A proposed artificial intelligence workflow to address application challenges leveraged on algorithm uncertainty

Dantong Li, Lianting Hu, Xiaoting Peng, ..., Yunfei Gao, Jian Wu, Huiying Liang

lianghuiying@hotmail.com (H.L.)
wujian2000@zju.edu.cn (J.W.)

Highlights

A COVID-19 artificial intelligence diagnosis system with uncertainty estimation
Reliability and optional reliability intervals at dataset level as references
A proposed workflow that could be expanded to other diseases in practice
INTRODUCTION

Enthusiasm around artificial intelligence (AI) in medicine has notably increased. AI systems are usually trained to detect the specific abnormalities in certain fields and then validated on datasets similar to their training environments, which do not accurately reflect the clinical practice as various abnormalities might fail. And the natural gap between training datasets and the in-reality data may lead to unexpected AI system malfunction. Knowing the underlying uncertainty is essential for improving system reliability. Therefore, we developed a COVID-19 AI system, utilizing a Bayesian neural network to calculate uncertainties in classification and reliability intervals of datasets. Validated with four multi-region datasets simulating different scenarios, our approach was proved to be effective to suggest the system failing possibility and give the decision power to human experts in time. Leveraging on the complementary strengths of AI and health professionals, our present method has the potential to improve the practicability of AI systems in clinical application.

SUMMARY

Artificial Intelligence (AI) has achieved state-of-the-art performance in medical imaging. However, most algorithms focused exclusively on improving the accuracy of classification while neglecting the major challenges in a real-world application. The opacity of algorithms prevents users from knowing when the algorithms might fail. And the natural gap between training datasets and the in-reality data may lead to unexpected AI system malfunction. Knowing the underlying uncertainty is essential for improving system reliability. Therefore, we developed a COVID-19 AI system, utilizing a Bayesian neural network to calculate uncertainties in classification and reliability intervals of datasets. Validated with four multi-region datasets simulating different scenarios, our approach was proved to be effective to suggest the system failing possibility and give the decision power to human experts in time. Leveraging on the complementary strengths of AI and health professionals, our present method has the potential to improve the practicability of AI systems in clinical application.

A proposed artificial intelligence workflow to address application challenges leveraged on algorithm uncertainty

Dantong Li,1,2,Lianting Hu,1,2,9 Xiaoting Peng,1,2 Ning Xiao,1 Hong Zhao,3 Guangjian Liu,1,2 Hongsheng Liu,4 Kuanrong Li,4 Bin Ai,4 Huimin Xia,4 Long Lu,4,6 Yunfei Gao,7,8 Jian Wu,5,10,* and Huiying Liang1,2,10,*

SUMMARY

Artificial Intelligence (AI) has achieved state-of-the-art performance in medical imaging. However, most algorithms focused exclusively on improving the accuracy of classification while neglecting the major challenges in a real-world application. The opacity of algorithms prevents users from knowing when the algorithms might fail. And the natural gap between training datasets and the in-reality data may lead to unexpected AI system malfunction. Knowing the underlying uncertainty is essential for improving system reliability. Therefore, we developed a COVID-19 AI system, utilizing a Bayesian neural network to calculate uncertainties in classification and reliability intervals of datasets. Validated with four multi-region datasets simulating different scenarios, our approach was proved to be effective to suggest the system failing possibility and give the decision power to human experts in time. Leveraging on the complementary strengths of AI and health professionals, our present method has the potential to improve the practicability of AI systems in clinical application.

INTRODUCTION

Enthusiasm around artificial intelligence (AI) in medicine has notably increased. AI systems are usually trained to detect the specific abnormalities in certain fields and then validated on datasets similar to their training environments, which do not accurately reflect the clinical practice as various abnormalities might fail. And the natural gap between training datasets and the in-reality data may lead to unexpected AI system malfunction. Knowing the underlying uncertainty is essential for improving system reliability. Therefore, we developed a COVID-19 AI system, utilizing a Bayesian neural network to calculate uncertainties in classification and reliability intervals of datasets. Validated with four multi-region datasets simulating different scenarios, our approach was proved to be effective to suggest the system failing possibility and give the decision power to human experts in time. Leveraging on the complementary strengths of AI and health professionals, our present method has the potential to improve the practicability of AI systems in clinical application.

Ideally, algorithms need to be trained and validated on the full spectrum of disease or datasets of different quality in a certain field, which could barely be accomplished due to the high variability in real-world clinical situations. For now, it seems that the best use of AI in medicine is to put it in the position of an assistant. In order to do that, we need to have a clear understanding of where tools can be used effectively. However, there are two main challenges lying ahead: (i) algorithm uncertainty (Gawlikowski et al., 2021; Roy et al., 2019): the lack of transparency of a deep neural network leads to the unknown reliability of their results, thus preventing users from knowing when the algorithms might fail; (ii) data uncertainty (Finlayson et al., 2020; Gawlikowski et al., 2021; Rabanser et al., 2018): datasets in the real world are naturally heterogeneous. Similar cases in different datasets may have different uncertainties. Not knowing the acceptable range of results, uncertainties leaves users unaware of the implicit AI malfunction caused by out-of-domain cases or datasets. Relying on uncertainty estimation to adapt the decision-making process might be key to preventing unintended errors (Ayhan et al., 2020). Therefore, we formulate the real-world application as a two-step problem and come up with a two-step method. Firstly, we aim to enable an algorithm to provide its uncertainty estimates to its human collaborators for each prediction, thereby reducing the opacity of the result by revealing whether it could be trusted. Secondly, based on the premise that one dataset shares the same quality and distribution, knowing the range of reliable uncertainties is crucial for further judgment.
Therefore, we propose the concepts of reliability and reliability intervals (RI). An RI is a range of uncertainties computed at a designated reliability level for the corresponding dataset. The uncertain predictions could be decided by experts to ignore or further evaluate based on whether the uncertainty is within the RI.

Bayesian inference is one of the algorithms that could estimate uncertainty by computing the conditional probabilities of multiple diagnoses/classes based on features and their prior probability (Gawlikowski et al., 2021), and Monte Carlo estimation is often used in variational inference to estimate the expected log-likelihood (Kendall, 2019). Bayesian inference has been widely used to provide probability-ranked differential diagnoses and to mimic the way of “clinical thinking” (Rauschecker et al., 2020; Rudie et al., 2020). Abhijit Guha Roy et al. envisaged that the introduced uncertainty metrics would help assess the fidelity of automated deep learning-based methods for large-scale population studies, as they enable automated quality control and group analyses in processing large data repositories (Roy et al., 2019). However, few of the previous studies explored the clinical use of uncertainty and its value in a real-world application. Instead of pursuing state-of-the-art performance, we aim to leverage the Bayesian neural network (BNN) to create practical tools for health professionals. The system was designed to follow the two-step method and could alert the users in practice by revealing itself when the inherent uncertainty exceeds the RI. Since the COVID-19 pandemic has been fast consumed, healthcare resources and a practical tool is desperately needed, we used COVID-19 as a case for the benchmark.

