischner Society. *Chest* **158**, 106–116 (2020).
9. 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).
10. Fourcade, A. & Khonsari, R. H. Deep learning in medical image analysis: A third eye for doctors. *J. of stomatology* **120**, 279-288 (2019).

11. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
12. Coudray, N. et al. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat. Med.* **24**, 1559–1567 (2018).
13. Nash, M. et al. Deep learning, computer-aided radiography reading for tuberculosis: a diagnostic accuracy study from a tertiary hospital in India. *Sci. Rep.* **10**, 1–10 (2020).
14. Qin, Z. Z. et al. Using artificial intelligence to read chest radiographs for tuberculosis detection: a multi-site evaluation of the diagnostic accuracy of three deep learning systems. *Sci. Rep.* **9**, 1–10 (2019).
15. 15S Xie, H., Yang, D., Sun, N., Chen, Z. & Zhang, Y. Automated pulmonary nodule detection in CT images using deep convolutional neural networks. *Pattern Recognit.* **85**, 109–119 (2019).
16. Ardila, D. et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat. Med.* **25**, 954–961 (2019).
17. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).
18. Zhang, K. et al. Clinically applicable AI system for accurate diagnosis, quantitative measurements and prognosis of COVID-19 pneumonia using computed tomography. *Cell* **181**, 1423–1433 (2020).
19. 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).
20. Mei, X. et al. Artificial intelligence–enabled rapid diagnosis of patients with COVID-19. Nat. Med. 26, 1224–1228 (2020).
21. Yousefzadeh, M. et al. AI-corona: Radiologist-assistant deep learning framework for COVID-19 diagnosis in chest CT scans. *medRxiv*. Preprint at https://doi.org/10.1101/2020.05.04.20082081 (2020).
22. Chen, J. et al. Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography. *Sci. Rep.* **10**, 19196 (2020).
23. Martin, A. et al. An artificial intelligence-based first-line defence against COVID-19: digitally screening citizens for risks via a chatbot. Sci. Rep. 10, 19012 (2020).
24. Warman, A. et al. Interpretable artificial intelligence for COVID-19 diagnosis from chest CT reveals specificity of ground-glass opacities. *medRxiv*. Preprint at https://doi.org/10.1101/2020.05.16.20103408 (2020).
25. Dhurgham et al. AI based chest X-ray (CXR) scan texture analysis algorithm for digital test of COVID-19 patients. *medRxiv*. Preprint at https://doi.org/10.1101/2020.05.05.20091561 (2020).
26. Nishio, M. et al. Automatic classification between COVID-19 pneumonia, non-COVID-19 pneumonia, and the healthy on chest X-ray image: combination of data augmentation methods. *Sci. Rep.* **10**, 17532 (2020).
27. Apostolopoulos, I. D., & Mpesiana, T. A. Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks. *Phys. Eng. Sci. Med.* **43**, 635–640 (2020).
28. Yeh, C.-F. et al. A cascaded learning strategy for robust covid-19 pneumonia chest x-ray screening. Preprint at https://arxiv.org/abs/2004.12786 (2020).
29. Ucar, F., & Korkmaz, D. COVID Diagnosis-net: deep bayes-squeezenet based diagnosis of the coronavirus disease 2019 (COVID-19) from x-ray images. *Med. Hypotheses.* https://doi.org/10.1016/j.mehy.2020.109761 (2020).
30. Jin, C. et al. Development and evaluation of an artificial intelligence system for COVID-19 diagnosis. *Nat. Commun.* **11**, 5088 (2020).
31. Peiffer-Smadja, N. et al. Machine learning for COVID-19 needs global collaboration and data-sharing. *Nat. Mach. Intelll.* **2**, 293–294 (2020).
32. Kaissis, G.A. et al. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat. Mach. Intell.* **2**, 305–311 (2020).
33. Liu, B., Yan, B., Zhou, Y., Yang, Y. & Zhang Y. Experiments of federated learning for COVID-19 chest X-ray images. Preprint at https://arxiv.org/abs/2007.05592 (2020).

