Deep Learning-Based Air Trapping Quantification Using Paired Inspiratory-Expiratory Ultra-low Dose CT

Sarah M. Muller[1,2,3](✉), Sundaresh Ram[4,5], Katie J. Bayfield[6], Julia H. Reuter[1,2,3], Sonja Gestewitz[1,2,3], Lifeng Yu[7], Mark O. Wielpütz[1,2,3], Hans-Ulrich Kauczor[1,2,3], Claus P. Heussel[2,3], Terry E. Robinson[8], Brian J. Bartholmai[7], Charles R. Hatt[5,9], Paul D. Robinson[6], Craig J. Galban[4,5], and Oliver Weinheimer[1,2,3]

[1] Department of Diagnostic and Interventional Radiology, University Hospital Heidelberg, Heidelberg, Germany
sarah.muller@med.uni-heidelberg.de
[2] Translational Lung Research Center Heidelberg (TLRC), German Center for Lung Research (DZL), University of Heidelberg, Heidelberg, Germany
[3] Department of Diagnostic and Interventional Radiology with Nuclear Medicine, Thoraxklinik at University of Heidelberg, Heidelberg, Germany
[4] Department of Biomedical Engineering, University of Michigan, Ann Arbor, MI, USA
[5] Department of Radiology, University of Michigan, Ann Arbor, MI, USA
[6] Department of Respiratory Medicine, The Children's Hospital at Westmead, Westmead, New South Wales, Australia
[7] Department of Radiology, Mayo Clinic, Rochester, MN, USA
[8] Center for Excellence in Pulmonary Biology, Department of Pediatrics, Stanford University Medical Center, Palo Alto, CA, USA
[9] Imbio LLC, Minneapolis, MN, USA

Abstract. Air trapping (AT) is a frequent finding in early cystic fibrosis (CF) lung disease detectable by imaging. The correct radiographic assessment of AT on paired inspiratory-expiratory computed tomography (CT) scans is laborious and prone to inter-reader variation. Conventional threshold-based methods for AT quantification are primarily designed for adults and less suitable for children. The administered radiation dose, in particular, plays an important role, especially for children. Low dose (LD) CT is considered established standard in pediatric lung CT imaging but also ultra-low dose (ULD) CT is technically feasible and requires comprehensive validation. We investigated a deep learning approach to quantify air trapping on ULDCT in comparison to LDCT and assessed structure-function relationships by cross-validation against multiple breath washout (MBW) lung function testing. A densely connected convolutional neural network (DenseNet) was trained on 2-D patches to segment AT. The mean threshold from radiographic assessments, performed by two trained radiologists, was used as ground truth. A grid search was conducted to find the best parameter configuration. Quantitative AT (QAT), defined as the percentage of AT in the lungs

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14222, pp. 432–441, 2023.
https://doi.org/10.1007/978-3-031-43898-1_42

detected by our DenseNet models, correlated strongly between LD and ULD. Structure-function relationships were maintained. The best model achieved a patch-based DICE coefficient of 0.82 evaluated on the test set. AT percentages correlated strongly with MBW results (LD: $R = 0.76$, $p < 0.001$; ULD: $R = 0.78$, $p < 0.001$). A strong correlation between LD and ULD ($R = 0.96$, $p < 0.001$) and small ULD-LD differences (mean difference $-1.04 \pm 3.25\%$) were observed.