The main contributions of our paper are as follows:

- We proposed the concept of reliability and RI as dataset-level references. Utilizing BNN, we developed a COVID-19 AI system that could report the prediction reliability by revealing if the prediction uncertainty is within the corresponding RI. What’s more, the expert collaborators could determine a reliability level and define the trustworthy RI based on their judgment of the data. In this way, the health professionals could leverage the strength of AI in rapid data processing for regular cases and take over the decision right for complicated cases where the system might be essentially guessing at random.

- We simulated four clinical scenarios with multicenter datasets to validate if the system could function in the real world. By reviewing the success and failure cases combined with their uncertainty and RI, we implement an initial “debug” for the system, which is crucial for AI’s actual application.

- We proposed a workflow that might be able to be expanded to other diseases and explored a way for future AI practical applications. The human-machine convergence might improve the overall performance and earn more faith in AI algorithms in real-world settings.

RESULTS
Datasets and the study design
Our study was based on the China Consortium of Chest CT Image Investigation (CC-CCII) dataset (Zhang et al., 2020) (consisted of normal, COVID-19, and common pneumonia (CP, other viral pneumonia, bacterial pneumonia, and mycoplasma pneumonia included), details see supplemental information) and four unique datasets named after their characteristics or different regions (tuberculosis (TB) and chronic obstructive pulmonary disease (COPD), Children, Linyi, and Zhejiang, Figure 1), where 4401 CT scans were employed (Table 1, Figure S1).

The proposed AI system consisted of four main steps including CT slices segmentation and reconstruction, fully convolutional network (FCN) training with the patch-wise training strategy, high-risk feature selection and prediction by BNN with uncertainty estimation, and RI as references. The system was trained and cross-validated using 80% of the patients in CC-CCII and tested on the other 20%, which was used as an internal validation that had never before seen in the training process. To verify the performance of the proposed method, we developed a traditional 3D convolutional neural network (CNN) and compared the performance of the two algorithms. We then forced the system directly to the four datasets, to address their regional variations and general applicability.

The AI system achieved good performance in internal and external validations
Figures 2 and S2 illustrated the diagnostic performance of the AI system. In the internal validation of CC-CCII, the system showed good performance in COVID-19 vs. others (AUC: 0.9759, accuracy: 0.9628, sensitivity: 0.9493, and specificity: 0.9716) and mild vs. severe (AUC: 0.9819, accuracy: 0.9538, sensitivity: 0.9333,
The performance of the proposed AI system was therefore superior to the traditional 3D CNN in COVID-19 vs. others (AUC: 0.9132, accuracy: 0.9026, sensitivity: 0.8911, and specificity: 0.9101) and in mild vs. severe (AUC: 0.8781, accuracy: 0.8571, sensitivity: 0.8621, and specificity: 0.8526) (Table S1).

The performance of the AI system was also validated with three independent datasets, Zhejiang, Linyi, and Children datasets (Figure 2, Table S2). Based on the Linyi dataset, the performance of the system in distinguishing COVID-19 pneumonia from CP and normal was excellent (AUC: 0.9249, accuracy: 0.9315, sensitivity: 0.9344, and specificity: 0.9167). The performance of the model declined in Zhejiang dataset (AUC: 0.8989, accuracy: 0.8357, sensitivity: 0.8382, and specificity: 0.8304). For the Children dataset, the AUC of the model was 0.9205, whereas the accuracy, sensitivity, and specificity of the model were 0.8947, 0.9167, and 0.8571, respectively. And for mild vs. severe in patients with COVID-19, the performance of the model for the Linyi dataset was good (AUC: 0.9611, accuracy: 0.9344, sensitivity: 0.8889, and specificity: 1.00) and again dropped a little bit in the Zhejiang dataset (AUC: 0.9145, accuracy: 0.9046, sensitivity: 0.8381, and specificity: 0.9559). This may be due to that some patients in Zhejiang dataset were also suffered from comorbidities which might affect the system diagnosis. The performance of the model for Children dataset was not evaluated because that no severe cases were included in the data collected. We worked with four junior radiologists with 4–8 years of clinical experience, and four senior radiologists with 10–15 years of clinical experience to diagnose and stratify COVID-19. We then compared the performance between our AI system and radiologists on COVID-19 vs. others and mild vs. severe in external validations (the performance of the radiologists seen in Table S3). Our AI system performance was overall superior to that of junior radiologists and comparable to senior radiologists.

Figure 1. The overview of the proposed artificial intelligence (AI) system

Step 1: the lung areas were segmented from the CT slices by a 2D U-net and then stacked as a 3D volume. Step 2: a 3D FCN was trained by repeated application to 3D patches of voxels that randomly sampled from the 3D volume and generated feature points for different classes by application of a softmax function. Step 3: the trained FCN then took the entire 3D volume as input and generated a feature map and based on which, the high-risk features were selected. Step 4: High-risk features were selected from the disease probability maps and then passed to the BNN for classification and uncertainty calculation. After processing the cases within one dataset, the system was able to generate a reliability curve and calculate the RI for the dataset as an additional reference.
The uncertainty estimations and reliability intervals in different datasets

The case uncertainty for different datasets was presented in a boxplot (Figure S3). The results suggested that the overall uncertainty level in pneumonia datasets (CC-CCII, Children, Linyi, and Zhejiang) was a lot lower than that in TB and COPD, indicating that the system was able to give high uncertainties for diseases that it had never before seen. And among the pneumonia datasets, the uncertainty level for Zhejiang was comparatively higher than the others, which was consistent with the dropped performance in the Zhejiang dataset. Moreover, the uncertainty for misclassified cases was higher than the correct ones. To confirm the relation between the predictive uncertainty and accuracy, we computed Spearman’s correlation coefficient between predictive entropy and prediction errors and the results suggested a close correlation (Table S4).

The above experiments showed that the prediction uncertainty correlates with accuracy, thus enabling the identification of false predictions or unknown cases. The high uncertainty could be used as a way for the system to say “I don’t know” and ask for reevaluation from the human experts.

With different uncertainty range for different datasets, it is crucial to know whether the uncertainty for a case is acceptable. Therefore, we calculated RI for each dataset as a basic reference. As shown in Figure 3, to maintain a reliability level at 95% for COVID-19 diagnosis, the corresponding RI for the CC-CCII dataset (0, 0.22), narrowed to (0.0, 0.13) in the Zhejiang dataset. The same trend was obtained for mild vs. severe, changing from (0.0, 0.43) in the CC-CCII dataset to (0.0, 0.33) in the Zhejiang dataset. The RI of other datasets is presented in Figure 3K. It should be noted the reliability could be changed at the experts’ request and the system could provide corresponding RI for reference.

