Prediction of Lung CT Scores of Systemic Sclerosis by Cascaded Regression Neural Networks

Jingnan Jia [ab], Marius Staring [ab], Irene Hernández-Girón [ab], Lucia J.M. Kroft [b], Anne A. Schouffoer [c], Berend C. Stoel* [ab]

[a] Division of Image Processing, [b] Department of Radiology, [c] Department of Rheumatology, Leiden University Medical Center (LUMC), P.O. Box 9600, 2300 RC, Leiden, The Netherlands.

ABSTRACT

Visually scoring lung involvement in systemic sclerosis (SSc) from CT scans plays an important role in monitoring progression, but its labor intensiveness hinders practical application. We proposed, therefore, an automatic scoring framework that consists of two cascaded deep regression neural networks. The first (3D) network aims to predict the craniocaudal position of five anatomically defined scoring levels on the 3D CT scans. The second (2D) network receives the resulting 2D axial slices and predicts the scores, which represent the extent of SSc disease. CT scans from 227 patients were used for both networks. 180 scans were split into four groups with equal number of samples to perform four-fold cross validation and an additional set of 47 scans constitute a separate testing dataset. Two experts scored all CT data in consensus and to obtain inter-observer variabilities they also scored independently 16 patients from the testing dataset. To alleviate the unbalance in training labels in the second network, we introduced a balanced sampling technique and to increase the diversity of the training samples, synthetic data was generated, mimicking ground glass and reticulation patterns. The four-fold cross validation results showed that our proposed score prediction network achieved an average MAE of 5.90, 4.66 and 4.49%, weighted kappa of 0.66, 0.58 and 0.65 for total score (TOT), ground glass (GG) and reticular pattern (RET), respectively. Our network performed slightly worse than the best human observation on TOT and GG prediction but it has competitive performance on RET prediction and has the potential to be an objective alternative for the visual scoring of SSc in CT thorax studies.

Keywords: Deep learning, Systemic sclerosis, Score prediction, Computed Tomography, Goh score

1. INTRODUCTION

Systemic sclerosis (SSc) is a rare immune-mediated connective tissue disease that affects different organs, with high mortality. Pulmonary disease is the leading cause of mortality in patients with SSc[1]. Because CT scans can provide accurate information on the lung, Goh[2] proposed a sensitive scoring system to quantify the extent of SSc disease from CT scans, further standardized by a CT reference atlas[3]. In this so-called Goh scoring system, 2D CT images are scored at five levels: 1) origin of the great vessels; 2) main carina; 3) pulmonary venous confluence; 4) halfway between the third and fifth level; 5) immediately above the right hemi-diaphragm. At each level, three patterns are scored as the percentage of the total lung area: total disease extent (TOT), ground-glass opacities (GG), and reticular patterns (RET). TOT is a powerful predictor for mortality of SSc[1], GG is a significant biomarker to differentiate SSc and idiopathic pulmonary fibrosis[4], and RET is a strong determinant of a decline in the forced vital capacity and progression-free survival of SSc[1]. Therefore, these three patterns are important markers to monitor disease progression.

The Goh scoring system is, however, laborious, subjective and dependent on the rater experience. Therefore, an automatic scoring tool is expected to overcome these limitations and benefit clinical application. Although several attempts have been made to automatically predict biomarkers[5,6], their common drawback is that they use segmentation networks to output the pixel labels as a basis for computing the final biomarkers, which requires time-consuming and laborious manual pixel-wise annotations as reference data.

The purpose of this paper is to build an automatic framework to identify the five levels and subsequently score the extent of disease from CT scans, without the need for segmentation annotations. A straightforward but effective data synthesis technique for the disease patterns (ground-glass opacities and reticulation) was introduced to alleviate the lack of training data and label unbalance.

2. METHODOLOGY

The proposed two-step framework is shown in Figure 1. In the first step, a *level selection network* (3D VGG11[7]) identifies five levels from the input 3D CT scans. Subsequently, a *Goh score prediction network* (2D VGG11[7]) predicts three scores (TOT, GG, and RET) for each input 2D slice. The 3D VGG11 has the same structure as the 2D VGG11 except that all 2D convolutions and pooling operations were replaced by the 3D convolutions and pooling operations, respectively. The pre-trained weights from ImageNet were applied to the 2D VGG11[8]. The loss function for both networks was the mean squared error (MSE) in predicting the manually annotated levels and manual Goh scores, respectively. The 3D VGG11 was trained on 3D CT scans for level selection, and the 2D VGG11 was trained on 2D slices of five levels for Goh score prediction. Both networks were executed on a machine equipped with a GPU NVIDIA GeForce RTX 2080TI with 11 GB memory. The source code is available at https://github.com/Jingnan-Jia/ssc_scoring. Datasets and the methodology details of the two networks are described below.

