Automatic pulmonary function estimation from chest CT scans using deep regression neural networks: the relation between structure and function in systemic sclerosis

Jia, J.N.; Marges, E.R.; Vries-Bouwstra, J.K. de; Ninaber, M.K.; Kroft, L.J.M.; Schouffoer, A.A.; ... ; Stoel, B.C.

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 9 November 2023, accepted 21 November 2023, date of publication 28 November 2023,
date of current version 6 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3337639

RESEARCH ARTICLE

Automatic Pulmonary Function Estimation From Chest CT Scans Using Deep Regression Neural Networks: The Relation Between Structure and Function in Systemic Sclerosis

JINGNAN JIA[1], EMIEL R. MARGES[2], JESKA K. DE VRIES-BOUWSTRA[3], MAARTEN K. NINABER[2], LUCIA J. M. KROFT[4], ANNE A. SCHOUFFOER[3], MARIUS STARING[1], AND BEREND C. STOEL[1]

[1]Division of Image Processing, Department of Radiology, Leiden University Medical Center (LUMC), 2300 RC Leiden, The Netherlands
[2]Department of Pulmonology, Leiden University Medical Center (LUMC), 2300 RC Leiden, The Netherlands
[3]Department of Rheumatology, Leiden University Medical Center (LUMC), 2300 RC Leiden, The Netherlands
[4]Department of Radiology, Leiden University Medical Center (LUMC), 2300 RC Leiden, The Netherlands

Corresponding author: Berend C. Stoel (b.c.stoel@lumc.nl)

This work was supported by the China Scholarship Council under Grant 202007720110.

This work involved human subjects in its research. 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.

ABSTRACT Pulmonary function tests (PFTs) play an important role in screening and following-up pulmonary involvement in systemic sclerosis (SSc). However, some patients are not able to perform PFTs due to contraindications. In addition, it is unclear how lung function is affected by changes in lung structure in SSc. Therefore, this study aims to explore the potential of automatically estimating PFT results from chest CT scans of SSc patients and how different regions influence the estimation of PFTs. Deep regression networks were developed with transfer learning to estimate PFTs from 316 SSc patients. Segmented lungs and vessels were used to mask the CT images to train the network with different inputs: from entire CT scan, lungs-only to vessels-only. The network trained on entire CT scans with transfer learning achieved an ICC of 0.71, 0.76, 0.80, and 0.81 for the estimation of DLCO, FEV$_1$, FVC and TLC, respectively. The performance of the networks gradually decreased when trained on data from lungs-only and vessels-only. Regression attention maps showed that regions close to large vessels were highlighted more than other regions, and occasionally regions outside the lungs were highlighted. These experiments show that apart from the lungs and large vessels, other regions contribute to PFT estimation. In addition, adding manually designed biomarkers increased the correlation (R) from 0.75, 0.74, 0.82, and 0.83 to 0.81, 0.83, 0.88, and 0.90, respectively. This suggests that that manually designed imaging biomarkers can still contribute to explaining the relation between lung function and structure.

INDEX TERMS Pulmonary lung function, deep learning, computerized tomography, systemic sclerosis.

I. INTRODUCTION

Systemic sclerosis (SSc) is a rare immune-mediated connective tissue disease that affects different organs. Interstitial lung disease (ILD) is, however, the leading cause of morbidity and mortality, and up to 90% of SSc patients have pulmonary function abnormalities [1]. To evaluate progression of SSc-ILD, various pulmonary function tests (PFTs) are used as key measures, such as the diffusion capacity for carbon monoxide (DLCO), forced expiratory volume in 1 second (FEV$_1$), forced vital capacity (FVC) and total lung capacity (TLC) [1], [2], [3]. In clinical practice, PFTs are

The associate editor coordinating the review of this manuscript and approving it for publication was Kumaradevan Punithakumar.

expressed either in absolute values or in percent predicted values (abbreviated as PFTs%pred, including DLCO%pred, FEV_1%pred, FVC%pred and TLC%pred). PFTs%pred are obtained by the standardization of the absolute values according to the patients' characteristics to avoid biases from sex, ethnicity and height [4]. A PFT%pred below 100% then represents a lung function that is lower than the average lung function in a population with the same age, gender, etc., with upper and lower limits of normal (usually 1.64 SD). Both absolute and percent predicted are commonly used clinically as outcome measures for progression of SSc-ILD [2]. PFTs can, however, not always be performed if there is a risk of disease transmission, e.g. in patients with COVID-19, active tuberculosis or other airborne infectious diseases [5], [6]. In addition, some patients, who have hemoptysis or had surgery in the past month, or other contraindications [7], [8], like aneurysmatic abnormalities and ischaemic stroke, are not able to perform PFTs because the forced exhalation during spirometry may increase the risk of complications [9]. Therefore, alternative methods to estimate PFT are of great interest. Because CT could provide high-resolution details of the lungs, it is regarded the gold standard for diagnosing SSc-ILD [10]. In previous research, quantitative biomarkers have been extracted from chest CT images of SSc patients, which correlate with PFTs [11]. Therefore, when PFTs are not possible and CT scans have been made for SSc patients, it is of great interest to see if CT could be used to estimate PFT.

Apart from being an alternative to PFTs, PFT estimation from CT scans can also be used to study the relation between structure and function as the lungs become affected by SSc-ILD. Initially, imaging biomarkers were designed for SSc to explicitly describe lung structure and subsequently determine their correlation with lung function. For SSc patients with fibrosis, Goh et al. [12] designed a visual fibrosis scoring system, which correlated with FVC ($R = -0.40$). For SSc patients without fibrosis visible on CT, Zhai et al. [13] found that two vascular tree-based biomarkers (α and β), which represent the lung vessel radius histogram, correlated with DLCO%pred ($R=-0.29$ and 0.32, respectively). For SSc patients with or without fibrosis, Ninaber et al. [3] found that lung density, measured by the 85th percentile density (Perc85) from CT scans, correlated significantly with DLCO%pred ($R=-0.49$) and FVC %predicted ($R=-0.64$).

Apart from these manually designed biomarkers, an altogether different approach would be to develop a deep learning model that is trained to output PFT prediction values directly, with or without fibrosis visible on CT scans. Subsequently, the trained model could be studied in detail to explore the relation between lung structure from CT and lung function from PFTs.

To the best of our knowledge, we are the first to estimate PFTs for SSc patients. There are no works to estimate PFTs for SSc patients previously. The most relevant and recent works on automatic estimation of PFTs from chest CT using deep learning [5], [11] are not for SSc patients. Choi, et al. [5] developed a network to estimate FEV_1 and FVC for patients before their first lung cancer surgery. Their network consisted of a ResNet-50 for feature extraction and a bidirectional long short-term memory (BiLSTM) network for PFT prediction. Park, et al. [11] trained two separate I3D networks to estimate FEV_1 and FVC, respectively, for subjects at risk of lung cancer. It is unclear if their models could be applied directly to SSc patients. In addition, both methods estimate FEV_1 and FVC only, lacking DLCO and TLC. For determining SSc-ILD progression, however, TLC and especially DLCO are important measurements, the latter of which is most predictive of adverse outcomes, including death [2]. Therefore, the aim of this study was to 1) develop a deep learning model to estimate DLCO, FEV_1, FVC and TLC for SSc patients from their CT scans; and 2) explore the contribution of different anatomical regions, and provide explanations from a clinical perspective.

The remaining paper is organized as follows. Section II describes the datasets and methods we used for the prediction of PFT. Detailed experiments and results are shown in Section III. Finally, section IV discusses the experiments, explains the results and concludes the paper.