Keywords: Deep Learning · Air Trapping Quantification · Cystic Fibrosis

1 Introduction

Cystic fibrosis (CF) lung disease is a progressive respiratory condition. It originates from a defect in the cystic fibrosis trans-membrane conductance regulator (CFTR) gene which causes a mucociliary dysfunction and airway mucus plugging, provoking chronic neutrophilic airway inflammation [9]. One of the early marker of CF detectable by imaging is the pathological retention of air in the lungs after exhalation. It is commonly designated as air trapping (AT). AT visually constitutes as low attenuation areas in the lung parenchyma observable on expiratory computed tomography (CT) scans. Conventional threshold-based methods for air trapping quantification (e.g. the −856 HU threshold on expiratory CT) are density-based and primarily designed for adults [7]. They depend on the CT protocol in use and the constitution of the patient. The CT dose, in particular, plays an important role which raises the question about the influence of dose reduction on AT quantification. More personalized thresholds have been defined by Goris et al. who use the median and 90 percentile of the inspiratory histogram of densities together with the difference in the 90 percentile values between expiration and inspiration [5]. Nowadays, low dose (LD) CT is considered established standard in pediatric lung CT imaging. Ultra-low dose (ULD) CT has been implemented recently and requires comprehensive validation.

To address the aforementioned issues concerning AT quantification in children with CF, we investigated a deep learning approach to quantify air trapping on ULDCT in comparison to LDCT. Structure-function relationships were assessed by comparison against multiple breath washout (MBW) lung function testing. The applied deep learning method is adopted from the one-channel approach proposed by Ram et al. [13] who trained a densely connected convolutional neural network (DenseNet) on LDCT. It achieved a good AT quantification with respect to the ground truth derived from an algorithm developed by Goris et al. [5] for generating subject-specific thresholds. The structure-function relationships for the deep learning approach were investigated by Bayfield et al. [4], on the same ULD-LDCT dataset, which we now used to train our model. The authors observed that the percentage of AT, detected by the model from Ram et al., did not correlate with pulmonary function test results. In a research letter published in the European Respiratory Journal, Bayfield et al. [3] address the urgency of reliable AT quantification on ULDCT. For this reason, we investigated the influence of dose reduction on AT quantification. We aimed to achieve a good AT segmentation on ULD as well as LDCT while maintaining structure-function relationships. In this context, we examined the usage of one or two input channels and the training on one or both scan protocols.

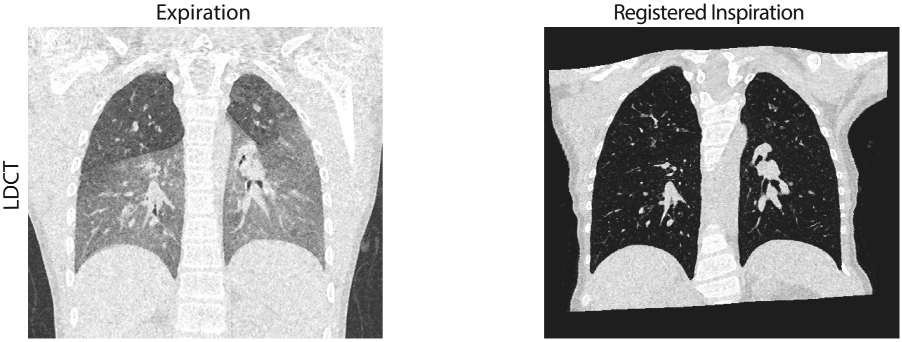

Fig. 1. Registration result for a representative LDCT case showing the registration of the inspiration to the corresponding expiration CT scan.

2 Methods

2.1 Data Acquisition

52 CF subjects with a mean age of 11.3 ± 3.6 years were included in this study. Paired spirometry-guided inspiratory-expiratory CT scans were acquired at LD (volume CT dose index ($\mathrm{CTDI_{vol}}$) 1.22 ± 0.56 mGy) and ULD ($\mathrm{CTDI_{vol}}$ 0.22 ± 0.05 mGy) in the same session. Between the two scan settings, the effective dose was reduced by 82%. Patients were scanned using a tube potential of 100 kVp with an added tin filter. The tin filter cuts out lower-energy photons and improves radiation dose efficiency. For the iterative reconstruction, a Br49d kernel with strength level 3 and a slice thickness of 0.6 mm with 0.3 mm incremental overlap were used. The lung clearance index (LCI) was determined by performing nitrogen-based multiple breath washout. It is determined as the number of lung volume turnovers required to clear the lung of the nitrogen [16]. Over the years, studies have indicated that pulmonary function test results obtained by spirometry do not always correlate well with the medical condition of children with CF, more precisely, that the forced expiratory volume in 1 s ($\mathrm{FEV_1}$) has a low sensitivity and the LCI is a better indicator of structural lung abnormalities [2,6].