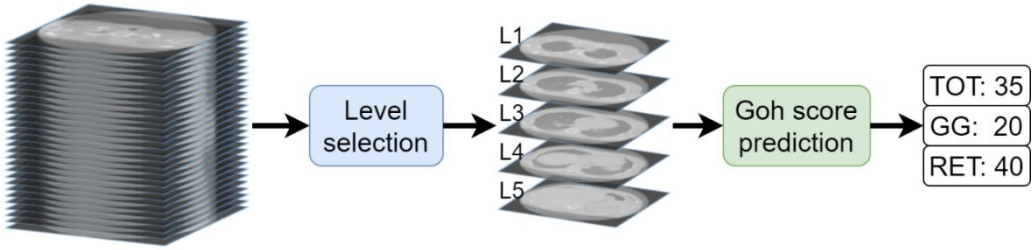


Figure 1. Proposed framework. 3D VGG11 outputs five numbers to select five axial slices. Subsequently, 2D VGG11 predicts three scores for each slice.

The 3D CT dataset consisted of 227 anonymized SSc patients, with a wide range of disease severity. All patients were scanned with the same CT scanner without contrast enhancement. All 3D CT scans were resized to a fixed size of 256×256×256 with a 1.2×1.2×1.2 mm voxel size. From the entire dataset, 180 scans were split into four groups with equal number of samples to perform four-fold cross validation and a separate set of 47 scans constitute the testing dataset. The craniocaudal world positions of the five levels and the Goh scores for each level were manually annotated by two experts in consensus. A rheumatologist (Obs1, over 5-year experience) and a radiologist (Obs2, over 20-year experience in chest CT) sat together and labeled each image. If they had different labels in mind for a particular image, they discussed their reasoning until they reached consensus. In addition, a subset of randomly selected 16 patients from the testing set was scored twice by Obs1 and Obs2 (with a one-month interval) and independently to estimate the inter-observer variation.

2.1 Level selection

We converted the world positions of the five levels to relative slice numbers in the resampled 3D CT scans (the bottom slice was regarded number 0). The relative slice numbers were used as the reference during training of the 3D VGG11. To increase the diversity of training samples, random crops with a fixed size of 256×256×192 (in *xyz*-direction) were applied on-the-fly. The patch always covered all five levels and was also suitable for the GPU memory constrains.

2.2 Goh score prediction

The 2D slices at five levels from the 227 patients were used for Goh score prediction. As shown in Figure 2, manual scores were estimated to the nearest five percent, following the protocol by Goh[9].

To alleviate the adverse effect of the label imbalance in the patient database, in which most labelled slices correspond to 0% scores (healthy) and very few to very high scores (Figure 2), we proposed two techniques: *balanced sampling* and *data synthesis*. *Balanced sampling* resamples the training images with a probability inversely proportional to the ratio of each score. It can make sure that the distribution of sampled scores is balanced, but its limitation is that there are a large number of images repeatedly sampled from the high-score categories. Therefore, a *data synthesis* method was proposed, by which new images are simulated with new scores (see Figure 3). The synthesis details are as follows. First, from selected slices without lesions (*a*), a binary lung mask (*b*) was made[10]. Then two binary masks c_1 and c_2 were generated by defining up to 3 ellipses with random centers, orientations and axial lengths, which were multiplied by the lung mask to obtain d_1 and d_2. The two masks were then filled with GG (e_1) and RET (e_2) patterns, respectively, to produce images f_1 and f_2. Two small patches e_1 and e_2, which fully contained two different patterns, were manually cropped from diseased slices in advance. Finally, f_1 and f_2 were inserted into the healthy slice *a* to obtain the synthetic image *h*. In this process,

edges were smoothed to avoid introducing high intensity gradients. The synthetic Goh scores were calculated by the ratio of the areas of the different patterns to the whole lung area. The whole synthesis was applied on-the-fly during training with a probability of 0.5. In other words, half of the training set was sampled from the original images and the other half of training samples were synthetic images.

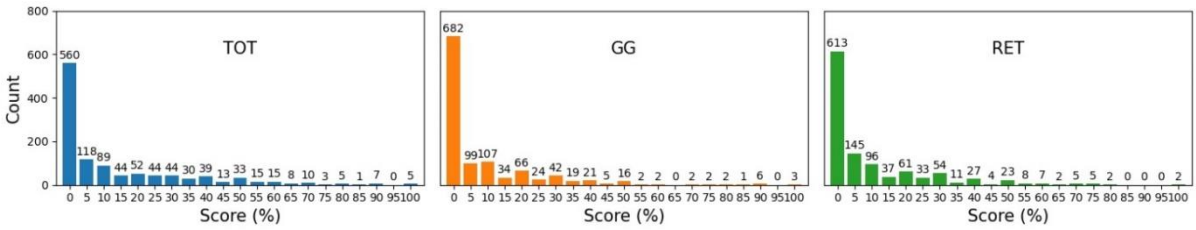


Figure 2. Label distribution of 1135 (5 × 227) slices for three patterns.

