nature machine intelligence

ARTICLES

https://doi.org/10.1038/s42256-022-00483-7

Check for updates

OPEN

An interpretable deep learning workflow for discovering subvisual abnormalities in CT scans of COVID-19 inpatients and survivors

Longxi Zhou[1,2,9], Xianglin Meng[3,9], Yuxin Huang[4,9], Kai Kang[3,9], Juexiao Zhou[1,2], Yuetan Chu[1,2], Haoyang Li[1,2], Dexuan Xie[5], Jiannan Zhang[3], Weizhen Yang[6], Na Bai[7], Yi Zhao[4], Mingyan Zhao[3], Guohua Wang[7] ✉, Lawrence Carin[2] ✉, Xigang Xiao[5] ✉, Kaijiang Yu[3] ✉, Zhaowen Qiu[4,7] ✉ and Xin Gao[1,2,8] ✉

Tremendous efforts have been made to improve diagnosis and treatment of COVID-19, but knowledge on long-term complications is limited. In particular, a large portion of survivors has respiratory complications, but currently, experienced radiologists and state-of-the-art artificial intelligence systems are not able to detect many abnormalities from follow-up computerized tomography (CT) scans of COVID-19 survivors. Here we propose Deep-LungParenchyma-Enhancing (DLPE), a computer-aided detection (CAD) method for detecting and quantifying pulmonary parenchyma lesions on chest CT. Through proposing a number of deep-learning-based segmentation models and assembling them in an interpretable manner, DLPE removes irrelevant tissues from the perspective of pulmonary parenchyma, and calculates the scan-level optimal window, which considerably enhances parenchyma lesions relative to the lung window. Aided by DLPE, radiologists discovered novel and interpretable lesions from COVID-19 inpatients and survivors, which were previously invisible under the lung window. Based on DLPE, we removed the scan-level bias of CT scans, and then extracted precise radiomics from such novel lesions. We further demonstrated that these radiomics have strong predictive power for key COVID-19 clinical metrics on an inpatient cohort of 1,193 CT scans and for sequelae on a survivor cohort of 219 CT scans. Our work sheds light on the development of interpretable medical artificial intelligence and showcases how artificial intelligence can discover medical findings that are beyond sight.

COVID-19 often causes pulmonary parenchyma lesions months after discharge, such as ground glass opacities, consolidations and long-term fibrosis[1,2]. Past studies quantified lesions in CT scans of COVID-19 inpatients and found that computerized tomography (CT) lesions are predictive indicators for COVID-19 inpatients' symptoms and short-term prognosis[3,4]. However, among COVID-19 survivors discharged from hospitals, both a recent study[1] and our survivor cohort show inconsistencies between survivor respiratory sequelae and their follow-up CT scans. First, survivors who had severe symptoms patients in general have much worse six-months follow-up lung function than the mild-symptom patients, whereas their six-month follow-up CT scans are very similar from almost all aspects[1]. Second, a large portion of COVID-19 survivors have respiratory sequelae six months after discharge. However, experienced radiologists and state-of-the-art (SOTA) artificial intelligence (AI) systems fail to detect any CT lesion on around half of the survivors, and can only detect negligible lesions (average volume $< 5\,cm^3$) on the remaining patients[1]. Such an inconsistency raises a key question towards understanding the prognosis and rehabilitation of COVID-19 patients, which is one the most critical questions at the post-pandemic era: are these respiratory sequelae caused by pulmonary lesions that are visually indiscernible on chest CT under the lung window, or are they caused by other reasons such as neurological impairments[5] and muscle weakness[1], whereas the patients' lungs are mostly recovered?

Artificial intelligence has shown the potential to solve the aforementioned question, as it has capabilities in mining subvisual image features[6–8]. To this end, existing methods train classifiers to distinguish the labelled classes (for example, CT scans from fully recovered survivors versus from survivors with sequelae), and then extract image features that contribute to the classification performance, such as indiscernible low-level textures, image intensity distributions, grey-level co-occurrence matrix, or local image patterns that correspond to filters in convolutional neural networks (CNNs). However, such subvisual features extracted by existing approaches have poor medical interpretability and are prone to false discoveries due to data bias[9]. These limitations consequently lead to difficulties in gaining pathological insights, understanding mechanisms, developing better treatment and driving scientific discoveries, which, unfortunately, are some of the most common criticisms of AI-based computer-aided detection methods.

[1]Computational Bioscience Research Center, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia. [2]Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia. [3]Department of Critical Care Medicine, The First Affiliated Hospital of Harbin Medical University, Harbin, China. [4]Heilongjiang Tuomeng Technology Company, Harbin, China. [5]Department of Computer Tomography, The First Affiliated Hospital of Harbin Medical University, Harbin, China. [6]Department of Radiology, The Affiliated Hongqi Hospital of Mudanjiang Medical University, Mudanjiang, China. [7]Institute of Information and Computer Engineering, NorthEast Forestry University, Harbin, China. [8]BioMap, Beijing, China. [9]These authors contributed equally: Longxi Zhou, Xianglin Meng, Yuxin Huang, Kai Kang. ✉e-mail: ghwang@nefu.edu.cn; larry.carin@kaust.edu.sa; xxgct_417@126.com; drkaijiang@163.com; qiuzw@nefu.edu.cn; xin.gao@kaust.edu.sa

To extract interpretable and predictive subvisual features from CT scans of COVID-19 inpatients and survivors, we propose the Deep-LungParenchyma-Enhancing (DLPE) method, which follows a different logic: instead of forcing AI models to extract features that have the best discriminative power on a given dataset, DLPE tries to help radiologists see the unseen by enhancing the previously visually indiscernible features to a discernible level. Radiologists can thus analyse the morphologies and the origins of previously invisible lesions, and can then provide good annotations for such lesions, which become the ground truth for further training automatic segmentation and quantification models. To this end, we first develop novel, accurate segmentation models for DLPE to exclude irrelevant tissues (such as airways and blood vessels) from the lung CT. We then calculate the scan-specific optimal window for observing pulmonary parenchyma, which removes patient–patient variation and system-specific bias, and substantially enhances parenchyma abnormalities compared with the lung window. With the enhanced images, radiologists can examine the detailed morphology for subvisual lesions and provide annotations. Deep-LungParenchyma-Enhancing then customizes our previously proposed SOTA deep learning model[10] to quantify interpretable radiomics for subvisual lesions, such as the lesion volume and the lesion severity. Finally, we study the predictive power of these DLPE-detected features on quantifying clinical metrics and sequelae of COVID-19 patients, based on which we further infer the pathological insights of these novel lesions.

Figure 1a shows the DLPE workflow, which consists of three steps: first, automatic segmentations of lungs, airways and blood vessels from CT scans. The segmentation models are trained over the dataset containing 3,644 CT scans collected from patients from five different hospitals. The backbone of our segmentation model is customized over our previously proposed SOTA 2.5D-based segmentation model[10], which combines the three-dimensional information of multiview two-dimensional models and thus achieves an effective tradeoff between the segmentation accuracy and model complexity. Based on the characteristics of airways and blood vessels, we further develop the feature-enhanced loss and the two stage-segmentation protocol, which achieve fast, robust and human-level segmentation, and make the segmentation of airways and blood vessels at different branching levels possible (see Methods). Second, removal of tissues other than pulmonary parenchyma and parenchyma enhancement. Tissues such as bronchiole, mediastinum and lymph glands are negligible in volume, thus inside lungs we only need to remove airways and blood vessels. Parenchyma enhancement needs an accurate estimation of the baseline CT value as well as the deviation of CT values for healthy parenchyma, as parenchyma voxels with outlier CT values imply abnormality. To this end, we first remove the known lesions using our previously proposed COVID-19 lesion segmentation model[10], and then calculate the scan-level baseline CT and the deviation for healthy parenchyma (here, healthy means that the parenchyma has no known lesion). With these scan-specific statistics, we can then considerably enhance the parenchyma lesions compared with the lung window (see Methods), which thus makes the previous visually indiscernible lesions visible. Third, discovery and quantification of novel subvisual lesions. During the discovery of the subvisual lesions, radiologists in this study compare the parenchyma-enhanced images from COVID-19 survivors with the normal CT scans of the healthy people (as control), and mark the regions that look different from healthy people. With these ground truths, we develop a segmentation model to gain pixel-level segmentations for the subvisual lesions, which is trained and tested over the dataset containing 1,412 COVID-19 chest CT scans (1,193 inpatient scans and 219 survivor scans) from five hospitals. Based on the segmentation, DLPE quantifies several interpretable radiomics by incorporating the knowledge of radiologists, which are then evaluated in terms of their predictive power for key COVID-19 clinical metrics and sequelae. More details for the three steps are described in the Methods.

Results

Segmentation of airways and pulmonary blood vessels. The power of DLPE not only relies on the scheme-level novelty, but also benefits from two technical issues solved within DLPE, that is, the segmentation for airways and for pulmonary blood vessels. Very recent studies on these two issues tried to detect more detailed structures and aimed at more robust segmentations[11,12]. However, the existing methods do not work well on our COVID-19 inpatient cohort, which contains many severe and critical cases. To accurately and robustly segment airways and blood vessels for COVID-19 inpatients, we proposed feature-enhanced loss and a two-stage segmentation protocol. The former can efficiently extract features from tissues with self-similarity, whereas the latter is designed to solve this large-scene-small-object problem. These technical novelties achieved SOTA performance for the segmentation of airways and pulmonary blood vessels for COVID-19 patients. Figure 2a shows a representative CT scan from a critically ill COVID-19 patient—for which the segmentation task is very challenging due to the strong lesion signals—and the segmentation results of DLPE and SOTA methods. Interestingly, although the segmentation model in DLPE was not specifically designed and trained for the two tasks separately, it considerably outperformed both recent SOTA methods on airway (average dice score of 0.75 versus 0.32) and blood vessel segmentation (average dice score of 0.88 versus 0.39) for critically ill COVID-19 inpatients, which demonstrates its robustness and generalization power.

When segmenting airways and blood vessels for CT scans with clear parenchyma, DLPE also achieved a SOTA dice score, especially for tiny structures. Figure 2b (left) shows the average dice score when segmenting CT scans from healthy people. Deep-LungParenchyma-Enhancing detected substantially more tiny structures for airways and blood vessels than recent SOTA methods. Figure 2b (right) shows representative segmentation results of DLPE for blood vessels and airways.

Quantification of subvisual lesions. We used DLPE to analyse our COVID-19 survivor follow-up dataset (including 69 survivors three or six months after discharge) and found substantial subvisual lesions: without DLPE, radiologists only found 3.5 cm^3 of lesions on average for each survivor, whereas after being enhanced by DLPE, they found 109 cm^3 of abnormalities on average. Figure 3a shows one example CT section of a survivor with severe respiratory sequelae (most metrics for lung functions are substantially lower than the reference value). However, the follow-up CT scan has nearly no visible lesion under the original lung window (Fig. 3a top panels). After being processed and enhanced by DLPE, there are easily visible lesions shown in the same CT section (Fig. 3a bottom panel).

We believe that the follow-up subvisual lesions reflect mild pulmonary fibrosis. These subvisual lesions have strong correlations with sequelae related to fibrosis: more subvisual lesions means lower lung capacity, less alveolar-capillary gas conductance and a worse St George's Respiratory Questionnaire (SGRQ) score (Extended Data Fig. 1), which are all typical consequences of pulmonary fibrosis[13]. Furthermore, pulmonary fibrosis provides good explanations for the morphologies and formations of the follow-up subvisual lesions. Similar to recent studies[14,15], we observed pulmonary fibrosis under the lung window in our cohort. However, these fibroses (visible under the original lung window) are actually enclosed by much more subvisual lesions (invisible without DLPE). Considering that fibrosis is caused by the accumulation of fibroblasts and collagen[16], it is likely that only the most severe accumulation can be seen in the lung window, while DLPE can unveil mild accumulation and provide much more information.