34. Vaid, A. Federated learning of electronic health records improves mortality prediction in patients hospitalized with COVID-19. *medRxiv*. Preprint at https://doi.org/10.1101/2020.08.11.20172809 (2020).
35. Xu, Y. A collaborative online AI engine for CT-based COVID-19 diagnosis, *medRxiv*. Preprint at https://doi.org/10.1101/2020.05.10.20096073 (2020).
36. Liu, Y., Kang, Y., Xing, C., Chen T. & Yang, Q. A secure federated transfer learning framework. *IEEE Intell. Sys.* **35**, 70-82 (2020).
37. Yang, H., He, H., Zhang, W., & Cao, X. FedSteg: A federated transfer learning framework for secure image steg-analysis. *IEEE Trans. Netw. Sci. Eng.* Preprint at https://ieeexplore.ieee.org/document/9099064 (2020)
38. Chen, Y., Qin, X., Wang, J., Yu, C. & Gao, W. FedHealth: A federated transfer learning framework for wearable healthcare. *IEEE Intell. Sys.* **35**, 83-93 (2020).
39. Maaten, Lvd & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
40. Rozantsev, A., Salzmann, M. & Fua, P. Beyond sharing weights for deep domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**, 801–814 (2018).
41. Long, M., Cao, Y., Wang, J. & Jordan M. I. Learning transferable features with deep adaptation networks. Preprint at https://arxiv.org/abs/2011.05574 (2015).
42. ACR. Recommendations for the use of chest radiography and computed tomography (CT) for suspected COVID-19 infection. https://wwwacrorg/ Advocacy-and-Economics/ACR-Position-Statements/ Recommendations-for Chest-Radiography-and-CT-for-Suspected-COVID19-Infection (2020).
43. Konečný, J. et al. Federated optimization: distributed machine learning for on-device intelligence. Preprint at https://arxiv.org/abs/1610.02527 (2016).
44. Sun, B., Feng, J. & Saenko K. Return of frustratingly easy domain adaptation. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2058–2065 (Springer, 2016).

Acknowledgements

This work is support by Nanyang Technological University (NTU) Startup Grant, Singapore Ministry of Education Academic Research Fund, Singapore National Research Foundation under its Strategic Capability Research Centres Funding Initiative: Strategic Centre for Research in Privacy-Preserving Technologies & Systems, Design Science and Technology for Secure Critical Infrastructure NSoE DeST-SCI2019-0012 and AI Singapore 100 Experiments (100E) programme.

Author contributions

H.L.Y conceived the ideal, wrote the paper, collected data, and designed experiment. J.Z., K.-Y.L., and Z.H.X. assisted to implement model, analyze the experiment data and results. L.X. and Q.Q.W contributed to drafting and revising the manuscript.

Competing interests

The authors declare no competing interests.

Figures and Tables

Table 1: The split of chest X-ray and CT scans in training, validation, and testing cohorts.

Table 1 Chest X-ray and CT cohorts used in system development and testing.

Scan type	Category	Training	Validation	Testing	In total
a Chest X-ray	COVID-19	609	173	488	1270
	Normal	1098	293	648	2039
	Pneumonia	1024	237	656	1917
	In total	2731	703	1792	5226
b Chest CT	COVID-19	991	295	495	1781
	Normal	1379	309	515	2203
	Pneumonia	1219	369	619	2207
	In total	3529	973	1629	6191

Note: the datasets of chest X-ray and CT scans are from different countries or cities, which are the publicly available databases (see data availability in Methods).