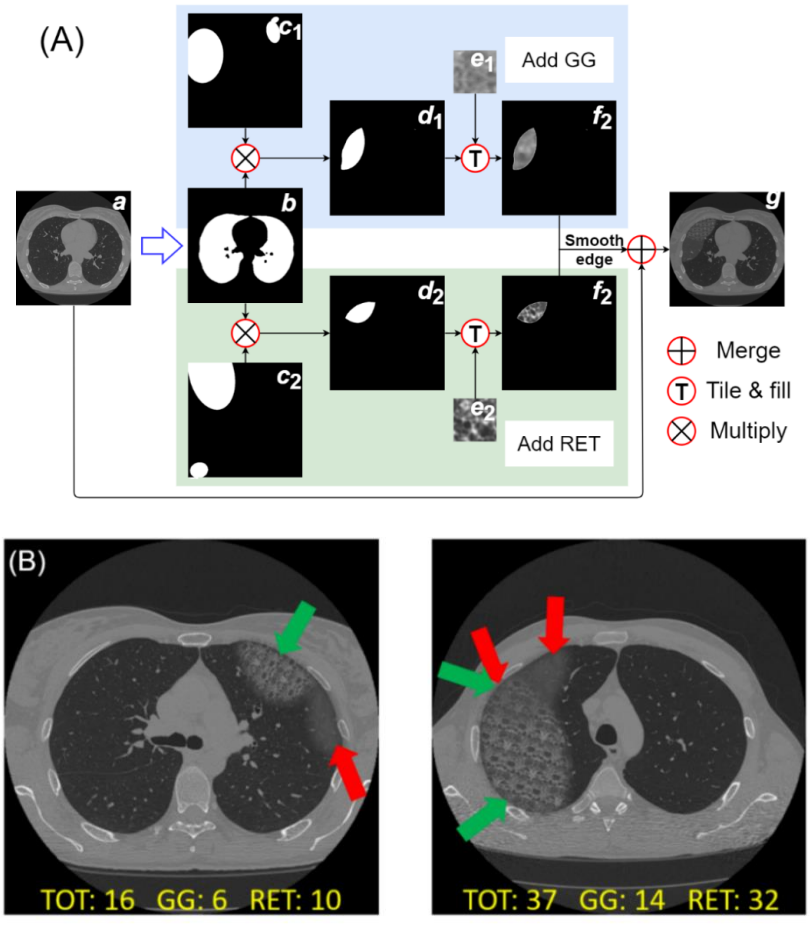


Figure 3. A) Flow chart to synthetize images; B) two synthetic examples (left: without overlap, right: with overlap), where the arrows indicate GG (red) and RET (green) areas. Goh scores of three patterns are indicated at the bottom of each image.

2.3 Evaluation metric

To have a complete evaluation of networks, apart from mean absolute error (MAE), weighted Kappa (κ) was calculated using *scikit-learn*[11] and the intra-class correlation coefficient (ICC) estimates were calculated using the *pingouin*[12] package in Python based on a single-rating, absolute-agreement, 2-way random-effects model. Statistical tests were performed using the Wilcoxon signed rank test.

3. EXPERIMENTAL RESULTS

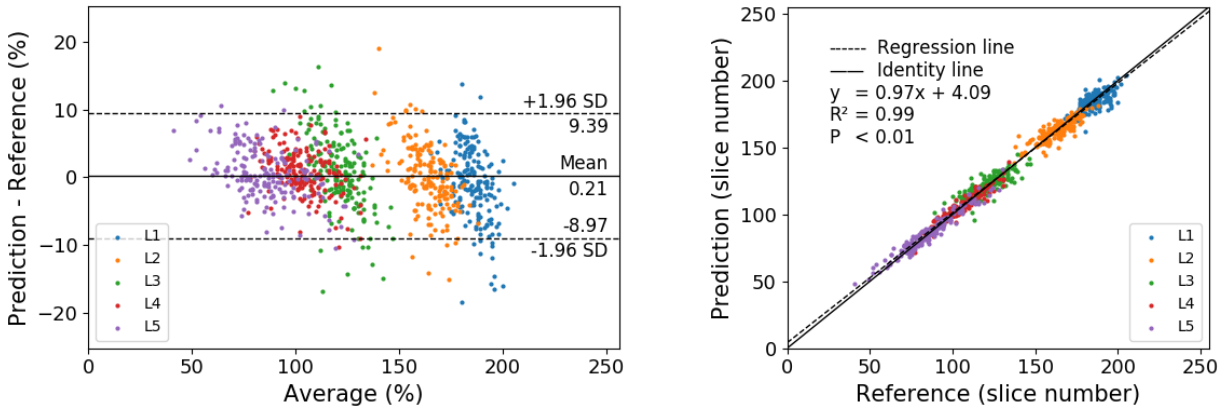
In this section, results are organized as follows: First the level selection output from 3DVGG11 (given by MAE and ICC) is shown. Then, the Goh score prediction from 2D VGG11 is shown for the two data augmentation techniques under study (balanced sampling and data synthesis). Finally, a comparison with human observer in a subset of testing dataset is provided.