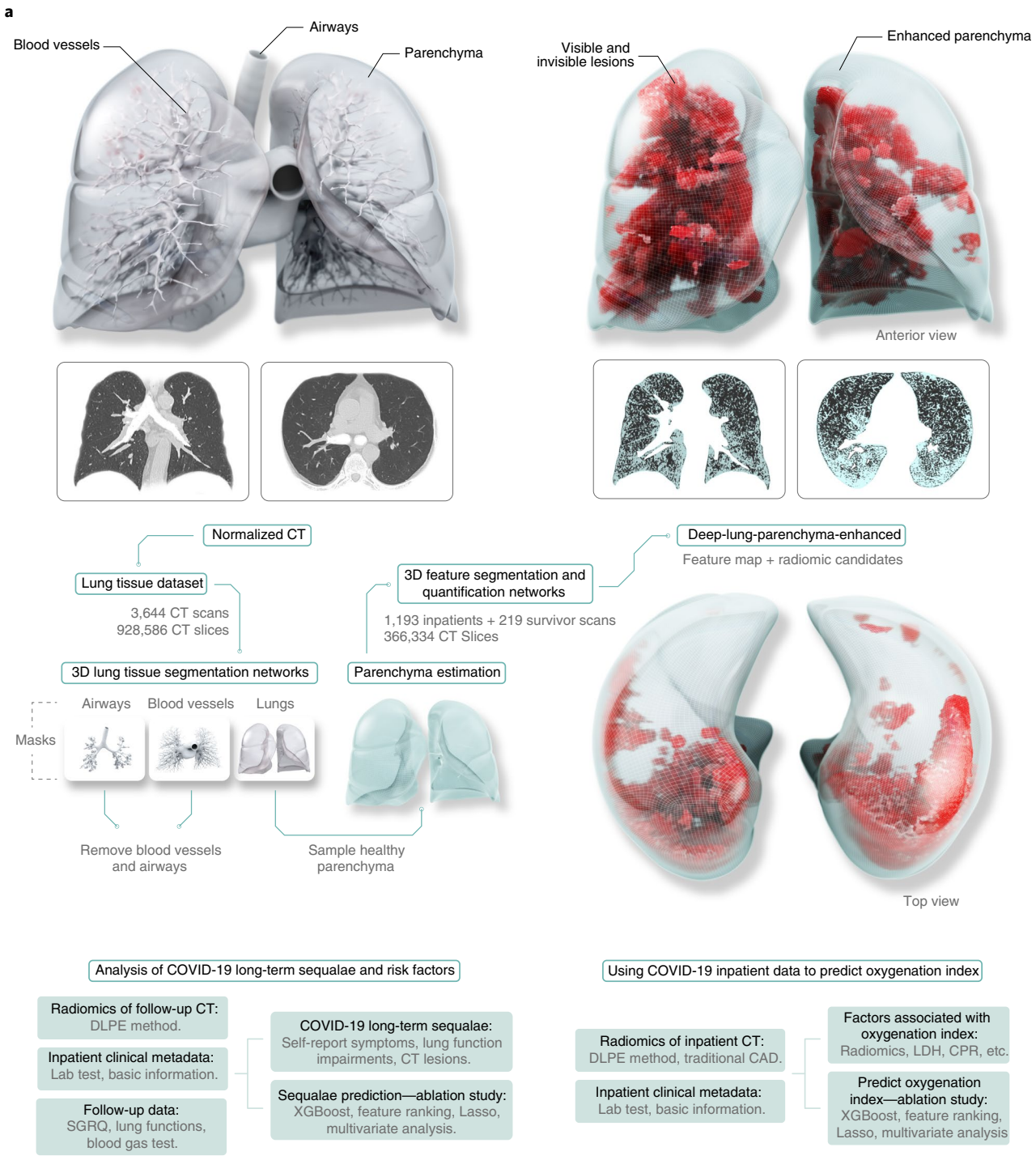
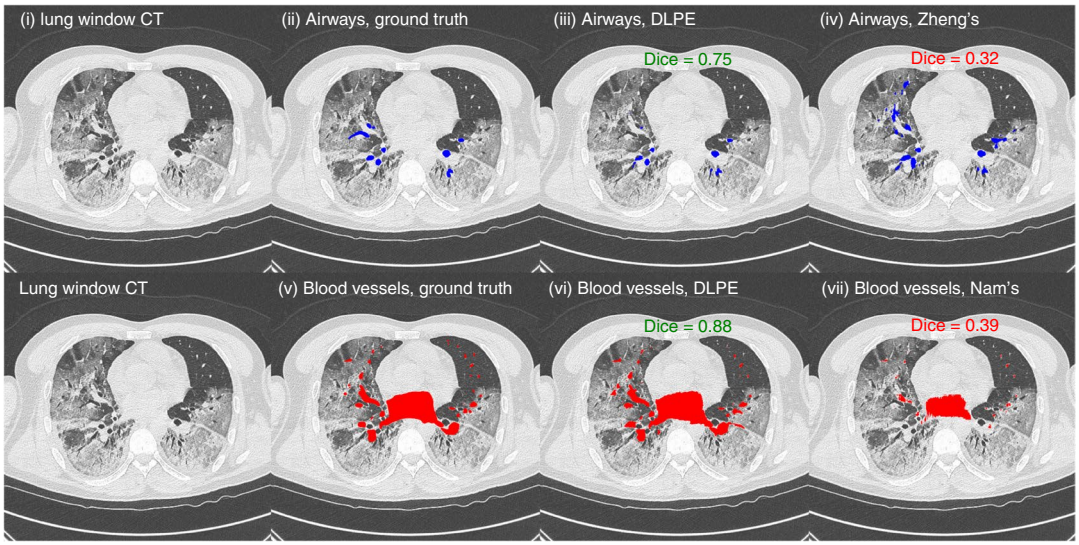


Fig. 1 | Workflow of the DLPE method and experimental design of the entire study. a, The workflow of the DLPE method. Step 1, segmentations of airways, blood-vessel and lung masks. Step 2, removal of irrelevant tissues (that is, airway, blood vessels and known lesion regions) and the sampling of the parenchyma to calculate the baseline CT value and the standard deviation of healthy parenchyma, which are then used to enhance the parenchyma dozens of times compared with the lung window. Step 3, radiologists provide annotations for abnormal regions from the enhanced CT, which is used as the ground truth for DLPE to train automatic pixel-level segmentation and quantification models for the subvisual lesions. Credit: Ivan Gromicho, King Abdullah University of Science and Technology (KAUST). **b**, We applied DLPE to two datasets: the COVID-19 inpatient dataset and the COVID-19 long-term follow-up dataset. Radiologists found novel subvisual lesions on both datasets, and gave pathological explanations for the lesion origins: follow-up subvisual lesions reflect mild pulmonary fibrosis, while inpatient subvisual lesions reflect mild plasma fluid leakages. These explanations guide and authenticate our findings. CAD: computer-aided detection.

In our cohort, the most prevalent sequela is the abnormal SGRQ score (see Supplementary Table 12 for the prevalence of the sequelae). St George's Respiratory Questionnaire score is the most frequently used and the most comprehensive[17] quantity of life assessment for respiratory sequelae; it has 50 items with 76 weighted responses and its score ranges from 0 to 100. A higher score

a



b

Segmentation performances for airways and blood vessels

Methods	Airway	BV	Airway*	BV*
DLPE	0.870	**0.911**	**0.782**	**0.847**
MPUnet (23)	0.801	0.844	0.339	0.668
3D U-net	0.784	0.866	0.135	0.554
Zheng's (11)	**0.880**	None	0.693	None
Nam's (12)	None	0.895	None	0.616

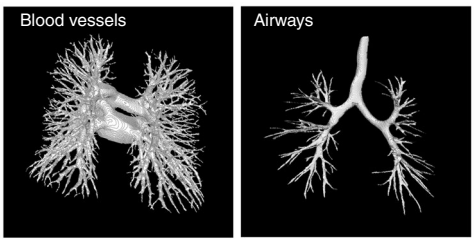
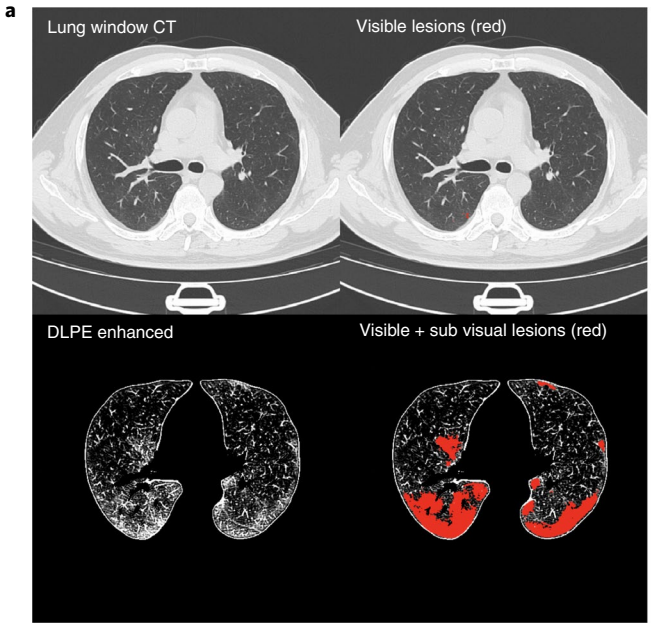


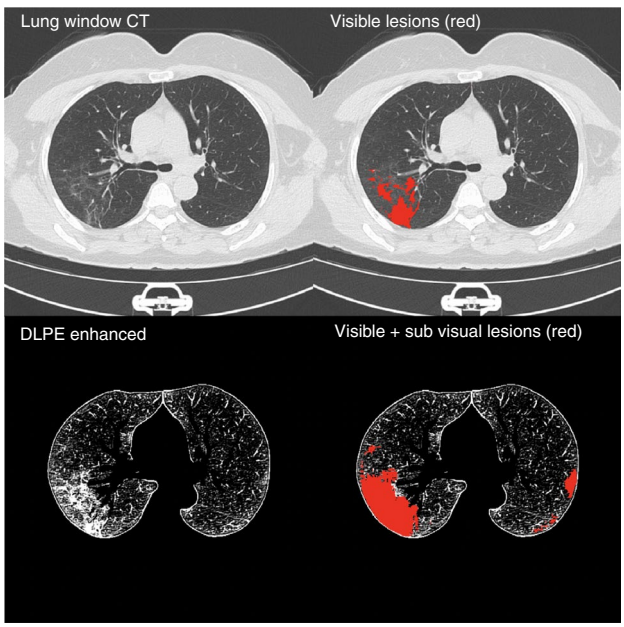
Fig. 2 | DLPE achieved SOTA segmentation for airways and pulmonary blood vessels for severely/critically ill COVID-19 inpatients and healthy people. **a**, Performance comparisons for the airway and pulmonary blood vessel segmentations on severely/critically ill COVID-19 inpatient scans. The test set contains 50 CT scans from severely/critically ill COVID-19 inpatients. The first row shows airway segmentation, whereas the second row shows blood vessel segmentation. **ai**, The chest CT scan in the lung window. **aii**, Ground-truth annotation for airways (in blue). **aiii**, Our segmentation model achieved an average dice score of 0.75 on the test set. **aiv**, Zheng and colleagues' model[11] achieved an average dice score of 0.32. **av**, Ground-truth annotation for blood vessels (in red). **avi**, Our segmentation model achieved an average dice score of 0.88 on the test set. **avii**, Nam, J. G. and colleagues' model[12] achieved an average dice score of 0.39. **b**, Segmentation performance for airways and pulmonary blood vessels on CT scans from healthy people. Left: performance is shown in a scan-level average dice score on the test set of 189 CT scans from healthy people, where BV refers to pulmonary blood vessels and asterisks indicate segmentation for tiny structures (branching level > 5, see Methods for a detailed definition). The best performer is in bold. Right: respresentative segmentation results of DLPE on the test set.

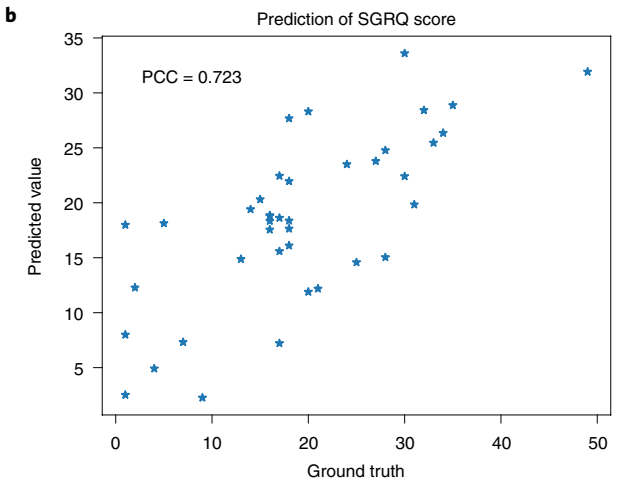
corresponds with a lower quality of life and the score should be less than 1 for healthy people. On the follow-up cohort, 46 survivors completed the SGRQ questionnaire, and among them 43 survivors self-reported respiratory sequelae that impacted their life quality, with an average SGRQ score of 18.6.