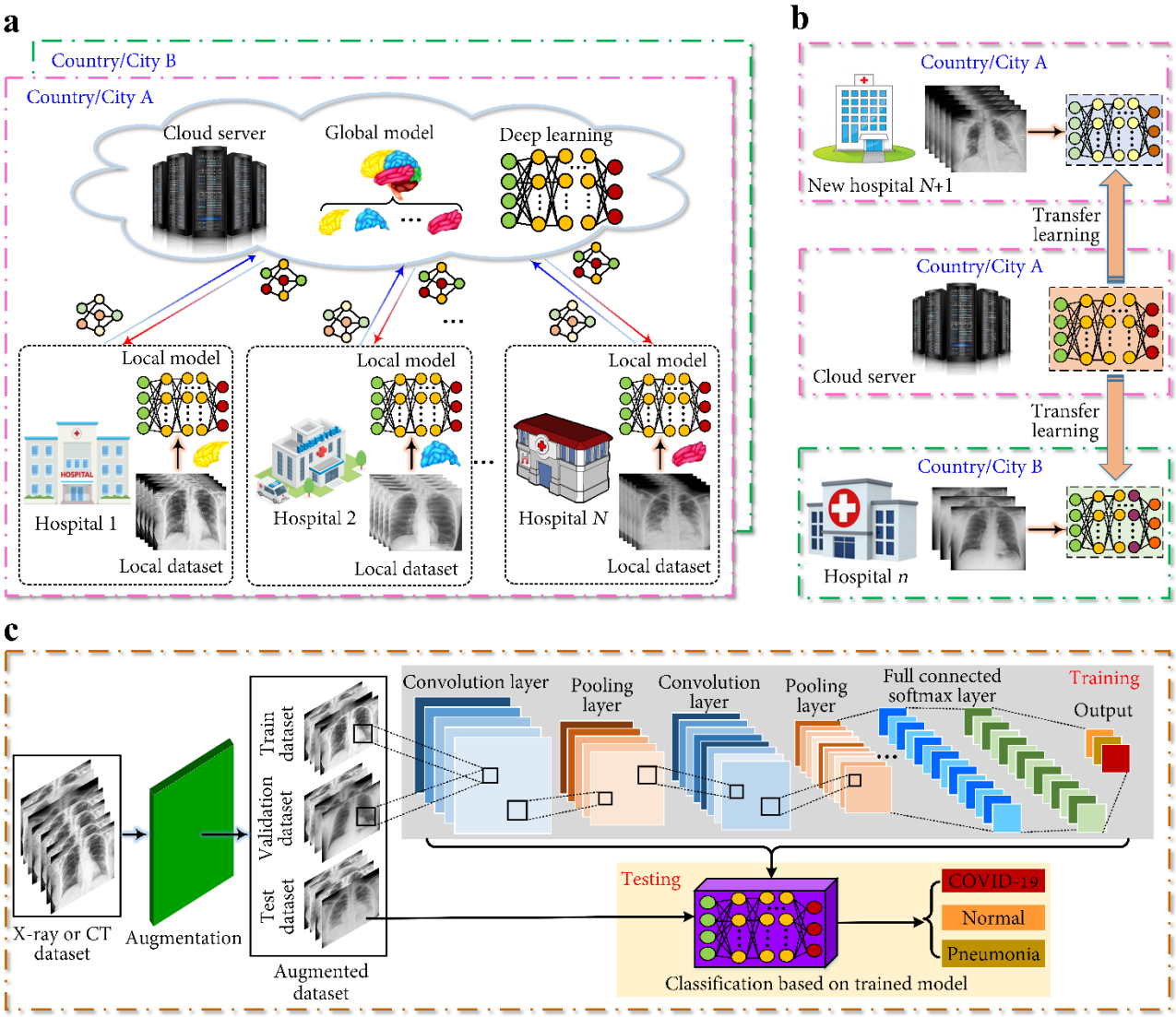


Fig. 1 Schematic diagram of the proposed privacy-preserving FTL system for COVID-19 diagnosis. a Schematic of federated learning framework for COVID-19 diagnosis, where the AI model is trained in a distributed scenario to preserve the private information of patients. Multiple participating hospitals locally train their learning-based diagnosis models and upload their AI models to a central cloud server instead of sending a huge amount of raw medical dataset. The cloud server incorporates the local AI models and broadcasts the parameters of the updated global model to participating hospitals. **b** Structure of transfer learning model for the federated deep neural network-based COVID-19 diagnosis, where each hospital can build its personally learning-based diagnosis model by using transferred knowledge from other medical institutions or the central server. **c** Architecture of the training and testing process of the deep neural network-based COVID-19 diagnosis model.

