scientific reports

OPEN

Explainable fully automated CT scoring of interstitial lung disease for patients suspected of systemic sclerosis by cascaded regression neural networks and its comparison with experts

Jingnan Jia[1], Irene Hernández-Girón[1], Anne A. Schouffoer[2], Jeska K. de Vries-Bouwstra[2], Maarten K. Ninaber[3], Julie C. Korving[4], Marius Staring[1], Lucia J. M. Kroft[4] & Berend C. Stoel[1]✉

Visual scoring of interstitial lung disease in systemic sclerosis (SSc-ILD) from CT scans is laborious, subjective and time-consuming. This study aims to develop a deep learning framework to automate SSc-ILD scoring. The automated framework is a cascade of two neural networks. The first network selects the craniocaudal positions of the five scoring levels. Subsequently, for each level, the second network estimates the ratio of three patterns to the total lung area: the total extent of disease (TOT), ground glass (GG) and reticulation (RET). To overcome the score imbalance in the second network, we propose a method to augment the training dataset with synthetic data. To explain the network's output, a heat map method is introduced to highlight the candidate interstitial lung disease regions. The explainability of heat maps was evaluated by two human experts and a quantitative method that uses the heat map to produce the score. The results show that our framework achieved a κ of 0.66, 0.58, and 0.65, for the TOT, GG and RET scoring, respectively. Both experts agreed with the heat maps in 91%, 90% and 80% of cases, respectively. Therefore, it is feasible to develop a framework for automated SSc-ILD scoring, which performs competitively with human experts and provides high-quality explanations using heat maps. Confirming the model's generalizability is needed in future studies.

Systemic sclerosis (SSc) is a rare autoimmune connective tissue disease affecting different organs with high mortality[1], of which interstitial lung disease (ILD) is the primary cause[2]. The extent of interstitial lung disease in systemic sclerosis (SSc-ILD) on lung CT images has been identified as an independent predictor of disease progression and mortality in patients with SSc[3]. Quantification of SSc-ILD extent is also needed for treatment initiation and evaluation of its efficacy[2]. Several scoring systems have been proposed to quantify SSc-ILD from chest CT scans[4]and the most useful and commonly used quantitative method in the clinical setting is proposed by Goh and colleagues[4,5]. In this scoring system, CT images are scored in five axial slices, corresponding to levels of: a) origin of the great vessels; b) main carina; c) pulmonary venous confluence; d) halfway between the third and fifth level; e) 1 cm above the right hemi-diaphragm[5]. At each level, three visual scores are estimated as the percentage of total lung area that is covered by: total disease extent (TOT), ground-glass opacities (GG) and reticular patterns (RET), as shown in Figure 1. TOT area is the union of the areas from GG and RET, making TOT scores less than or equal to the sum of GG and RET scores. Each of these scores is used in risk

[1]Division of Image Processing, Department of Radiology, Leiden University Medical Center (LUMC), P.O. Box 9600, 2300 RC Leiden, The Netherlands. [2]Department of Rheumatology, Leiden University Medical Center (LUMC), P.O. Box 9600, 2300 RC Leiden, The Netherlands. [3]Department of Pulmonology, Leiden University Medical Center (LUMC), P.O. Box 9600, 2300 RC Leiden, The Netherlands. [4]Department of Radiology, Leiden University Medical Center (LUMC), P.O. Box 9600, 2300 RC Leiden, The Netherlands. ✉email: b.c.stoel@lumc.nl

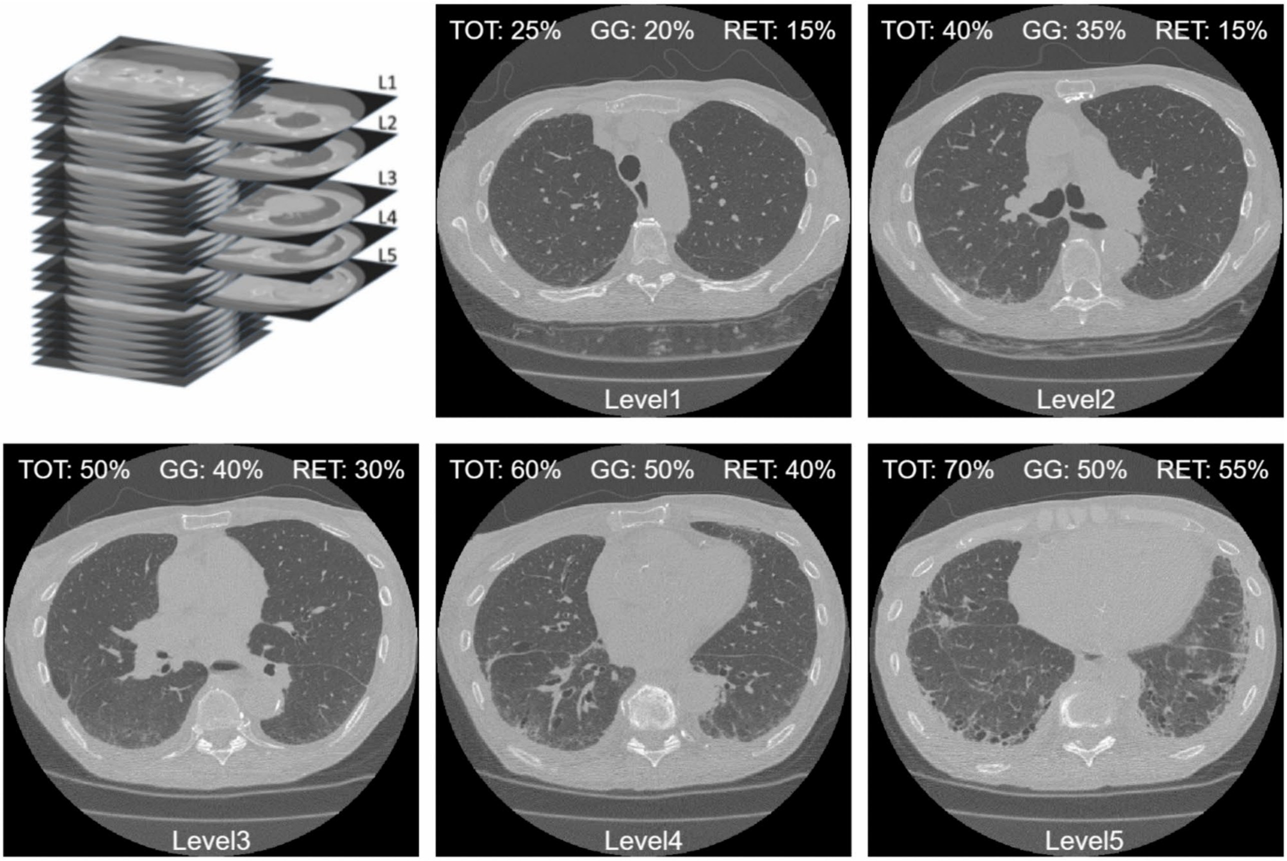

L1
L2
L3
L4
L5
TOT: 25% GG: 20% RET: 15%
Level1
TOT: 40% GG: 35% RET: 15%
Level2
TOT: 50% GG: 40% RET: 30%
Level3
TOT: 60% GG: 50% RET: 40%
Level4
TOT: 70% GG: 50% RET: 55%
Level5

Fig. 1. Illustration of five levels in the same CT volume from one patient of systemic sclerosis. Interstitial lung disease scores from human experts are annotated on the top of each image. The level numbers are annotated at the bottom. TOT = total disease extent, GG = ground glass opacity, RET = reticular pattern.

stratification, where TOT and RET are associated with mortality[5], while GG can differentiate SSc-ILD from idiopathic pulmonary ILD[6].

Nonetheless, visual scoring remains a challenging task, because of difficulties in recognizing different patterns and estimating its ratio to the whole lung. From Figure 1it is conceivable how difficult it is to identify different patterns and estimate their ratios for each level, especially when GG and RET overlap. Therefore, an atlas was published to provide a guiding consensus document to reduce inter-observer variability[7]. Despite this, ILD scoring is still laborious and dependent on rater experience. Therefore, an automatic scoring tool is needed to overcome these limitations[8,9]and to make the scoring available for clinical practice. An automated scoring tool would consist of two steps: 1) selecting the five levels (axial CT slices) according to anatomical landmarks; and 2) estimating the score for each selected slice by recognizing the different patterns and estimating their proportion to the total lung area. Recently, deep neural networks have been proposed for anatomical level localization[10]and quantification of imaging features[3,11,12], which are closely related to the two steps needed for automated ILD scoring. While several methods combined the two steps together to estimate other imaging biomarkers[13–16], there are few published works applied on SSc-ILD scoring.

The purpose of this study was, therefore, to build a fully automated framework to select the five anatomical levels and score the extent of SSc-ILD for each level directly, without needing manual segmentations. The main contributions of this paper are as follows:

- To the best of our knowledge, we are the first to propose an automated framework for SSc-ILD scoring without pixel-wise fibrosis annotation.
- We introduced a data synthesis method to generate training images with exact SSc-ILD scores, leading to significant improvement in the SSc-ILD scoring.
- We proposed replacement-based heat maps, which can visually explain the network's output, making the framework more easily acceptable for clinicians. The reliability of heat maps was evaluated by an automatic evaluation method and by two human experts.
- Our framework performed competitively to experienced human experts, while costing only several seconds.

Related work

An automated scoring framework may consist of two networks: 1) level selection (i.e. localization) from 3D medical images; and 2) scoring of the resulting 2D image slices. This section therefore reviews related studies of these two topics and their application on SSc-ILD.