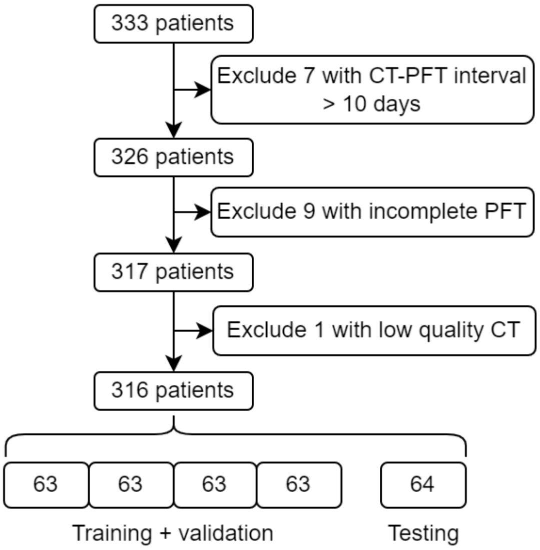
333 patients
Exclude 7 with CT-PFT interval > 10 days
326 patients
Exclude 9 with incomplete PFT
317 patients
Exclude 1 with low quality CT
316 patients
63
63
63
63
64
Training + validation
Testing

FIGURE 1. Flowchart of the dataset inclusion and partition.

II. MATERIALS AND METHODS

A. DATASET

In this study, we retrospectively selected 333 patients who were referred to our targeted outpatient health care program (combined care in systemic sclerosis) between April 2009 and October 2015 in Leiden University Medical Center. Because of the diagnosis of SSc according to the referring rheumatologist, or a strong suspicion for SSc, they underwent high-resolution CT scans, followed by pulmonary function tests. As shown in Figure 1, we excluded seven patients with a CT-PFT interval greater than ten days, nine patients with

TABLE 1. **Dataset characteristics of systemic sclerosis patients. STD: standard deviation.**

Characteristic	Patients (n=316)
Age [years], mean ± STD	53.4 ± 14.6
Female (%)	258 (81.4)
Interstitial lung disease detected on CT (%)	132 (41.8)
Pulmonary arterial hypertension (%)	88 (27.8)
Disease Subset	
Non-cutaneous (%)	40 (12.7)
Diffuse cutaneous (%)	85 (26.9)
Limited cutaneous (%)	183 (57.9)
Others (%)	8 (2.5)
Pulmonary function, mean ± STD	
DLCO [mL/min/mm Hg]	5.55 ± 1.92
FEV_1 [L]	2.62 ± 0.77
FVC [L]	3.32 ± 0.97
TLC [L]	4.91 ± 1.23
DLCO%pred [%]	71.95 ± 20.10
FEV_1%pred [%]	89.32 ± 17.69
FVC%pred [%]	90.58 ± 18.97
TLC%pred [%]	85.62 ± 17.07

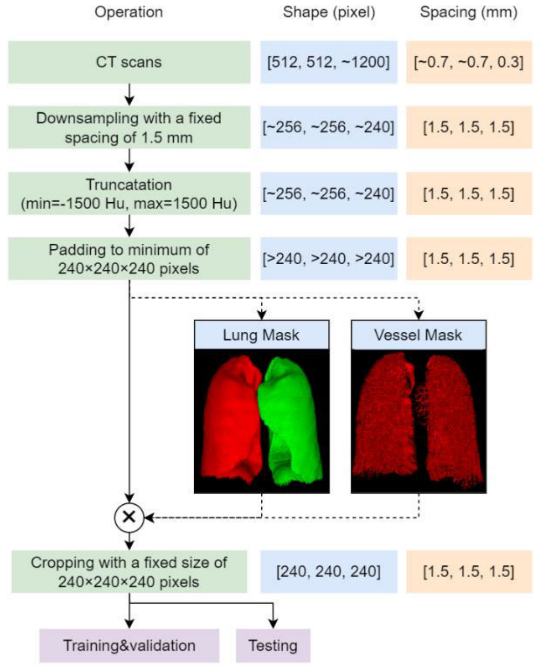

FIGURE 2. **CT scan preprocessing procedure.**

incomplete PFTs, and one patient with a low-quality CT scan, resulting in 316 CT-PFT pairs. The dataset was split into two disjoint groups: 252 for four-fold training and cross-validation, and 64 for testing. The research protocol was granted approval by the local Medical Ethics Committee and written informed consent was provided by all patients.

1) CT SCANNING

All subjects underwent scanning at full inspiration without contrast enhancement using an Aquilion 64 CT scanner (Canon Medical Systems), configured at 120 kVp, a median tube current of 140 mA, a rotation time of 0.4 seconds, a collimation of 64 × 0.5 mm and a helical beam pitch of 0.8; leading to a median $CTDI_{vol}$ of 8.2 mGy. The images were reconstructed with filtered back projection and an FC86 kernel, with a median pixel spacing of 0.64 mm × 0.64 mm, with a slice thickness and increment of 0.5 and 0.3 mm, respectively.

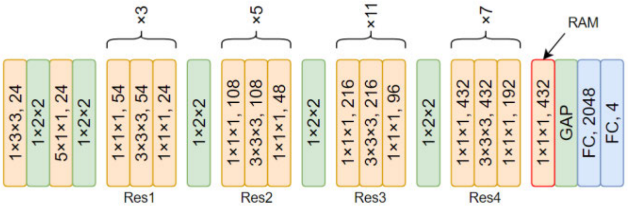

FIGURE 3. **X3D_M structure. The whole network consists of 3D convolution layers (orange boxes), max-pooling layer (green boxes) and two fully connected layer (blue box). Kernel size (xyz) and channel number are denoted by the first three numbers and the last number, respectively.**

2) PFT MEASUREMENTS

PFTs were performed by an experienced technologist using a spirometer under ERS/ATS guidelines [14], [15] including single-breath diffusion capacity for carbon monoxide corrected for haemoglobin concentration (DLCO), forced expiratory volume in 1 second (FEV_1), forced vital capacity (FVC) and total lung capacity (TLC). While DLCO was measured in units of mm/Hg/min, FEV_1, FVC, TLC were measured in units of liter. The PFT percent predicted values (PFTs%pred) were calculated with the latest official conversion equations and reference values [16], [17], [18]. Clinical characteristics of the 316 patients are shown in Table 1.

3) DATA PREPROCESSING AND AUGMENTATION

Because of GPU memory limitations, we first down-sampled all CT scans to an isotropic spacing of 1.5 mm, as illustrated in Figure 2, resulting in a median image size of 256 × 256 × 240 voxels. Next, we performed intensity truncation to clip voxel values between −1500 and 1500 HU to remove some artifacts. Then we applied padding, if necessary, to guarantee a minimum image size of 240 × 240 × 240 voxels. To subsequently augment the training data, a random 3D patch of a fixed size (240 × 240 × 240 voxels, which was ensured to cover the whole lung area) were cropped from each volume as they are fed into the model. In different epochs, different 3D patches were cropped from each CT for training. The epoch number is the number of 3D patches cropped from each CT. In the validation and testing phase, we used 3D patches of 240 × 240 × 240 voxels at the center position, from (x0-120, y0-120, z0-120) to (x0+120, y0+120, z0+120) where (x0, y0, z0) is the coordinates of the center point of each validation and testing CT image. To investigate the contribution of different chest regions, we masked the CT images using various masks. Lung masks were obtained by a multi-atlas based method [13], while vessel masks were acquired using a graph-cut based vessel segmentation network [19]. The segmentation of lung and vessel masks was obtained by an in-house script in MeVisLab 2.7.1 (VC12-64). The implementation details could be found at the online document which were released along with the original paper [13] (http://links.lww. com/JTI/A114). The source code of our in-house script for the segmentation could be found at https://github.com/Zhiwei-Zhai/Lung-Vessel-Segmentation-Using-Graph-cuts. No additional data augmentation was performed.

B. NETWORK DESIGN

The network was adapted from X3D [20], which was originally designed for video recognition. The original paper proposed a series of networks with different capacities. An X3D of medium size (X3D_M) was selected as the architecture of our network, to account for limited GPU memory. As illustrated in Figure 3, the network consists of several convolution and max-pooling layers, followed by four ResNet blocks with max-pooling layers between each of them, and finally one global average-pooling layer (GAP) and two fully connected (FC) layers. The output of the last FC layer has four values, representing the four (absolute) PFT parameters, simultaneously estimated in one network. We also developed four separate networks with 1-class outputs for each of them, estimating the different PFT parameters, separately. The comparison between these 1-class and 4-class networks will be shown later in Q4 of Section Experiments and Results.