Case-based system evaluation in different datasets

We reviewed the uncertainty for cases with correct predictions. For each CT scan, the class with the highest softmax output for predictive distribution mean was considered as the system prediction and presented by violin plot. The less concentrated the output probability distributions, the less reliability of the system in its’ prediction. The uncertainty of each case was calculated based on the predictive entropy of the output distributions (measured as in Equation 9 of the STAR Methods). Figure 4 showed a few cases from the four datasets.
Figure 2. Performance of COVID-19 AI system for COVID-19 vs. others (normal and common pneumonia included) and mild vs. severe, and compared with radiologist’s judgment

(A–D) The area under the receiver operating characteristic curve (AUC) of COVID-19 versus other two classes, including CP and normal controls (Normal) in (A) China Consortium of Chest CT Image Investigation (CC-CCII), (B) Zhejiang, (C) Linyi, and (D) Children.

(E–G) AUC of COVID-19 status stratification (mild vs. severe) in (E) CC-CCII, (F) Zhejiang, and (G) Linyi but not in Children validation as no severe cases were found in it. The graphs showed the scores for individual patients in three external datasets, performed by radiologists (4 senior and 4 junior). CI (Table S2).
datasets with uncertainty and the corresponding predictive distributions generated by BNN. For cases that are similar to training data, such as normal or CP in CC-CCII or Zhejiang dataset, the uncertainty was relatively lower. For unknown diseases that were not included in the training set, such as TB and COPD, the system was able to report a high uncertainty, which could be served as a misdiagnosis signal. For COVID-19 cases of different ages (Children dataset), the uncertainties were also acceptable. This might indicate that the system was able to identify the radiological features (red pixels) that are similar to those in the training environment despite the different age groups (Figure 4B). Considering system application, take the presented cases in Figure 4A as examples, the uncertainties of the cases were all within the corresponding RI, therefore the predictions are more likely to be correct and its collaborators could put more faith in them.

Radiographic images represent a field of probabilities. Diseases usually present in dynamic and continuous forms, and there may be no clear boundaries between disease conditions where uncertainty might reach its

Figure 3. The reliability of the AI system and corresponding reliability intervals in different datasets
(A–D) The number of correct and wrong classifications of the system under different uncertainties for COVID-19 vs. others in CC-CCII (A), Zhejiang (B), Linyi (C), and Children dataset (D).
(E–H) The number of correct and wrong classifications of the system under different uncertainties for mild vs. severe in CC-CCII (E), Zhejiang (F), Linyi (G), and Children dataset (H).
(I and J) The reliability curve of the system in different datasets for COVID-19 vs. others (I) and mild vs. severe (J).
(K) The 95% reliability intervals for different datasets
peak. Take a set of serial CT scans as examples (Figures 4C and S4). The patient was diagnosed as “COVID-19/mild” after admission and the corresponding CT scan was correctly classified as “mild” with an uncertainty of 0.1679, which was within the Linyi RI. As the disease progressed, the area of ground-glass opacities (GGO, a cardinal hallmark of COVID-19) became larger (red circles) and the reticular patterns (a common manifestation in longer diseases course) became more apparent (red frames). The uncertainty rose to 0.3025 and 0.5929 for the second and third scans. This might be due to the CT manifestations for the above cases being near or in a transitional zone from mild to severe, and when the features for severe got more pronounced in the fourth scan, the system uncertainty dropped. However, the uncertainty of the latter three cases all exceeded the Linyi RI. Under such circumstances, human experts could require additional diverse data for further evaluation.

Figure 4. The typical cases with their corresponding uncertainties in different datasets
(A) The uncertainty for cases in internal (CC-CCII) and external (Zhejiang) validations.
(B) The uncertainty for unknown diseases (TB/COPD dataset) and patients of different ages (Children dataset).
(C) The uncertainty for serial CT scans of a patient in the Linyi dataset.
(D) The uncertainty for patients with or without comorbidity in the Zhejiang dataset. The red pixels on CT slices indicated the features of high risk. The violin plots stand for probabilities of different classes. The “Un” in the blue box means the value of calculated uncertainty based on the probabilities above. The CT manifestations such as ground-glass opacities and reticular patterns were indicated by red circles and red frames, respectively.
As patients with comorbidities were collected in the Zhejiang dataset, we compared the uncertainties between the two groups (with or without comorbidities) to find out if comorbidities affected system uncertainty. The results showed no difference (Figures 4D and S5). We then compared the accuracy between the two groups. No effect was observed on the accuracy of the AI system in COVID-19 vs. others (0.8364 in patients with comorbidities vs. 0.8394 for those without comorbidities, Figure S6, Table S5). However, the presence of comorbidity reduced the system performance in mild vs. severe. The accuracy of the system for mild vs. severe was 0.8841 for individuals with comorbidities and 0.9300 for those without comorbidities (Figure S6, Table S5). Therefore, we reviewed the misclassified cases in the Zhejiang dataset for further explanation.

The exceptions that cannot be warned by uncertainty

Of all the misclassified cases, those with low uncertainties within the Zhejiang RI caught our extra attention. We found that such errors mainly happened during the recovery period (a period that began with the better clinical manifestations on medical records and ended with discharge). Comparing the serial CT scans with information of corresponding medical records in the recovery period, we found a time segment that began with better manifestations in medical records and ended with a healing of the lesions on CT, i.e., the patient’s overall condition became better while unmatched injuries in the lung could still be seen in CT slices (Figures 5A, SB, S7, and S8). It seems that the “time lag” for healing CT manifestations caused the malfunction of the CT-only-based system. For example, the CT scans obtained for a patient after admission (1) showed severe injuries which progressed to a worse condition (2). The system correctly classified the cases with low uncertainties. With treatment, the patient’s condition got better on day 12 and was clinically stratified as mild. The CT scan (3) obtained on the same day, however, still presented with serious lesions. Based on CT only, the system gave a wrong “severe” prediction with low uncertainty. The same patterns could be observed in patients with comorbidities as well; only the “time lag” seems longer for patients with comorbidities. This might be one of the reasons for the declined accuracy in the Zhejiang dataset as previously described in the system performance section. To prove the assumption, we selected the serial CT scans with disease status changes in Zhejiang and calculated the system accuracy during the “time lag” and the rest of the recovery period (recover period – time lag). Figure 5D showed that the system accuracy dropped in “time lag”.

Such errors should be avoided by multimodal data that could comprehensively represent the patient’s condition. As information was limited in the CC-CCII dataset, we only added age and gender (two available features for part of the patients with COVID-19 in the CC-CCII dataset) to the system, considering that they would partially influence the interpretation of the patients’ condition. After full training, the system achieved better performance in the Zhejiang dataset (Table S6).

There was another exceptional case that could not be warned by uncertainty. Clinically, patients with comorbidities were more likely to be labeled as severe despite minor lung damage, as they often suffered from bad overall conditions (Figure 5C), and the system unsurprisingly made wrong classifications with low uncertainty.