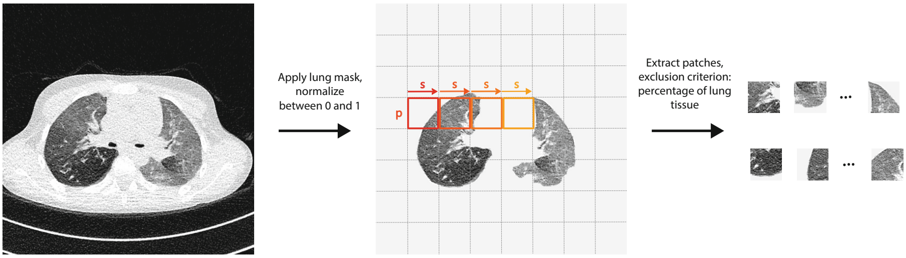

Fig. 2. Post-processing pipeline after the registration step. Patch creation using sliding window with patch size p and stride s for an exemplary choice of $p = s$.

2.2 Post-processing

The inspiratory CT scan was registered to the expiratory CT scan by applying 3-D deformable image registration using the image registration tool Elastix (version 5.0.1) [8,15]. A representative registration result is presented in Fig. 1. The subsequent post-processing steps are outlined in Fig. 2. Lung masks were generated using the CT analysis software YACTA [18]. The generated lung masks were applied to the CT images to permit an undeflected focus on the lungs. All CT images were normalized between zero and one. The mean threshold from radiographic assessments, performed separately by two trained radiologists, was used as ground truth. To generate the ground truth segmentation, the radiologist loaded the inspiratory and corresponding expiratory CT scan in our inhouse software. After loading, the scans were displayed next to each other where the radiologist could go through each of them individually. The segmentation was not drawn manually by the radiologist. Instead, we used a patient-specific threshold T. An AT map was generated by classifying all expiratory CT voxels < T as AT. Using an integrated slider functionality, the radiologist was asked to choose T for each patient such that the AT map best describes the trapped air. Since a manual AT assessment is very time-consuming, the slider-based approach provides a good trade-off between time consumption and accuracy. With this technique, we are able to guarantee a high ground truth quality since two trained radiologists selected a personalized threshold for each patient and no generic method was used. 2-D patches were created from the 2-D axial slices of the expiratory, the corresponding registered inspiratory CT scan and the ground truth segmentation. A sliding window with a patch size p of 32 and stride s was used as demonstrated in Fig. 2. The generated patches were then utilized as one- or two-channel input to a densely connected convolutional neural network which was trained to segment AT.

2.3 DenseNet Architecture and Training

A sketch of the DenseNet architecture is presented in Fig. 3. It has a common u-net structure [14] and is adopted from Ram et al. [13]. Each dense block contains four dense block layers where each dense block layer consists of a batch

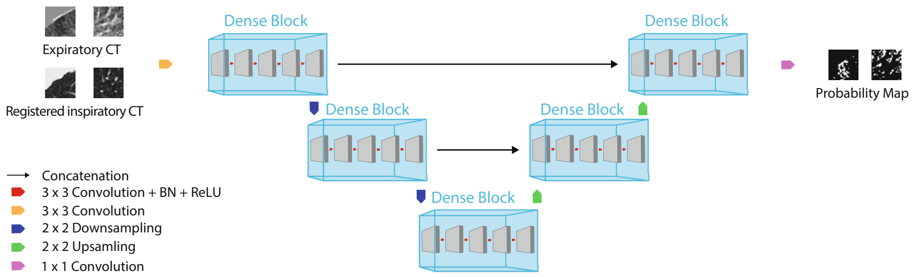

Fig. 3. Sketch of the proposed DenseNet architecture adapted from Ram et al. [13]. 2-D patches of the expiration and the corresponding registered inspiration CT scan have been used as input to the network. The mean of the radiographic assessments performed by two trained radiologists was used as ground truth.