To increase network performance, we introduced transfer learning (TL), in which the network was initialized by the weights trained from another domain. Although it may achieve better performance if the source domain is similar with the target domain, the lack of large annotated lung CT dataset makes it impractical to apply pre-trained weights from lung CT dataset. However, TL has been widely used in deep learning because it was reported to improve network performance significantly even if the source domain is different [20]. Therefore, our X3D_M network was pre-trained on Kinetics, a human action video dataset [21], [22] (pytorch.org/hub/facebookresearch_pytorchvideo_x3d).

Although there are other 3D networks which may also works on our task, X3D_M is the network which was released recently, achieved the SOTA performance, publish their pre-trained weights from Kinetics dataset, and could be fit into our GPU with memory of 11 GB.

We studied two ways to estimate PFTs%pred: 1) directly, by one network that is trained to estimate PFTs%pred directly; and 2) indirectly, where the absolute PFTs are obtained by a network, from which the PFTs%pred are subsequently calculated by the official conversion equations [16], [17], [18].

C. RELATION BETWEEN LUNG STRUCTURE AND FUNCTION

After we obtained the optimized network and training method, we performed two strategies to understand how the network derived the estimation and how each chest region (such as muscle, lung, vessel, etc.) contributed to the PFT estimation. The first strategy was to train multiple networks with various inputs: whole CT image, lungs-only (by excluding the volume outside the lungs), left or right lung-only, vessels-only and the binarized version of vessels-only. The difference in performance between the different networks implies the contribution of these different regions. The second strategy was 3D regression activation mapping (RAM-3D), which is a variant of the Grad-CAM [23] on 3D regression tasks. The original Grad-CAM was designed for 2D image classification [23], which could generate heat maps to highlight the important regions for classification by convolutional neural networks (CNNs). Inspired by that, Wang et al. proposed a RAM for 2D image regression [24]. In this work, we extended this RAM from 2D to 3D to highlight areas of interest in the 3D CT volumes for the PFT estimation. To capture detailed regional information, we computed the gradient for the linear output layer with respect to the feature maps of the convolution layer right before the GAP layer (marked in Figure 3).

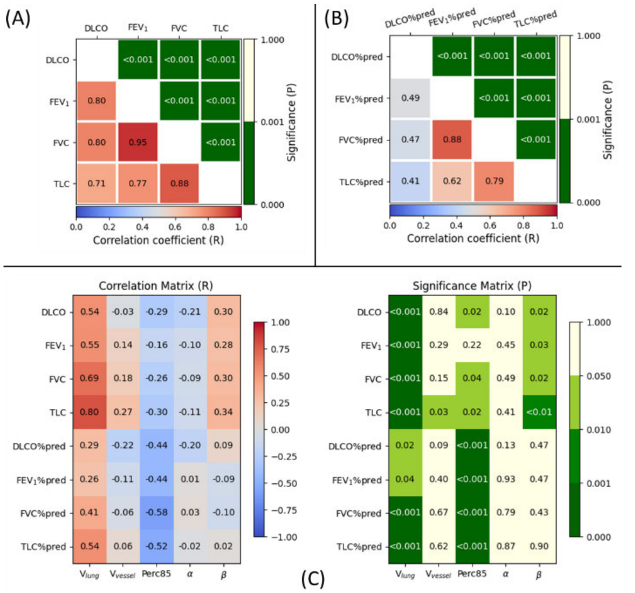

FIGURE 4. Pearson correlation coefficients and corresponding significance levels: (A) between different absolute PFTs; (B) between different PFTs%pred; and (C) between manually designed biomarkers and PFTs/PFTs%pred, from the testing dataset.

D. EVALUATION METRICS AND STATISTICAL ANALYSIS

The performance of the proposed deep learning networks was evaluated on two separate datasets: a four-fold cross-validation dataset and a separate testing set. The optimization of network structure and training strategy was based on the four-fold cross-validation results. The testing dataset was used only for the final performance assessment, and for comparison our network's performance with standard repeatability criteria of PFT measurements.

We used various metrics to evaluate the agreement between our network output and measured values (from spirometry). The mean absolute error (MAE) was used to reflect the

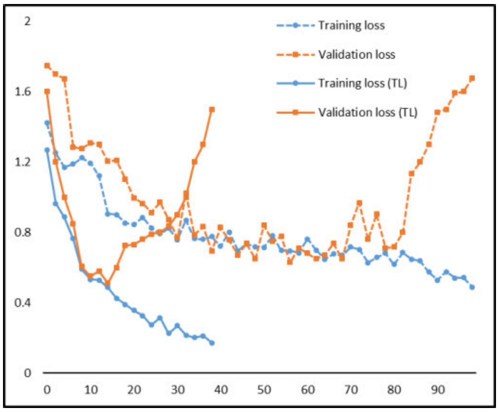

FIGURE 5. Comparison of training and validation curves with or without transfer learning (TL) on the same fold.

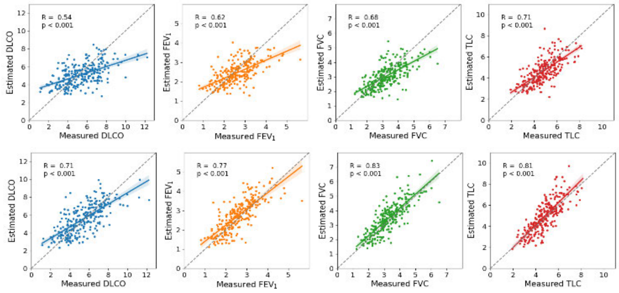

FIGURE 6. Scatter plots comparison between networks without (upper) and with (lower) transfer learning. Each image shows the identify line (dot line), regression line (solid line) and the 95% confidence intervals (shaded areas).

absolute agreement. Because the unit and scale of the four PFTs are different, we used the mean absolute percentage error (MAPE), which is the ratio of MAE to the real measurements, to reflect the relative uncertainty of prediction. MAE and MAPE were calculated as follows:

$$\mathrm{MAE} = \frac{1}{N}\sum\nolimits_{i=1}^{N} |y_e - y_m| \tag{1}$$

$$\mathrm{MAPE} = \frac{100\%}{N}\sum\nolimits_{i=1}^{N} \frac{|y_e - y_m|}{y_m} \tag{2}$$

where i is the index of samples and N represents the total number of samples, y_e is the network's estimated value, and y_m the measured PFT value.

We used the Pearson correlation coefficient (R) to indicate the linear correlation. An absolute value of R below 0.1 indicates a negligible correlation, a value between 0.1 and 0.39 indicates a weak correlation, between 0.4 and 0.69 a moderate correlation, between 0.7 and 0.89 a strong correlation, and over 0.9 indicates a very strong correlation [25]. The intra-class correlation coefficient (ICC) is a measure of reliability, which represents not only the absolute agreement but also the linear correlation. ICC was calculated by Pingouin 0.4.0 [26] based on a single-rating, absolute-agreement, 2-way mixed-effects model [24]. ICC values below 0.5 indicate poor reliability, between 0.5 and 0.75 moderate reliability, between 0.75 and 0.9 good reliability, and any value above 0.9 indicates excellent reliability [27].