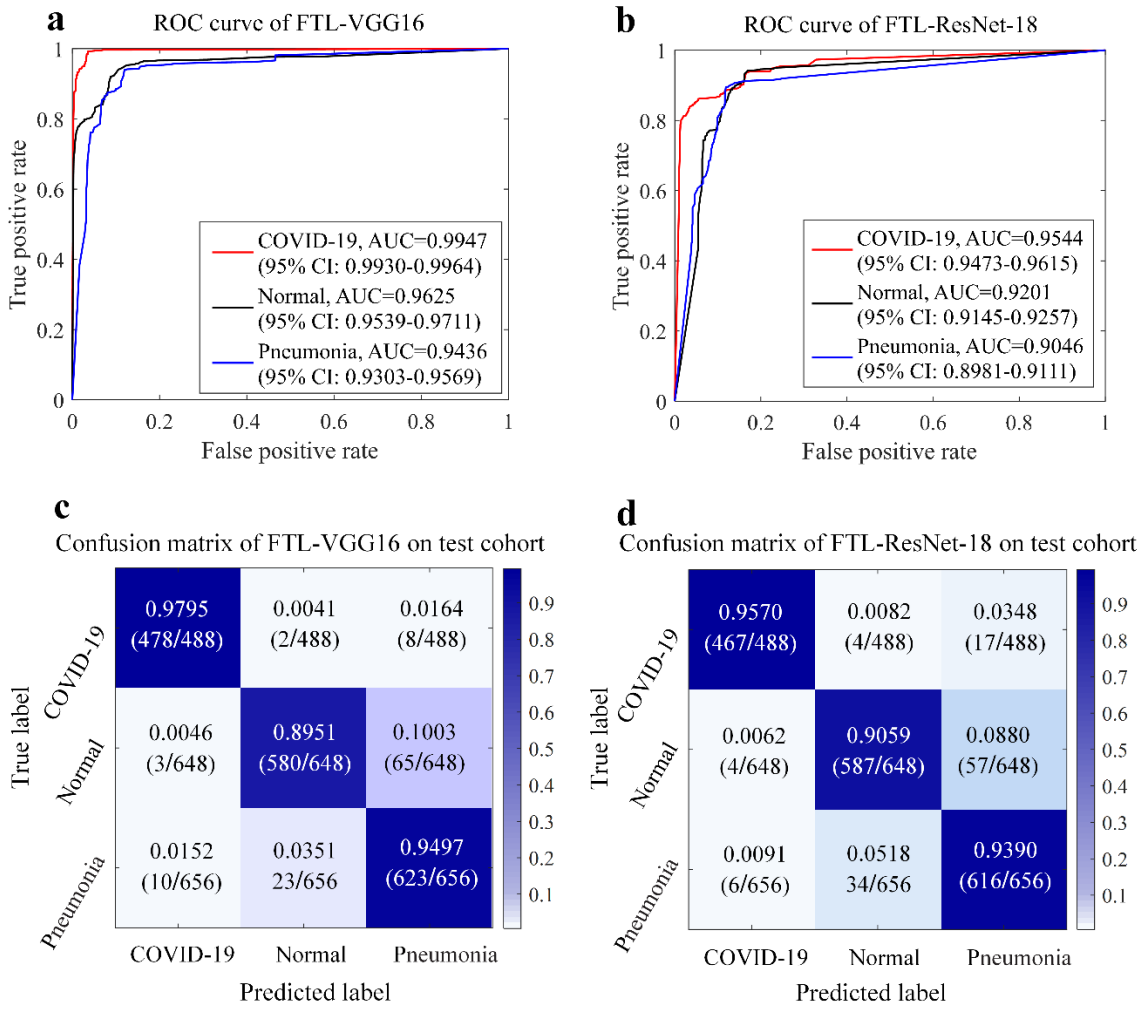


Fig. 2 Diagnostic performances of the FTL system on chest X-ray scans. a, b Evaluations of ROC curves for VGG16 model and ResNet-18 model on the three-category training-validation dataset, respectively. True positive rate (TPR) is plotted on the y-axis, while false positive rate (FPR) is plotted on the x-axis. **c, d** Confusion matrix on the three diagnosis categories for VGG16 model and ResNet-18 model, respectively, where the trained models are loaded to perform COVID-19 diagnosis on the test cohort.

Table 2: Diagnostic performances of FTL system on chest X-ray test cohorts in terms of VGG19 model and ResNet-18 model. The performance metrics (i.e., precision, sensitivity, specificity, F1-score, and AUC) for COVID-19 diagnosis, normal diagnosis, and pneumonia diagnosis were listed in this table at a 95% CI.

Table 2 Diagnostic performances of FTL system on chest X-ray test cohorts

Model	Category	Precision (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	F1-score (95% CI)	AUC (95% CI)
a VGG16	COVID-19	0.9874 (0.9765-0.9983)	0.9495 (0.9379-0.9611)	0.9211 (0.9064-0.9357)	0.9680 (0.9619-0.9742)	0.9974 (0.9963-0.9985)
	Normal	0.8751 (0.8502-0.9001)	0.9612 (0.9520-0.9704)	0.9051 (0.8893-0.9210)	0.9161 (0.9011-0.9312)	0.9752 (0.9742-0.9762)
	Pneumonia	0.9512 (0.9410-0.9615)	0.8875 (0.8644-0.9106)	0.9574 (0.9493-0.9655)	0.9182 (0.9041-0.9323)	0.9613 (0.9599-0.9627)
b ResNet-18	COVID-19	0.9322 (0.9057-0.9586)	0.9496 (0.9286-0.9707)	0.9186 (0.9066-0.9305)	0.9407 (0.9269-0.9546)	0.9814 (0.9811-0.9817)
	Normal	0.9052 (0.8883-0.9222)	0.9422 (0.9310-0.9534)	0.9117 (0.8981-0.9253)	0.9233 (0.9123-0.9344)	0.9474 (0.9464-0.9485)
	Pneumonia	0.9386 (0.9239-0.9532)	0.8972 (0.8821-0.9121)	0.9444 (0.9330-0.9558)	0.9174 (0.9043-0.9305)	0.9348 (0.9332-0.9364)