Automated level selection

Existing methods show that, despite trained on global image-level labels, convolution neural networks (CNNs) have a remarkable ability to localize objects-of-interest[17]. Level selection aims to localize the anatomical level or anatomical landmarks[4,5] from 3D CT images. In recent years, there are a great number of works on landmark localization in 3D medical images (see Table 1), e.g. localization of the upper and lower edge of lungs[18]in chest CT images, the lumbar vertebra[10,19,20]in spine CT images, the anatomical landmarks in cardiac ultrasound images[21], anatomical structure localization in CT images[22,23], probe localization in liver CT images[17]. The network design of these works all consist of a feature extractor followed by several fully connected layers. The feature extractors consist of several CNN blocks (a CNN layer, a normalization layer and an activation layer) with residual connections[20]. Although these works are all developed for non-SSc patients, the success of regression networks on the aforementioned works shows its potential on level selection of SSc patients.

Automated scoring

A large number of deep neural networks has been proposed for scoring tasks in general medical imaging, which were not focused on scoring SSc-ILD. An indirect method is to develop a segmentation network and score images based on lesion segmentation results, such as idiopathic pulmonary fibrosis[24]. The limitation of such an indirect scoring method is that it requires pixel-wise segmentation labels. A direct method is to develop a network to output the score directly without any segmentation. If the scoring output contains less than 5 categories, researchers usually treat it as a classification task, such as Gleason scoring of prostate cancer in histopathology images[25–27], grading of ulcerative colitis in endoscopic images[28], grading of abnormalities in knee MRI[29], diabetic retinopathy grading in eye fundus images[30], osteoarthritis severity grading in knee MRI[31], fibrosis estimation[32]. When the scoring is a real (floating point) number or contains more than 5 categories, regression neural networks are preferred, e.g. Agatston scoring in chest CT images[33], ventricle function indices estimation in ultrasound images[34], coronary calcium scoring in chest CT scans[35,36], bone mineral density (BMD) and the estimation of percentage of lung emphysema from CT scans[37]. Because we aimed to estimate the ratio of fibrosis to the total lung area without segmentation, a regression neural network was adopted in our work.

Automated scoring for SSc-ILD

To the best of our knowledge, there are no automated level selection methods published for SSc scoring. In addition, there are only two automated scoring frameworks developed for SSc patients (see Table 1). Chassagnon et al[3]. developed networks, which could output the fibrosis areas and severity quantification for SSc patients. However, their work used segmentation networks to output the pixel labels as a basis for computing the final biomarkers, which is time-consuming and requires laborious manual pixel-wise annotations for training. Since pixel-wise annotations for GG and RET are even more difficult to obtain due to unclear boundaries between the two patterns, only TOT patterns have been segmented to assess SSc-ILD. In contrast to only segmenting TOT pattern, Su et al[38]. developed a severity scoring model for connective tissue disease associated ILD (CTD-ILD, including SSc-ILD) that could segment GG, RET and honeycombing patterns, separately. This also requires

Patients	Task	Network	Target	Dataset
Non-SSc	Localization	Regression	Lung upper and lower edge[18]	Chest CT
			Lumbar vertebra[20]	Spine CT
			Anatomical structures[23]	Body CT
			Anatomical plane landmarks[21]	Cardiac ultrasound
			Probe localization[17]	Liver CT
	Scoring	Segmentation	Idiopathic pulmonary fibrosis[24]	Chest CT
		Classification	Grading of ulcerative colitis[28]	Endoscopic
			Grading of abnormalitie[29]	Knee MRI
			Diabetic retinopathy grading[30]	Eye fundus
			Osteoarthritis severity grading[31]	Knee MRI
			FIbrosis estiation[32]	Chest CT
		Regression	Ventricle function indices[34]	Cardiac ultrasound
			Coronary calcium scoring[35,36]	Chest CT
			Percentage emphysema[37]	Chest CT
			Agatston scoring[33]	Chest CT
SSc included	Scoring	Segmentation	SSc-ILD assessing[3]	Chest CT
			CTD-ILD assessing[38]	Chest CT

Table 1. Summary of related works for automated scoring frameworks on medical imaging.

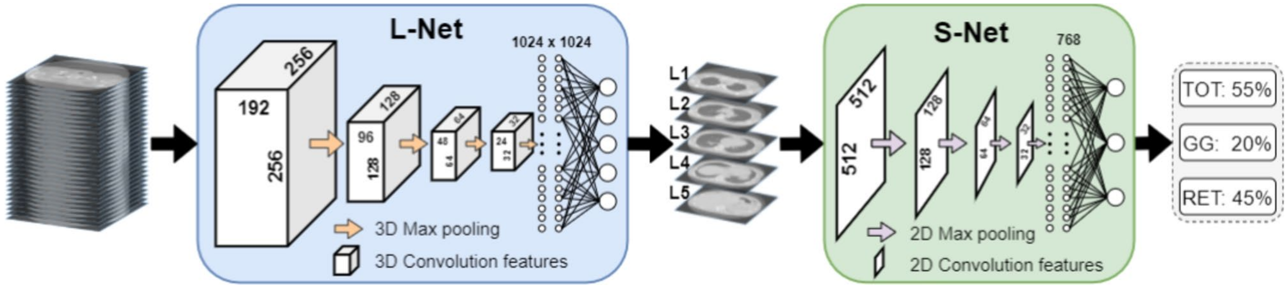


Fig. 2. Proposed framework for SSc-ILD scoring based on two cascaded neural networks. L-Net outputs five values of anatomical levels. S-Net outputs three values for automatic SSc-ILD scoring. TOT = total disease extent, GG = ground glass pattern, RET = reticular pattern.

Characteristic	Patients, (n=230)	
Age [years] (standard deviation)	54	(14.0)
Female (%)	191	(83.0)
Interstitial lung disease detected on CT (%)	96	(41.7)
Anti-centromere antibodies positive (%)	88	(38.3)
Anti-topoisomerase antibodies positive (%)	56	(24.3)
Pulmonary arterial hypertension (%)	8	(3.5)
Disease Subset:		
Non-cutaneous (%)	29	(12.6)
Diffuse cutaneous (%)	62	(27.0)
Limited cutaneous (%)	133	(57.8)
Alternative diagnosis* (%)	6	(2.6)

Table 2. Dataset properties of patients suspected of systemic sclerosis. *Morphea scleroderma, undifferentiated connective tissue disease (UCTD), UCTD with Sjögren's syndrome, puffy fingers without systemic disease, and two cases of very early diagnosis of systemic sclerosis (VEDOSS).

laborious pixel-wise annotations. As far as we know, there are no published methods on automated scoring of SSc-ILD without the need for pixel-wise fibrosis annotations.

Materials and methods

The proposed two-step framework is shown in Figure 2. A level selection network (L-Net) selects the five anatomical levels from the input 3D CT scans. Subsequently, five 2D slices were selected according to the five level positions and an SSc-ILD scoring network (S-Net) estimates three scores (TOT, GG and RET) for each input 2D slice.

Dataset

The dataset was collected retrospectively and consisted of de-identified high-resolution CT scans of 230 patients suspected of SSc, who were referred to our multidisciplinary healthcare program[39] for suspected SSc (Table 2). The CT scans were performed with an Aquilion 64 scanner (Canon Medical Systems), with 120 kVp, median tube current 140 mA, median CTDIvol 8.2 mGy, rotation time 0.4 seconds, collimation 64×0.5mm and 0.8 helical beam pitch[40]. Images were reconstructed with filtered back projection and using an FC86 kernel, 0.64×0.64 mm median pixel spacing, and a slice thickness and increment of 0.5 and 0.3 mm, respectively. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board of the LUMC under Protocol Nos. P09.003/SH/sh, REU 036/SH/sh, REU 043/SH/sh, and B19.008/KB/kb. All patients have given informed consent for the use of data in this prospective Combined Care in Systemic Sclerosis (CCISS) cohort study. All methods have been performed in accordance with the relevant guidelines and regulations. CT data were accessed for this retrospective study on 18-03-2021. The data was anonymized previously, therefore there was no access to data that could identify individual participants during or after data collection. The dataset was randomly split into two disjoint groups with 180 and 50 patients for training and testing, respectively. The 180 training patients were randomly divided into four groups of equal size for four-fold cross-validation (45 patients for validation in each fold).

Before training the L-Net, all CT scans were resized to a fixed size of $256 \times 256 \times 256$-pixel matrix. After resizing, the pixel spacing along the *x* and *y*axis was 1.26 mm. The average spacing along the z-axis was 1.2 mm. CT values were truncated between -1500 HU and 1500 HU. The world positions of five levels for each CT scan were manually selected by a rheumatologist (Observer-A, 5-year experience) and a radiologist (Observer-B, 20-

year experience) in consensus. Subsequently, we converted the world positions of the five levels to relative slice numbers in the resampled 3D CT scans (the bottom slice was regarded as number 0, corresponding to the base of the lung)[41]. These slice numbers were used as the ground truth. To increase the image diversity for training the L-Net, we randomly cropped patches with a fixed size of $256 \times 256 \times 192$ (ordered by *xyz*) on-the-fly during training. These patches always covered all five levels and could also be fitted into the GPU memory of 11 GB.

While L-Net was trained and tested using the down-sampled CT volumes, S-Net used the 2D axial slices of five levels with the original in-plane resolution (512×512) from the aforementioned 230 scans. High-resolution images include fine spatial details, which can help to distinguish and grade the three fibrosis patterns. All 2D slices were scored in consensus by two experts to obtain the ground truth. Additionally, to evaluate inter/intra-observer agreement 16 patients (80 axial slices) were randomly selected from the testing dataset and the two experts scored them independently. Then they independently scored the same 80 axial slices again after six weeks. The TOT, GG and RET scores can range from 0% to 100%, and were estimated with a precision of five percent (Appendix Figure A1), following the protocol by Goh et al[5]. To augment our dataset, two neighboring slices (above and below the chosen slice of each level) were added for training. Because the spacing of neighboring slices is only 0.3 mm, we assumed that these represent the same score. In addition, the 2D training images were augmented on the fly by random rotation $(\pm 30^{\circ})$, scale (95% - 105%) and shift (±10 pixels) along the horizontal and vertical direction.