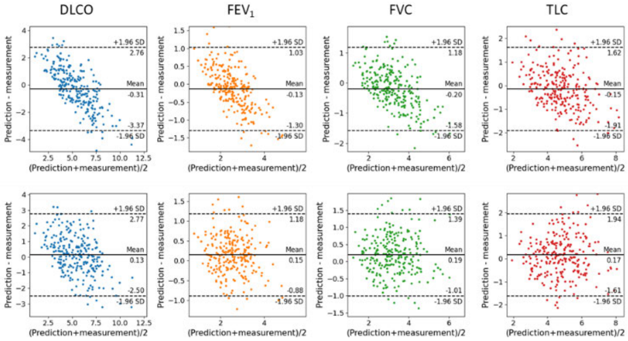

FIGURE 7. Bland-Altman plots comparison between networks without (upper) and with (lower) transfer learning. The mean difference and the limits of agreement (mean ± 1.96 × SD, where SD is the standard deviation of the differences) are also shown on the plots.

To statistically test differences between groups, a Wilcoxon signed-rank test was 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. Bland-Altman plots were used to analyze the mean differences (bias) and limits of agreement. These statistical analyses were performed by an in-house python 3.8 script with corresponding libraries.

In addition, we applied multiple variable regression analysis using IBM SPSS Statistics version 27 software (IBM, Armonk, USA), to determine if manual biomarkers could contribute to the prediction from the developed networks.

III. EXPERIMENTS AND RESULTS

We sequentially conducted a series of experiments to answer the following questions and optimize our method, based on the answers to these questions: **Q1:** How well can traditional manually designed features predict PFTs in our dataset? **Q2**: Does our network benefit from transfer learning? **Q3**: For PFTs%pred estimation, is the direct estimation better than the indirect estimation? **Q4**: How does a 1-class network perform compared to a 4-class network? **Q5**: How much do the different chest regions contribute to the PFT estimation? **Q6**: How does our method perform compared to standard repeatability criteria for PFTs? **Q7**: Are manual biomarkers still valuable for SSc patients given our automatic method?

A. EXPERIMENT SETTING

Our neural networks were implemented using PyTorch 1.11.0 (https://pytorch. org). The loss function was the mean squared error (MSE), and a batch size of 1 was used. The Adam optimizer was used with a learning rate of 1e-4 and a weight decay of 1e-3. Multithreading was used to accelerate the on-the-fly data augmentation. The training will stop when the validation loss does not decline in 25 consecutive epochs or once 100 epochs have been completed. The workstation for training and validation was equipped with an Intel(R) Xeon(R) CPU Gold 6126 2.6GHz with 90 GB memory

TABLE 2. Four-fold validation results comparison between networks trained with or without transfer learning. TL: transfer learning.

TL	Out	Metrics	DLCO					FEV_1					FVC					TLC				
			1	2	3	4	All	1	2	3	4	All	1	2	3	4	All	1	2	3	4	All
No	4-class	R	0.72	0.55	0.57	0.47	0.54	0.77	0.54	0.78	0.63	0.62	0.80	0.59	0.78	0.71	0.68	0.86	0.68	0.77	0.76	0.71
		ICC	0.64	0.51	0.40	0.44	0.48	0.73	0.52	0.61	0.52	0.58	0.77	0.59	0.68	0.60	0.65	0.85	0.64	0.68	0.69	0.70
		MAE	0.96	1.27	1.36	1.21	1.25	0.39	0.49	0.39	0.42	0.46	0.45	0.63	0.45	0.48	0.56	0.50	0.86	0.63	0.56	0.70
		MAPE	20%	25%	26%	29%	26%	16%	22%	18%	15%	19%	17%	24%	13%	15%	17%	13%	20%	13%	14%	15%
Yes	4-class	R	0.77	0.74	0.73	0.80	0.71	0.85	0.74	0.81	0.75	0.77	0.86	0.82	0.91	0.81	0.83	0.85	0.85	0.88	0.68	0.82
		ICC	0.76	0.73	0.68	0.66	0.71	0.76	0.74	0.82	0.72	0.76	0.82	0.81	0.90	0.71	0.8	0.82	0.85	0.82	0.67	0.81
		MAE	0.90	1.00	1.14	0.99	1.04	0.41	0.38	0.38	0.42	0.41	0.44	0.41	0.38	0.56	0.49	0.60	0.48	0.71	0.79	0.63
		MAPE	19%	21%	23%	19%	22%	19%	16%	15%	20%	18%	15%	15%	13%	20%	16%	12%	10%	16%	17%	13%
		P†	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
	1-class	R	0.68	0.79	0.75	0.66	0.71	0.85	0.70	0.82	0.75	0.76	0.86	0.80	0.87	0.77	0.78	0.91	0.89	0.88	0.87	0.85
		ICC	0.67	0.65	0.73	0.62	0.70	0.80	0.70	0.81	0.74	0.76	0.80	0.77	0.86	0.67	0.76	0.90	0.84	0.85	0.79	0.84
		MAE	0.99	0.93	1.07	1.10	1.07	0.32	0.38	0.32	0.46	0.38	0.39	0.42	0.42	0.52	0.51	0.43	0.56	0.50	0.62	0.55
		MAPE	22%	21%	22%	25%	23%	15%	17%	14%	22%	16%	16%	13%	17%	19%	17%	10%	13%	12%	16%	12%
		P‡	0.51	0.68	0.57	0.74	0.60	0.22	< 0.01	0.10	0.61	0.13	0.15	< 0.01	0.26	0.39	0.18	0.80	0.75	0.66	0.34	0.61

† Significance of the difference in mean error between the networks with or without transfer learning.
‡ Significance of the difference in mean error between the 4-class and 1-class network with transfer learning.

and a NVIDIA GPU GeForce RTX 2080TI with 11 GB memory. Our code and trained models are publicly available via GitHub (https://github.com/Jingnan-Jia/PFT) for the convenience of reproducing our method or applying our model to other datasets.

B. MANUALLY DESIGNED BIOMARKERS (Q1)

First, the correlation between different PFTs and PFTs%pred are shown in Figure 4 (A) and (B), respectively. Consistent with the literature, high correlations with $p<0.001$ were found among the four PFTs with R ranging from 0.71 to 0.95, and lower correlations with still $p<0.001$ among the four PFTs%pred with R from 0.41 to 0.88. We applied previously developed manual quantification methods on our CT dataset to obtain various imaging biomarkers including lung volume (V_{lung}), vessel volume (V_{vessel}), Perc85 [3], α and β [13]. The correlation between manually designed biomarkers and measured PFTs and PFTs%pred are presented in Figure 4 (C). V_{lung} was significantly correlated with PFTs and PFTs%pred with $p<0.05$. V_{vessel} showed no significant correlation with any PFTs except TLC values ($p=0.03$). Perc85 correlated significantly with all PFTs%pred ($p<0.001$), which is consistent with a previous report [3]. α and β showed no significant correlations with any of the PFTs%pred. With the absolute PFT measures, β still showed a significant correlation, with R ranging from 0.28 to 0.34, but α did not show any significant correlations.

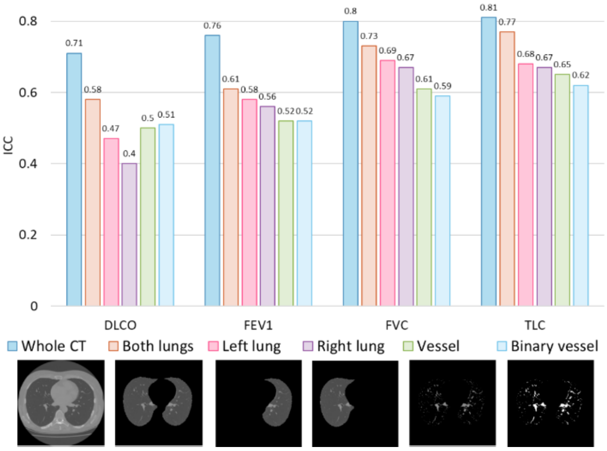

FIGURE 8. Performance of the networks trained on different regions of the chest. The bars with different colors represent networks trained by different regions, which are illustrated at the bottom.

TABLE 3. PFTs%pred estimation, compared between two methods.