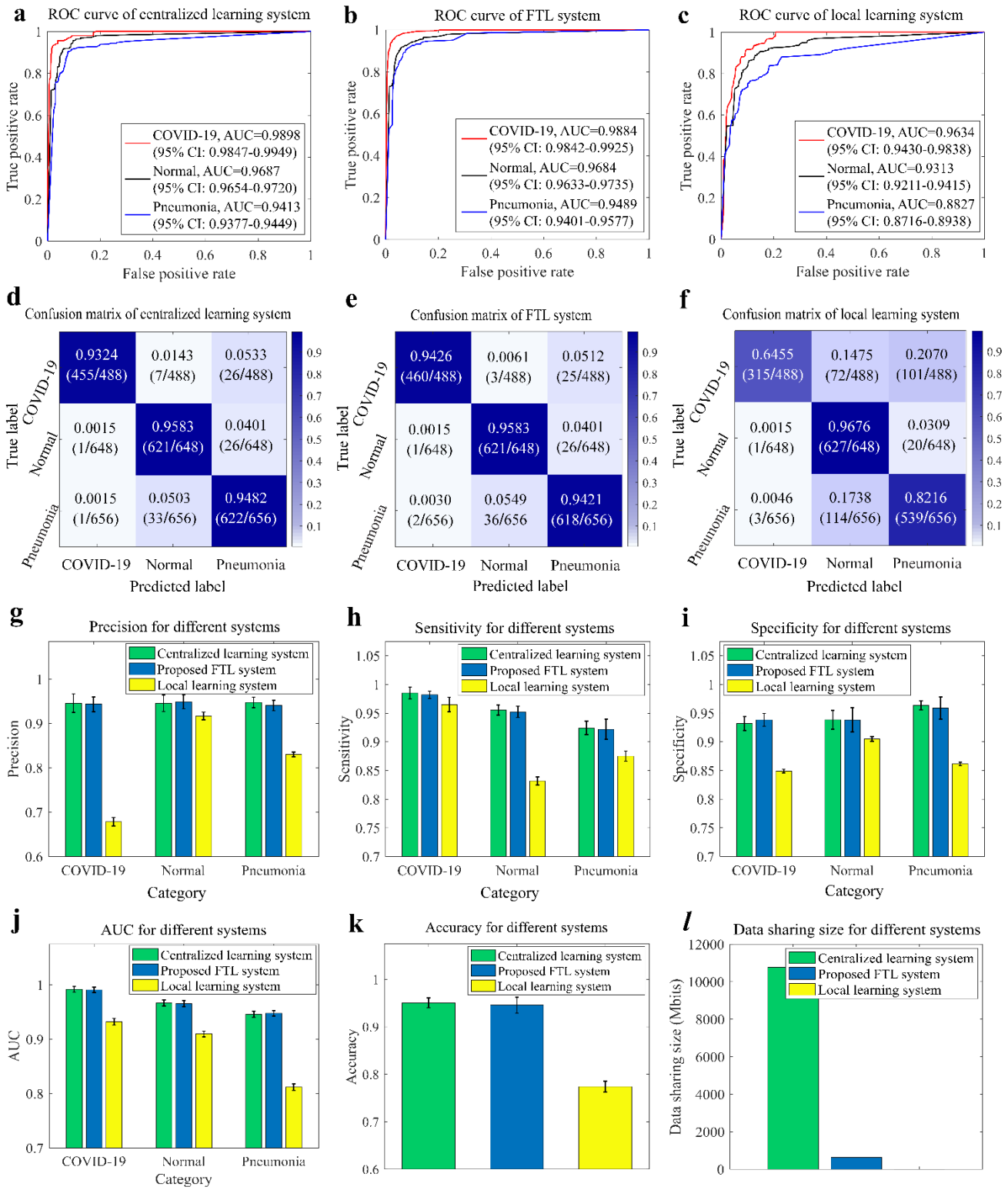
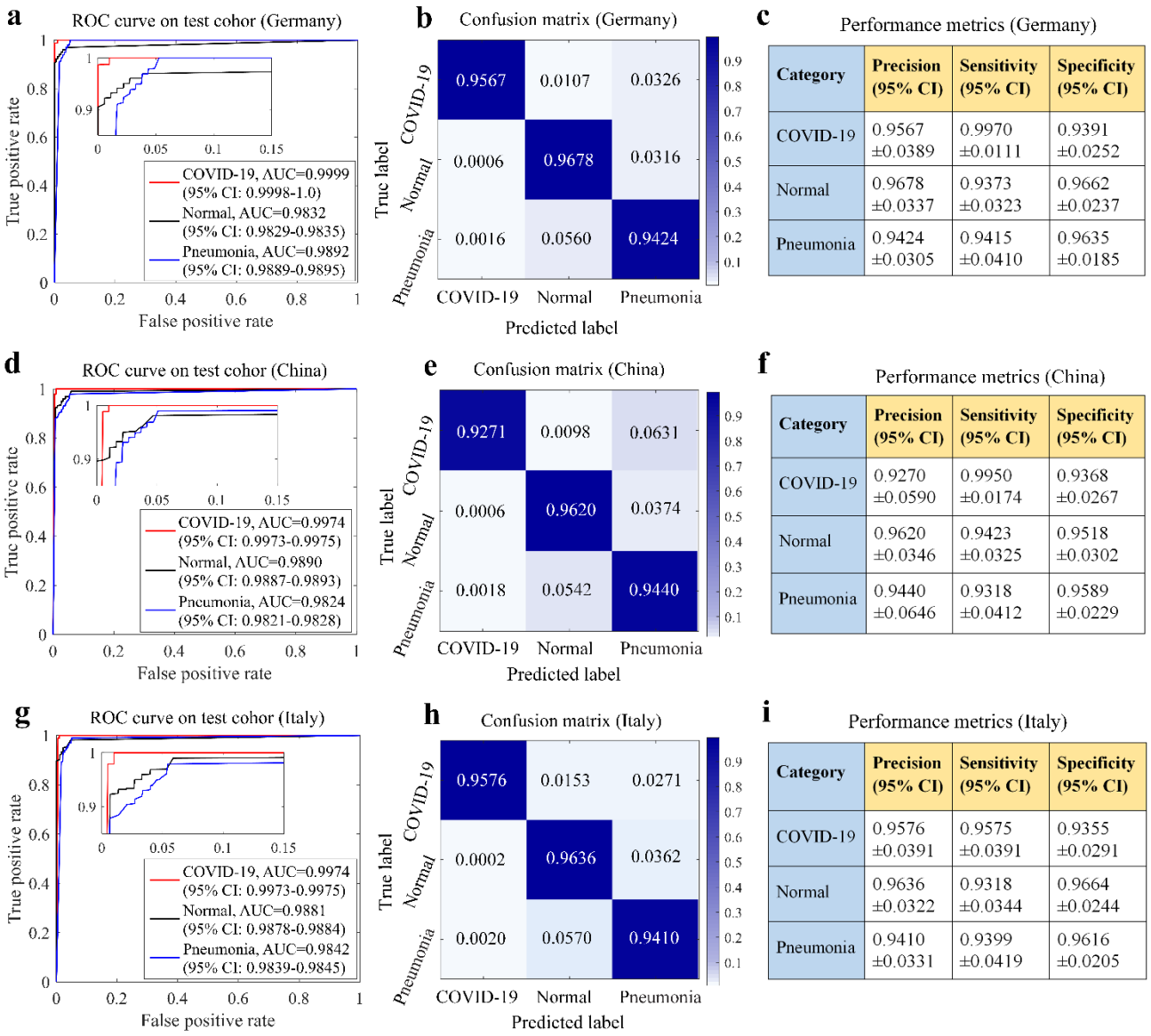