After we completed the separate training of L-Net and S-Net, we cascaded them together to build a whole framework for inference. The relative slice number prediction from L-Net was converted to absolute physical height (mm) in the original CT. We then selected the slice which has the closest distance to the converted physical height. This slice (512×512) was the input to the S-Net to obtain the final score estimation.

Network design

Inspired by[10], we experimented with different 3D VGG[42] networks as the structure L-Net (Figure 3-A), including VGG11, VGG16 and VGG19. Deeper networks like 3D ResNet50[43]would lead to GPU memory problems with the same input patch size, so no deeper networks were tested further. Therefore, VGG11 was finally selected for L-Net. As for the S-Net, we evaluated different 2D networks with different capacities including SqueezeNet[44], VGG11,16, and 19[42], ResNet18[43], ResNet50, ResNeXt50[45], DenseNet[46], ShuffleNet[47], ConvNeXt[48], and finally decided to adopt ConvNeXt for S-Net due to its state of the art performance (Figure 3-B). Compared with the original VGG11 proposed in[42], L-Net extends all convolutional and max-pooling layers from 2D to 3D. The feature extractor (convolutional layers) of the S-Net was initialized by pre-trained weights from ImageNet[48], while the fully connected layers were initialized randomly using a normal distribution. The architecture and training details of L-Net and S-Net are shown in Table 3.

Techniques to overcome label imbalance

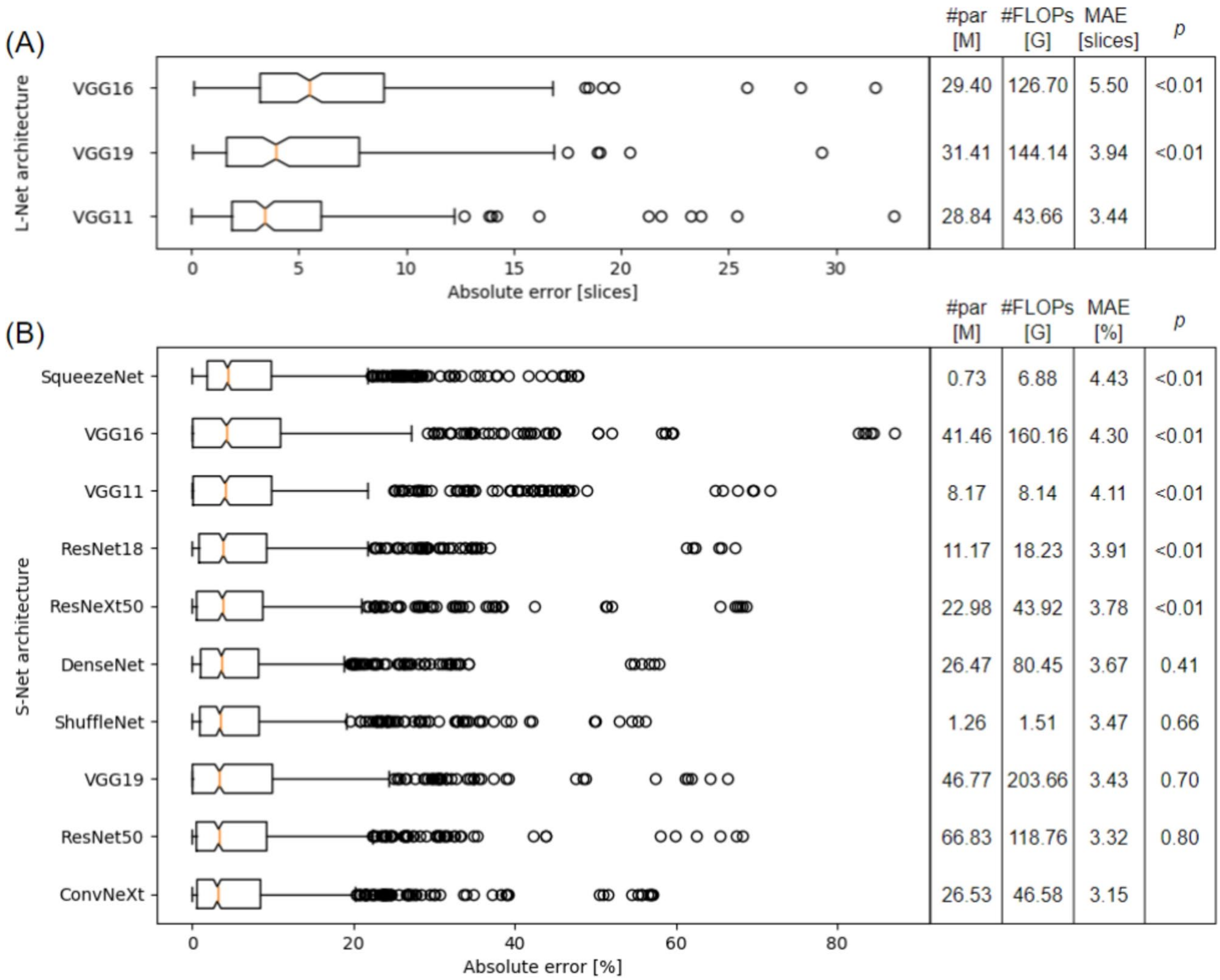
From Appendix Figure A1, we could find that the score distribution is highly askew-some high scores even do not exist in the training dataset. The noticeable score imbalance with 50% of TOT scores being 0 would limit the networks' performance. Therefore, to ensure a balanced score distribution during training, we developed a sampling method that randomly selects training images with a probability inversely proportional to the ratio of each TOT score[49]. In this way, the scores that appear less frequently (i.e. higher scores) would be used for training more frequently. To further address the label imbalance and to increase data diversity, we synthesized training images with SSc-ILD scores that are lacking in the original dataset, by digitally inserting GG and RET patterns into healthy training images.

The flowchart of data synthesis is shown in Figure 4-A. First, we created two patches full of different patterns, one for GG and one for RET, by manually extracting local CT patches from training images with high scores in these two patterns separately. Subsequently, the healthy training images (TOT=0) were augmented by the digital insertion of these patterns. The candidate lesion regions for the pattern insertion were randomly generated by defining up to three ellipses with random centers, orientations and axes lengths (lengths range from 1 to 150 pixels), which were then cropped by the lung mask to ensure the patterns will be inserted in the lung area only. The lung mask was automatically generated by a multi-atlas based automatic lung segmentation algorithm[50]. To avoid introducing unrealistic borders during pattern insertion, the inserted patterns gradually fade out at the edge, by a Gaussian decay in intensity. To increase the pattern diversity of synthetic data, we always applied random rotation $(\pm 180^{\circ})$ and scale (95%-105%) to the patterns before each pattern insertion. The disease severity scores were then adapted according to the area of inserted patterns. Some synthetic image examples and their scores were shown in Figure 4-B. The synthetic data constitute half of the training dataset, while validation and testing were performed on real patient data only. Although some generative models like GANs[51]or diffusion models[52] may output more realistic images, however, we did not introduce them in this work because these generative models cannot provide precise scores for the generated images. To overcome the label imbalance in our small dataset, we need not only synthetic images but also their corresponding ground truth scores.

Heat map visualization: network explainability

Application in clinical practice is limited, if the output of a deep learning network is difficult to explain. Therefore, inspired by the occlusion-based visualization method[53], we developed a replacement-based method to generate heat maps, indicating which areas were important for the S-Net in recognizing different disease patterns. The method details are as follows.

A rectangular patch full of healthy lung tissue, in advance cropped from a lung region of healthy slices, was used to cover the test image from top-left to bottom-right step-by-step, separately. The patch size was 64×64 pixels with a step size of 16. The output score from the S-Net of the original test image was regarded baseline. Each time we slide the healthy patch, the original image was occulted by the healthy patch at a different position.



L-Net architecture	#par [M]	#FLOPs [G]	MAE [slices]	p
VGG16	29.40	126.70	5.50	<0.01
VGG19	31.41	144.14	3.94	<0.01
VGG11	28.84	43.66	3.44	

S-Net architecture	#par [M]	#FLOPs [G]	MAE [%]	p
SqueezeNet	0.73	6.88	4.43	<0.01
VGG16	41.46	160.16	4.30	<0.01
VGG11	8.17	8.14	4.11	<0.01
ResNet18	11.17	18.23	3.91	<0.01
ResNeXt50	22.98	43.92	3.78	<0.01
DenseNet	26.47	80.45	3.67	0.41
ShuffleNet	1.26	1.51	3.47	0.66
VGG19	46.77	203.66	3.43	0.70
ResNet50	66.83	118.76	3.32	0.80
ConvNeXt	26.53	46.58	3.15	

Fig. 3. Network architecture selection for L-Net (**A**) and S-Net (**B**). #par: number of trainable parameters; #FLOPs: number of floating point operations; MAE: mean absolute error. VGG11 and ConvNeXt achieve the lowest MAE for L-Net and S-Net, respectively. p-values in (**A**) were obtained by the Wilcoxon signed rank test between each network and VGG11; p-values in (**B**) were obtained by the Wilcoxon signed rank test between each network and ConvNeXt.