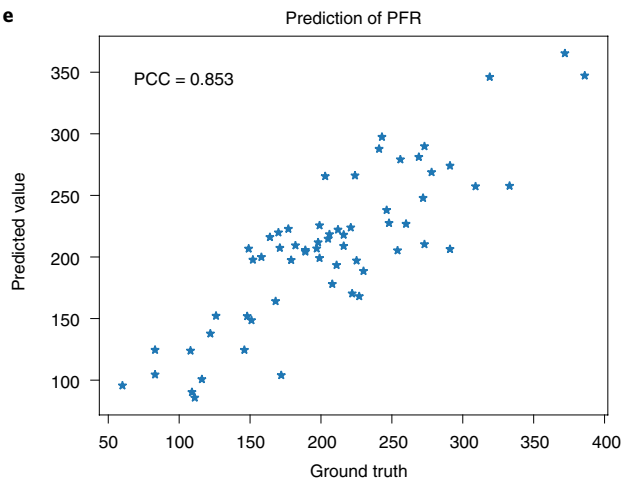
Radiomics quantified by DLPE predict the SGRQ score with high accuracy, and the subvisual lesions provide nearly all the dominant features in the prediction. Deep-LungParenchyma-Enhancing quantified six interpretable radiomics (see Methods) and we used XGBoost to predict the SGRQ score based on these features. As shown in Fig. 3b, the Pearson correlation coefficient (PCC) between the predicted and the ground-truth SGRQ score is 0.723 ($P < 0.0001$), and DLPE radiomics explain 52.3% of the variance of the SGRQ score. To our knowledge, only few methods reported their performance for SGRQ prediction, but the SOTA model for predicting chronic obstructive pulmonary disease assessment test (a good surrogate for SGRQ[17]) only explained <50% of its variance with their features[18]. As shown in Fig. 3c, if the six radiomics are calculated by visible lesions only, the PCC is 0.243 ($P = 0.130$), which means that DLPE plays a critical role in extracting subvisual radiomics that are essential for COVID-19 follow-up CT analysis. We found that two radiomics of DLPE detected lesions are crucial for predicting the SGRQ score: the median signal difference between lesions and baseline (R1, or the median lesion severity), and the ratio between the lesion volume and the lung volume (R2). The mean absolute error (MAE) will significantly increase if either R1 ($P < 0.001$) or R2 ($P < 0.0001$) is removed from the predictive model. In addition, when predicting most of the other follow-up sequelae, DLPE radiomics consistently have one of the best discriminative powers among all features (Extended Data Fig. 1). Altogether, these results strongly suggest that the subvisual lesions identified by DLPE are not artefacts, but true characteristics of long-term sequelae of COVID-19, whose radiomics can be effective indicators for quantitative analysis of COVID-19 sequelae.

To evaluate the generalization power of DLPE on other tasks, we further trained and tested DLPE on the COVID-19 inpatient cohort, which contains 1,193 CT scans. On the inpatient CT dataset, DLPE found novel subvisual lesions (Fig. 3d) that resemble fainter ground-glass opacities, which may reflect mild plasma fluid leakages due to disruption of the epithelium of alveolar[19]. Plasma fluid leakages usually decrease the PaO2/FiO2 ratio (PFR)[19], which is the definitive metric when classifying COVID-19 inpatients[20,21]. We used clinical metrics and radiomics to predict PFR (see Methods), and Fig. 3f shows that subvisual lesions provide important









c

Ablation study when predicting SGRQ score

Feature	PCC	MAE	RMSE
DLPE	**0.723**	**5.83**	**7.33**
Visible	0.243	11.97	15.80
–R1	0.498	7.17	9.21
–R2	0.588	7.05	8.68

f

Ablation study when predicting PFR

Feature	PCC	MAE	RMSE
DLPE	**0.853**	**29.20**	**35.48**
Visible	0.760	36.80	44.84
–CT lesion	0.618	44.85	57.73
–LDH	0.663	42.73	52.48
–CRP	0.692	39.88	49.64

Fig. 3 | Subvisual lesions discovered by DLPE, and their relationship with sequelae and clinical metrics. a, One section of the chest CT scan from a COVID-19 survivor with an SGRQ score of 27 (that is, probably severe chronic obstructive pulmonary disease in clinical diagnosis). Upper left, a CT scan under the original lung window; lower left, its DLPE-enhanced counterpart; upper and lower right, the corresponding segmentation of lesions from the CT image in the upper and lower left, respectively. Red parts are the detected lesions, from which it can be seen that the CT under the original lung window only contains negligible lesions, whereas there are substantial lesions after being enhanced by DLPE. **b**, A scatter plot showing the predicted SGRQ score by using radiomics quantified by DLPE versus the true SGRQ score. **c**, The ablation study showing the prediction performance when only using visible radiomics, using radiomics extracted by DLPE but without using R1 (the median signal difference between lesions and baseline), and without using R2 (the ratio between the lesion volume and the lung volume) to predict the SGRQ score. RMSE, root-mean-square error. The best performer is in bold. **d**, Same as **a**, but for one section of the chest CT scan from a COVID-19 inpatient with a PFR of 274. **e**, The scatter plot showing the predicted PFR by using radiomics quantified by DLPE versus the true PFR. **f**, The ablation study showing the prediction performance when only using visible radiomics, using radiomics extracted by DLPE but without using CT lesions, without using LDH, and without using CRP to predict PFR. The best performer is in bold.

a

Segmentation performances for sub visual lesions

Methods	Original CT	DLPE enhanced
2.5D Model (10)	**0.612 ± 0.165**	**0.886 ± 0.117**
2D U-net	0.588 ± 0.181	0.819 ± 0.144
MPUnet (23)	0.601 ± 0.177	0.869 ± 0.139
3D U-net	0.598 ± 0.153	0.830 ± 0.168

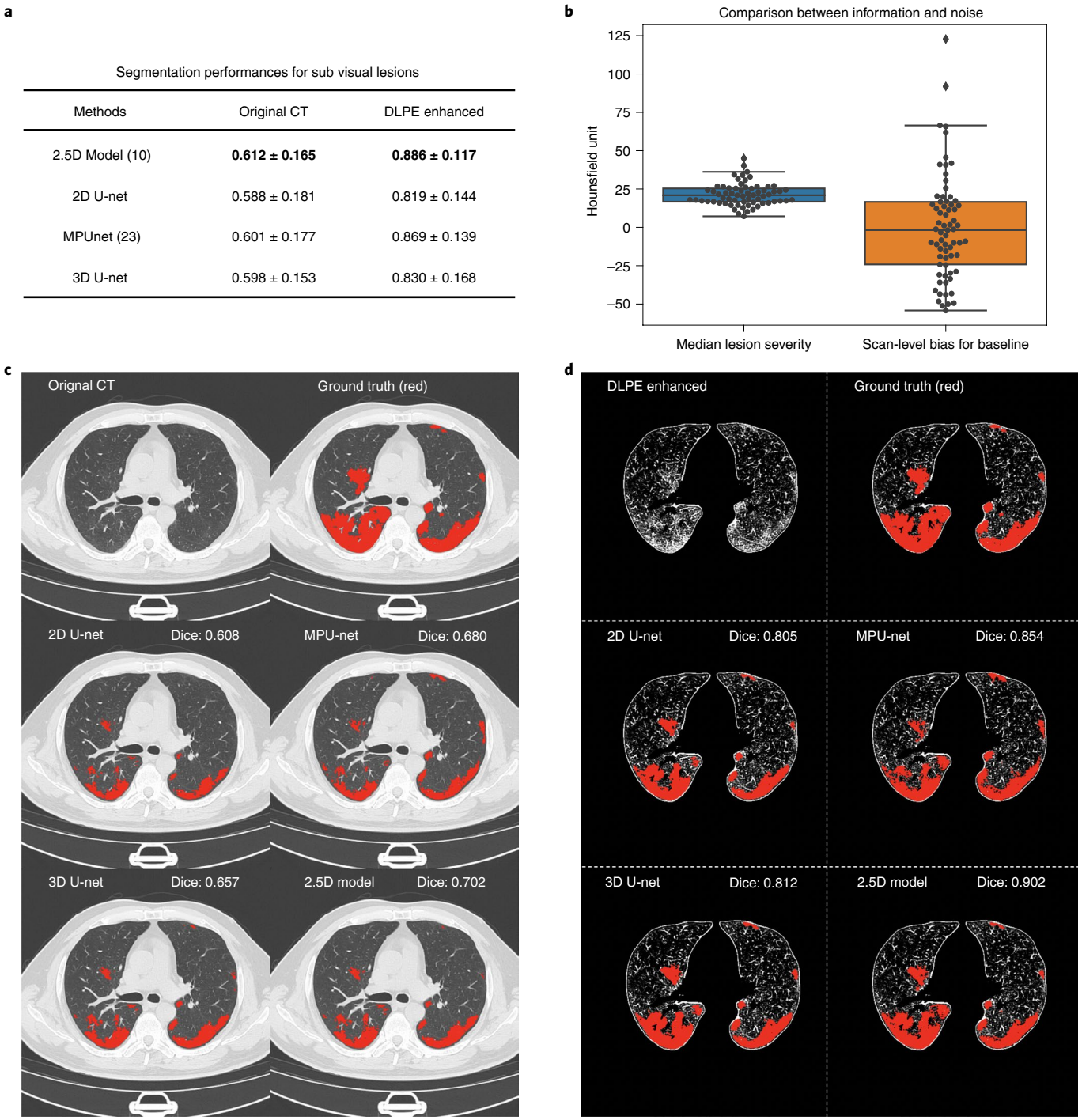


Fig. 4 | The DLPE scheme is crucial for the quantification of subvisual lesions. a, Segmentation performances for subvisual lesions without and with the DLPE scheme. Performance is shown in scan-level average dice score ± s.d. of the fivefold cross-validation on the follow-up cohort of 69 survivors. In the second column, models were inputted with the original CT, while in the third column models were inputted with the DLPE-enhanced CT. All models were pre-trained on the 1,193 COVID-19 inpatient CT scans. The best performers are in bold. **b**, Comparisons between the median lesion severity (the radiomic feature R1) with scan-level bias for baseline (a scan-level bias removed by DLPE) on the follow-up cohort. Each data point on the left shows R1 of a scan, whereas each data point on the right shows the baseline CT value of a scan subtracting the average baseline CT values of all scans. On the follow-up cohort, the variation of the baseline (noise) is 22.28 times larger than that of the R1 radiomic. This implies that without DLPE, many features for subvisual lesions may be concealed by the scan-level bias. **c**, Visualizations for models in **a**, second column. The segmentations for subvisual lesions are in red, and the dice scores indicate the performance of the models for this scan. **d**, Visualizations for models in **a**, third column. The segmentations for subvisual lesions are in red, and the dice scores indicate the performance of the models for this scan.

information for predicting PFR: if we quantify radiomics only with lesions visible under the lung window, the PCC between the predicted and ground-truth PFR decreases from 0.853 (Fig. 3e) to 0.760, and the mean absolute error (MAE) increases significantly, from 29.2 to 36.8 ($P = 0.0040$). Using DLPE, the MAE is only 29.2, which is an outstanding performance; by comparison, the MAE

between the SOTA minimally invasive PFR measurement and the invasive PFR ground truth is 26.4 (ref. [22]). Our results also found that the lactate dehydrogenase (LDH) and C-reactive protein (CRP) greatly decrease the MAE during PFR prediction ($P<0.0001$) (Fig. 3f), which conforms with previous studies[20,21].

Ablation studies. The DLPE enhancement removes scan-level bias (see equation (9)) and thus enables precise quantification of subvisual lesions. Figure 4b compares the median lesion severity (R1, an important radiomic) with the baseline CT signal (a scan-level bias, removed by DLPE) on the follow-up cohort. Without DLPE enhancement, R1 (the left distribution of Fig. 4b) is dominated by the scan-level bias of the baseline (the right distribution of Fig. 4b). On the follow-up cohort, the variation of the baseline CT signal is 22.28-times greater than that of median lesion severity, which justifies the necessity of removing the scan-level bias.

We further carried out ablation studies to show how scan-level bias can hamper the quantification of subvisual lesions. Even with the ground-truth annotation of subvisual lesions, the DLPE enhancement is still crucial for their segmentation: we compared the main segmentation model used in DLPE, that is, the 2.5D model[10], with other SOTA models such as MPU-net[23] and 3D U-net, and found that their performance difference is much smaller than the one caused by whether or not to use the DLPE enhancement to remove scan-level bias. Specifically, after DLPE enhancement, almost all existing segmentation models can segment the subvisual lesions (Fig. 4a, third column, best average dice score of 0.886), whereas without the DLPE enhancement, the best average dice score for all segmentation models is only 0.612 (Fig. 4a, second column). Figure 4c,d visualize the segmentation results for subvisual lesions without and with DLPE scheme: all models were trained with the same ground-truth annotations, but only models that were inputted with DLPE-enhanced CT scan can accurately segment subvisual lesions. Furthermore, without the DLPE enhancement to remove scan-level bias and noise for the radiomics, the PCCs of predicting several key respiratory sequelae significantly decreases (Extended Data Fig. 1). This means when radiomics are quantified without DLPE enhancement, their explanatory power significantly decreased.