Method	Metrics	DLCO%pred	FEV_1%pred	FVC%pred	TLC%pred
Direct	R	0.32	0.47	0.51	0.54
	ICC	0.30	0.47	0.50	0.53
	MAE	17.30	14.88	14.98	13.59
	MAPE	29%	18%	18%	17%
Indirect	R	0.60	0.65	0.74	0.76
	ICC	0.60	0.60	0.69	0.75
	MAE	13.85	14.79	13.76	11.04
	MAPE	22%	18%	16%	13%

C. TRANSFER LEARNING (TL) VERSUS TRAINED FROM SCRATCH (Q2)

The performance of the network based on TL was compared with the network trained from scratch, see Table 2. It is shown that the R and ICC values increased and MAE values decreased after the introduction of TL. The standard deviation also decreased, which means that the networks with TL were more stable than those trained from scratch. This finding was verified by the scatter plots of the two networks (Figure 6), where the regression lines of the network with TL were closer to the identity line than the networks without TL. Figure 7 shows the Bland-Altman plots of networks without or with TL. The plots display the differences between the automatically estimated PFTs and measured PFTs against their mean. From Figure 6 and 7, we can observe that the network trained from scratch tended to give conservative estimations: close to the mean value of measurements.

Therefore, the images with lower PFTs were overestimated and higher PFTs were underestimated. After the introduction

TABLE 4. Performance of the networks trained from different inputs. The units of MAE are mL/min/mm Hg for DLCO and liter for FEV_1/FVC/TLC.

Input	Metrics	DLCO	FEV_1	FVC	TLC	Mean
Whole CT	R	0.71	0.77	0.83	0.82	0.78
	MAE	1.04±0.86	0.41±0.36	0.49±0.42	0.63±0.48	0.64±0.53
	MAPE	22%±24%	18%±18%	16%±13%	13%±9%	17%±16%
Both lungs	R	0.59	0.62	0.73	0.78	0.68
	MAE	1.19±0.98	0.45±0.42	0.49±0.44	0.57±0.49	0.68±0.58
	MAPE	27%±32%	20%±22%	16%±15%	13%±12%	19%±20%
Left lung	R	0.47	0.58	0.69	0.68	0.61
	MAE	1.29±1.02	0.48±0.43	0.56±0.45	0.74±0.58	0.77±0.62
	MAPE	28%±30%	20%±19%	18%±14%	16%±13%	21%±20%
Right lung	R	0.45	0.59	0.69	0.69	0.61
	MAE	1.31±1.02	0.45±0.41	0.52±0.44	0.69±0.55	0.74±0.61
	MAPE	30%±39%	19%±20%	17%±17%	15%±14%	20%±23%
Vessels	R	0.51	0.53	0.62	0.66	0.58
	MAE	1.28±1.01	0.53±0.46	0.58±0.49	0.73±0.55	0.78±0.63
	MAPE	27%±30%	23%±22%	19%±16%	16%±12%	21%±20%
Binarized vessels	R	0.57	0.55	0.64	0.68	0.61
	MAE	1.28±1.04	0.52±0.45	0.62±0.50	0.76±0.56	0.80±0.64
	MAPE	27%±29%	22%±22%	19%±18%	16%±13%	21%±21%

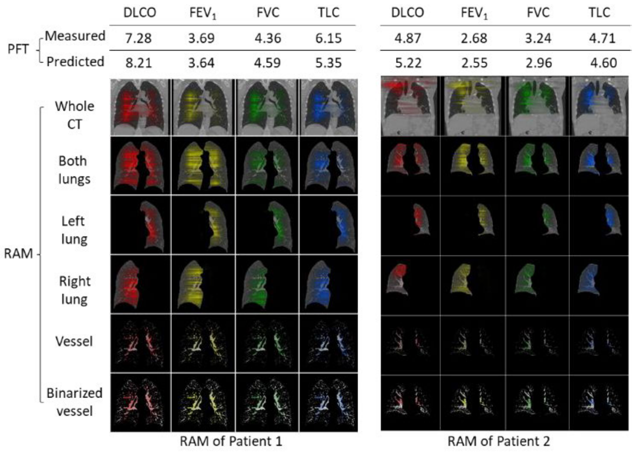

FIGURE 9. RAMs of two patients for different networks (coronal view). RAMs of different rows are generated from the networks trained by different regions. Red, yellow, green and blue highlight the attention of networks on DLCO, FEV_1, FVC and TLC, respectively.

of TL, these pattern disappeared in FEV_1, FVC and TLC, whereas in DLCO a similar but less prominent pattern remained. This indicates that the network trained with TL achieves a better agreement to measured PFTs. In addition, Figure 5 shows that TL could speed up the training: decreasing the training epochs from 68 to 14. Therefore, we used TL in all the following experiments.

D. ESTIMATION OF PFTS%PRED: DIRECT VERSUS INDIRECT (Q3)

Table 3 shows the performance of estimating PFTs%pred for the two methods. The indirect estimation achieved ICC values of 0.60, 0.60, 0.69 and 0.75 for DLCO%pred, FEV_1%pred, FVC%pred and TLC%pred, respectively. These ICC values were higher than those of the direct method (ICC=0.30, 0.47, 0.50 and 0.53). The indirect method also achieved higher R and lower MAE values. Therefore, all the following networks were trained to estimate the absolute PFTs first.

E. 1-CLASS VERSUS 4-CLASS (Q4)

Table 2 shows that the ICC values of the 4-class network (ICC=0.71, 0.76, 0.80, and 0.81) were similar to the ICC values of tsshe four 1-class networks (ICC=0.70, 0.76, 0.76, and 0.84). The R and MAE values for the two network designs were also similar. p-values of 0.60, 0.13, 0.18 and 0.61 indicate that the results of the four networks with 1-class output did not show a significantly difference compared to the 4-class network. Because the 4-class network can output four PFTs at a same time, which saves training & inference time and GPU memory, all the following networks were trained with a 4-class output.

F. CONTRIBUTION OF THE DIFFERENT CHEST REGIONS (Q5)

The PFT estimation performance of our proposed networks are summarized in Figure 8 and Table 4. DLCO was always the most difficult parameter to estimate, followed by FEV_1 and then FVC and TLC. The network trained on the whole CT volume achieved the best performance (highest ICC and R, as well as lowest MAE values). The network trained on both lungs performed slightly worse. The performances for the left and right lung were similar, which implies similar contribution of left and right lung to the PFT estimation. The network trained on pulmonary vessels performed worse than the previous networks for FEV_1, FVC and TLC estimation, but better for DLCO estimation. The network trained on binarized vessels (1 as foreground and 0 as background) achieved similar ICC and MAE values and slightly higher R values, as compared to the network trained on gray scale vessels. The networks trained on gray scale vessels performed the worst compared to the other networks, but still better than the manually designed vessel based biomarkers α (R=−0.29) and β (R=0.32).

We generated heat maps by RAM-3D for various networks trained by different regions of lung CT scans. The coronal and axial views are shown in Figure 8, respectively. If we look at the RAMs generated from the network trained on whole CT, for some patients, the highlights are limited to the lungs, see Figure 8 (left, row of Whole CT). For some other patients, the highlights also appeared outside the lungs (i.e. in the chest wall), see Figure 8 (right, row of Whole CT). For all networks, the two center/hilar regions of the two lungs, where the larger pulmonary vessels are located, were highlighted the most. This pattern applied to the RAMs of all networks. The coronal views of RAMs were vertically discontinuous; this is because the X3D_M network only applied pooling layers along the x and y axes, while leaving the z axis free of pooling layers, before the layer where our RAM-3D was applied. That led to a narrow reception field along the z axis.

G. COMPARISON STANDARD REPEATABILITY CRITERIA (Q6)

After comparing our method to previous works, putting the results of our network into clinical perspective is still needed.

TABLE 5. Comparison between official repeatability criteria and the relative error of our method.

	DLCO (%)	FEV_1 (%)	FVC (%)	TLC (%)
Repeatability criteria	10	6	5	10
Our method (MAPE±STD)	20 ± 18	19 ± 14	15 ± 12	13 ± 11

TABLE 6. Multivariable stepwise linear regression analysis for DLCO, FEV_1, FVC and TLC.