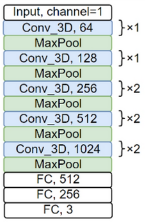
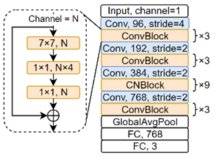
Net	L-Net (VGG-11)	S-Net (ConvNeXt)
Structure		
Dimension	3D	2D
Batch size	4	10
Training time (h)	2.8	4.9
Parameters (M)	28.84	22.53
Training epochs	500	500
Learning rate	0.0001	0.0001

Table 3. Details of network design and training scheme for L-Net and S-Net.

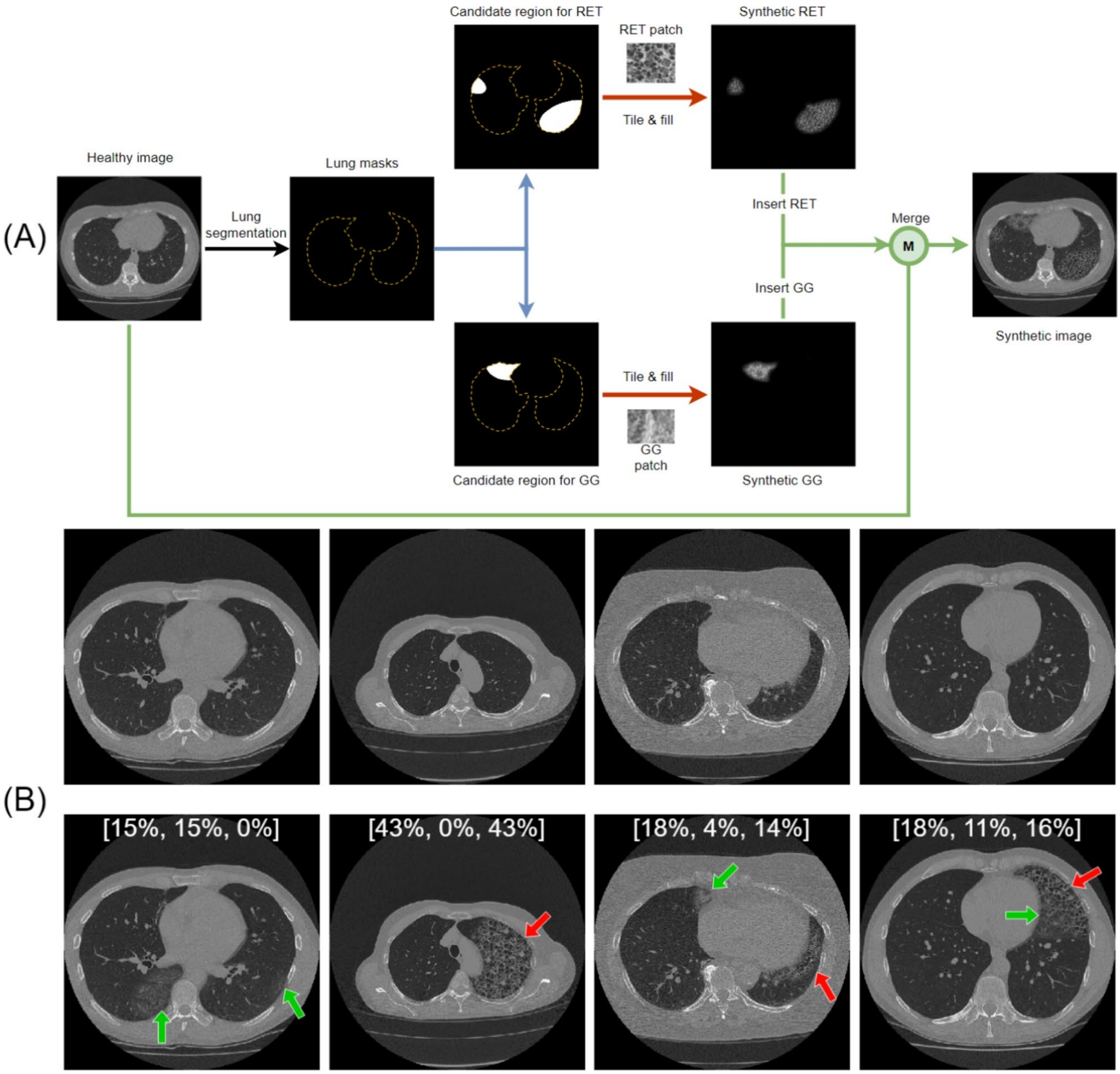


Fig. 4. Data synthesis flowchart and examples. (**A**) Flowchart to synthesize images with different disease patterns. Blue arrows indicate the generation of random candidate lesion regions; blue arrows indicate the generation of candidate regions, red arrows indicate the filling of patterns; green arrows indicate the insertion of patterns. (**B**) Four pairs of synthetic examples. The upper row shows the original images; the lower row shows the corresponding synthetic images. Green arrows point to GG; red arrows point to RET. Different pattern combinations are shown from left to right: only GG, only RET, GG and RET without overlap, GG and RET with overlap. The scores of these synthetic images are shown on the image in the order of [TOT, GG, RET]. TOT = total disease extent, GG = ground glass pattern, RET = reticular pattern.

We could obtain the new score from S-Net and record its change at each position. A heat map of the image is then generated using the magnitude of the score change (ΔP). A negative score change ($\Delta P < 0$) implies that the network regarded the original patch as diseased since the score decreased after concealing the area with healthy tissue. If the output score remains unchanged ($\Delta P = 0$), the original patch was already considered healthy. A score increase ($\Delta P > 0$) means that the network produced a false positive, since the inserted healthy patch was apparently classified as diseased. To make sure that the replaced pixels are in the lung area, the rectangular healthy patch was cropped by the lung mask before each replacement. The patch edge fades gradually by linearly increasing transparency to make it more natural.

This replacement-based heat map was inspired by the occlusion-based visualization method[53]. The difference is that the occlusion-based method would cover the original image using a patch with a constant value, which would introduce artifacts, while our replacement-based method covers the original image using a patch cropped from a healthy CT scan, which still includes lung texture and makes the generated image more natural.

In order to evaluate the performance of the heat maps, blinded to the network's output, two human experts independently rated their agreement with the heat maps using a Likert scale, with five labels (1-5): "Strongly disagree", "Disagree", "Neutral", "Agree" and "Strongly agree", using dedicated software (Appendix Figure A2).

Additionally, we developed an automatic method to evaluate the heat map's explainability. By thresholding the heat map, the different patterns were segmented and their areas were divided by the total lung area, to obtain a derived SSc-ILD score. Subsequently, we tested the network's consistency by the correlation between the derived SSc-ILD score and the network's output. The optimal threshold was obtained from the validation dataset by varying the threshold from -4% to 0% and selecting the one with the smallest mean absolute error (MAE) between the derived SSc-ILD score and S-Net output.

Statistical analysis and evaluation

To evaluate our networks, the following statistical analyses were performed by an in-house Python 3.8 script with corresponding libraries.

The MAE and standard deviation (STD) are reported. MAE were calculated as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i|, \tag{1}$$

where $i \in N$ is the index of samples, N represents the total number of samples, $\hat{Y}_i$ is the network's estimated value, and Y_i the measured PFTs value. To evaluate the inter-observer agreement, Cohen's linearly weighted kappa (κ)[54]and intra-class correlation coefficient (ICC)[55] were used. Weighted κ, a measure to assess the agreement level between two raters when evaluating ordering categorical items, was calculated by scikit-learn 0.24.2[56] based the equation:

$$\kappa = \frac{P_o - P_c}{1 - P_c} \tag{2}$$

where P_o is the proportion of observed agreements and $P_c$is the proportion of agreements expected by chance. ICC, a measure reflecting both degree of correlation and agreement between measurements, was calculated by pingouin 0.4.0[57]based on a single-rating, absolute-agreement, 2-way mixed-effects model[55]. To statistically test differences between groups, a paired T-test and Wilcoxon signed rank test were performed, as implemented by scikit-learn 0.24.2. A P value of less than 0.05 was considered to indicate a statistically significant difference. All metrics were calculated based on the testing dataset unless stated otherwise.

Network implementation details

Our neural networks, L-Net and S-Net, were implemented using PyTorch 1.7.1 (https://pytorch.org). For both networks, the loss function was the mean squared error (MSE). The Adam optimizer was used with a learning rate of 1e-4, a weight decay of 1e-4 and 500 epochs. Multithreading was used to accelerate the on-the-fly data augmentation. The workstation was equipped with an Intel(R) Xeon(R) CPU Gold 6126 @ 2.6GHz with 90 GB memory and a GPU RTX 2080TI with 11 GB memory. The source code and trained models are published at https://github.com/Jingnan-Jia/ssc to facilitate reproduction of results.

Experiments and results

SSc-ILD scoring performance

First, we trained and evaluated the L-Net (Figure 5) and S-Net (Table 4), separately. Figure 5 shows that the ICC of the five consecutive levels was 0.72, 0.84, 0.81, 0.96 and 0.97. No significant bias was observed among the five levels (P=0.20, 0.93, 0.42, 0.49, and 0.76, respectively. Subsequently, an end-to-end framework was built as a cascade of the trained L-Net and S-Net (called L&S-Net), in which the input slices for S-Net were automatically selected by the L-Net (Table 5). For none of the levels, the automatic scoring results of L&S-Net showed any significant differences as compared to solely S-Net which received the manually selected slices (Table 5).

Comparison with human experts

The inter- and intra-observer agreement in the sub-group of 16 patients (80 axial slices) from the testing dataset were compared with our proposed method (Table 6). The inter-observer agreement was higher during the second scoring session. The intra-observer agreement of Observer-B was higher than Observer-A, and the inter- and intra-observer agreement in GG scoring was always lower than in TOT and RET scoring.