Conclusion

The DLPE method combines the strengths of medical experts and AI through a human-in-the-loop training scheme to extract fully interpretable subvisual CT features for pulmonary parenchyma. Deep-LungParenchyma-Enhancing can help radiologists discover, annotate and quantify novel parenchyma lesions under many scenarios, by customizing the known lesion segmentation model in the second step of DLPE workflow for different tasks (Fig. 1). For example, we applied the DLPE scheme to the segmentation task of seven different lung diseases, including different pneumonia, tuberculosis, pulmonary nodules and lung cancers. As shown in Extended Data Fig. 2, DLPE can make robust enhancement and critical segmentation for various lung diseases, which demonstrates its generalization power and potential clinical usefulness.

In this work we applied DLPE on COVID-19 inpatient and follow-up datasets, and discovered interpretable subvisual lesions. The pathological explanations of these novel COVID-19 lesions mutually authenticate with analyses between radiomics and key clinical metrics. On our follow-up cohort, 97% of lesions are subvisual, which is one of the most important culprits of COVID-19 respiratory sequelae. More studies and more follow-up cases are needed to unveil the origin, relationship and treatment for these long-term CT lesions.

Methods

Dataset description

Training dataset for the DLPE method. We trained and cross-validated the DLPE method on a training dataset with 3,644 CT scans, for lung segmentation, airway segmentation, blood vessel segmentation, heart segmentation and parenchyma baseline CT value estimation. These data were collected from five hospitals and provided by Heilongjiang Tuomeng Technology Company. The slice thickness ranges from 1.0 mm to 5.0 mm. All of these CT scans are not acquired from COVID-19 inpatients or survivors.

COVID-19 cohort analysed by the DLPE method. We applied the DLPE method to COVID-19 survivors and inpatients. The survivor cohort contains 69 participants who were under the severe or critical condition during their inpatient period (that is, they were placed in intensive care unit). All involved survivors gave informed consent before their participation in the study. For each participant, we recorded the inpatient clinical metrics, inpatient CT scans, follow-up CT scans, follow-up lung functions and follow-up laboratory tests. These survivors provided 219 CT scans collected by one of the two commercial CT scanners: Philips iCT 256 and UIH uCT 528. The slice thicknesses range from 1.0 mm to 2.5 mm. Inpatient metics and follow-up laboratory tests are listed in Supplementary Section 2.3.9. Follow-up lung functions are listed in Supplementary Table 7. The inpatient cohort contains 1,193 COVID-19 inpatient CT scans (from 633 patients) from five hospitals. The slice thickness ranges from 1.0 mm to 2.5 mm. These patients were infected during January, 2020 and August, 2020 in Heilongjiang, China.

Data inclusion/exclusion criteria during analysis. For the training of the DLPE method, a small number of CT scans (148 CT scans) were collected after the injection of the contrast agents. However, as our DLPE method does not require a contrast-agent-injected CT to estimate the pulmonary parenchyma, we excluded these 148 CT scans during training. We instead used these 148 CT scans with contrast agents as an independent validation set to evaluate the generalization power of the DLPE method (see Supplementary Section 1.1.4).

For inpatient data of the COVID-19 cohort, we analysed the relationship between CT features and PFR. Not all inpatient CT scans have the corresponding PFR, we therefore excluded these CT scans. Finally, we have 63 inpatient CT scans with the corresponding PFR.

For follow-up data of the COVID-19 cohort, some survivors missed certain metrics. We discarded a sample if the number of missing metrics exceed 30% of the total metrics (excluded 6 out of 69).

More detailed inclusion and exclusion criteria are in Supplementary Sections 2.1 and 2.2.

DLPE method. DLPE is an interpretable and powerful method, which removes irrelevant tissues from the perspective of pulmonary parenchyma and intensifies the parenchyma lesions considerably compared to the lung window. Deep-LungParenchyma-Enhancing can thus help radiologists discover and quantify interpretable subvisual lesions. This ability is based on precise three-dimensional segmentations of airways, blood vessels, lungs and known COVID-19 lesions, as these segmentations provide landmarks when DLPE samples healthy parenchyma, and reduces noise during lesion quantification.

DLPE develops and integrates many SOTA methods, and uses multiple datasets. Here we describe key components for DLPE methods: (1) CT data normalization, (2) segmentation models, (3) parenchyma enhancement and (4) quantification of lesions.

CT data normalization. Chest CT data from different scanners have different pixel spacing, slice thickness and optimal lung windows. We therefore apply spatial and signal normalizations (Supplementary Section 1.1) to cast the data into a same, standard space, which has proven to be able to greatly improve both the robustness and accuracy in our previous research[10]. During the spatial normalization, we use the Lanczos interpolation to scale each voxel of the chest CT scan to the standard resolution of $\frac{334}{512} \times \frac{334}{512} \times 1.00\,\mathrm{mm}^3$, then pad the data into the standard shape of $512\times512\times512$. Note that the spatially normalized data correspond to the standard volume of $334\times334\times512\,\mathrm{mm}^3$, which is big enough for almost all patients in practice. During the signal normalization, we linearly rescale the original data which cast the lung window to the range of $[-0.5,0.5]$. Note that the optimal lung windows for different scanners have some differences, thus the signal normalization alleviates the system-specific bias in the datasets.

Segmentation models. Deep-LungParenchyma-Enhancing requires fast and precise segmentations for lungs, airways, blood vessels and COVID-19-visible lesions. We have developed a SOTA COVID-19 lesion segmentation model[10], which uses a 2.5D segmentation algorithm. For the segmentation of lungs, we customized the 2.5D segmentation algorithm. To segment the airways and blood vessels, we developed a two-stage segmentation protocol, which is based on our 2.5D segmentation algorithm but uses a specifically designed loss function (feature-enhanced loss) and a two-stage training and inference procedure. These approaches make our segmentations greatly exceed existing methods, especially for tiny structures, which enables the sampling of healthy parenchyma and removes irrelevant tissues.

2.5D segmentation algorithm. The 2.5D segmentation algorithm combines the two-dimensional segmentation results from *XY*, *YZ* and *XZ* planes, and then outputs the final three-dimensional segmentation. We used the 2D U-net to get

the two-dimensional segmentation results, and we further used ensemble learning to combine the results from different views. We used this algorithm to segment lungs for the DLPE method, and the segmentation reaches SOTA performance. The detailed workflow for the 2.5D segmentation algorithm is explained in Supplementary Section 1.2.

It is true that the off-the-shelf three-dimensional models such as 3D U-net and 3D V-net may extract more information, but they are orders of magnitude slower than the 2.5D algorithm: in the lung segmentation task, using two V100 GPU, the 2.5D algorithm takes 4.5 s, whereas 3D U-net needs 530 s. However, the F1-score (the harmonic mean of precision and recall) for the 2.5D segmentation algorithm and SOTA three-dimensional models are very similar. We thus used the 2.5D segmentation algorithm as the basis of DLPE segmentation models, as it is fast and accurate, which is suitable for large dataset analysis and clinical applications.

Feature-enhanced loss. In our 2.5D segmentation algorithm, the inputs of the 2D U-nets are cross-sections of the human chest. In these cross-section images, the masks of airways and blood vessels are presented as disconnected regions. The size of these regions varies greatly: cross-sections for aortas are hundreds of pixels, whereas cross-sections for tiny blood vessels are only of a few pixels. However, traditional loss functions that are based on voxel-wise performance (like voxel-level cross-entropy loss, dice loss and so on) give too little focus for tiny regions, as the area summation of all tiny regions are far less than that of big ones, which will lead to misdetections for tiny structures. We thus proposed the feature-enhanced loss that helps the 2D U-nets extract features of tiny structures.

Feature-enhanced loss is a voxel-level balanced cross-entropy loss. It is the summation of all voxel loss. For each voxel, the loss is defined as:

$$\text{voxel loss} = -w \times \ln(p) \times p\prime - \ln(1-p) \times (1-p\prime), \tag{1}$$

where p is the predicted probability that the voxel is positive (inside the structures to be segmented), and p' is the ground truth probability that the voxel is positive, which is a binary value; w is the weight indicating the penalty for the false negative prediction of this voxel, and penalty weights for the false positives are always 1 (we also tried other values which are discussed in Supplementary Section 1.3.6). Every positive voxel ($p'=1$) has a specific w, which quantifies the focus for the voxel: with higher w, the model will put more focus on the voxel.

The idea to calculate w is intuitive: airways and blood vessels have affine self-similarity, thus we require the summation of w for each branching level to be the same, because the features from different branching levels are similar while different in the scale (they are associated by affine transformations). Here we formulized the branching level: the biggest tube (level 0) splits into several (for example, 2) big tubes (level 1), and the level 1 tubes further split into a number of (for example, 4) level 2 tubes. In practice, CT images allow experienced radiologists to distinguish up to levels 7–9 for airways and levels 10–12 for blood vessels. We used the cross-section pixel number of the tube to approximate its branching level: we found that A_l, the average cross-section pixel number for the branching level l, roughly satisfies the relationship $A_i = A_0\alpha^i$, for example, for blood vessels, $A_0 = 589$, $\alpha = 0.411$. In other words, the regions which have the area (number of pixels) within $[A_{i+1}, A_i)$ are considered from branching level i, and the summation of w for all regions from branching level i is required to be a constant.

Let A be the region area which is an integer equals to the number of pixels of the region, f be the number of regions with area A in our dataset, and r be the Pearson coefficient score. We found that the the power law function is a good fit for the f–A relationship, that is, $f = c_0A^{-\gamma}$, as their log–log plot can be considered as a straight line (Extended Data Fig. 3)[24]:

For blood vessels, we analysed 1,594,446 regions and found $\gamma = 1.92$:

$$\ln(f) = -1.92\ln(A) + 18.1,\ r = -0.9944. \tag{2}$$

For airways, we analysed 420,667 regions and found $\gamma = 1.75$:

$$\ln(f) = -1.75\ln(A) + 16.0,\ r = -0.9961. \tag{3}$$

The cross-section area between the branching level i and $i+1$ belongs to $[A_0\alpha^{i+1}, A_0\alpha^i)$; the total area for the branching level i to $i+1$ is therefore given by:

$$\text{total area for level } i = \int_{A_0\alpha^{i+1}}^{A_0\alpha^i} f(A)\,A\,dA = \int_{A_0\alpha^{i+1}}^{A_0\alpha^i} c_0A^{-\gamma}A\,dA \tag{4}$$

If $\gamma = 2$, we have:

$$\text{total area for level } i = \int_{A_0\alpha^{i+1}}^{A_0\alpha^i} c_0A^{-1}dA = -c_0 ln(\alpha) \tag{5}$$

The physical meaning of γ is: $\gamma < 2$ means that the total area for tiny regions is small; $\gamma > 2$ means that the total area for big regions is small; and $\gamma = 2$ means that the total area is a constant for each branching level.

Denote the average w of a region with area A as $\bar{w}(A)$, and then the sum of $\bar{w}A$ for $A \in [A_i, A_{i+1}]$ is given by:

$$\text{focus level } i = \int_{A_0\alpha^{i+1}}^{A_0\alpha^i} f(A)\bar{w}(A)A\,dA = \int_{A_0\alpha^{i+1}}^{A_0\alpha^i} c_0A^{-\gamma}\bar{w}(A)A\,dA. \tag{6}$$

We want the focus for each branching level to be a constant, thus a simple solution is to set:

$$\bar{w}(A) = c_1A^{\gamma-2}, \tag{7}$$

where c_1 is any positive constant. We then have:

$$\text{focus level } i = \int_{A_0\alpha^{i+1}}^{A_0\alpha^i} c_0A^{-\gamma}c_1A^{\gamma-2}A\,dA = -c_0c_1\ln(\alpha). \tag{8}$$

Using equations (2), (3) and (7), $\bar{w}$ for airways follows $\bar{w} = c_1A^{-0.25}$, and for blood vessels $\bar{w}$ follows $\bar{w} = c_1A^{-0.08}$. The total focus for one region is $\bar{w}A$, and considering that the boundary pixels contain more information than insider pixels for the segmentation task (explained in Supplementary Section 1.3.6), we set the boundary pixels to have higher w: first, half of $\bar{w}A$ is equally allocated to all pixels and then the other half of $\bar{w}A$ is added equally to boundary pixels.

Finally, for class balance consideration, all w is multiplied by a constant to make the total focus for positives (sum of w) equal to the total focus for negatives (sum of penalty for negatives, which equals to the number of negatives).