DLCO

Parameter	R
NetDLCO	0.75
NetDLCO, V_{lung}	0.78
NetDLCO, V_{lung}, β	0.81

FEV_1

Parameter	R
$NetFEV_1$	0.74
$NetFEV_1$, Perc85	0.83

FVC

Parameter	R
NetFVC	0.82
NetFVC, Perc85	0.87
NetFVC, Perc85, V_{lung}	0.88

TLC

Parameter	R
NetTLC	0.83
NetTLC, Perc85	0.88
NetTLC, Perc85, V_{lung}	0.90

Estimating PFTs from CT scans by human experts is impractical for obvious reasons, therefore we could not compare our method with human observations. However, we could compare our results with the theoretically best obtainable result, as determined by the officially recommended repeatability criteria for spirometric measurements. The PFT measures are normally obtained by means of three repetitions of the measurements [9]. According to the most recent official standard on pulmonary function testing [9], the repeatability for DLCO and TLC obtained by the helium dilution technique between technically acceptable measurements should be within 10% of the average value. The repeatability criterion for FEV_1 and FVC is that differences should be lower than 0.15 L [9]. To have a fair comparison between different PFTs, the acceptable errors of 0.15 L for FEV_1 and FVC were divided by the mean measured values in Table 1, obtaining a percentage error of 6% ($MAPE_{FEV_1} = 0.15/\bar{y}_{FEV_1} = 0.15/2.62 \approx 6\%$) and 5% ($MAPE_{FVC} = 0.15/\bar{y}_{FVC} = 0.15/3.32 \approx 5\%$). As shown in Table 5, the repeatability criteria is 10%, 6%, 5% and 10% for DLCO, FEV_1, FVC and TLC, respectively.

H. MULTIPLE VARIABLE REGRESSION ANALYSIS (Q7)

A multivariable regression analysis was performed to evaluate if manual biomarkers could still contribute to the estimation of PFTs, in addition to the estimation of our method. Multivariable stepwise linear regression was performed with DLCO as the dependent variable and the network-estimated DLCO (NetDLCO), Perc85, α, β, V_{lung}, and V_{vessel} as independent variables. We performed similar analyses for FEV_1, FVC and TLC (Table 6). The multivariable stepwise regression analysis showed that the inclusion of V_{lung} and β could significantly improve the regression coefficient of DLCO (from R = 0.75 to R = 0.81). Similarly, by including V_{lung} and Perc85, the estimation of FEV_1, FVC and TLC could also be significantly improved. Therefore, although we have developed automatic networks that outperformed manually designed biomarkers as single predictors, these manual biomarkers could still contribute further to the estimation of PFTs.

TABLE 7. Comparison between our method and previous automatic methods for the estimation on PFTs. Because previous methods did not estimate DLCO and TLC, the corresponding results are not included. NR: not reported.

Method	Study population	Backbone network	#Subjects	MAE		R	
				FEV_1	FVC	FEV_1	FVC
Choi, *et al.* [5]	Lung cancer	ResNet [28]	546	0.33	0.37	0.73	0.82
Park, *et al.* [11]	Risk of lung cancer	I3D [29]	16148	0.22	0.22	NR	NR
Our method	SSc	X3D [20]	316	0.41	0.49	0.77	0.83

IV. DISCUSSION AND CONCLUSION

This paper demonstrated that it is feasible to automatically estimate comprehensive PFTs and PFTs%pred from chest CTs, using deep learning. Our results indicate that CT scans can assist in estimating PFTs with considerable predictive accuracy.

To the best of our knowledge, there is no currently published work on estimating PFT values for SSc patients. The only two works [5], [11] that aimed to automatically estimate PFTs from CT were developed based on 546 subjects diagnosed with lung cancer [5] or 16148 subjects with a risk of developing lung cancer [11], as shown in Table 7. The R values of our method is slightly higher than Choi, et al. [5], while the MAE of our method is higher than the other two works. Because we have already applied X3D with transfer learning, which achieved state-of-the-art performance on video recognition, possible reasons of the performance gap may include: 1) Dataset sizes are different. Considering the best MAE was achieved by the network trained on the largest number of subjects (16148 patients), there is still potential benefit from increasing the training set size for our network. 2) Different disease has different pathogenesis, leading to different difficulties in learning the relation between function and structure. In SSc patients, for example FVC may remain stable while DLCO significantly decline over time [28]. Therefore, compared with previous work, which only estimate FVC and FEV_1, our work estimates a more comprehensive set of PFT measurements (DLCO, FEV_1, FVC and TLC) for SSc patients, rendering it more clinically relevant for SSC patients, that is likely of additional clinical value. The comparison is for reference only, since it is based on different datasets sizes, different networks and different diseases. Implementing the two methods on our dataset to have an absolute fair comparison is impractical because the other two methods did not have public available pre-trained weights as what we have for X3D from Kinetics dataset.

The observed correlation between α, β and PFT in our study differs from the original report [13]. This is because patients with lung fibrosis were excluded in the original report, whereas our dataset comprised 80% CT images with various degrees of fibrosis. In patients with lung fibrosis, fibrotic areas led to over-segmentation of vessels, decreasing the correlation between the α and β calculated and PFT.

Estimating TLC was consistently more successful than for the other three PFT measurements. The MAE and prediction uncertainty in percentage (represented by MAPE) of TLC are always lower than the others. This could be explained by the fact that lung volume calculated by simply counting the number of voxels in both lungs is already strongly correlated with TLC, as measured by spirometry [29]. The estimation of DLCO consistently underperformed compared to the other three measures, since gas exchange is less correlated with TLC.

While the agreement between estimated PFTs and the measured PFTs ranged from moderate to good, the agreement between directly estimated PFTs%pred and measured PFT%pred ranged from poor to moderate. This finding is consistent with a previous report [11]. This can be attributed to the challenge of estimating reference equations for diverse population groups. Therefore we proceeded our research on estimating absolute PFTs, because 1) estimating PFTs%pred indirectly was more accurate than a direct estimation; and 2) this approach is more flexible as other PFT biomarkers, such as FEV_1/FVC [1] and FVC/DLCO [30], can then also be derived from the estimated absolute PFT values.

From the comparison between networks trained by different regions of CT scans, we found that networks trained on the whole CT image could achieve the best performance. CT masked by both lungs produced slightly inferior results, suggesting that tissue outside the lung area still contribute to the estimation of PFTs to some extent. This observation could be verified by Figure 8, where some regions outside lungs are highlighted for Patient 2 in the first row (network trained on the whole CT) while regions outside lungs are not highlighted for Patient 1. This suggests that the interaction between the chest wall and intercostal muscles contribute to PFTs in some patients. This is consistent with the clinical knowledge that stronger intercostal muscles combined with a compliant chest wall will have a positive effect on PFTs [31]. In contrast, chest wall stiffness, as sometimes observed in patients with SSc, may negatively influence PFTs. RAMs of different networks trained on different regions of CT have similar patterns: the entire lung is highlighted to different extents while the center regions of lungs are highlighted mostly. This implies that the networks for estimating PFTs need global information of the whole CT, while focusing more on the center regions where the largest vessels are located. This is consistent with findings in the previous study [11]. Apparently, it would be of greater clinical value if we could further extract what the contributors are. However, limited by the low resolution of current RAM techniques, we could not give more detailed contributors. Because a visualization centered on model interpretability would bridge the divide between AI-driven analyses and clinical practitioners, we will research more detailed visualization methods in our future work.

It is surprising that networks trained solely on grayscale vessels or binarized vessel masks still achieved R and ICC values over 0.5 for all four PFT measures. This implies that, in addition to vessel radius histogram information (used by α and β), the spatial structure of the vascular tree plays a more significant role in estimating PFTs.