For scoring TOT, our automatic method was close to the first rating by Observer-A ($\mathrm{Obs\text{-}A_{T1}}$), but Observer-B was closer to the consensus than our method. For GG the model had a fair agreement with human consensus, while the observers agreed moderately, and for RET the model's agreement was moderate, but moderate/substantial for observers. Except for the second GG scoring by Observer-B ($\mathrm{Obs\text{-}B_{T2}}$, P < 0.05), other human observations did not perform significantly better than our method. The Bland-Altman plots (Appendix Figure A3) illustrate the performance of an individual human score ($\mathrm{Obs\text{-}B_{T2}}$) and our automatic network.

The average time of fully automated scoring for the five levels is less than ten seconds per patient, while human experts need around ten minutes (around 2.5 minutes to identify five levels and another 7.5 minutes to

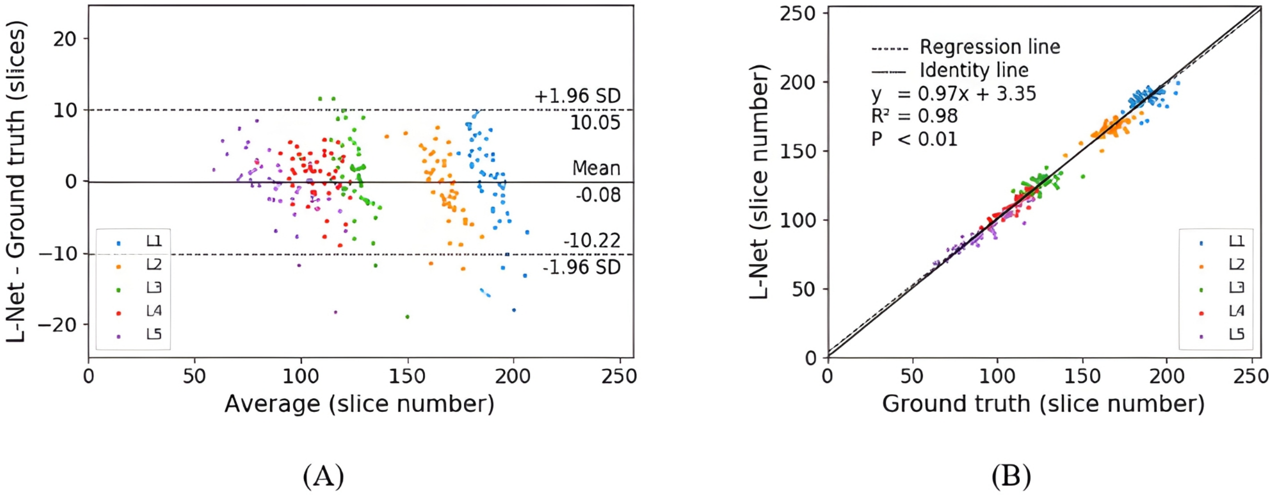


Fig. 5. Testing results of the L-Net in selecting slices on the five levels, L1-L5. (**A**) Bland-Altman plot and (**B**) Correlation plot. The average spacing between slices was 1.2 mm.

score three patterns of five levels). The comparison to related work and corresponding discussion is presented in Appendix-2.

Heat map explanation and its evaluation

The replacement-based heat maps of the automatic scoring for the three different patterns are shown in Figure 6 of different patients from the testing subset. The proposed visualization method can show areas of different patterns and display the severity with different colours. The yellow and red areas in the heatmaps (Figure 6) denote the negative score change after the area has been covered by a healthy patch, which means that the original patch is "diseased". A red area means more severe and more obvious patterns than a yellow area. The green and blue areas mean that the network produced a false positive, since the inserted healthy patch was apparently classified as diseased. The heat maps can also help to find the cause of errors, as shown in the last row in Figure 6, where the GG scoring result (30%) is far lower than the ground truth (90%). From the heat maps, we can see only about 30% of the whole lung was activated (yellow and red area) and the GG pattern was missed in around 1/3 of the right lung (blue area). Alternative heat maps are presented in Appendix Figure A4 to indicate false negatives and positives.

From the semi-quantitative evaluation of the heat maps, Observer-A rated the heat maps with "Strongly Agree" or "Agree" in 97.0%, 94.2% and 89.8% of cases for TOT, GG and RET, respectively (Figure 7, upper row). Ratings of "Strongly agree" or "Agree" by Observer-B occurred in 84.0%, 85.8% and 70.2% of cases for TOT, GG and RET, respectively. Thus, on average they agreed in 90.5%, 90.0% and 80.0% with the heat maps, respectively.

After applying an optimized threshold value (Appendix Figure A5) to the heat maps on the testing dataset, a significant linear correlation was found between the heat map-derived SSc-ILD score and the L&S-Net's output (Figure 7, lower row). For TOT, GG and RET, 84%, 87% and 83% of the S-Net's output variation can be explained by the heat maps, respectively.

Discussion

In this study, we developed a deep learning framework to perform fully automated SSc-ILD scoring in chest CT scans. By cascading two separate networks, the framework was able to select the five anatomical levels from 3D CT scans and then quantify the extent of three different disease patterns for each level. The training of the framework only needs visual scores as the ground truth without the requirement of prior manual segmentations. Heat maps can intuitively explain the network's output, and can be used to derive coarse segmentations of the different patterns that are consistent with the network's output. Our framework has the potential to serve as an alternative to visual SSc-ILD scoring of lung involvement in systemic sclerosis.

Explanation and discussion on results

Our framework consists of two networks, trained independently: L-Net for automatic level selection and S-Net for automatic ILD scoring. For the L-Net, the selection of the first level is more difficult (ICC = 0.72) than other levels because indicating the origin of the great vessels is variable as it was not defined precisely. Nevertheless, the automatically selected levels were accurate enough, because the ultimate scoring did not show significant differences compared with the single S-Net's performance with manually annotated slices. This may be due to the fact that disease patterns appear and disappear only gradually from one slice to the other.

In our paper we demonstrated that different network structures with different capacities did not necessarily show a significant difference (see Figure 3). That implies that network design is not the bottleneck for our task. Our further investigation showed that the key issue, hindering the network performance, is the low quality of

Experiments				TOT			GG			RET		
(i)	DS*	BS	PT	MAE [%]↓	κ↑	ICC↑	MAE [%]↓	κ↑	ICC↑	MAE [%]↓	κ↑	ICC↑
	-	-	-	8.87 (12.55)	0.42	0.59	6.62 (11.64)	0.31	0.38	6.59 (9.30)	0.42	0.58
	-	-	✓	7.85 (10.71)	0.53	0.72	5.96 (10.03)	0.45	0.60	5.81 (8.42)	0.53	0.72
(ii)	-	✓	-	8.30 (11.35)	0.46	0.62	5.89 (10.88)	0.40	0.48	5.96 (8.70)	0.48	0.65
	✓	-	-	6.74 (9.59)	0.59	0.77	5.07 (9.38)	0.54	0.67	5.08 (7.76)	0.57	0.75
	-	✓	✓	6.87 (9.61)	0.59	0.77	5.09 (9.50)	0.54	0.66	5.11 (7.45)	0.59	0.78
(iii)	✓	✓	-	7.98 (11.31)	0.48	0.64	5.43 (9.82)	0.47	0.60	5.52 (8.56)	0.52	0.69
	✓	-	✓	6.26 (8.67)	0.63	0.82	4.76 (8.81)	0.58	0.70	4.76 (6.70)	0.62	0.82
(iv)	✓	✓	✓	**5.90 (8.77)**	**0.66**	**0.83**	**4.66 (8.83)**	**0.58**	**0.71**	**4.49 (6.70)**	**0.65**	**0.84**

Table 4. SSc-ILD scoring performance from S-Net with different technique combinations. *DS = Data synthesis, BS = Balanced sampling, PT = Pre-training, TOT = total disease extent, GG = ground glass pattern, RET = reticular pattern, Mean absolute error (MAE) is followed by the standard deviation (STD) between parentheses,κ= Cohen's weighted kappa, ICC = Intra-class correlation coefficient.↓= lower is better,↑= higher is better. **Bold** numbers indicate the best performance. (i). Baseline established by training the S-Net from scratch without balanced sampling or synthesized data. (ii). Introducing either pre-trained weights from ImageNet, balanced sampling or data synthesis. (iii). Combination of two of the three techniques. (iv). Combination of all three techniques which obtained the best performance. Therefore, the proposed method contains all three techniques.

	MAE of TOT*			MAE of GG			MAE of RET		
Level	LS-Net [%]	S-Net [%]	P	LS-Net [%]	S-Net [%]	P	LS-Net [%]	S-Net [%]	P
1	4.20 ± 7.94	6.28 ± 11.57	0.33	2.34 ± 6.88	2.81 ± 6.71	0.57	3.49 ± 6.25	5.14 ± 9.67	0.42
2	4.61 ± 6.89	4.74 ± 7.46	0.06	3.11 ± 7.67	3.33 ± 7.62	0.88	3.35 ± 5.67	3.72 ± 6.15	0.49
3	5.61 ± 8.00	6.17 ± 8.63	0.64	4.65 ± 8.42	4.82 ± 8.19	0.56	4.33 ± 5.24	4.69 ± 6.13	0.74
4	7.12 ± 8.87	6.88 ± 9.23	0.17	5.88 ± 9.58	5.98 ± 9.59	0.10	4.99 ± 6.75	4.93 ± 6.65	0.90
5	8.12 ± 10.00	7.58 ± 9.99	0.11	7.13 ± 10.49	6.75 ± 10.41	0.40	6.15 ± 8.26	5.74 ± 7.78	0.11
ALL	5.86 ± 8.46	5.90 ± 8.77	0.21	4.56 ± 8.80	4.66 ± 8.83	0.13	4.40 ± 6.55	4.49 ± 6.70	0.28

Table 5. MAE comparison of SSc-ILD scoring between the whole framework (L&S-Net) and sole S-Net. * Mean absolute error (MAE) ± standard deviation (STD) is presented expressed as %, TOT = total disease extent, GG = ground glass pattern, RET = reticular pattern, ALL = Calculated based on the results from all the five levels.