The human-machine convergence application and a proposed workflow

To evaluate the practical value of the uncertainty and RI, we randomly selected 50 cases from the external datasets and calculated the initial accuracy of the four junior radiologists. Given (i) the prediction result on each patient by AI system, and (ii) prediction uncertainty and corresponding RI together with (i), the same four junior radiologists were required to diagnosis those 50 cases again. To avoid a potential memorization bias, (ii) was performed 2 weeks after (i). The performance of (ii) was improved compared to the initial one (Figures 6A and 6B).

Based on the above results, we proposed a clinical workflow for the AI system application (Figure 7). During the system deployment, radiologists and clinicians could first use the historical cases to test the system. If the system gave acceptable reliability curves, they could then choose the reliability value and find the corresponding reliability intervals based on their knowledge of the hospital data. They could also review the wrong predictions with low uncertainties to find out the discrepancy in CT readings and exceptions in different datasets caused by the direct application of the trained systems. In this way, cases that are beyond the capabilities of the system might be found in time and be reevaluated to avoid reckless diagnosis. And it would be better for the system to process diverse cases, thus giving us more
opportunities to rule out the potential problems. Still, it should be noted that it is impossible to identify all the problems during the system deployment. We have discussed with the clinicians and radiologists and designed an extra discussion module. When abnormal cases pop out, the module provides an easy way for patient information exchange and timely reexamination. The process should also be given back to the system for dynamic iteration. In this case, the system could become a better tool along with its increasing use.

Figure 5. The typical cases for exceptions that cannot be warned by uncertainty
(A–C) Serial CT images and the corresponding uncertainty for patients in the Zhejiang dataset. The red pixels on CT slices indicated the features of high risk. The violin plots stand for probabilities of different classes. The “Un” in the blue box means the value of calculated uncertainty based on the probabilities above. The orange and yellow boxes represent the disease course for the patients. And the numbers on it (1, for example) represent the time points of the CT scan during the course of the disease.
(D) Histograms of the accuracy for AI system in “time lag” or “recovery period–time lag” (the rest of the recovery period).
DISCUSSION

In this study, we developed an AI system that could robustly classify COVID-19 from normal and CP, and stratify mild and severe in COVID-19 cases. Our unique approach was to report underneath uncertainty for each prediction and generate a trustworthy RI for the current dataset at the time the diagnosis is made. We proposed a workflow that first assesses the system performance by processing part of the data and generating a basal RI. The misclassified cases with low uncertainties within the RI can also be used for debugging a system and for retrospective analysis of system failures. Combining the uncertainty and RI, the human expert could decide if extra information or further evaluation is needed for the case. We believe that the system can be used as a fast triage tool and facilitate COVID-19 detection and disease evaluation.

It should be emphasized that AI systems were only the assistant of the health professionals. Therefore, understanding if a model is uncertain can help make a better tool. A variety of methods have been developed for quantifying predictive uncertainty. Among them, BNNs have the ability to combine the scalability, expressiveness, and predictive performance of neural networks with Bayesian theory and endows the ability to compute the principled predictive uncertainty. Wilson and Izmailov argued that a key advantage of BNNs lies in its marginalization step, which particularly can improve both the accuracy and calibration of modern deep neural networks (Wilson and Izmailov, 2020). We thus utilized BNN for classification and used predictive entropy for uncertainty estimation (Ghoshal et al., 2021; Ghoshal and Tucker, 2020). However, estimating the predictive uncertainty only is not sufficient for safe decision-making. Furthermore, it is crucial to assure that the uncertainty estimates are reliable. To this end, our system calculated the RI as a reference for the human experts. The radiologists reported that the uncertainties for each case and RI for the corresponding dataset were beneficial additions to evaluate the CT scans. They think such indicators reduce the opacity of the model to some extent and help earn more faith for AI predictions.

Our proposed system and workflow might be able to be expanded to other diseases. For different diseases, however, there may be different types for the exceptional cases that are misclassified with low uncertainties, which might pose interesting challenges and need help from human experts to rule out the malfunction situations in the deployment section.

In summary, we developed an AI system that could inform the users of its prediction uncertainty and provide an RI as a reference, so that the users could have the chance to decide when to take over. The proposed approach integrated the advantages of AI in rapid data processing with the experts' ability to...
handle complex cases. The challenges caused by algorithm uncertainty or data uncertainty during AI application were thus mitigated to some extent by the convergence of human and artificial intelligence. Its clinical significance might be that health professionals could actually leverage the AI system as a triage tool to improve operational efficiency in a landscape of increasing clinical volumes.

Limitations of the study
Our study has several limitations, which we hope to address in the future. First, we only employed a CT scan for system establishment to show the relevance between input and uncertainty variation. It should be noted that a system with multimodal data could provide a more comprehensive view and a closer application to the real world. Second, we only tested the system in four simulated clinical scenarios, which could not represent the real-world application. Third, we only randomly selected 50 cases from all datasets to verify the performance of the human-machine convergence application. With a small sample size, inconsistencies in one or two cases can make a difference in the final figure. The system’s actual application value should be further validated in multicenter prospective clinical trials.

Ethics approval
CT images of external datasets were collected from the First Affiliated Hospital-Zhejiang University School of Medicine, Linyi People’s Hospital, Guangzhou Women and Children’s Medical Center, and the First Affiliated Hospital of Guangzhou Medical University, and the Guangzhou Chest Hospital. Institutional Review Board (IRB)/Ethics Committee approvals were obtained in all the institutions and consents were obtained from all participants or parents of the included children.

STAR METHODS
Detailed methods are provided in the online version of this paper and include the following:

- **KEY RESOURCES TABLE**
- **RESOURCE AVAILABILITY**
  - Lead contact
  - Materials availability
METHOD DETAILS

Datasets details and preprocessing
Methods

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analysis
The sensitivity of the system for COVID-19 diagnosis relative to radiologists

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.103961.

ACKNOWLEDGMENTS

This work was supported by The National Key Research and Development Plan (2019YFB1404803), The Excellent Young Scientists Fund (82122036), and The General Program of National Natural Science Foundation of China (62076076). The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

AUTHOR CONTRIBUTIONS

D.L. and H.L. conceived and designed the study. N.X., H.Z., K.L., L.L., and J.W. collected data, B.A., H.L., and G.L. performed imaging annotation. L.H. performed the experiments. D.L. and L.H. conducted the data analysis and mapping. D.L., H.L., X.P., Y.G., and H.L. accomplished manuscript writing. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

Ayhan, M.S., Kuhlewein, L., Aliyeva, G., Inhoffen, W., Ziemsen, F., and Berens, P. (2020). Expert-validated estimation of diagnostic uncertainty for deep neural networks in diabetic retinopathy detection. Med. Image Anal. 64, 101724. https://doi.org/10.1016/j.media.2020.101724.

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppelman, P., Moore, S., Phillips, S., Maffitt, D., and Pringle, M. (2013). The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J. Digit. Imag. 26, 1045–1057.