normalization (BN), a rectified linear unit (ReLU) activation function and a 3 × 3 convolution. For downsampling, BN, a ReLU activation, a 1 × 1 convolution and a 2 × 2 max pooling operation with a stride of two are used. Upsampling is implemented using a 3 × 3 transposed convolution with a stride of two. In the last layer, a 1 × 1 convolution and a softmax operation are performed to obtain the output probability map. The network was trained to minimize the Dice loss [10,17]. Weights were optimized using stochastic gradient decent (Momentum) [12] with a momentum term of $\beta = 0.9$ and a learning rate of $\alpha = 0.001$. The implementation is done in PyTorch [11] (version 1.12.1) using Python version 3.10 (Python Software Foundation, Python Language Reference, http://www.python.org). No augmentation has been used. In contrast to Ram et al. [13], the registered inspiratory CT scan was added as second input channel to the network and the model was trained on 2-D patches instead of slices. We argue that, as for the radiologist, using the inspiration CT as second input can add useful information in the form of inspiration expiration differences. Patches were selected over slices to increase the number of training samples. The subsequent experiments were executed on a high performance computing cluster. On the cluster, the method ran on a NVIDIA Tesla V100 HBM2 RAM with 32 GB of memory. A batch size of 512 was used. The dataset was divided into a training and test set based on a 80 : 20 ratio using 41 patients for training/validation and 11 for testing. A grid search was performed on the training set using the hyperparameter optimization framework Optuna [1], aiming to find the best hyperparameter combination. Training was performed in 5-fold cross-validation where each model was trained for 100 epochs. Following parameters were varied during grid search: stride, number of channels and the scan protocols included in the training analyzing if using LDCT scans only, or both, LD and ULDCT scans, should be included in the training. Only patches containing at least 50% of lung tissue are considered to only include the most informative patches.

2.4 Model Evaluation

The DenseNet AT percentages used to compute the correlations presented in Table 1 are computed as the normalized sum of the DenseNet output probabilities using the best model in terms of DICE coefficient evaluated on the validation set. Since the network outputs a probability map, for computing the DICE coefficient over the patches, the output probabilities were converted to binary maps classifying all probabilities ≥ 0.5 as 1, and 0 otherwise. In the same way, the so-called quantitative AT (QAT) values which can be found in the upper right of the overlayed ground truth and DenseNet output images of Fig. 4, were computed as the percentage of AT over the entire lung using the binary segmentation maps.

3 Results

For each model of each parameter combination obtained from cross-validation, the DICE coefficient was computed over the patches from the test set. Resulting mean DICE coefficients and standard deviations over the 5-folds for the different parameter combinations are presented in Table 1. The highest DICE coefficient was achieved when training the network on two channels using a stride of 16. Here, the best model achieved a score of 0.82. It was trained on both, LD and ULD. The scores did not differ noticeably (third decimal) when training on LD only (mean DICE coefficient 0.806 ± 0.213) compared to including both scan protocols, LD and ULD (0.809 ± 0.216).

Analyzing the correlations of the percentage of AT in the lungs detected by the best DenseNet model of the five folds between LD and ULD, strong correlations and small ULD-LD differences become apparent for all tested parameter

Table 1. Grid search results for the different parameter combinations evaluated on the test set. Mean DICE coefficent and standard deviation (evaluated on patches) computed over the five models. Mean ULD-LD difference of the percentage of AT obtained from the best DenseNet model of the five folds, presented together with the correlations (Pearson's R) between LD and ULD and with the LCI for LD and ULD ($p < 0.001$ for all R).