Two-stage segmentation protocol. Using the 2.5D segmentation algorithm with the feature-enhanced loss, we achieved SOTA dice score (0.86 for airway segmentations and 0.89 for blood vessel segmentations). However, the segmentations for small tubes are not very natural: the segmented boundaries may zigzag, and are not smooth or continuous, and the dice score for tiny structures that have branching level > 5 is not satisfactory: 0.52 for tiny airways and 0.80 for tiny blood vessels. We thus proposed the two-stage segmentation protocol to further refine the results of the 2.5D segmentation algorithm, which dramatically improves the segmentations for tiny structures.

This protocol includes two 2.5D segmentation models using the feature-enhanced loss function: the first-stage model and the second-stage model. The first-stage model takes the normalized CT as input, and outputs a high recall mask (recall = 0.95) and a high precision mask (precision = 0.93) separately, which narrow down the search space of the second-stage model by thousands of times (Extended Data Figs. 4 and 6). The second-stage model takes the normalized CT, the high recall mask and the high precision mask as inputs, and outputs the final segmentation results.

When segmenting tiny structures, the second-stage model only needs to search in a very small search space guided by the high recall and the high precision masks. Thus, the second-stage model gives better segmentation performance, especially for tiny structures, which looks natural and very similar to human segmentations. The two-stage segmentation protocol reaches mean dice score of 0.87 for airways and 0.91 for blood vessels. For tiny structures that have branching level > 5, the dice score improvements are substantial: mean dice improves to 0.78 for tiny airways and 0.85 for tiny blood vessels.

In addition, the two-stage protocol considerably improves the robustness of the segmentations for airways and blood vessels. See Supplementary Section 1.4 for detailed discussions and visual interpretations.

Visual interpretations of airway and blood vessel segmentation models. The Extended Data Fig. 4 gives illustrations for the feature-enhanced loss, the high precision mask and the high recall mask.

We modified the Grad-Cam[25–27] to visualize the discriminative regions (Grad-Cam on the bottleneck layer) and feature-importance map (Grad-Cam on the last convolutional layer) for the segmentation models (see Supplementary Section 1.2.8 for detailed methods). As shown in Extended Data Figs. 5 and 6, the first-stage model searches on a wide range of regions that may contain discriminative features, whereas the second-stage model only focuses on regions that contain airways or blood vessels.

Parenchyma enhancement. Based on the segmentations, we could remove irrelevant tissues other than pulmonary parenchyma as well as regions with known lesions, we then randomly sampled 20,000 voxels from the remaining parenchyma of each scan.

The baseline CT value is defined as the median CT signal of the sampled voxels. Note that when blood vessels, airways and known lesions are removed from pulmonary parenchyma, there remain few tissues such as bronchiole, mediastinum, lymph glands and so on; however, they are negligible in the volume, which means that the median of the sampled CT signals can efficiently remove such noise and provide the baseline CT value for healthy parenchyma.

During the calculation of standard deviation of healthy parenchyma CT, we discarded 20% of the largest and the lowest CT values to remove outliers and potential subvisual lesions.

The optimal window centre and window width for inspecting the subtle parenchyma lesions are determined by baseline CT and the standard deviation, σ. For every CT image, we clipped the CT signal and gave radiologists in our study

two enhanced versions: one clipped with [baseline, baseline + 3σ], and the other one clipped with [baseline − 3σ, baseline]. Extended Data Fig. 7 gives illustrations of how medical experts view the enhancements. On the other hand, the enhanced CT for AI systems is the linear rescale of the original CT signals, which effectively removed scan-level bias:

$$\text{enhanced CT} = (\text{original CT} - \text{baseline})/\sigma. \quad (9)$$

On the healthy people dataset, we found that the scan-level σ is 39.5 ± 6.2 under the Hounsfield scale, and the scan-level σ on the follow-up CT dataset is 40.6 ± 6.9, which is close to that of healthy people. This consistency further supports the calculation of σ. In addition, on the follow-up dataset, scan-level standard deviations for CT signals of subvisual parenchyma lesions are 154.6 ± 47.2, which is much bigger than that of healthy parenchyma ($P < 0.0001$).

Quantification of lesions. The parenchyma enhancement removes irrelevant tissues and enhances the lesions for dozens of times compared to the lung window. Thus, some previously invisible lesions can be identified by radiologists, and radiologists give annotations for regions that look very different from the same type of images of healthy people. D.X., X.M. and X.X. were the radiologists responsible for the annotations of novel lesions in this study, and they all have more than twenty years of experience in interpreting chest CT scans.

The quantification of novel lesions has two steps: (1) training of the voxel-level novel lesion segmentation model and (2) quantification of interpretable radiomics. For step 1, the lesion segmentation model takes input of the enhanced CT (defined by equation (9)) and outputs the masks for both known and novel subvisual lesions. To this end, we used our 2.5D segmentation algorithms, but with a human-in-the-loop procedure: first, we trained an initial model on the existing COVID-19 lesion dataset of 1,193 inpatients scans, which means that the initial model can give SOTA segmentations for known lesions. Second, aided by DLPE, the radiologists in our study provided region-of-interest level annotations for regions that are likely to contain subvisual lesions for 34 CT scans. The initial model was trained on these region-of-interest level annotations and had basic abilities to detect subvisual lesions, which then gives coarse segmentations for all of the 1,412 COVID-19 CT scans. Finally, radiologists further refined 201 out of the 1,412 coarse segmentations, which are used as ground truth annotations. We thus obtained a powerful segmentation model that automatically gives voxel-level segmentations for both visible and subvisual lesions.

In step 2, we quantified the lesion volume ratio, the median lesion severity and the total lesion severity for all parenchyma and for lower respiratory parenchyma, respectively. We thus had six interpretable radiomics. The lesion volume ratio is defined as the volume of lesions divided by the volume of parenchyma; the median lesion severity is the median of the CT signal difference of lesions and baseline CT (the two kinds of novel subvisual lesions always have higher CT signals than healthy parenchyma as they should be caused by plasma fluid leakages and fibroproliferation); and the total lesion severity is the integral of CT signal difference of lesions and baseline CT. In addition, we approximated the 'lower respiratory parenchyma' with the blood vessel mask: if the nearest blood vessel for a parenchyma voxel has branching level > 7, we considered this voxel to be 'lower respiratory parenchyma' (see Extended Data Fig. 8 for lower respiratory parenchyma).

Data analysis. When the novel subvisual lesions are discovered and quantified, we evaluated how these subvisual lesions correlate with symptoms and clinical metrics, which helped us understand the origins and consequences of these novel lesions. We thus needed to answer two questions: (1) whether main symptoms and clinical metrics can be explained by the novel lesions; and (2) how much information is provided by the novel lesions in the regression or classification tasks to predict clinical metrics. To this end, we used several efficient data analysis methods including Lasso regression, XGBoost and multivariable analysis. Lasso is a light and efficient regression model. XGBoost is a powerful regressor that can rank feature importance and select features. We also used an algorithm based on neural networks for multivariable analysis.

Feature and target selection. We needed to evaluate whether main symptoms and clinical metrics can be predicted by subvisual lesions. During the multivariable analysis, the neural network tries to form mappings between input features with symptoms and metrics that minimize the loss function, which quantifies the overall prediction error. The loss function forces the network to give more attention to symptoms and metrics that can be predicted, and also quantifies the predictabilities for these symptoms and metrics. More details about this multivariable analysis method are in Supplementary Sections 2.3.2–2.3.6.

For respiratory sequelae, there are 16 metrics that measure the life quality decrease and lung function impairments. We selected 53 informative features (Supplementary Table 9): 21 statistics of key clinical metrics and 3 radiomics during hospitalization, 5 basic information features, 6 follow-up CT radiomics and 18 key follow-up blood test features. We thus tried to map the 16 sequelae metrics with these 53 features. We found that the SGRQ score is the most predictable sequela, which draws 32.6% of the model's total focus. Other predictable metrics are total lung capacity, expiratory reserve volume and so on (see Supplementary Tables 10 and 11 for more details). These results conform with past knowledge of pulmonary fibrosis sequelae.

For inpatient clinic metrics, we focused on their correlations with PFR. We selected 12 informative features (Supplementary Table 8), including 3 radiomics (the median lesion severity, the lesion volume ratio and the total lesion severity of all parenchyma); 6 clinic metrics (LDH, CRP, lymphocyte count, neutrophil count, D-Dimer, and the ratio between the lymphocyte count and the neutrophil count); and 3 basic information features (sex, age and body mass index).

Feature importance ranking. Both XGBoost and multivariable analysis can quantify the discriminative power of features and select optimal feature combinations. Both methods show that subvisual features from DLPE are one of the most important for predicting key COVID-19 clinical metrics and sequelae (Supplementary Tables 10 and 11). These results indicate that the DLPE method can extract important information that has the best discriminating power for various key clinical metrics and COVID-19 respiratory sequelae.

Regression models. We tried Lasso regression, XGBoost and multivariable analysis. Among them XGBoost outperforms the other two in most cases. We therefore used XGBoost to predict PFR and respiratory sequelae. We used leave-one-out cross-validation to predict the target value, and we used the PCC between the predicted and the ground-truth to evaluate the regression performance. Extended Data Fig. 1 presents the detailed performance and ablation results without the DLPE method. As it is very hard to collect a large amount of follow-up data, the performance of the regression models may be improved with more data.

Ethics statement. The patient data were collected from The First Affiliated Hospital of Harbin Medical University following the approval from the Institutional Review Board. The study was also approved by the Institutional Biosafety and Bioethics Committee at King Abdullah University of Science and Technology. Informed consent was waived on the training cohort and the inpatient cohort due to the retrospective nature of the study. All involved participants in the survivor cohort gave informed consent before their participation in the study.

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

Data availability

The anonymized follow-up CT data that support the findings of this study are attached publicly with the trained models. See https://github.com/LongxiZhou/DLPE-method (ref. [28]) for step-by-step guidance for downloading the CT data and the trained models. The datasets for the training of the DLPE method are available from the corresponding author on reasonable request. Detailed manuals for the replication of our study are in the Supplementary Information.

Code availability

The source code and the trained models for a working version of DLPE is available at https://github.com/LongxiZhou/DLPE-method (ref. [28]).

Received: 19 September 2021; Accepted: 4 April 2022;
Published online: 23 May 2022

References

1. Huang, C. et al. 6-Month consequences of COVID-19 in patients discharged from hospital: a cohort study. *Lancet* **397**, 220–232 (2021).
2. Han, X. et al. Six-month follow-up chest CT findings after severe COVID-19 pneumonia. *Radiology* **1**, 299 (2021).
3. 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).
4. Francone, M. et al. Chest CT score in COVID-19 patients: correlation with disease severity and short-term prognosis. *Eur. Radiol.* **30**, 6808–6817 (2020).
5. Chen, X. et al. A systematic review of neurological symptoms and complications of COVID-19. *J. Neurol.* **268**, 392–402 (2021).
6. Mannil, M. et al. Texture analysis and machine learning for detecting myocardial infarction in noncontrast low-dose computed tomography: unveiling the invisible. *Invest. Radiol.* **53**, 338–343 (2018).
7. Savadjiev, P. et al. Demystification of AI-driven medical image interpretation: past, present and future. *Eur. Radiol.* **29**, 1616–1624 (2019).
8. Pesapane, F. et al. Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine. *Eur. Radiol. Exp.* **2**, 35 (2018).
9. Oren, O. et al. Artificial intelligence in medical imaging: switching from radiographic pathological data to clinically meaningful endpoints. *Lancet Digital Health* **2**, e486–e488 (2020).