Douglas, P.S., De Bruyne, B., Pontone, G., Patel, M.R., Norgaard, B.L., Byrne, R.A., Curzen, N., Purcell, I., Guthbert, M., and Rioufol, G. (2016). 1-year outcomes of FFRT-guided care in patients with suspected coronary disease: the PLATFORM study. J. Am. Coll. Cardiol. 68, 435–445.

Finlayson, S.G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., Kohane, I.S., and Saria, S. (2020). The clinician and dataset shift in artificial intelligence. N. Engl. J. Med. 385, 283–286.

Gal, Y. (2016). Uncertainty in Deep Learning (University of Cambridge, UK).

Gawlinski, J., Tassi, C.R.N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., and Roscher, R.J.a.p.a. (2021). A survey of uncertainty in deep neural networks. Preprint at arXiv, 2107.03342.

Ghoshal, B., Tucker, A., Sanghera, B., and Lup Wong, W. (2021). Estimating uncertainty in deep learning for reporting confidence to clinicians in medical image segmentation and diseases detection. Comput. Intel. 37, 701–734.

Ghoshal, B., and Tucker, A.J.a.p.a. (2020). Estimating uncertainty and interpretability in deep learning for coronavirus (COVID-19) detection. Preprint at arXiv, 2003.10769.

Harmon, S.A., Sanford, T.H., Xu, S., Turkbey, E.B., Roth, H., Xu, Z., Yang, D., Myronenko, A., Anderson, V., Amalou, A., et al. (2020). Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinationals datasets. Nat. Commun. 11, 4080. https://doi.org/10.1038/s41467-020-17971-2.

Jin, C., Chen, W., Cao, Y., Xu, Z., Tan, Z., Zhang, X., Deng, L., Zheng, C., Zhou, J., Shi, H., and Feng, J. (2020). Development and evaluation of an artificial intelligence system for COVID-19 diagnosis. Nat. Commun. 11, 5088. https://doi.org/10.1038/s41467-020-18685-1.

Jun, M., Cheng, G., Yuxin, W., Xingle, A., Jiantao, G., Ziqi, Y., Minqing, Z., Xin, L., Xueyuan, D., and Shucheng, C. (2020). COVID-19 CT lung and infection segmentation dataset (Zenodo).

Kendall, A. (2019). Geometry and Uncertainty in Deep Learning for Computer Vision (University of Cambridge, UK).

Ma, J., Wang, Y., An, X., Ge, C., Yu, Z., Chen, J., Zhu, Q., Dong, G., He, J., and He, Z. (2020). Towards efficient COVID-19 CT annotation: a benchmark for lung and infection segmentation. Preprint at arXiv, 2004.12537.

Qiu, S., Joshi, P.S., Miller, M.I., Xue, C., Zhou, X., Karjadi, C., Chang, G.H., Joshi, A.S., Dwyer, B., and Zhu, S.J.B. (2020). Development and Validation of an Interpretable Deep Learning Framework for Alzheimer’s Disease Classification. Brain 143, 1920–1933.

Rabanser, S., Gunnemann, S., and Lipton, Z.C. (2018). Failing loudly: an empirical study of methods for detecting dataset shift. Preprint at arXiv, 1810.11953.

Rauschecker, A.M., Rudie, J.D., Xie, L., Wang, J., Duong, M.T., Bozolakas, E.J., Kovalovich, A.M., Egan, J., Cook, T.C., Bryan, R.N., et al. (2020). Artificial intelligence system Approaching...
neuroradiologist-level differential diagnosis accuracy at brain MRI. Radiology 295, 626–637. https://doi.org/10.1148/radiol.2020190283.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional Networks for Biomedical Image Segmentation (Springer International Publishing), pp. 234–241.

Roy, A.G., Conjeti, S., Navab, N., and Wachinger, C.; Alzheimer’s Disease Neuroimaging Initiative (2019). Bayesian QuickNAT: model uncertainty in deep whole-brain segmentation for structure-wise quality control. Neuroimage 195, 11–22.

Rudie, J.D., Rauschecker, A.M., Xie, L., Wang, J., Duong, M.T., Botzolakis, E.J., Kovalovich, A., Egan, J.M., Cook, T., Bryan, R.N., et al. (2020). Subspecialty-level deep gray matter differential diagnoses with deep learning and bayesian networks on clinical brain MRI: a pilot study. Radiol. Artif. Intell. 2, e190146. https://doi.org/10.1148/ryai.2020190146.

Singh, K., Valley, T.S., Tang, S., Li, B.Y., Kamran, F., Spaulding, M.W., Wiers, J., Otles, E., Donnelly, J.P., Wei, M.Y., et al. (2020). Evaluating a widely implemented proprietary deterioration index model among hospitalized COVID-19 patients. Preprint at medRxiv. https://doi.org/10.1101/2020.04.24.20079012.

Subbaswamy, A., Adams, R., and Saria, S. (2021). Evaluating Model Robustness and Stability to Dataset Shift (PMLR), pp. 2611–2619.

Subbaswamy, A., and Saria, S.J.B. (2020). From development to deployment: dataset shift, causality, and shift-stable models in health AI. Biostatistics 21, 345–352.

Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V.X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., et al. (2019). Do no harm: a roadmap for responsible machine learning for health care. Nat. Med. 25, 1337–1340. https://doi.org/10.1038/s41591-019-0548-6.

Wilson, A.G., and Izmailov, P.J.a.p.a. (2020). Bayesian deep learning and a probabilistic perspective of generalization. Preprint at arXiv, 2002.08791.

Yang, J., Sharp, G., Veeraraghavan, H., van Elmpt, W., Dekker, A., Lustberg, T., and Gooding, M. (2017). Data from Lung CT Segmentation Challenge (The cancer imaging archive).

Yang, J., Veeraraghavan, H., Armato, S.G., III, Farahani, K., Kirby, J.S., Kalpathy-Kramer, J., van Elmpt, W., Dekker, A., Han, X., and Feng, X. (2018). Autosegmentation for thoracic radiation treatment planning: a grand challenge at AAPM 2017. Med. Phys. 45, 4568–4581.

Zhang, K., Liu, X., Shen, J., Li, Z., Sang, Y., Wu, X., Zha, Y., Liang, W., Wang, C., and Wang, K. (2020). Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography. Cell 181, 1423–1433.e11.
## STAR METHODS

### KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---------------------|--------|------------|
| **Software and algorithms** | | |
| Artificial intelligence system | This paper | https://github.com/xiangxiangzhuyi/COVID-19-diagnosis |
| U-Net | (Ronneberger et al., 2015) | http://lmb.informatik.unifreiburg.de/people/ronneber/u-net |
| **Other** | | |
| CC-CCII dataset | (Zhang et al., 2020) | http://ncov-ai.big.ac.cn/download?lang=en |
| External validation datasets | This paper | https://www.kaggle.com/ideahu/covid19-diagnosis |
| LCTSC2017 dataset | (Yang et al., 2018) | https://wiki.cancerimagingarchive.net/display/Public/Lung+CT+Segmentation+Challenge+2017 |
| COVID-19-CTLIS dataset | (Ma et al., 2020) | https://zenodo.org/record/3757476#.X8eAxs37RPb |
| Kaggle-CT dataset | K Scott Mader | https://www.kaggle.com/kmader/finding-lungs-in-ct-data |

### RESOURCE AVAILABILITY

#### Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Huiying Liang (lianghuiying@hotmail.com).