Stride	# channels	LD only	DICE (patches)	DenseNet AT ULD-LD		DenseNet AT-LCI	
				difference [%]	R	R (LD)	R (ULD)
32	1	true	0.773 ± 0.215	0.137 ± 3.066	0.977	0.755	0.751
32	2	true	0.785 ± 0.219	-0.876 ± 2.983	0.972	0.769	0.745
32	1	false	0.790 ± 0.209	-0.476 ± 3.090	0.963	0.729	0.737
32	2	false	0.801 ± 0.215	-0.670 ± 3.186	0.963	0.766	0.781
16	1	true	0.793 ± 0.213	-0.324 ± 3.217	0.956	0.722	0.706
16	2	true	$\mathbf{0.806 \pm 0.213}$	-0.715 ± 3.322	0.956	0.752	0.748
16	1	false	0.793 ± 0.216	-0.682 ± 3.200	0.959	0.729	0.747
16	2	false	$\mathbf{0.809 \pm 0.216}$	-1.040 ± 3.254	0.962	0.757	0.776

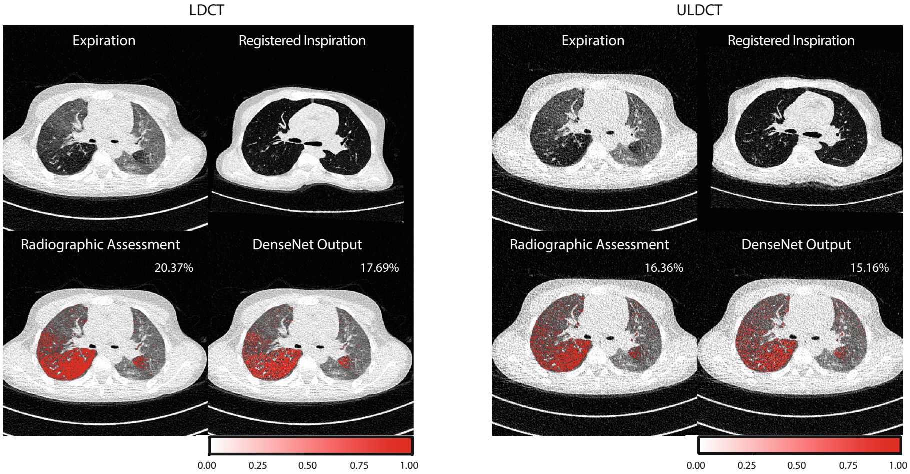

Fig. 4. Representative slices of the expiration, corresponding registered inspiration, the segmentation ground truth (averaged radiographic assessments), and the DenseNet output probability map for a LD and the corresponding ULD CT scan. The DenseNet was trained on LD only using a two-channel input and a stride of 32. QAT values are displayed in white in the upper right of the corresponding images. Window level: −400 HU, window width: 1100 HU.

combinations (Table 1). The table shows that the mean ULD-LD difference is lower when training with one compared to two channels. Comparing the correlation of the AT percentage in the lungs detected by the best DenseNet models with the LCI for LD and ULD (Table 1), the strongest correlations were achieved when training LD and ULD together using a 2-channel input. Applying a stride of 32 results in a strong correlation for LD ($R = 0.77$, $p < 0.001$) and ULD ($R = 0.78$, $p < 0.001$) which can be also achieved using overlapping patches with a stride of 16 (LD ($R = 0.76$, $p < 0.001$) and ULD ($R = 0.78$, $p < 0.001$)). A strong correlation between LD and ULD ($R = 0.96$, $p < 0.001$) and small ULD-LD differences (mean difference $-1.04 \pm 3.25\%$) can be observed. Figure 4 shows the inputs, the corresponding ground truth segmentation, and the resulting DenseNet output probability map for a representative LD scan (left) and its corresponding ULD scan (right). The model used for evaluation was trained on LD only using a two-channel input and a stride of 32. Comparing the DenseNet output with the ground truth, a good AT segmentation can be observed for LD and ULD. The segmentation results demonstrate that although only trained on LD, a good AT quantification can also be obtained for ULD. The deep learning method detected less AT than the ground truth. This becomes apparent by looking at the QAT values, displayed in white, in the upper right of the overlayed images. However, the difference between the DenseNet and the ground truth segmentations is small for LD (QAT value difference: −2.68%) as well as ULD (QAT value difference: −1.2%).