Currently, there are no established guidelines for the level of precision required to implement new techniques in clinical practice for predicting PFTs. The repeatability criteria to measure PFTs is the standard for spirometry, which is the upper limit of any methods which aim to replace spirometry. At the current stage, our method could not perform competitively with spirometry if we compare our MAE with the repeatability criteria of spirometry. In addition, our method has not been prospectively validated, so it can only be used in research at present. Nevertheless, our method still 1) verified the possibility to estimate PFT, especially DLCO, from CT scans for patients of systemic sclerosis. 2) paved the way for more accurate methods and foster medical community to establish standards and regulations for such methods in the future. It would be beneficial to witness its integration into the clinical (randomized) trials in the future.

The multiple variable regression analysis showed that previous manually designed biomarkers could further explain variation in PFTs. This observation implies that if we add manually designed biomarkers as extra input to the networks, we might improve networks further in future research.

There are some limitations to our research. Because of the lack of public available 3D network weights pre-trained by lung CT images, we applied TL from Kinetics dataset, which may not optimal for PFTs estimation. In the future, we will explore the potential of network weights pre-trained from lung CT scans once we have the access to large annotated lung CT datasets. In addition, due to the need to protect healthy individuals from radiation exposure, it is not feasible to design a prospective experiment to collect CT and PFT pairs for a healthy control group. As a result, the retrospectively collected CT-PFT pairs in our study do not include healthy participants. Consequently, it remains uncertain whether our trained network can be applied to distinguish lung-structure relations in SSc patients from those in healthy individuals. Moreover, all SSc patients in this study were scanned with the same scanner at the same center. Therefore, additional experiments involving other patients and scanners are necessary to verify the generalizability of our deep learning method in the future. To achieve optimal performance for new scanners, we may need to fine-tune our model based on new datasets. If more image modalities are available in the future, we can explore the potential scalability of our methods on other modalities. Therefore, external validation is needed to be imbedded in clinical (randomized) trials. For now, the method can only be used for clinical research.

In conclusion, our method can automatically and comprehensively estimate PFTs for SSc patients. This can help to estimate lung function for patients who are unable to perform these tests, while there are CT scans available. The method can form a basis for studying the relation between function and structure, since we found for example that regions outside the lungs also contribute to the estimation of PFTs. For future work, we will investigate how to extract the contributors outside the lungs in more detail, which would be of great clinical value.

ACKNOWLEDGMENT

This work was performed using the computational resources from the Academic Leiden Interdisciplinary Cluster Environment (ALICE) provided by Leiden University.

REFERENCES

[1] J. Behr and D. E. Furst, "Pulmonary function tests," *Rheumatology*, vol. 47, no. 5, pp. v65–v67, Oct. 2008, doi: 10.1093/rheumatology/ken313.

[2] M. Caron, S. Hoa, M. Hudson, K. Schwartzman, and R. Steele, "Pulmonary function tests as outcomes for systemic sclerosis interstitial lung disease," *Eur. Respiratory Rev.*, vol. 27, no. 148, Jun. 2018, Art. no. 170102, doi: 10.1183/16000617.0102-2017.

[3] M. K. Ninaber, J. Stolk, J. Smit, E. J. Le Roy, L. J. M. Kroft, M. E. Bakker, J. K. de Vries Bouwstra, A. A. Schouffoer, M. Staring, and B. C. Stoel, "Lung structure and function relation in systemic sclerosis: Application of lung densitometry," *Eur. J. Radiol.*, vol. 84, no. 5, pp. 975–979, May 2015, doi: 10.1016/j.ejrad.2015.01.012.

[4] B. L. Graham, I. Steenbruggen, M. R. Miller, I. Z. Barjaktarevic, B. G. Cooper, G. L. Hall, T. S. Hallstrand, D. A. Kaminsky, K. McCarthy, M. C. McCormack, C. E. Oropez, M. Rosenfeld, S. Stanojevic, M. P. Swanney, and B. R. Thompson, "Standardization of spirometry 2019 update. An official American thoracic society and European respiratory society technical statement," *Amer. J. Respiratory Crit. Care Med.*, vol. 200, no. 8, pp. e70–e88, Oct. 2019, doi: 10.1164/rccm.201908-1590st.

[5] Y. S. Choi, J. Oh, S. Ahn, Y. Hwangbo, and J.-H. Choi, "Automated pulmonary function measurements from preoperative CT scans with deep learning," in *Proc. IEEE-EMBS Int. Conf. Biomed. Health Informat. (BHI)*, Sep. 2022, pp. 1–4.

[6] A. McGowan et al., "International consensus on lung function testing during the COVID-19 pandemic and beyond," *ERJ Open Res.*, vol. 8, no. 1, p. 602, Jan. 2022, doi: 10.1183/23120541.00602-2021.

[7] B. G. Cooper, "An update on contraindications for lung function testing," *Thorax*, vol. 66, no. 8, pp. 714–723, Aug. 2011, doi: 10.1136/thx.2010.139881.

[8] H. Meng, Y. Liu, X. Xu, Y. Liao, H. Liang, and H. Chen, "A machine learning approach for preoperatively assessing pulmonary function with computed tomography in patients with lung cancer," *Quantum Imag. Med. Surg.*, vol. 13, no. 3, pp. 1510–1523, Mar. 2023.

[9] K. P. Sylvester, N. Clayton, I. Cliff, M. Hepple, A. Kendrick, J. Kirkby, M. Miller, A. Moore, G. F. Rafferty, L. O'Reilly, J. Shakespeare, L. Smith, T. Watts, M. Bucknall, and K. Butterfield, "ARTP statement on pulmonary function testing 2020," *BMJ Open Respiratory Res.*, vol. 7, no. 1, Jul. 2020, Art. no. e000575, doi: 10.1136/bmjresp-2020-000575.

[10] J. Jia, M. Staring, I. H. Girón, L. J. Kroft, A. A. Schouffoer, and B. C. Stoel, "Prediction of lung CT scores of systemic sclerosis by cascaded regression neural networks," *Proc. SPIE*, vol. 12033, pp. 837–843, Apr. 2022, doi: 10.1117/12.2602737.

[11] H. Park, J. Yun, S. M. Lee, H. J. Hwang, J. B. Seo, Y. J. Jung, J. Hwang, S. H. Lee, S. W. Lee, and N. Kim, "Deep learning–based approach to predict pulmonary function at chest CT," *Radiology*, vol. 307, no. 2, pp. 1–12, Apr. 2023, doi: 10.1148/radiol.221488.

[12] N. S. L. Goh, S. Veeraraghavan, S. R. Desai, D. Cramer, D. M. Hansell, C. P. Denton, C. M. Black, R. M. Du Bois, and A. U. Wells, "Bronchoalveolar lavage cellular profiles in patients with systemic sclerosis–associated interstitial lung disease are not predictive of disease progression," *Arthritis Rheumatism*, vol. 56, no. 6, pp. 2005–2012, Jun. 2007, doi: 10.1002/art.22696.

[13] Z. Zhai, M. Staring, M. K. Ninaber, J. K. D. Vries-Bouwstra, A. A. Schouffoer, L. J. Kroft, J. Stolk, and B. C. Stoel, "Pulmonary vascular morphology associated with gas exchange in systemic sclerosis without lung fibrosis," *J. Thoracic Imag.*, vol. 34, no. 6, pp. 373–379, Nov. 2019, doi: 10.1097/rti.0000000000000395.

[14] M. R. Miller, "Standardisation of spirometry," *Eur. Respiratory J.*, vol. 26, no. 2, pp. 319–338, Aug. 2005, doi: 10.1183/09031936.05.00034805.

[15] B. L. Graham, V. Brusasco, F. Burgos, B. G. Cooper, R. Jensen, A. Kendrick, N. R. MacIntyre, B. R. Thompson, and J. Wanger, "2017 ERS/ATS standards for single-breath carbon monoxide uptake in the lung," *Eur. Respiratory J.*, vol. 49, no. 1, Jan. 2017, Art. no. 1600016, doi: 10.1183/13993003.00016-2016.