#### Materials availability

This study did not generate new unique reagents.

#### Data and code availability

- The CC-CCII dataset can be accessed via http://ncov-ai.big.ac.cn/download?lang=en. The three external validation datasets can be accessed via https://www.kaggle.com/ideahu/covid19-diagnosis. The LCTSC2017 dataset can be accessed via https://wiki.cancerimagingarchive.net/display/Public/Lung+CT+Segmentation+Challenge+2017. The COVID-19-CTLIS dataset can be accessed via https://zenodo.org/record/3757476#.X8eAxs37RPb. The Kaggle-CT dataset can be accessed via https://www.kaggle.com/kmader/finding-lungs-in-ct-data.
- All code associated with this paper can be freely accessed and downloaded via https://github.com/xiangxiangzhuyi/COVID-19-diagnosis.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

### METHOD DETAILS

#### Datasets details and preprocessing

**CC-CCII dataset.** CC-CCII is a large dataset of 395,910 CT slices from 2,742 subjects with normal control, COVID-19 pneumonia, or CP (viral pneumonia, bacterial pneumonia, and mycoplasma pneumonia included) (Zhang et al., 2020). The size of most CT slices was 512 x 512, and slices with other abnormal sizes were resized to the common size. The range of the number of slices in a scan was from 16 to 690. All subjects were diagnosed as one of three categories: normal control, CP, and COVID-19.

In the CC-CCII dataset, all slices were stored in the image file format, such as PNG, TIFF, BMP, JPEG. Comparing to the standard medical image format: DICOM, the scanning information was not available in those image file format. The acquisition number of CT slices in a scan could only be determined by the image file name, which was an uncertain factor in the 3D reconstruction of the lung area volume. Therefore, we performed manual quality control on all CT scans and removed unusable scans and slices. The following 6 kinds of quality problems mainly occurred. 1) Multiple series of CT slices were contained in a single CT scan. We kept the longest series and removed other series. 2) The name of image files was sort in the order opposite to the stacking order of CT slices. In this case, the
order was reversed in the 3D reconstruction of the lung area. 3) The order of image files was completely confusing. We tried to sort those images manually. However, some scans with dense slices were unable to be sorted again and then be removed. 4) The lung area in the scan was incomplete. This kind of scan was removed. 5) Slices from other organs or other scanning directions were mingled to some scans. These foreign slices were removed. 6) Slices with significant resolution reduction were removed.

After the exclusion for the above quality problem, 351,850 CT slices belonging to 3,780 scans of 2,430 subjects were retained. In the CC-CCII dataset, 289,133 slices belonging to 3,065 scans of 1,736 subjects were not segmented and kept the original appearance. However, the remaining 62,717 slices belonging to 715 scans of 699 subjects have been segmented. In the segmented slices, only the lung area was kept, and other areas were filled by pixels with zero value.

In the COVID-19 group, 408 patients were further diagnosed as mild or severe. After quality control mentioned above, only 302 patients including 158 mild patients (89 males, average age 45.38 ± 15.24 years), and 144 severe patients (75 males, average age 51.86 ± 21.44 years) were qualified. However, the information on other subjects in the CC-CCII dataset was not available.

An additional 750 CT slices with corresponding segmentation masks were also provided. The 750 slices belonged to 150 scans and each scan contained 5 slices. According to the pixel value in the corresponding mask, each pixel in the slices could be classified as one of the backgrounds, the lung area, the ground-glass opacity, and the consolidation.

In the CC-CCII dataset, 20% of the scans were selected as holdout dataset for internal validation. The remaining 80% of the scans were equally divided into 5 subsets (Table S7) in order to cross-validate the AI diagnostic system. The division of the whole dataset was random, but it also met two premises: 1) the scans of the same subject can’t be divided into different subset; and 2) the proportions of different type (Normal control/CP/COVID-19, mild/severe) of scans should be similar across all subsets.

**External validation datasets.** A total of 184 subjects were enrolled in the Zhejiang, Linyi and Guangzhou children datasets. Several subjects performed multiple CT scans in their developing health statuses which could be developed from COVID-19 to normal control or severe COVID-19 to mild COVID-19. Therefore, the same subject could have various scans labeled as different categories on different health statuses. In these external validation datasets, the size of all slices was 512 × 512, and the range of the number of slices in a scan was from 71 to 619. All slices were stored in DICOM format.

The Zhejiang dataset was collected at the First Affiliated Hospital, Zhejiang University School of Medicine from February 14, 2020 to March 8, 2020. The Zhejiang dataset contained 140 subjects, 93 of whom had detailed clinical information. In addition to age and gender, the subject’s comorbidities information was also included in the clinical information. Eight comorbidities were recorded: hypertension (37 subjects), diabetes (11 subjects), coronary heart disease (6 subjects), hepatopathy (18 subjects), COPD (3 subjects), nephropathy (2 subjects), cancer (2 subjects), and other comorbidities (6 subjects). However, the clinical information of the remaining 47 subjects was missing. All subjects were diagnosed as normal control, CP, or COVID-19. All COVID-19 patients were further diagnosed as mild COVID-19 and severe COVID-19.

The Linyi dataset was collected at the Linyi People’s Hospital from January 22, 2020 to March 1, 2020. There were totally 33 subjects included in this dataset: 10 normal controls and 23 COVID-19 patients. Multiple CT scans were obtained for 23 subjects during the period of disease development. Finally, 12 CT scans labeled as normal and 61 scans labeled as COVID-19 were collected. The 61 CT scans of the COVID-19 patients were further classified as 25 mild and 36 severe ones, according to their pathological conditions of scans and their medical records. The age and gender information of the 33 subjects were available.

The Guangzhou children dataset was collected at the Guangzhou Women and Children’s Medical Center from February 5, 2020 to February 28, 2020. There was a total of 11 subjects in the Guangzhou dataset. The age range of all subjects was from 2 months to 15 years. All subjects were diagnosed as normal control, or COVID-19, and all COVID-19 patients were mild.
Disease status stratification for external datasets was performed independently by two experienced radiologists. Any discrepancies were resolved by a third more experienced radiologist. The classification was based on the Novel Coronavirus Pneumonia Diagnosis and Treatment protocol developed by the National Health Commission of the People’s Republic of China.