Table 2. Mean DICE coefficent and standard deviation (evaluated on the 512 × 512 CT scan slices) for different AT quantification methods. Mean ULD-LD difference of the percentage of AT detected by the corresponding method presented together with the correlations (Pearson's R) between LD and ULD and the LCI for LD and ULD ($p < 0.001$ for all R).

Method	DICE (slices)	Method AT ULD-LD		Method AT-LCI	
		difference [%]	R	R (LD)	R (ULD)
−856 HU	0.379 ± 0.230	1.649 ± 0.661	0.989	0.892	0.885
Goris et al. [5]	0.645 ± 0.171	9.126 ± 4.947	0.908	0.909	0.852
Radiographic assessment	-	1.761 ± 2.837	0.960	0.877	0.92
DenseNet	0.837 ± 0.130	−1.040 ± 3.254	0.962	0.757	0.776

As presented in Table 2, the conventional −856 HU threshold has only a small overlap with the ground truth radiographic assessment resulting in a low slice-based DICE coefficient of 0.379 ± 0.230, suggesting that this threshold is less suitable for children. The subject-specific threshold method from Goris et al. [5], on the other hand, achieves a noticeably higher DICE coefficient of 0.645 ± 0.171. The best DenseNet model achieves the highest DICE coefficient of the compared methods but less strong correlations with the LCI.

4 Discussion

In this study, we trained a densely connected convolutional neural network to segment AT using 2-D patches of the expiratory and corresponding registered inspiratory CT scan slices. We wanted to evaluate the best settings and the effect of a noticeable dose reduction on AT quantification.

Using a smaller stride and respectively more patches only resulted in a slightly higher patch-based DICE coefficient evaluated on the test set (Table 1). No noticeable increase in DICE coefficient could be observed for the models trained on both scan protocols or two input channels, compared to the reference. Only small differences were observed regarding the correlations of the DenseNet AT percentage between LD and ULD, and the DenseNet AT with the LCI for LD and ULD. Since furthermore only small ULD-LD differences were observed, the study indicates that the ULD scan protocol allows a comparable air trapping quantification, in comparison to the standard. Good correlations and small ULD-LD differences are also obtained with the models trained on LD only which proposes that training the model on both, LDCT and ULDCT is not necessarily needed. Adding the registered inspiratory scan as a second input channel to the network shows slight improvements in AT detection compared to a 1-channel approach. Future work will investigate to what extent an improved image registration can further improve the performance of the 2-channel approach. The comparison with other AT quantification methods, more precisely, the largest agreement with the radiographic ground truth segmentation in combination with a less strong correlation with the LCI might eventually indicate that the deep learning method is more sensible and detects structural impairment when function test results are still in a normal range (Table 2). Furthermore, it highlights the importance to distinguish AT severities when comparing structure-function relationships.

It is important to note the limitations of this study. First of all, it should be mentioned that generating ground truth is a difficult task even for experienced radiologists, which is not always clearly solvable. In addition, the results presented are limited to the available number of patients. Children were scanned at inspiration and expiration, with two different scan protocols, without leaving the CT table. This results in four scans for each patient and explains the limited availability of patients to be included in the study. The particularity of the dataset clarifies why the model could not easily be tested on an independent test dataset since there is none available obtained in a comparable manner.

5 Conclusion

We were able to show that similar QAT indices can be calculated on ULD CT images despite an 82% reduced dose. QAT values were comparable for ULD and LD across all parameter combinations. The relationship to the LCI was retained. AT is not only an early sign of incipient pulmonary dysfunction in patients with CF, but also in other diseases such as COPD or asthma. We want to investigate how our DenseNet performs on other data sets of patients with CF, COPD or asthma and, if necessary, expand the amount of training data.