Fig. 3 Performance comparisons of different AI systems in the chest X-ray dataset. a, b, c Comparisons of ROC curves for the centralized learning system, the proposed FTL system, and the local learning system on the training-validation cohort. TPR is plotted on the y-axis, while FPR is plotted on the x-axis. The AUC values of COVID-19 diagnosis, normal diagnosis, and common

pneumonia diagnosis are also presented in Fig. 3a–c. **d, e, f** Confusion matrices (of three-category classification task) of the three compared AI systems on the test cohort, respectively. **g, h, i, j, k** Comparisons of the precision, sensitivity, specificity, AUC, and overall accuracy of the three compared AI systems on the test cohort, respectively. ***l*** Data sharing overhead in Mbits using different AI systems.



Category	Precision (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
COVID-19	0.9567 ±0.0389	0.9970 ±0.0111	0.9391 ±0.0252
Normal	0.9678 ±0.0337	0.9373 ±0.0323	0.9662 ±0.0237
Pneumonia	0.9424 ±0.0305	0.9415 ±0.0410	0.9635 ±0.0185

Category	Precision (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
COVID-19	0.9270 ±0.0590	0.9950 ±0.0174	0.9368 ±0.0267
Normal	0.9620 ±0.0346	0.9423 ±0.0325	0.9518 ±0.0302
Pneumonia	0.9440 ±0.0646	0.9318 ±0.0412	0.9589 ±0.0229

Category	Precision (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
COVID-19	0.9576 ±0.0391	0.9575 ±0.0391	0.9355 ±0.0291
Normal	0.9636 ±0.0322	0.9318 ±0.0344	0.9664 ±0.0244
Pneumonia	0.9410 ±0.0331	0.9399 ±0.0419	0.9616 ±0.0205

Fig. 4 Performance evaluation of the FTL System on test chest X-ray cohort in three different countries. a, b, c ROC curves, normalized confusion matrix, and performance metrics on a COVID-19 chest X-ray dataset from Germany. **d, e, f** ROC curves, normalized confusion matrix, and performance metrics on a COVID-19 chest X-ray dataset from China. **g, h, i** ROC curves, normalized confusion matrix, and performance metrics on a COVID-19 chest X-ray dataset from Italy.