The TB and COPD dataset was collected from the First Affiliated Hospital of Guangzhou Medical University and the Guangzhou Chest Hospital respectively, to evaluate the performance of the AI diagnostic system when dealing with CT scans with categories never learned. There were 10 patients in the COPD dataset, including 3 moderate patients, 3 severe patients, and 4 critical patients. The TB dataset contained 16 patients, including 2 endobronchial tuberculosis patients, 1 tuberculous pleuritis patient, 11 secondary tuberculosis patients, and 2 disseminated tuberculosis patients. All patients in the two datasets only performed 1 CT test. All slices had the same size of 512×512 and were all stored in DICOM format.

Public segmentation datasets. Three public segmentation datasets were employed in our study to train the segmentation network: LCTSC2017 dataset (Clark et al., 2013; Yang et al., 2017, 2018), COVID-19-CTLIS Dataset (Jun et al., 2020; Ma et al., 2020), and Kaggle-CT dataset. The LCTSC2017 dataset was provided in association with a challenge competition and related conference session conducted at the AAPM 2017 Annual Meeting. There were 9,533 slices belonging to 60 CT scans in The LCTSC2017 dataset. The COVID-19-CTLIS dataset contained 20 COVID-19 CT scans with labeling. Left lung, right lung, and infections were labeled by two radiologists and verified by another experienced radiologist. In our study, only 2,581 slices belonging to 10 CT scans were used. The Kaggle-CT dataset was created by K Scott Mader. There was no complete lung CT scan in the dataset, only 265 CT slices and corresponding lung masks were included in the dataset. All slices in the three datasets had the same size of 512×512 and were all stored in DICOM format.

Methods

Dicom slices preprocessing. The slices in the CC-CCII dataset were converted and shown in the lung window used to view lung parenchyma. The converted slice was stored in the image file format, and its pixel value range was from 0 to 255. The slices in the other dataset were not converted to the lung window view and were stored in the DICOM format with a pixel value range of about -1,000 to 1,000. In order to make our AI diagnostic system trained on the CC-CCII dataset effective on other datasets, the slices in other datasets should also be converted to the lung window view. The window level and width were unknown in the conversion of the CC-CCII dataset. Therefore, the commonly used lung window level (-500 HU) and width (1,500 HU) were applied to our conversion (Figure S9). The conversion could be described by the following three equations:

\[
\begin{align*}
    p^D_{xy} &= -1250, \text{ if } p^D_{xy} < -1250 \\
    p^D_{xy} &= 250, \text{ if } p^D_{xy} > 250 \\
    p^C_{xy} &= 255 \times \frac{p^D_{xy} - \min(p^D)}{\max(p^D) - \min(p^D)}
\end{align*}
\]  

(Equation 1)

\(p^D_{xy}\) and \(p^C_{xy}\) indicate the pixel value of the DICOM slice and the converted slice respectively in the xth row and the yth column. \(p^D\) indicates the set of all \(p^D_{xy}\) in the DICOM CT scan.

Lung area segmentation. Zhang, Kang, et al. reported the performance of multiple segmentation networks on the CC-CCII dataset (Zhang et al., 2020). Among them, U-Net achieved the best average dice coefficient of 0.959 ± 0.003 in the segmentation of the lung area. Zhang, Kang, et al. also reported the best segmentation network could only get the average dice coefficient of 0.587 ± 0.012 in the segmentation of multiple lesions (Zhang et al., 2020). Based on this report, we believed that the existing methods could not segment the lesion well. Therefore, we only performed the segmentation of the lung area by U-Net in our study. The additional 750 annotated CT slices in the CC-CCII dataset were used to train a U-Net model to segment the CC-CCII dataset. The difference in appearance between the CC-CCII dataset and other datasets has been reduced but not completely eliminated by the conversion method. The three public segmentation datasets LCTSC2017, COVID-19-CTLIS, and Kaggle-CT were also stored in the DICOM format and had a similar pixel value range. Therefore, three public segmentation datasets, instead of the
additional 750 annotated CT slices in the CC-CCII dataset, were used to segment the lung area of external validation datasets, the COPD dataset, and the TB dataset. After the segmentation, the segmented mask was used to cover the converted slice to get the lung area image.

The 750 annotated CT slices with corresponding segmentation masks in the CC-CCII dataset were divided into 700 slices and 50 slices, which were used to train and validate the segmentation network respectively. In the segmentation mask, except for the background pixels, all other kinds of pixels were reset as the lung area. Several data augmentation technologies: cropping, padding, Gaussian blur, and affine transformation were performed on the 700 slices to augment the sample size from 700 to 7,000. The input size of the U-Net was 512 × 512. The learning rate was 0.0001 and the batch size was 4. After fully trained, the U-Net could get the average dice coefficient of 0.9930 on the validation dataset. Then all slices in the CC-CCII dataset were segmented by the trained U-Net. The three public segmentation datasets were mixed together as a dataset with 12,379 slices. The mixed dataset was divided into the training dataset (95%) and the validation dataset (5%). The configuration of U-Net was the same as that of the above. The training phase was stopped when the average dice coefficient achieved 0.9803. Then external validation datasets, the COPD dataset, and the TB dataset were segmented by the trained U-Net. All segmented slices were further processed by the closing operation to fill holes, by the opening operation to eliminate isolated points, and by the labeled component analysis to remove isolated components far away from the slice center (Figure S10).

We stacked all slices in the third dimension and calculated the bounding box of the segmented lung area in the scan. The bounding box was slightly larger than the lung area. Then the lung area in all scans was cropped and resized to the size of 254 × 373×60 which was the average size of bounding boxes in all scans.

**3D fully convolutional network development.** The 3D FCN network was adopted to identify important features for COVID-19 diagnosis and assess the disease status. The detailed configuration of our 3D FCN framework was shown in Table S8. 7 convolutional blocks and 3 max-pooling layers were included in the framework. Each convolutional block was comprised of a 3 × 3×3 convolution, a batch normalization layer, and a Relu nonlinearity. The number of filters in the convolutional block was doubled after two convolutional blocks. The number of filters in the first convolutional block was set as the hyperparameter to control the complexity of the 3D FCN framework. The number of the channel in the output volume of the 3D FCN framework depended on the number of classified categories.

The 3D FCN framework was trained by 3D patches randomly sampled from the lung area 3D volume. The size of the 3D patch was 47 × 47×47. The label of lung area 3D volume was assigned to each 3D patch sampled from it. After passing the 3D FCN framework, an output volume with the size of 1 × 1×1×C (C is the number of categories) was produced, which was also the predicted label of the patch. The learning rate was 0.0005 and the batch size was 50. The training phase was stopped when the accuracy of the validation dataset was not increasing anymore. After the 3D FCN framework fully trained, the whole lung area 3D volume was inputted to the 3D FCN to get a feature map with the size of 27 × 42×3×C.