Acknowledgements. The authors acknowledge support by the state of Baden-Württemberg through bwHPC. This study was supported by the Australian Cystic Fibrosis Research Trust - 2018 Innovation Grant (PDR), the National Institutes of Health and Cystic Fibrosis Foundation (CJG) and grants from the German Federal Ministry of Education and Research to OW and MOW (82DZL004A1).

References

1. Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: a next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, pp. 2623–2631. Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3292500.3330701
2. Aurora, P., et al.: Multiple breath inert gas washout as a measure of ventilation distribution in children with cystic fibrosis. Thorax **59**(12), 1068–1073 (2004)
3. Bayfield, K.J., et al.: Implementation and evaluation of ultra-low dose CT in early cystic fibrosis lung disease. Eur. Respir. J. **62**, 2300286 (2023). https://doi.org/10.1183/13993003.00286-2023
4. Bayfield, K., et al.: Deep learning improves the detection of ultra-low-dose CT scan parameters in children with cystic fibrosis. In: TP125: TP125 Structure and Function in the Pediatric Lung, pp. A4628–A4628. American Thoracic Society (2021)

5. Goris, M.L., Zhu, H.J., Blankenberg, F., Chan, F., Robinson, T.E.: An automated approach to quantitative air trapping measurements in mild cystic fibrosis. Chest **123**(5), 1655–1663 (2003). https://doi.org/10.1378/chest.123.5.1655. https://www.sciencedirect.com/science/article/pii/S0012369215337028
6. Gustafsson, P.M., De Jong, P.A., Tiddens, H.A., Lindblad, A.: Multiple-breath inert gas washout and spirometry versus structural lung disease in cystic fibrosis. Thorax **63**(2), 129–134 (2008)
7. Hersh, C.P., et al.: Paired inspiratory-expiratory chest CT scans to assess for small airways disease in COPD. Respir. Res. **14**, 42 (2013)
8. Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P.W.: elastix: a toolbox for intensity-based medical image registration. IEEE Trans. Med. Imaging **29**(1), 196–205 (2010)
9. Mall, M.A., Hartl, D.: CFTR: cystic fibrosis and beyond. Eur. Respir. J. **44**(4), 1042–1054 (2014). https://doi.org/10.1183/09031936.00228013
10. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571. IEEE (2016)
11. Paszke, A., et al.: Automatic differentiation in PyTorch. In: 31st Conference on Neural Information Processing Systems (NIPS 2017) (2017)
12. Qian, N.: On the momentum term in gradient descent learning algorithms. Neural Netw. **12**(1), 145–151 (1999). https://doi.org/10.1016/S0893-6080(98)00116-6
13. Ram, S., et al.: Improved detection of air trapping on expiratory computed tomography using deep learning. PLoS ONE **16**(3), e0248902 (2021)
14. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
15. Shamonin, D.P., Bron, E.E., Lelieveldt, B.P., Smits, M., Klein, S., Staring, M.: Fast parallel image registration on CPU and GPU for diagnostic classification of Alzheimer's disease. Front. Neuroinform. **7**(50), 1–15 (2014)
16. Subbarao, P., et al.: Multiple-breath washout as a lung function test in cystic fibrosis. A cystic fibrosis foundation workshop report. Ann. Am. Thorac. Soc. **12**(6), 932–939 (2015)
17. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Jorge Cardoso, M.: Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Cardoso, M.J., et al. (eds.) DLMIA/ML-CDS -2017. LNCS, vol. 10553, pp. 240–248. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67558-9_28
18. Weinheimer, O., Achenbach, T., Heussel, C.P., Düber, C.: Automatic lung segmentation in MDCT images. In: Fourth International Workshop on Pulmonary Image Analysis, Toronto, Canada, 18 September 2011, vol. 2011, pp. 241–255. CreateSpace (2011)