10. Zhou, L. et al. A rapid, accurate and machine-agnostic segmentation and quantification method for CT-based COVID-19 diagnosis. *IEEE Trans. Med. Imaging* **39**, 2638–2652 (2020).
11. Zheng, H. et al. Alleviating class-wise gradient imbalance for pulmonary airway segmentation. *IEEE Trans. Med. Imaging* **40**, 2452–2462 (2021).
12. Nam, J. G. et al. Automatic pulmonary vessel segmentation on noncontrast chest CT: deep learning algorithm developed using spatiotemporally matched virtual noncontrast images and low-keV contrast-enhanced vessel maps. *Eur. Radiol.* **31**, 9012–9021 (2021).
13. Swigris, J. J. et al. The psychometric properties of the St George's Respiratory Questionnaire (SGRQ) in patients with idiopathic pulmonary fibrosis: a literature review. *Health Qual. Life Outcomes* **12**, 124 (2014).
14. Vasarmidi, E. et al. Pulmonary fibrosis in the aftermath of the COVID-19 era (review). *Exp. Ther. Med.* **20**, 2557–2560 (2020).
15. Grillo, F. et al. Lung fibrosis: an undervalued finding in COVID-19 pathological series. *Lancet Infect. Dis.* **21**, E72 (2021).
16. Peter, M. et al. Pulmonary fibrosis and COVID-19: the potential role for antifibrotic therapy. *Lancet Resp. Med.* **8**, 807–815 (2020).
17. Ringbaek, T. et al. A comparison of the assessment of quality of life with CAT, CCQ, and SGRQ in COPD patients participating in pulmonary rehabilitation. *COPD* **9**, 12–15 (2012).
18. Karloh, M. et al. The COPD assessment test: what do we know so far?: A systematic review and meta-analysis about clinical outcomes prediction and classification of patients into GOLD stages. *Chest* **149**, 413–425 (2016).
19. Burnham, E. L. et al. The fibroproliferative response in acute respiratory distress syndrome: mechanisms and clinical significance. *Eur. Resp. J.* **43**, 276–285 (2014).
20. Ranieri, V. M. Acute respiratory distress syndrome: the Berlin definition. *JAMA* **307**, 2526–2533 (2012).
21. Fu, L. et al. Influence factors of death risk among COVID-19 patients in Wuhan, China: a hospital-based case-cohort study. Preprint at https://www.medrxiv.org/content/10.1101/2020.03.13.20035329v1 (2020)
22. Brown, S. M. et al. Non-linear imputation of PaO2/FIO2 from SpO2/FIO2 among mechanically ventilated patients in the intensive care unit: a prospective, observational study. *Crit. Care Med.* **45**, 1317–1324 (2017).
23. Mathias, P. et al. One network to segment them all: a general, lightweight system for accurate 3D medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019* 30–38 (Springer, 2019).
24. Jeong, H. et al. The large-scale organization of metabolic networks. *Nature* **407**, 651–654 (2000).
25. Selvaraju, R. R. et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. In *Proc. IEEE International Conference on Computer Vision (ICCV)* 618–626 (IEEE, 2017).
26. Vinogradova, K. et al. Towards interpretable semantic segmentation via gradient-weighted class activation mapping. *Proceedings of the AAAI Conference on Artificial Intelligence* 13943–13944 (AAAI 2020).
27. Wickstom, K. et al. Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps. *Med. Image Anal.* **60**, 101619 (2020).
28. Zhou, L. et al. *1.0 LongxiZhou/DLPE-method: DeepLungParenchymaEnhancement* (Zenodo, 2022); https://doi.org/10.5281/ZENODO.6387701

Acknowledgements

This work was supported by grants from the National Natural Science Foundation of China to G.W. (grant no. 62072095), from KAUST to X.G. (grant nos. BAS/1/1624-01, FCC/1/1976-18-01, FCC/1/1976-23-01, FCC/1/1976-25-01, FCC/1/1976-26-01, URF/1/4098-01-01, REI/1/0018-01-01, REI/1/4216-01-01, REI/1/4437-01-01, REI/1/4473-01-01, URF/1/4352-01-01 and REI/1/4742-01-01), from the Application Technology Research and Development Program of Heilongjiang Province to K.Y. (grant no. GA200001), and from the Science and Technology Program of Suzhou (grant nos. ZXL2021431 and RC2021130) and the Fundamental Research Funds for the Central Universities (grant no. 2572020DR10) to Z.Q. We thank X. Chen at North Carolina Central University, and S. Garantziotis at the National Institute of Environmental Health Sciences, for their advice during the data collection of the COVID-19 follow-up cohort.

Author contributions

X.G. and L.Z. conceived this study. X.G., Z.Q. and K.K initiated the study. Y.H., N.B., Y.Z. and L.Z. extracted and prepared the CT datasets for COVID-19 inpatient lesions, airways, pulmonary blood vessels, lungs and the heart. X.M., D.X. and X.X. prepared the CT datasets for COVID-19 survivors and gave medical guidance for the project. J.N.Z., M.Z. and Y.C. prepared the datasets for clinical metrics. L.Z. implemented the DLPE method. L.Z., J.X.Z. and H.L. completed the data analysis. L.Z. wrote the manuscript under supervision of X.G. All authors are involved in discussion and finalization of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s42256-022-00483-7.

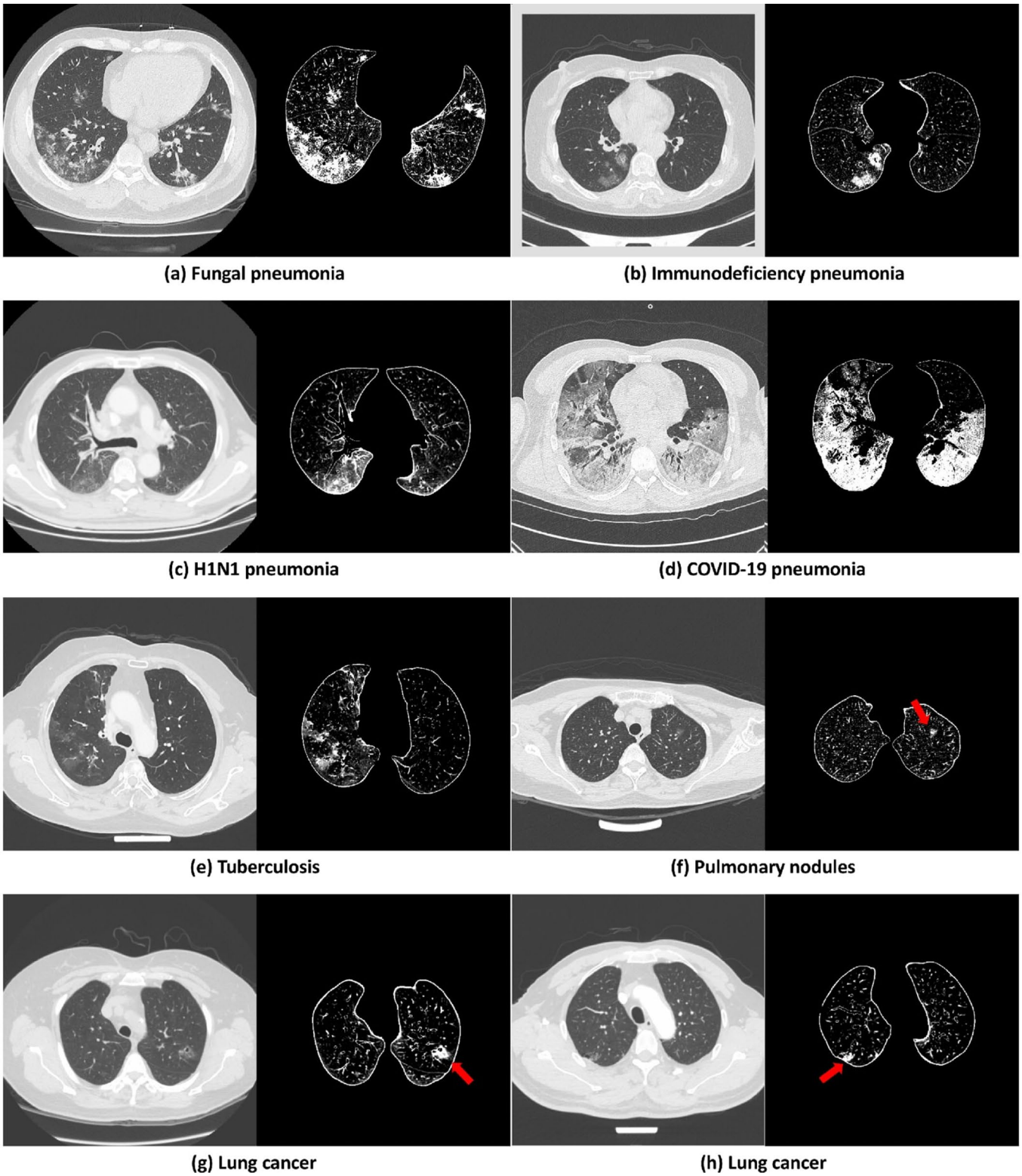
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-022-00483-7.

Correspondence and requests for materials should be addressed to Guohua Wang, Lawrence Carin, Xigang Xiao, Kaijiang Yu, Zhaowen Qiu or Xin Gao.

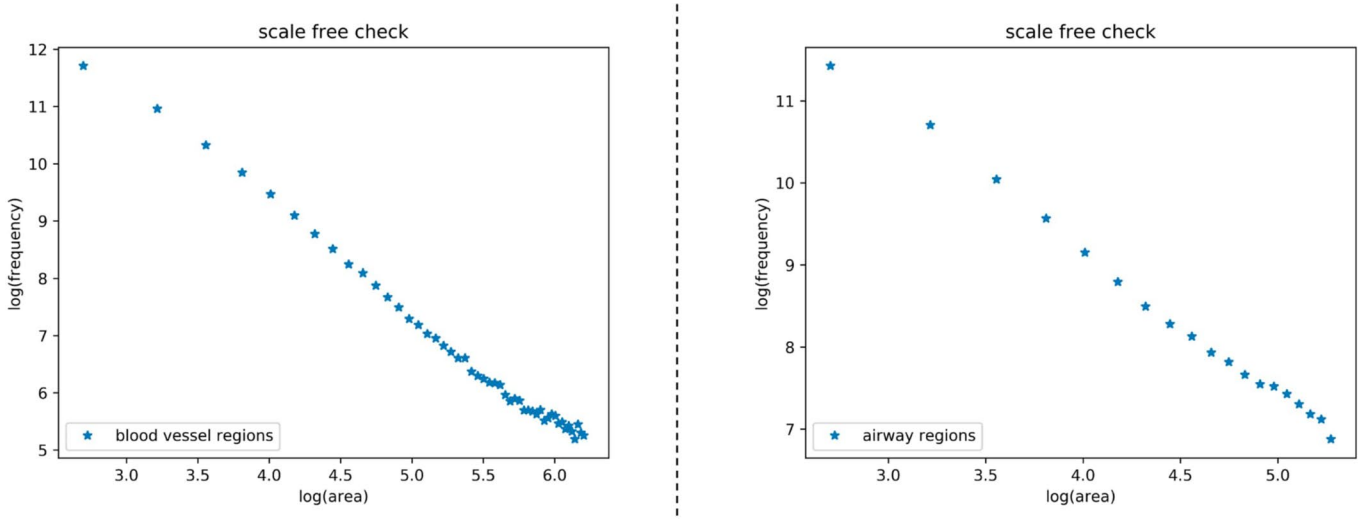
Peer review information *Nature Machine Intelligence* thanks Andrey Rzhetsky, Chaofan Chen and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Target	PCC with DLPE	Top three informative	PCC without DLPE	DLPE Sensitive Radiomics
SGRQ	**0.732**	R2 total, R1 total, R1 lower	0.243	R2 total***, R1 total**
DLCO/VA	**0.550**	ctBil, R2 Total, pH	0.327	R2 total*
pO2(A-a)	**0.718**	Leu, ctBil, R2 total	0.532	R2 total*
FEV1	**0.623**	ctBil, FO2Hb, R1 lower	0.471	R1 lower*
FVC	**0.605**	R2 total, ctBil, cLac	0.325	R2 total**, R2 average*
FEV1/FVC	**0.649**	R1 lower, FO2Hb, CK	0.387	R1 lower**
MEF50	**0.462**	CRP, pO2, PT	0.411	None
TLC	**0.705**	R1 total, pO2, R1 lower	0.391	R1 total**, R1 lower*, R2 lower*
FRC	**0.434**	R2 total, cLac, CK	0.395	None
RV	0.413	P50, ALT, Age	**0.420**	None
ERV	**0.729**	Height, R2 total, Cr	0.546	R2 total*
IC	0.444	CRP, ctBil, Leu	**0.449**	None
PEF	**0.450**	pO2, ALT, R1 lower	0.375	None
VT	**0.386**	BMI, Cr, R1 lower	0.361	None
MV	**0.483**	R2 total, pH, LDH	0.423	None
mMRC	**0.675**	Age, Cr, LDH	0.613	None