We set four options for the hyperparameter: 4, 8, 16, 32. Four 3D FCNs were trained under the four options. Through comparing the performance of the four 3D FCNs, we found the optimized hyperparameter. The hyperparameter optimization was performed in the normal/CP/COVID-19 classification and the mild/severe classification separately. Based on the experimental result (Table S9), the optimized numbers of filters in the first convolutional block were 32, and 4 in normal/CP/COVID-19 classification and the mild/severe classification separately.

**Feature selection.** After feature maps of all CT scans were generated, the feature selection was performed on the feature map. This selection was based on observation of the overall performance of the 3D FCN as estimated using the accuracy on the CC-CCII training dataset. Specifically, we selected features from multiple fixed locations that were indicated to have high accuracy values (Qiu et al., 2020). In this step, two hyperparameters were introduced: the number of fixed locations, and the feature type in each fixed location. The options of the first hyperparameter were: 20, 50, 100, 200, 300, 400, 500, 600, 700, 800. The second hyperparameter had two options: the probability distribution, and the integer label with the maximum probability (normal → 0, CP → 1, COVID-19 → 2, mild → 0, severe → 1). A BNN framework was used as the classifier to make the final prediction and calculate the uncertainty of each
study, variational estimation was used to approximate the input features and the corresponding label of a sample in the dataset, respectively, where $y \in \{1, 2, \ldots, C\}$. The neural network is used to model the predicted distribution $p(Y|X)$, which can be defined by the following equation:

$$P(Y|X) = \int p(Y|X, W)p(W|D)dW$$

(Equation 2)

where given weight $W$ and input $X$, $p(Y|X, W)$ represents the distribution of $Y$. We only need to fit the weight distribution $p(W|D)$ based on the dataset $D$. According to the Monte Carlo method, $m$ samples that follow the distribution of $p(W|D)$ are sampled, and calculate $\frac{1}{m} \sum_{i=1}^{m} p(Y|X, W_i)$ to get $p(Y|X)$. In our study, variational estimation was used to approximate $p(W|D)$ by $q(W|\theta)$, where $\theta$ is the distribution of $W$. Therefore, our ultimate optimization goal is minimizing the following formula:

$$E_{q(W|\theta)}[\log q(W|\theta)] - E_{q(W|\theta)}[\log q(Y|X, W)] - E_{q(W|\theta)}[\log q(W)]$$

(Equation 3)

Given a training sample $(x_i, y_i)$, the Monte Carlo method was used for (Equation 3) to get the following formula:

$$\sum_{j=1}^{m} \log q(W_j|\theta) - \sum_{j=1}^{m} \log q(y_i|W_j, x_i) - \sum_{j=1}^{m} \log q(W_j)$$

(Equation 4)

Suppose $W$ obeys Gaussian distribution and is independent. In the case of $m = 1$, we can get the loss function of the sample $(x_i, y_i)$:

$$L = \sum_{k=1}^{K} [\log q(w_k|\theta) - \log p(w_k)] - \log p(y_i|W, x_i)$$

(Equation 5)

In the BNN, $q(w_k|\theta)$, $p(w_k)$, and $p(y_i|W, x_i)$ are supposed to obey Gaussian distribution.

**Ultimate prediction.** The expectation over multiple BNN iterations is used as the ultimate prediction:

$$P(y|x) = \frac{1}{I} \sum_{i=1}^{I} p^i(y|x)$$

(Equation 6)

where $p^i$ is the predicted probability of the BNN in $i$th iteration and can be solved by Equation 2, $I$ is the number of iterations and was set as 100 in our study. Then the ultimate predicted label $\hat{y}$ can be converted by the following equation:

$$\hat{y} = \arg\max_C (P(y|x))$$

(Equation 7)

**Uncertainty calculation.** The uncertainty is also calculated based on the predictive distribution (Ghoshal and Tucker, 2020). First, the predicted label of each iteration should be calculated:

$$\hat{y}^i = \arg\max_C (p^i(y|x))$$

(Equation 8)

where $\hat{y}^i$ is the predicted label in the $i$th iteration. Then the uncertainty $u(x)$ can be calculated by the entropy of the distribution of the predicted label in $I$ iterations.

$$u(x) = -\sum_{C} p(\hat{y}^i = c|x) \log p(\hat{y}^i = c|x)$$

(Equation 9)

Where $p(\hat{y}^i = c|x)$ represents the proportion of iterations where the predicted label is $c$. 
Reliability computation. The reliability $R(u)$ of the system outcome can be defined as the prediction accuracy of the sample under the corresponding uncertainty, which can be expressed by the following equation:

$$R(u) = \frac{\sum_{n=1}^{N} \delta_{\varepsilon} \left( \left( u(x_n), u \right) \right) \hat{y}_n = y_n}{\sum_{n=1}^{N} \delta_{\varepsilon} \left( u(x_n), u \right)} \quad (0 \leq u \leq 1) \quad (Equation\ 10)$$

where $x_n$, $y_n$, and $\hat{y}_n$ represent the input features, the truth label, and the ultimate predicted label of the $n$th samples in $D$. $u(x_n)$ represents the uncertainty value of $x_n$. $\delta(\cdot)$ is the Dirac delta function, and used as the counter in our study, which is:

$$\delta_{\varepsilon}(u(x_n), u) = \begin{cases} 
1, & u(x_n) \in \left[ u - \frac{\varepsilon}{2}, u + \frac{\varepsilon}{2} \right] \\
0, & \text{others} 
\end{cases} \quad (Equation\ 11)$$

where $\varepsilon$ was set as 0.1 in our study. For the high accuracy we believe the system was trustworthy, we sought the corresponding lowest and highest uncertainty and thus formed an interval, i.e., when the system gave an outcome with uncertainty within the interval, we ranked the result as highly trusted.

General 3D CNN. In our study, a general 3D CNN framework was also developed to classify normal/CP/COVID-19 and mild/severe. The detailed configuration was shown in Table S10. The learning rate was 0.0005 and the batch size was 50. The training phase was stopped when global loss did not decrease any more. The general 3D CNN achieved 91.29% and 90.26% accuracy on the training dataset (5-fold) and validation dataset respectively for classifying normal/CP/COVID-19. For classifying mild/severe, the general 3D CNN got 85.08% and 85.71% accuracy on the training dataset (5-fold) and validation dataset respectively.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analysis

Segments of the lung area were evaluated using the average dice coefficient in the internal test set. The COVID-19 diagnostic performance of the AI was evaluated using the receiver operating characteristic (ROC) curves and area under the ROC curve (AUC). The optimal threshold for AI was determined using Youden’s index. The accuracy of the AI system at a 95% confidence interval was calculated using non-parametric bootstrapping with 1,000 iterations.

The sensitivity of the system for COVID-19 diagnosis relative to radiologists

The sensitivity and specificity of the AI system relative to clinical examination were assessed using the three independent datasets (Children, Linyi, and Zhejiang). Clinical diagnosis was performed by four radiologists with an average of 6 years professional experience (range 4–8 years) and four senior radiologists with an average practice experience of 12 years (range 10–15 years).