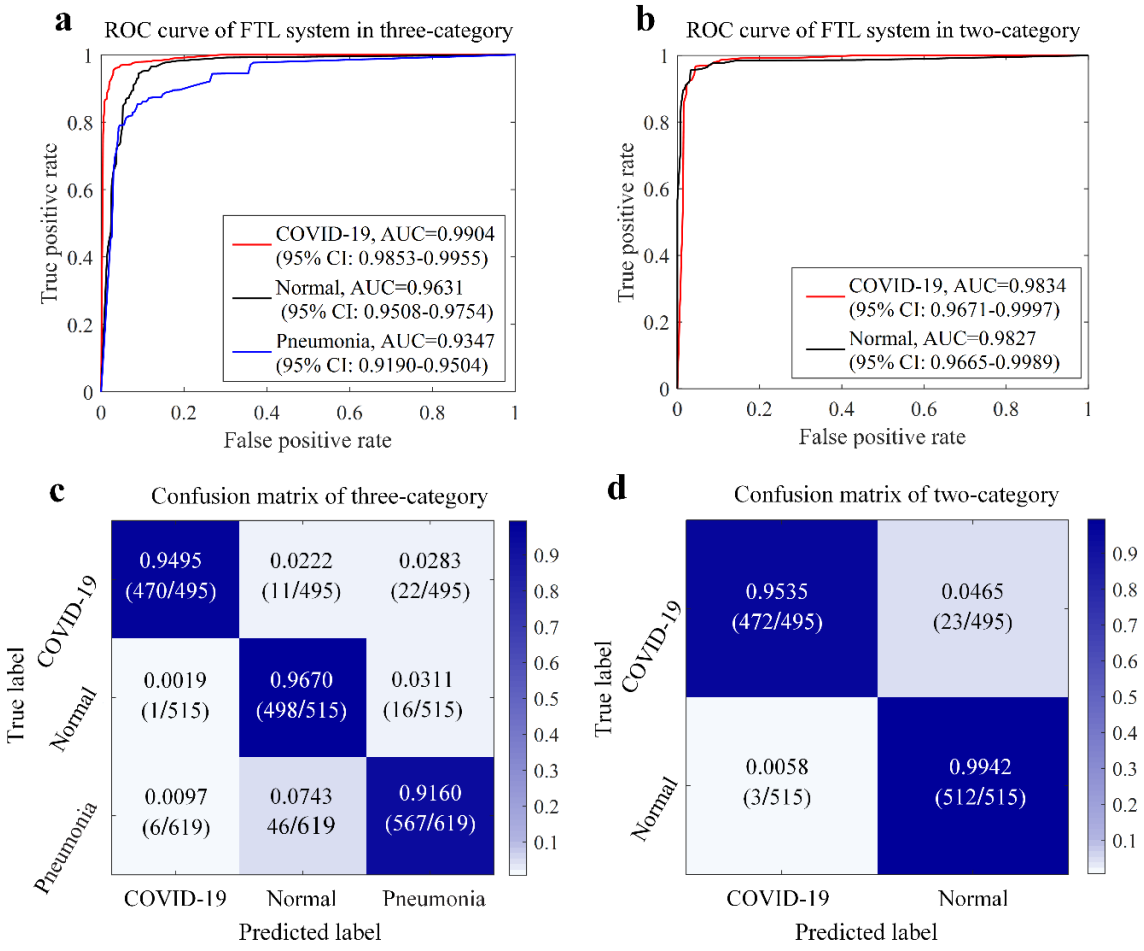


Fig. 5 Diagnostic performances of the FTL system in chest CT dataset. a Evaluation of ROC curve for the FTL system on the three-category training-validation dataset. **b** Evaluation of ROC curve for the FTL system on the two-category training-validation dataset. Note that the three-category diagnosis task (Fig.4a) includes COVID-19, normal, and common pneumonia categories, while the two-category diagnosis task (Fig.4b) contains COVID-19 and normal categories. **c** Confusion matrix of the three diagnostic categories on the test cohort. **d** Confusion matrix of the two diagnostic categories on the test cohort.

Table 3: Diagnostic performances of FTL system on chest CT test cohorts in terms of three-category and two-category tasks. The performance metrics were listed in this table at a 95% CI.

Table 3 Diagnostic performances of FTL system on chest CT test cohorts

Subject	Category	Precision (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	F1-score (95% CI)	AUC (95% CI)
a Three-category	COVID-19	0.9554 (0.9440-0.9668)	0.9468 (0.9280-0.9655)	0.9240 (0.9109-0.9369)	0.9510 (0.9406-0.9613)	0.9721 (0.9713-0.9729)
	Normal	0.9321 (0.9066-0.9577)	0.9205 (0.9084-0.9325)	0.9358 (0.9204-0.9513)	0.9262 (0.9114-0.9410)	0.9548 (0.9540-0.9556)
	Pneumonia	0.9103 (0.8943-0.9262)	0.9270 (0.9050-0.9490)	0.9332 (0.9211-0.9454)	0.9185 (0.9044-0.9326)	0.9492 (0.9501-0.9483)
b Two-category	COVID-19	0.9492 (0.9372-0.9612)	0.9838 (0.9723-0.9954)	0.9528 (0.9422-0.9634)	0.9662 (0.9587-0.9737)	0.9772 (0.9767-0.9777)
	Normal	0.9850 (0.9740-0.9959)	0.9528 (0.9422-0.9634)	0.9838 (0.9723-0.9954)	0.9686 (0.9616-0.9755)	0.9768 (0.9757-0.9778)