[16] G. L. Hall, N. Filipow, G. Ruppel, T. Okitika, B. Thompson, J. Kirkby, I. Steenbruggen, B. G. Cooper, and S. Stanojevic, "Official ERS technical standard: Global lung function initiative reference values for static lung volumes in individuals of European ancestry," *Eur. Respiratory J.*, vol. 57, no. 3, Mar. 2021, Art. no. 2000289, doi: 10.1183/13993003.00289-2020.

[17] S. Stanojevic, B. L. Graham, B. G. Cooper, B. R. Thompson, K. W. Carter, R. W. Francis, and G. L. Hall, "Official ERS technical standards: Global lung function initiative reference values for the carbon monoxide transfer factor for Caucasians," *Eur. Respiratory J.*, vol. 50, no. 3, Sep. 2017, Art. no. 1700010, doi: 10.1183/13993003.00010-2017.

[18] P. H. Quanjer, S. Stanojevic, T. J. Cole, X. Baur, G. L. Hall, B. H. Culver, P. L. Enright, J. L. Hankinson, M. S. M. Ip, J. Zheng, and J. Stocks, "Multi-ethnic reference values for spirometry for the 3–95-yr age range: The global lung function 2012 equations," *Eur. Respiratory J.*, vol. 40, no. 6, pp. 1324–1343, Dec. 2012, doi: 10.1183/09031936.00080312.

[19] Z. Zhai, M. Staring, and B. C. Stoel, "Lung vessel segmentation in CT images using graph-cuts," *Proc. SPIE*, vol. 9784, Mar. 2016, Art. no. 97842K, doi: 10.1117/12.2216827.

[20] V. Cheplygina, "Cats or CAT scans: Transfer learning from natural or medical image source data sets?" *Current Opinion Biomed. Eng.*, vol. 9, pp. 21–27, Mar. 2019, doi: 10.1016/j.cobme.2018.12.005.

[21] C. Feichtenhofer, "X3D: Expanding architectures for efficient video recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 200–210.

[22] W. Kay. *The Kinetics Human Action Video Dataset*. Accessed: Apr. 27, 2023. [Online]. Available: http://deepmind

[23] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626. Accessed: Sep. 19, 2021. [Online]. Available: http://gradcam.cloudcv.org

[24] Z. Wang and J. Yang. *Diabetic Retinopathy Detection via Deep Convolutional Networks for Discriminative Localization and Visual Explanation*. Accessed: Oct. 11, 2022. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs312/en/

[25] P. Schober, C. Boer, and L. A. Schwarte, "Correlation coefficients: Appropriate use and interpretation," *Anesthesia Analgesia*, vol. 126, no. 5, pp. 1763–1768, May 2018, doi: 10.1213/ane.0000000000002864.

[26] R. Vallat, "Pingouin: Statistics in Python," *J. Open Source Softw.*, vol. 3, no. 31, p. 1026, Nov. 2018, doi: 10.21105/joss.01026.

[27] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *J. Chiropractic Med.*, vol. 15, no. 2, pp. 155–163, Jun. 2016, doi: 10.1016/j.jcm.2016.02.012.

[28] N. Le Gouellec, A. Duhamel, T. Perez, A.-L. Hachulla, V. Sobanski, J.-B. Faivre, S. Morell-Dubois, M. Lambert, P.-Y. Hatron, E. Hachulla, H. Béhal, R. Matran, D. Launay, and M. Remy-Jardin, "Predictors of lung function test severity and outcome in systemic sclerosis-associated interstitial lung disease," *PLoS ONE*, vol. 12, no. 8, Aug. 2017, Art. no. e0181692, doi: 10.1371/journal.pone.0181692.

[29] S. Iwano, T. Okada, H. Satake, and S. Naganawa, "3D-CT volumetry of the lung using multidetector row CT," *Academic Radiol.*, vol. 16, no. 3, pp. 250–256, Mar. 2009, doi: 10.1016/j.acra.2008.09.019.

[30] E. A. Vallejos, J. Martinez, F. Cabrera, N. Mastandrea, M. Pertuz, Z. Marengo, M. J. L. Meiller, V. Volberg, and R. G. Tejada, "Association of FVC/DLCO with pulmonary hypertension risk and interstitial disease in systemic sclerosis patients," *Eur. Respir. J.*, vol. 56, no. 64, p. 302, Sep. 2020, doi: 10.1183/13993003.CONGRESS-2020.302.

[31] S. J. Park, "Effects of inspiratory muscles training plus rib cage mobilization on chest expansion, inspiratory accessory muscles activity and pulmonary function in stroke patients," *Appl. Sci.*, vol. 10, no. 15, p. 5178, Jul. 2020, doi: 10.3390/app10155178.

JINGNAN JIA received the bachelor's degree in applied physics from the Taiyuan University of Technology, China, and the master's degree in electromagnetic and microwave technology from Xidian University. He is currently pursuing the Ph.D. degree with the Division of Image Processing (LKEB), Leiden University Medical Center (LUMC). His current research interests include deep learning on medical image analysis, CT segmentation, and biomarker regression.

EMIEL R. MARGES received the degree in pulmonologist from the Leiden University Medical Center (LUMC), Leiden, The Netherlands, in 2022, where he is currently pursuing the Ph.D. degree in early detection of systemic sclerosis-interstitial lung disease collaboration with the Department of Rheumatology. He is a Pulmonologist with the Department of Respiratory Medicine, LUMC.

JESKA K. DE VRIES-BOUWSTRA is currently a Rheumatologist and an Associate Professor in systemic sclerosis with the Department of Rheumatology, Leiden University Medical Center (LUMC). She is also the Head of the Care Pathway in Systemic Sclerosis and a PI of the Ongoing Prospective Cohort Study: Combined Care in Systemic Sclerosis (CCISS), LUMC. Her research interest includes increasing the understanding of the clinical heterogeneity in SSc, to improve patient stratification, and enabling timely treatment of disease complications.

MAARTEN K. NINABER is currently a Chest Physician and the Director of the Pulmonary Residency Program. He is also involved in the development of management recommendations and conduction of clinical trials. His main research interests include the diagnosis (including invasive diagnostic techniques, such as EBUS, EUS, and lung cryobiopsy), treatment, and clinical research related to systemic sclerosis, and other CTDs with pulmonary involvement (interstitial and vascular compartment of the lung).

LUCIA J. M. KROFT received the medical and Ph.D. degrees from Leiden University, in 1994 and 1999, respectively. She has been registered as a Radiologist and a Staff Member with the Leiden University Medical Center (LUMC), since 2004. She is scientifically active and authored/coauthored 245 peer-reviewed scientific articles and 20 book chapters. Her sub-specializations are cardiovascular and thoracic radiology.

ANNE A. SCHOUFFOER received the Ph.D. degree from Leiden University, in 2014. She has been trained as a Rheumatologist with the Leiden University Medical Centre (LUMC), The Netherlands. She is currently with the Haga Teaching Hospital, The Hague, The Netherlands, and LUMC.

MARIUS STARING is currently a Professor of machine learning for medical imaging with the Department of Radiology, Leiden University Medical Center (LUMC), and the Vice Director of the Division of Image Processing (Dutch abbreviation LKEB). He is a member of the Program Committees of MICCAI, IEEE ISBI, SPIE Medical Imaging, and WBIR, and an Associate Editor of IEEE Transactions on Medical Imaging.

BEREND C. STOEL received the M.Sc. degree in medical informatics from Leiden University, Leiden, The Netherlands, in 1989, and the Ph.D. degree in objective assessment of X-ray image quality from the Leiden University Medical Center (LUMC), Leiden, in 1996. He is currently an Associate Professor with the Division of Image Processing, Department of Radiology, LUMC, where he is the Head of the Section of Orthopaedics and Pulmonology. His current research interests include pulmonology, particularly quantification of emphysema, pulmonary fibrosis and embolisms, and bronchial tree quantification, with a focus on clinical validation and applications in drug evaluation trials.