		TOT*			GG			RET		
Agreement	Comparison	MAE [%]↓	κ↑	ICC↑	MAE [%]↓	κ↑	ICC↑	MAE [%]↓	κ↑	ICC↑
Inter-observer	$\mathrm{Obs\text{-}A_{T1}}vs\mathrm{Obs\text{-}B_{T1}}$	5.25	0.59	0.76	4.00	0.54	0.66	3.81	0.61	0.82
	$\mathrm{Obs\text{-}A_{T2}}vs\mathrm{Obs\text{-}B_{T2}}$	4.25	0.67	0.88	3.94	0.58	0.80	3.44	0.63	0.86
Intra-observer	$\mathrm{Obs\text{-}A_{T1}}vs\mathrm{Obs\text{-}A_{T2}}$	4.38	0.63	0.83	3.50	0.56	0.73	3.06	0.67	0.84
	$\mathrm{Obs\text{-}B_{T1}}vs\mathrm{Obs\text{-}B_{T2}}$	3.50	0.74	0.89	3.69	0.62	0.74	2.69	0.72	0.90
vs GT	$\mathrm{Obs\text{-}A_{T1}}$vs GT	7.06 (0.41)	0.51	0.73	5.63 (0.51)	0.44	0.68	4.94 (0.37)	0.56	0.76
	$\mathrm{Obs\text{-}A_{T2}}$vs GT	6.19 (0.14)	0.58	0.82	5.38 (0.78)	0.46	0.59	4.75 (0.13)	0.58	0.78
	$\mathrm{Obs\text{-}B_{T1}}$vs GT	6.56 (0.42)	0.58	0.80	5.38 (0.32)	0.48	0.63	4.63 (0.18)	0.61	0.84
	$\mathrm{Obs\text{-}B_{T2}}$vs GT	4.94 (0.40)	0.67	0.86	4.94 (0.001)†	0.55	0.75	4.19 (0.18)	0.63	0.80
	L&S-Net vs GT	6.40	0.54	0.79	6.13	0.39	0.55	4.44	0.61	0.84

Table 6. SSc-ILD scoring performance of human experts (Observer-A and Observer-B) in two scoring sessions (T1 and T2, with 6 weeks interval) and our proposed network in a subset of 16 patients from the testing dataset. *TOT = total disease extent, GG = ground glass pattern, RET = reticular pattern, GT = Ground truth (human consensus score), MAE = mean absolute error,κ= Cohen's weighted kappa, ICC = Intra-class correlation coefficient. Data between parentheses are P values from the Wilcoxon signed rank test comparing observers with our proposed method. $\mathrm{Obs\text{-}A_{T1/T2}}$ and $\mathrm{Obs\text{-}B_{T1/T2}}$ denote the observations from Observer-A and Observer-B at the first or second scoring session, respectively. † Significantly better than our method (P<0.05).

dataset. Therefore, we improved the training method by introducing synthetic training images that significantly improved the network's performance.

For the ILD scoring network, the pre-trained weights, balanced sampling and the proposed data synthesis all helped to steadily improve the network's performance for all three patterns (see Table 4). Comparing experiment (ii) and (iii), we can find that adding "Balance sampling" lead to a decrease in the effectiveness of "Data synthesis" for S-Net without pre-training. This is because "Data synthesis" already alleviated the data imbalance to a great extent. Then "Balance sampling" only introduces a lot of repeated samples for training. Compared to the combination, single "Data synthesis" will not introduce any repeated samples for training. Therefore, the combination of "Balanced sampling" and "Data synthesis" perform worse than single "Data synthesis" for the S-Net trained from scratch. Our random ILD insertion method is very effective and easy to implement, which only requires two small patches fully covered by GG and RET. Generally, ILD in SSc has a specific distribution, e.g. classical subpleural sparing earlier in the disease[58], which was not simulated by our synthesis method. The scoring results were however not affected by this limitation, since the neural network only needs to estimate the ratio of ILD, irrespective of the location of ILD. Nevertheless, there is still some space for improvement in GG scoring. Also for human experts, GG is more difficult to define and identify than RET patterns, because of the limited spatial resolution of CT and consequential 'partial volume' effect. Moreover, some GG patterns resemble noise from image acquisition or reconstruction. Conversely, reticular lesions are larger than voxel size and can be visually or automatically identified as structures, such as thickened interlobular septa or thickened airways causing pathological reticular patterns. Our proposed network may help in distinguishing noise from actual pathological ground glass lesions when noise patterns can be identified.

With the help of the replacement-based heat map, we visualized which areas contribute to which scores respectively. Two experts evaluated the heat maps independently and both gave very satisfactory ratings. After we applied a threshold to the heat map, the ratio of different patterns to the total lung area was consistent with the automatic ILD scores by L&S-Net. The quantitative measurement shows that our proposed heat maps can

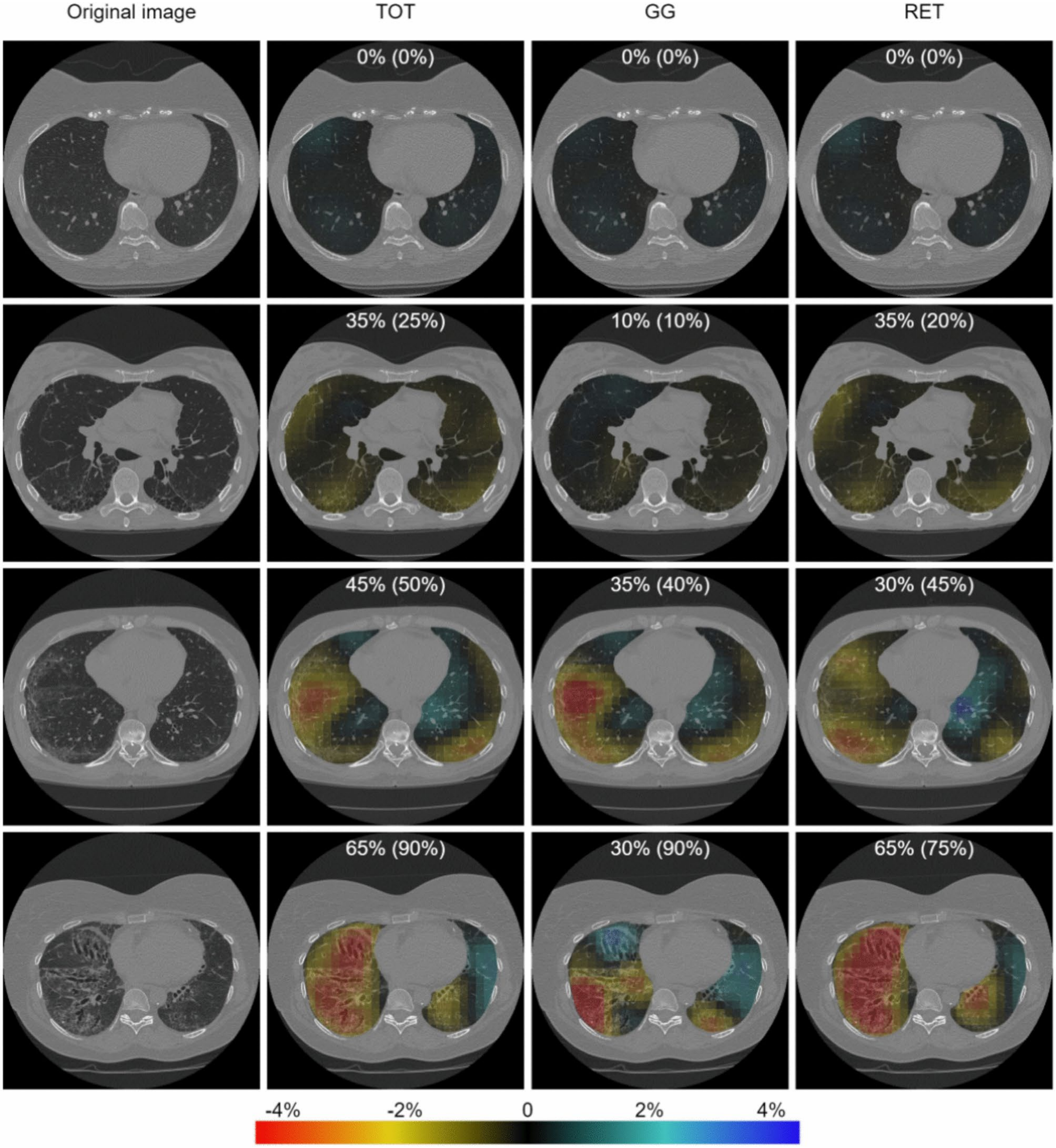


Fig. 6. Heat map visualization for various test images. The percentages on the scale now indicate how much the output score (in percent points) changes after replacing an area with a healthy patch. Each row represents one axial slice from a different patient. The first column is the original image and the subsequent three columns show the heat maps of the three disease patterns. Different colours represent the magnitude of score change. From top to bottom, the images show increasing disease severity. The automatic ILD score is shown on the top of each image followed by the ground truth (human consensus) between parentheses. TOT = total disease extent, GG = ground glass pattern, RET = reticular pattern.