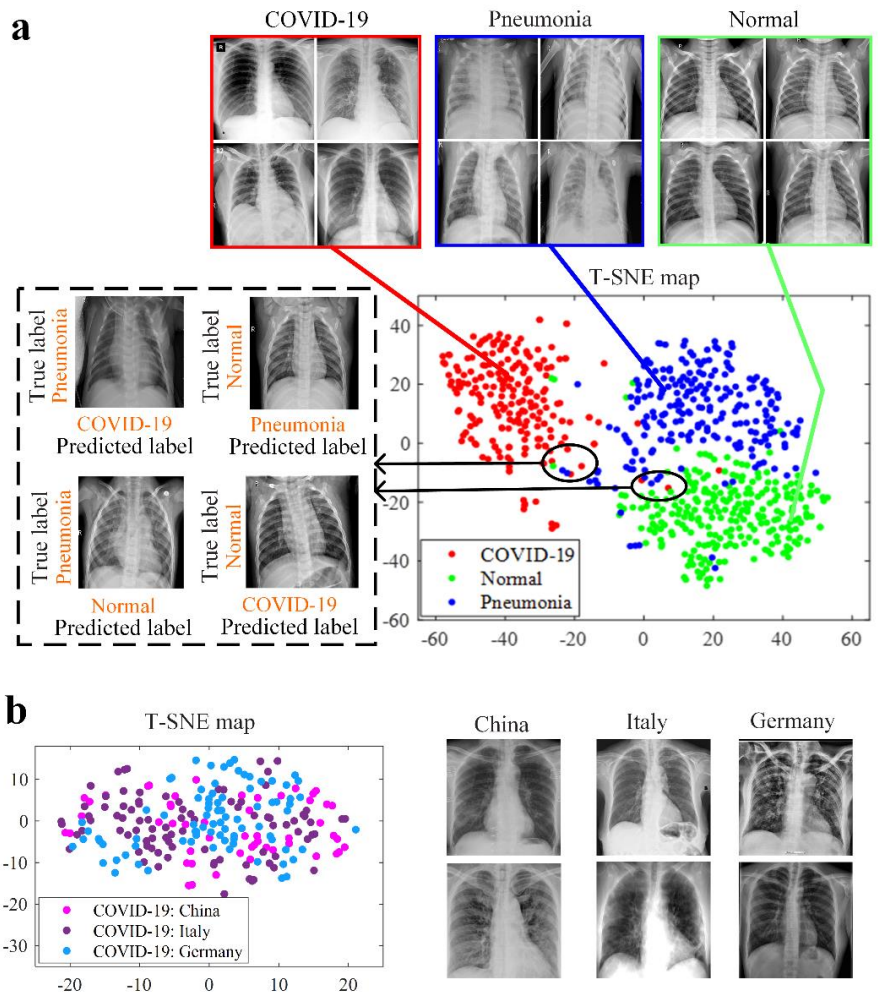


Fig. 6 Interpreting the FTL system on chest X-ray. a Visualizing chest scan features using *t-SNE*. It also provides some examples of incorrect diagnosis in three-category classification test using the FTL system. **b** Visualizing features of different countries in term of COVID-19 X-ray using *t-SNE*. Here we take the chest scans from China, Italy, and Germany as an example.

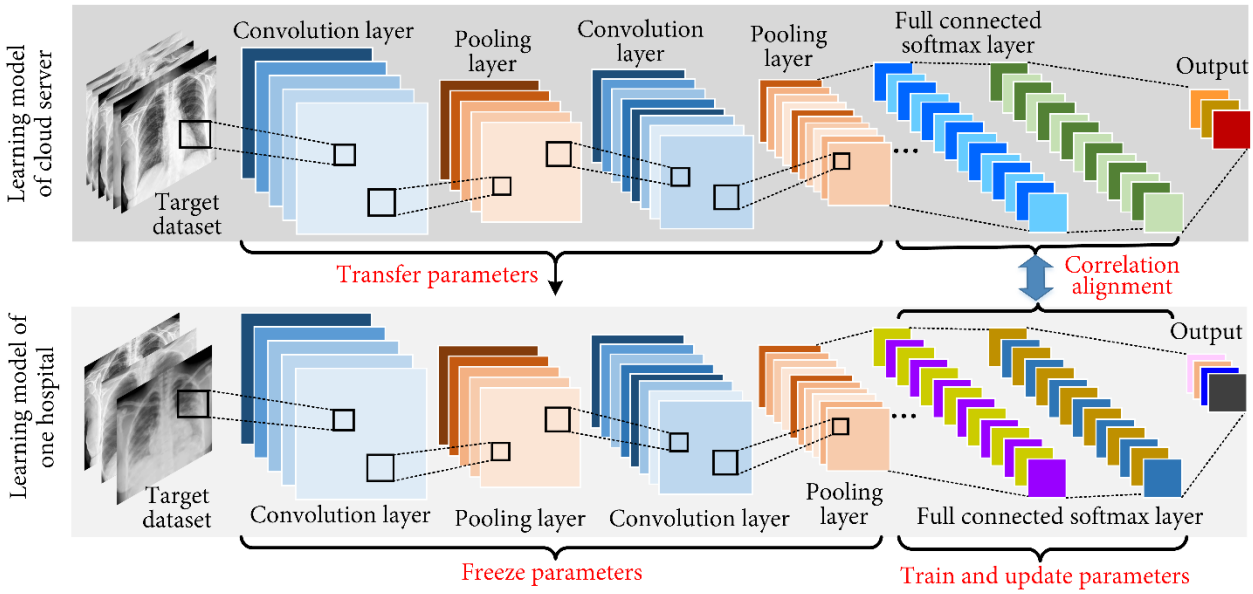


Fig. 7 Transfer learning process of the proposed federated CNN-based COVID19 diagnosis model.