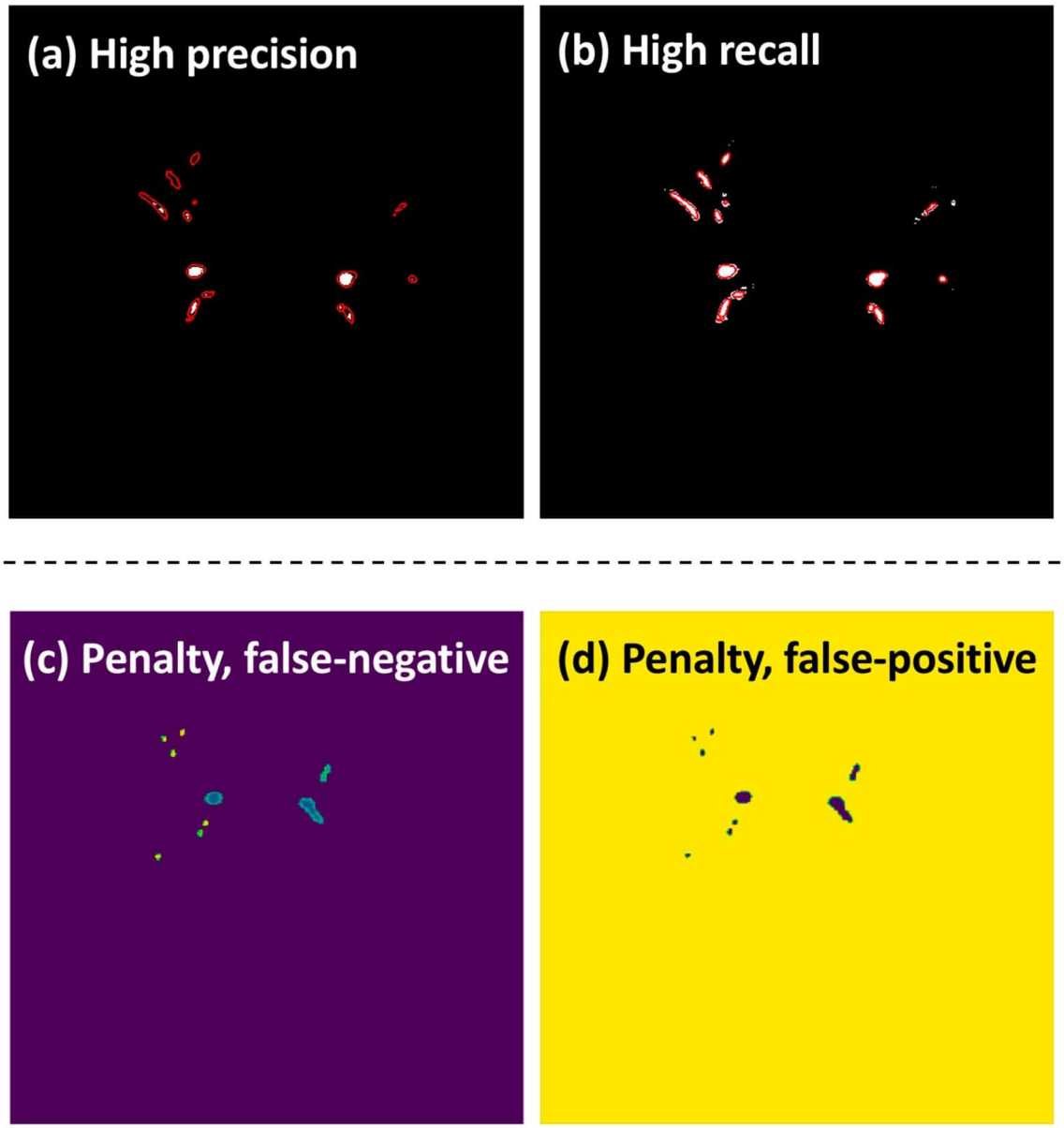
Extended Data Fig. 1 | PCC when predicting COVID-19 long-term respiratory sequelae. Using inpatient and follow-up clinical metrics and radiomics to predict the COVID-19 long-term respiratory sequelae. See Supplementary Table 7 for the descriptions of these sequelae (the first column). See Supplementary Table 9 for the description of these targets (the third column). The prediction model is XGBoost. First column, the 16 sequelae. Second column, the PCC between the predicted value and the ground truth value. Third column, the top three most informative features ranked by the XGBoost. Fourth column, replace the radiomics with the lesion quantification of the state-of-the-art methods for COVID-19 lesions, but without using the DLPE scheme, and re-train the model. Fifth column, radiomics that without DLPE to remove bias will significantly decrease the PCC: * means $p < 0.01$, ** means $p < 0.001$ and *** means $p < 0.0001$. We compare the second column and the fourth column, and the best performer is in bold.



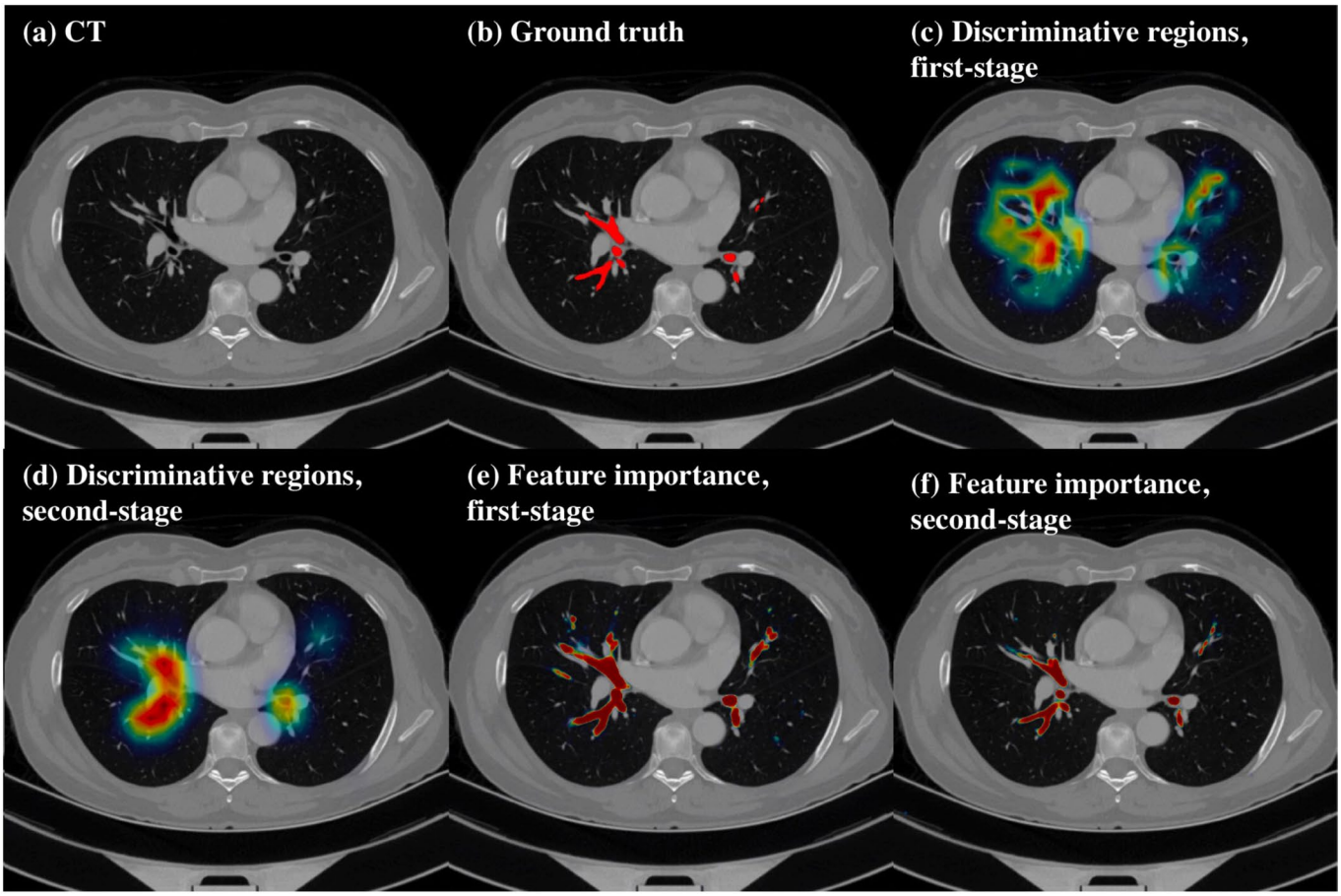
Extended Data Fig. 2 | DLPE enhancements for other lung diseases. DLPE enhancements for other lung diseases. (**a**-**d**) are different kinds of pneumonia, and DLPE can make robust enhancements for various data quality and lesion severity: (**a**) fungal pneumonia; (**b**) immunodeficiency pneumonia; (**c**) H1N1 pneumonia; (**d**) COVID-19 pneumonia. (**e**-**h**) are examples of DLPE enhancements that may help radiologists observe the lesions: (**e**) tuberculosis; (**f**) a high risk nodule that may develop to cancer (red arrow); (**g**-**h**) lung cancers (red arrow).



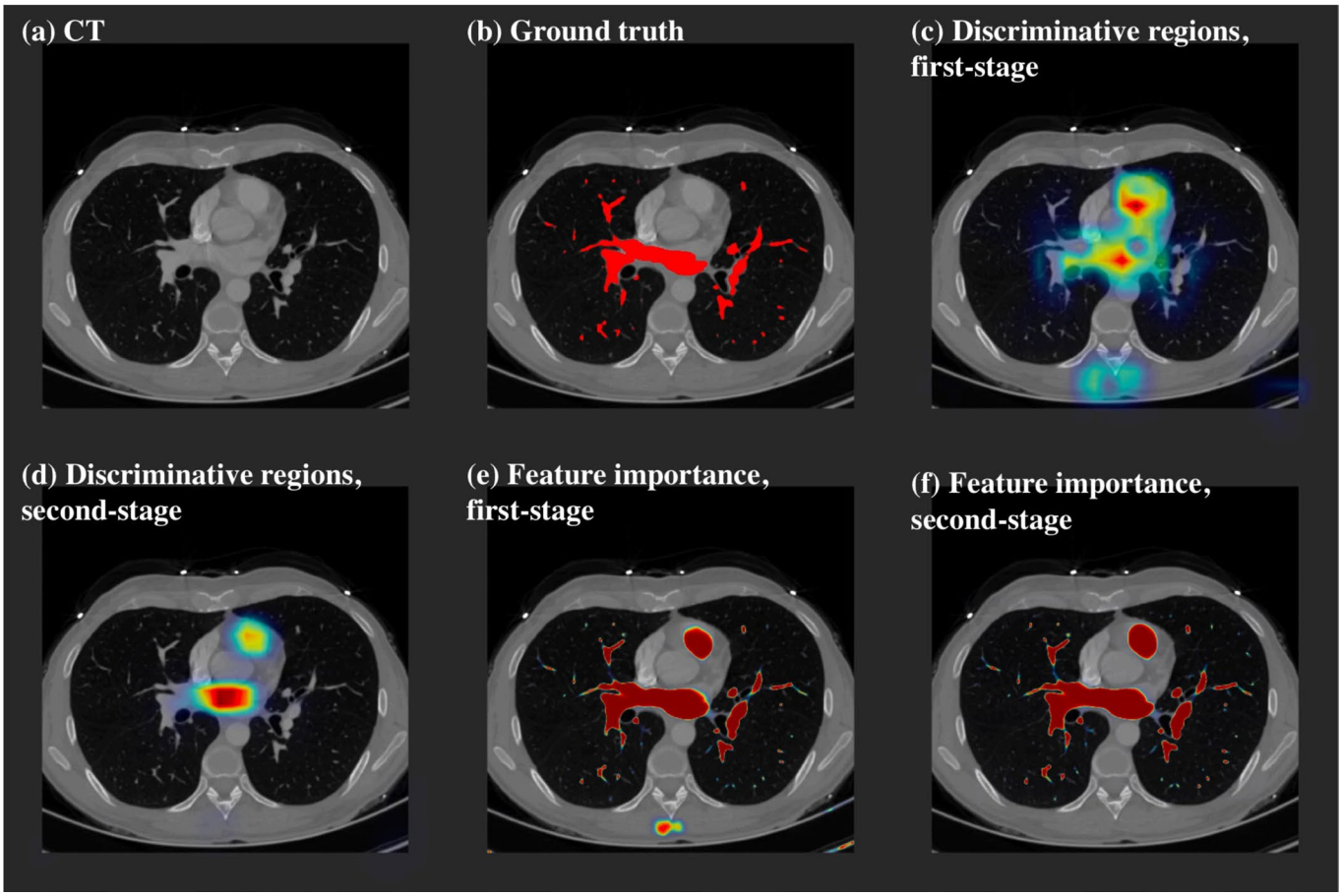
Extended Data Fig. 3 | The log-log plot for the *f*~*A* relationship for blood vessels and airways. The log-log plot for the *f*~*A* relationship for blood vessels (the left panel) and airways (the right panel). The linear regression for the log-log plot of the blood vessels results in $\ln(f) = -1.92\ln(A) + 18.1$, $r = -0.9944$. The linear regression for the log-log plot of the airways results in $\ln(f) = -1.75\ln(A) + 16.0$, $r = -0.9961$. Thus, the power law function is a good fit for the *f*~*A* relationship for blood vessels and airways.