accurately explain the L&S-Net's output. This can increase the clinicians' confidence in the network's output. A heat map "highlights" the detected pathology that may help the physician with a quick image interpretation. Exploring the heat map can also be used to check the quality of the automatic score. The heat map could be regarded as a coarse segmentation of TOT, GG and RET. Normally it is not practical to have large datasets of SSc ILD pattern segmentations because it is very time-consuming and laborious. The heat maps can act as an initial step to obtain manual segmentation reducing annotation time. From this perspective, we successfully

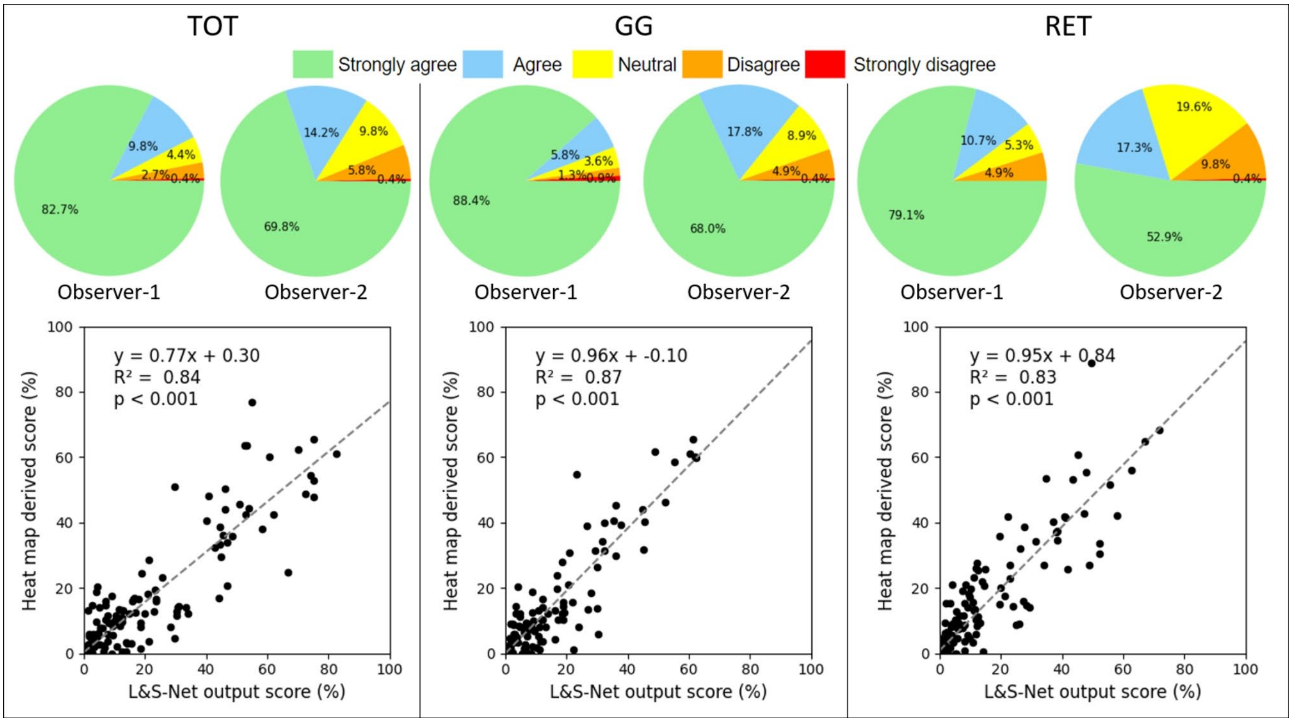


Fig. 7. Heat map performance evaluated by the two observers (pie charts in the upper row) and the association between the heat map derived ILD scoring and L&S-Net's output (scatter plots in the lower row), to indicate its explainability. TOT = total disease extent, GG = ground glass pattern, RET = reticular pattern.

obtained a coarse ILD pattern segmentation network for patients suspected of SSc without the requirement of a segmentation ground truth. Compared with the normal binary segmentation[3], the advantage of our heat map is that it also gives an indication of the severity of a disease pattern, as shown by the colour, instead of a binary classification into either healthy or disease.

We observed that human experts gave higher ratings to the GG heat maps than RET, although the MAE of automatic GG scoring was actually consistently higher than RET. This can be explained by the fact that human experts have more confidence in recognizing RET patterns, so they use more strict criteria for RET heat maps. Since they were less confident in their GG recognition, reflected by the lower inter-/intra-observer agreement, this resulted in more tolerance for GG heat maps.

In the visual ILD scoring system, the use of only five anatomical levels has been a compromise, for clinical practice. It is already very time-consuming and laborious to manually select the five levels and score three patterns for each level (taking around ten minutes). Our method, however, could automatically complete the level selection and ILD scoring in several seconds. In addition, our method can be easily extended to score all slices of the entire CT volume, which is practically not feasible for humans.

Limitations

Our method has some limitations. The L-Net was initialized with a random distribution instead of pre-trained weights. This may be improved if pre-trained weights from a large 3D medical image dataset are available. The quality of data synthesis could clearly be improved further. The current pattern insertion method may distort the structure of airways and vessels and introduce some periodic artefacts. In future research, how to generate more realistic synthetic patterns with accurate labels can be a research direction to explore. The data used in this study is from a single model CT scanner within a single healthcare programme with tightly-controlled acquisition and reconstruction parameters. Because of the lack of publicly available independent testing dataset, whether this method could be used across a range of CT scanners, sites and protocols still needs to be verified. The ILD scores of our synthetic training images were obtained by the ratio of different patterns, while the ILD scores of the real images were estimated by the human observer without any pattern segmentation or contours. Therefore, there may be a systematic bias between them, which could contribute to the disagreement between our framework and experts.

Conclusion

In conclusion, we proposed the first fully automated framework to estimate scores for ground glass opacities, reticular patterns and total disease extent from 3D CT scans, specific for systemic sclerosis. The output scores can be clearly explained by the replacement-based heat maps. The results show its potential as an objective alternative for visual scoring of systemic sclerosis and could be extended to other applications where a diagnosis is based on scores at different anatomical levels.

Data availability

All code used to develop and verify the deep neural networks in this study has been published at https://github.com/Jingnan-Jia/ssc_scoring. All data and materials used in the analysis can be available upon request for the purposes of reproducing or extending the analysis via the corresponding author, in accordance with local and institutional guidance and legal requirements.

Received: 6 May 2024; Accepted: 30 October 2024
Published online: 04 November 2024

References

1. Denton, C. P. & Khanna, D. Systemic sclerosis. *The Lancet* **390**, 1685–1699 (2017).
2. Wells, A. U. Interstitial lung disease in systemic sclerosis. *La Presse Médicale* **43**, e329–e343. https://doi.org/10.1016/J.LPM.2014.08.002 (2014).
3. Chassagnon, G. et al. Deep Learning-based Approach for Automated Assessment of Interstitial Lung Disease in Systemic Sclerosis on CT Images. *Radiology: Artificial Intelligence* **2**, e190006. https://doi.org/10.1148/ryai.2020190006 (2020).
4. Assayag, D., Kaduri, S., Hudson, M., Hirsch, A. & Baron, M. High Resolution Computed Tomography Scoring Systems for Evaluating Interstitial Lung Disease in Systemic Sclerosis Patients. *Rheumatology, an open access journal Assayag et al. Rheumatology* **1**, 3, https://doi.org/10.4172/2161-1149.S1-003 (2012).
5. Goh, N. S. et al. Interstitial lung disease in systemic sclerosis: a simple staging system. *American journal of respiratory and critical care medicine* **177**, 57–59. https://doi.org/10.1164/RCCM.200706-877OC (2008).
6. Desai, S. R. et al. CT features of lung disease in patients with systemic scerosis: Comparison with idiopathic pulmonary fibrosis and nonspecific interstitial pneumonia. *Radiology* **232**, 560–567. https://doi.org/10.1148/radiol.2322031223 (2004).
7. Williamson, L. New reference atlas for pulmonary fibrosis severity score in systemic sclerosis. *The Lancet Respiratory Medicine* **9**, 130–131. https://doi.org/10.1016/S2213-2600(20)30565-8 (2021).
8. Collins, C. D. et al. Observer variation in pattern type and extent of disease in fibrosing alveolitis on thin section computed tomography and chest radiography. *Clinical Radiology* **49**, 236–240. https://doi.org/10.1016/S0009-9260(05)81847-1 (1994).
9. Sverzellati, N. et al. Method for minimizing observer variation for the quantitation of high-resolution computed tomographic signs of lung disease. *Journal of computer assisted tomography* **35**, 596–601. https://doi.org/10.1097/RCT.0B013E3182277D05 (2011).
10. Belharbi, S. et al. Spotting L3 slice in CT scans using deep convolutional network and transfer learning. *Computers in Biology and Medicine* **87**, 95–103. https://doi.org/10.1016/j.compbiomed.2017.05.018 (2017).
11. Gonzalez Serrano, G., Washko, G. R. & San José Estépar, R. Deep learning for biomarker regression: application to osteoporosis and emphysema on chest CT scans. In *Proceedings of SPIE–the International Society for Optical Engineering*, vol. 10574, 52, https://doi.org/10.1117/12.2293455 (SPIE-Intl Soc Optical Eng, 2018).
12. González, G., Washko, G. R., Estépar, R. S. J., Cazorla, M. & Cano Espinosa, C. Automated Agatston score computation in non-ECG gated CT scans using deep learning. *In Proceedings of SPIE-the International Society for Optical Engineering* **10574**, 91. https://doi.org/10.1117/12.2293681 (SPIE-Intl Soc Optical Eng (2018).
13. Wang, Y. et al. A two-step convolutional neural network based computer-aided detection scheme for automatically segmenting adipose tissue volume depicting on CT images. *Computer Methods and Programs in Biomedicine* **144**, 97–104. https://doi.org/10.1016/j.cmpb.2017.03.017 (2017).
14. Dabiri, S. et al. Deep learning method for localization and segmentation of abdominal CT. *Computerized Medical Imaging and Graphics* **85**, 101776. https://doi.org/10.1016/j.compmedimag.2020.101776 (2020).
15. Bridge, C. P. *et al.* Fully-automated analysis of body composition from CT in cancer patients using convolutional neural networks. In *OR 2.0 Context-aware operating theaters, computer assisted robotic endoscopy, clinical image-based procedures, and skin image analysis*, vol. 11041 LNCS, 204–213, https://doi.org/10.1007/978-3-030-01201-4_22 (Springer, Cham, 2018). arXiv:1808.03844.
16. Shadmi, R., Mazo, V., Bregman-Amitai, O. & Elnekave, E. Fully-convolutional deep-learning based system for coronary calcium score prediction from non-contrast chest CT. *Proceedings - International Symposium on Biomedical Imaging* **2018-April**, 24–28, https://doi.org/10.1109/ISBI.2018.8363515 (2018).
17. Wei, D. et al. Slir: Synthesis, localization, inpainting, and registration for image-guided thermal ablation of liver tumors. *Medical image analysis* **65**, 101763 (2020).
18. Proskurov, V., Kurmukov, A., Pisov, M. & Belyaev, M. Fast lung localization in computed tomography by a 1d detection network. In *2021 Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT)*, 0173–0176 (IEEE, 2021).
19. Chen, H. *et al.* Automatic localization and identification of vertebrae in spine ct via a joint learning model with deep neural networks. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part I 18*, 515–522 (Springer, 2015).
20. Cheng, P., Yang, Y., Yu, H. & He, Y. Automatic vertebrae localization and segmentation in ct with a two-stage dense-u-net. *Scientific Reports* **11**, 22156 (2021).
21. Jafari, M. H. et al. Automatic biplane left ventricular ejection fraction estimation with mobile point-of-care ultrasound using multi-task learning and adversarial training. *International journal of computer assisted radiology and surgery* **14**, 1027–1037 (2019).
22. De Vos, B. D. et al. Convnet-based localization of anatomical structures in 3-d medical images. *IEEE transactions on medical imaging* **36**, 1470–1481 (2017).
23. Humpire-Mamani, G. E., Setio, A. A. A., Van Ginneken, B. & Jacobs, C. Efficient organ localization using multi-label convolutional neural networks in thorax-abdomen ct scans. *Physics in Medicine & Biology* **63**, 085003 (2018).
24. Nam, J. G. et al. Prognostic value of deep learning-based fibrosis quantification on chest ct in idiopathic pulmonary fibrosis. *European Radiology* **33**, 3144–3155 (2023).
25. Nagpal, K. et al. Development and validation of a deep learning algorithm for improving gleason scoring of prostate cancer. *NPJ digital medicine* **2**, 48 (2019).
26. Linkon, A. H. M. et al. Deep learning in prostate cancer diagnosis and gleason grading in histopathology images: An extensive study. *Informatics in Medicine Unlocked* **24**, 100582 (2021).
27. Bulten, W. et al. Automated deep-learning system for gleason grading of prostate cancer using biopsies: a diagnostic study. *The Lancet Oncology* **21**, 233–241 (2020).
28. Stidham, R. W. et al. Performance of a deep learning model vs human reviewers in grading endoscopic disease severity of patients with ulcerative colitis. *JAMA network open* **2**, e193963–e193963 (2019).
29. Astuto, B. et al. Automatic deep learning-assisted detection and grading of abnormalities in knee mri studies. *Radiology: Artificial Intelligence* **3**, e200165 (2021).
30. Araujo, T. et al. Dr| graduate: Uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images. *Medical Image Analysis* **63**, 101715 (2020).
31. Chen, P., Gao, L., Shi, X., Allen, K. & Yang, L. Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss. *Computerized Medical Imaging and Graphics* **75**, 84–92 (2019).

32. Walsh, S. L. et al. Deep Learning-based Outcome Prediction in Progressive Fibrotic Lung Disease Using High-Resolution Computed Tomography. *American journal of respiratory and critical care medicine* **206**, 883–891. https://doi.org/10.1164/RCCM.202112-2684OC (2022).
33. Cano-Espinosa, C., González, G., Washko, G. R., Cazorla, M. & Estépar, R. S. J. Automated agatston score computation in non-ecg gated ct scans using deep learning. In *Medical Imaging 2018: Image Processing*, vol. 10574, 673–678 (SPIE, 2018).
34. Luo, G. et al. Commensal correlation network between segmentation and direct area estimation for bi-ventricle quantification. *Medical image analysis* **59**, 101591 (2020).
35. De Vos, B. D. et al. Direct automatic coronary calcium scoring in cardiac and chest ct. *IEEE transactions on medical imaging* **38**, 2127–2138 (2019).
36. Mu, D. et al. Calcium scoring at coronary ct angiography using deep learning. *Radiology* **302**, 309–316 (2022).
37. González, G., Washko, G. R. & Estépar, R. S. J. Deep learning for biomarker regression: application to osteoporosis and emphysema on chest ct scans. In *Medical Imaging 2018: Image Processing*, vol. 10574, 372–378 (SPIE, 2018).
38. Su, N. et al. Computed tomography-based deep learning model for assessing the severity of patients with connective tissue disease-associated interstitial lung disease. *Journal of computer assisted tomography* **47**, 738–745 (2023).
39. Meijs, J. *et al.* Original article: Therapeutic and diagnostic outcomes of a standardised, comprehensive care pathway for patients with systemic sclerosis. *RMD Open* **2**, https://doi.org/10.1136/RMDOPEN-2015-000159 (2016).
40. Ninaber, M. K. et al. Lung structure and function relation in systemic sclerosis: Application of lung densitometry. *European Journal of Radiology* **84**, 975–979. https://doi.org/10.1016/J.EJRAD.2015.01.012 (2015).
41. Jia, J. *et al.* Prediction of lung CT scores of systemic sclerosis by cascaded regression neural networks. In *Medical Imaging 2022: Computer-Aided Diagnosis*, vol. 12033, 837—-843, https://doi.org/10.1117/12.2602737 (SPIE, 2022).
42. Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*[SPACE]https://doi.org/10.48550/arxiv.1409.1556 (2014).
43. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).
44. Iandola, F. N. *et al.* SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. arXiv preprint arXiv:1602.07360 (2016).
45. Xie, S., Girshick, R., Dollár, P., Tu, Z. & He, K. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1492–1500 (2017).
46. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely Connected Convolutional Networks. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* **2017-Janua**, 2261–2269, https://doi.org/10.48550/arxiv.1608.06993 (2016).
47. Zhang, X., Zhou, X., Lin, M. & Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6848–6856 (2018).
48. Liu, Z. *et al.* A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11966–11976, https://doi.org/10.1109/cvpr52688.2022.01167 (2022). arXiv:2201.03545.
49. Zhang, J., Liu, L., Wang, P. & Shen, C. To balance or not to balance: A simple-yet-effective approach for learning with long-tailed distributions. arXiv preprint arXiv:1912.04486 (2019).
50. Zhai, Z. et al. Pulmonary vascular morphology associated with gas exchange in systemic sclerosis without lung fibrosis. *Journal of thoracic imaging* **34**, 373–379 (2019).
51. Goodfellow, I. *et al.* Generative adversarial nets. *Advances in neural information processing systems* **27** (2014).
52. Croitoru, F.-A., Hondru, V., Ionescu, R. T. & Shah, M. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).
53. Zeiler, M. D. & Fergus, R. Visualizing and understanding convolutional networks. In *European conference on computer vision*, 818–833 (Springer, 2014).
54. Sim, J. & Wright, C. C. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical therapy* **85**, 257–268 (2005).
55. Koo, T. K. & Li, M. Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine* **15**, 155–163 (2016).
56. Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
57. Vallat, R. Pingouin: statistics in python. *J. Open Source Softw.* **3**, 1026 (2018).
58. Herzog, E. L. *et al.* Interstitial Lung Disease Associated With Systemic Sclerosis and Idiopathic Pulmonary Fibrosis: How Similar and Distinct? *Arthritis & rheumatology (Hoboken, N.J.)* **66**, 1967, https://doi.org/10.1002/ART.38702 (2014).

Acknowledgements

This work was performed using the computational resources from the Academic Leiden Interdisciplinary Cluster Environment (ALICE) provided by the Leiden University. This work was supported by the China Scholarship Council under Grant 202007720110.

Author contributions

All authors were involved in analysing and interpreting the data. JJ drafted the manuscript and IH-G, AAS, JCK, JKV, MKN, MS, LJMK and BCS critically revised the manuscript. AAS and LJMK built the database. JJ and BCS directly accessed and verified the underlying data reported in the manuscript. JJ, MS and BCS contributed to the study design. JJ performed the statistical analysis. JJ obtained funding. BCS and MS supervised the study. The corresponding author attests that all listed authors meet the authorship criteria and that no other researchers meeting the criteria have been omitted. The corresponding author had full access to the data and the final responsibility to submit it for publication.

Declarations

Competing interests

The authors declare no conflicts of interest. The corresponding author is responsible for submitting a competing interests statement on behalf of all authors of the paper. This statement must be included in the submitted article file.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/1

0.1038/s41598-024-78393-4.

Correspondence and requests for materials should be addressed to B.C.S.

Reprints and permissions information is available at www.nature.com/reprints.