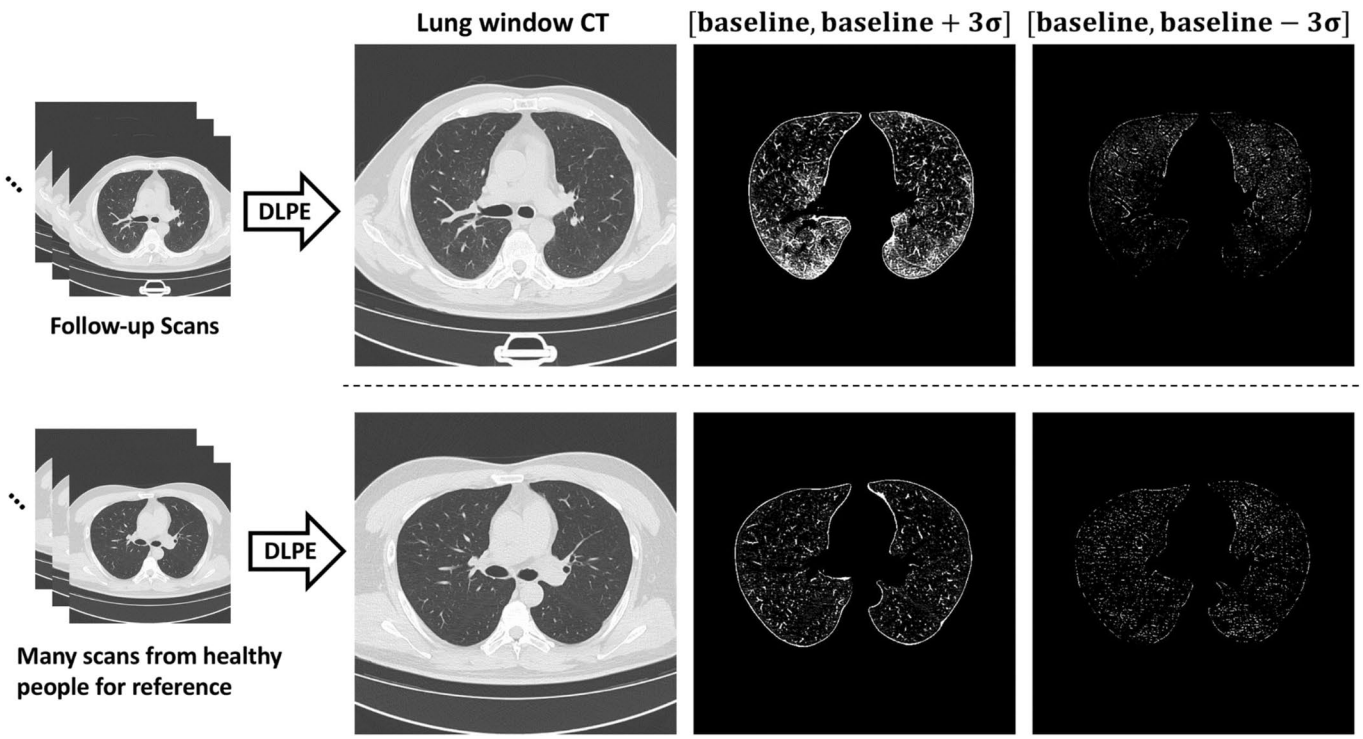
Extended Data Fig. 4 | Illustration for the two-stage protocol and feature-enhanced loss. Illustration for the two-stage protocol and feature-enhanced loss. (**a**,**b**) high precision and high recall masks; (**c**,**d**) the feature-enhanced loss. For (**a**,**b**), the red lines give the boundary of the ground truth of the airways, while the white regions are the high precision mask. (**c**) Gives the penalty weight (yellow means higher, blue means lower) for the false negative prediction, and the penalty weight distribution is determined by the branching level. (**d**) Gives the penalty weight (yellow means higher, blue means lower) for the false positive prediction: here is to get the airway segmentation, and the penalty weight for each non-airway voxel is a constant that ensures the total weights for positives and negatives are class balanced.



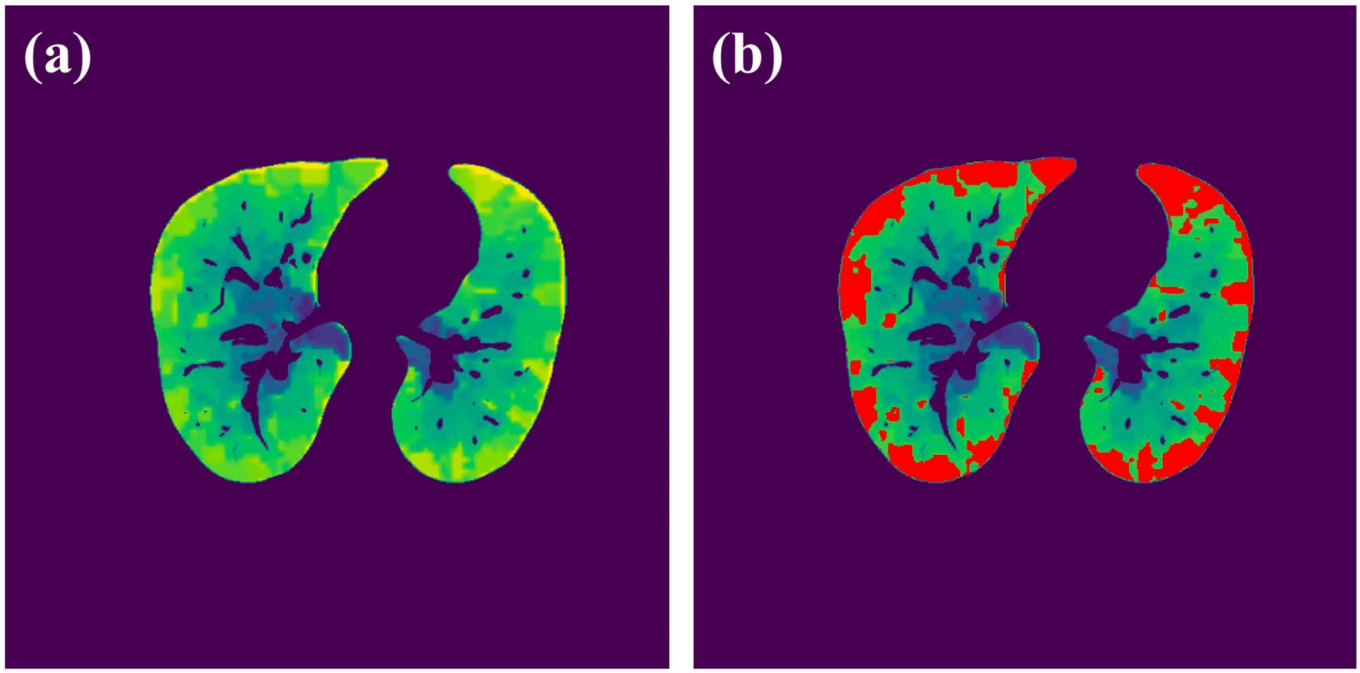
Extended Data Fig. 5 | Visual interpretation for the first-stage and the second-stage models when segmenting the airways. Visual interpretation for the first-stage and the second-stage models when segmenting the airways. (**a**) The spatial rescaled CT from the x-y plane. (**b**) Red regions give the ground truth for the airways. (**c**) The discriminative regions for the first-stage model. We can see that when segmenting the airways, the model searches on a wide range of regions that may contain discriminative features. (**d**) The discriminative regions for the second-stage model. We can see that the second-stage model only searches on the regions that contain airways. (**e**) The feature importance map for the first-stage model. We can see that the model focuses on the region broader than the airways, as the model needs to find out where the tracheal walls are. In addition, the first-stage model also focuses on some wrong regions. (**f**) The feature importance map for the second-stage model. We can see that the second-stage model only focuses on the airway regions.



Extended Data Fig. 6 | Visual interpretation for the first-stage and the second-stage models when segmenting the blood vessels. Visual interpretation for the first-stage and the second-stage models when segmenting the blood vessels. (**a**) The spatial rescaled CT from the x-y plane. (**b**) Red regions give the ground truth for the blood vessels. (**c**) The discriminative regions for the first-stage model. We can see the model looks on wide regions, for example, the model searches on the bottom part where there are no blood vessels. (**d**) The discriminative regions for the second-stage model. We can see that the discriminative regions for the second-stage model are much more concentrated. (**e**) The feature importance map for the first-stage model. We can see that the model falsely focuses on the bottom regions. (**f**) The feature importance map for the second-stage model. We can see that the second-stage model only focuses on the blood vessels, and puts more focus on tiny vessels.



Extended Data Fig. 7 | Illustration for how medical experts view the enhancements. Illustration for how medical experts view the enhancements. For a CT scan aimed to find subvisual lesions, each CT slice will be presented as: the lung window CT, [baseline,baseline + 3σ] and [baseline,baseline−3σ]. At the same time, medical experts can view CT slices from thousands of healthy people for reference.



Extended Data Fig. 8 | Illustration of the lower respiratory regions. Illustration of the lower respiratory regions. We use the branching level of the nearest blood vessels to approximate the lower respiratory regions. (**a**) Branching level of the nearest blood vessels. Brighter means higher branching level, and the brightest is of branching level around 12. (**b**) Red regions give the estimated lower respiratory regions. If the nearest blood vessel is of branching level > 7 for a parenchyma voxel, we classify the voxel as a lower respiratory voxel.

nature research

Corresponding author(s): Xin Gao

Last updated by author(s): Mar 27, 2022

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection	The medical imaging data is the raw output of the CT machines, which is in the Digital Imaging and Communication in Medicine (DICOM) format. Clinical metrics and follow-up questionnaires did not include any software for the data collection.
Data analysis	The data was analyzed with Python 3.6.7. The deep-learning module in this work was written with Pytorch 1.9.0. Our program is publicly available at https://github.com/LongxiZhou/DLPE-method

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

Data

Policy information about availability of data

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

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

We declare that all the data supporting the novel lesions for COVID-19 inpatient and survivors are available at https://github.com/LongxiZhou/DLPE-method; Figure 1-4 in the main text and Supplementary Figure 4, 5, 8, 9, 16 are associated with the raw data.
The dataset for training the DLPE method is owned by Heilongjiang Tuomeng Technology Co. Ltd., Harbin, China, and is available upon request.
The trained DLPE models are available at https://github.com/LongxiZhou/DLPE-method, which can convert CT scans in DICOM into enhanced arrays (remove airways, blood vessels, provide the optimal window), or can be used as state-of-the-art segmentation models for the segmentation of COVID-19 lesions, lungs, airways and blood vessels.

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	The DLPE method is based on accurate segmentations for the lungs, airways and blood vessels. The segmentation models were trained over the dataset containing 3644 CT scans collected from patients from 5 different hospitals. We tested DLPE method on COVID-19 inpatient and survivors, in total 219 CT scans. The DLPE method generated satisfactory enhancement effects for all these scans (especially for patients with faint lesions).
Data exclusions	Data inclusion/exclusion criteria were described in the main text and discussed in details in Supplementary Section 1.1.1. In brief, we excluded data if the CT quality is very low, such as extensive artifacts and noises.
Replication	We tested the robustness of the DLPE method on CT scans collected under varies conditions, e.g., with contrast agents, extensive nosies, etc. The detail information is listed in Supplementary Section 1.1.4, 1.2.10 and 1.4.6.
Randomization	We conducted multi-fold cross validation during our model training.
Blinding	Blinding is not relevant as our study does not involve new clinical trials or group allocation of experimental subjects.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

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

Methods

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

Human research participants

Policy information about studies involving human research participants

Population characteristics	COVID-19 patients of severe or critical conditions (requiring supplemental oxygen). These patients were infected from Feb to Apr, 2020. All clinical data were collected in Heilongjiang, China.
Recruitment	COVID-19 patients of severe or critical condition (requiring supplemental oxygen). Time: Feb to Apr, 2020. All clinical data were collected in Heilongjiang, China. Potential biases: we did not include light or mild patients, as these patients did not sample the clinical metrics we wanted to analyze, e.g., PaO2/FiO2 ratio, blood-gas data, etc. The participates were all Chinese and they were infected before the occurrence of the major SARS-CoV-2 variants.
Ethics oversight	The First Affiliated Hospital of Harbin Medical University, China; King Abdullah University of Science and Technology, Saudi Arabia

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

Clinical data

Policy information about clinical studies

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

Clinical trial registration	The study does not include new clinical trials. Our study analyzed the existing clinical data.
Study protocol	N/A. The study does not include new clinical trials.
Data collection	Recruitment: COVID-19 patients with severe or critical conditions (requiring supplemental oxygen). Time: Feb to Apr, 2020. All clinical data were collected in Heilongjiang, China.
Outcomes	N/A