3.1 Level selection

Table 1 shows the slice prediction performance based on the results of four-fold cross validation. The ICC values indicate moderate reliability on the first level, good reliability on the second and third level, and excellent reliability on the last two levels. The reason that the first level is more difficult to predict may be because it requires more anatomical knowledge to locate the origin of great vessels. This is illustrated in Figure 4.A, where most of the points outside of overall 95% CI are from the first three levels. Since the mean error of the prediction was only 0.21 slices, there was no systematic bias. Figure 4.B shows that the regression line (slope = 0.97, intercept = 4.09) nearly coincided with the identity line, which shows outstanding agreement between the predictions and labels.

Table 1. Slice prediction results of each level. MAE values are followed by standard deviations (STD).

Level	1.Origin great vessels	2.Main Carina	3.Venous Confluence	4.Halfway	5.Right hemi-diaphragm
MAE (#slices)	4.21 (3.58)	3.70 (3.17)	4.23 (3.55)	2.49 (2.15)	2.92 (2.46)
ICC	0.72	0.84	0.81	0.96	0.97



(A) (B)

Figure 4. A) Bland-Altman plot and B) Correlation plot of level selection network. In the right plot, dash line is the regression line and solid line is the identify line.

3.2 Score prediction

Table 2 compares the four-fold validation results of our Goh score prediction networks using different techniques. It shows that both balanced sampling and synthetic data augmentation can improve performance significantly.

Table 2. Goh score prediction results from 2D VGG11. ↓: lower is better, ↑: higher is better. †: significantly better than the upper row (P<0.05). All experiments were trained based on pre-trained weights from ImageNet[8].

Balanced sampling	Data synthesis	TOT			GG			RET		
		MAE[%]↓	κ↑	ICC↑	MAE[%]↓	κ↑	ICC↑	MAE[%]↓	κ↑	ICC↑
-	-	7.85 (10.71)	0.53	0.72	5.96 (10.03)	0.45	0.60	5.81 (8.42)	0.53	0.72
√	-	6.87 (9.61)†	0.59	0.77	5.09 (9.50)†	0.54	0.66	5.11 (7.45)†	0.59	0.78
√	√	**5.90 (8.77)†**	**0.66**	**0.83**	**4.66 (8.83)†**	**0.58**	**0.71**	**4.49 (6.70)†**	**0.65**	**0.84**

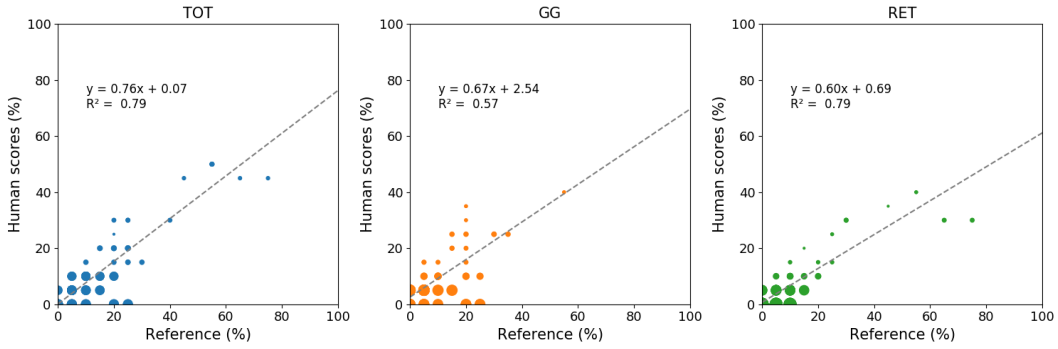
3.3 Comparison with human experts

Error! Not a valid bookmark self-reference. shows the performance comparison on a subset (from 16 patients) of the testing dataset between our method and each human observer with experts' consensus as the reference. The difficult samples account for a larger proportion in these 16 patients, therefore the results of our method in **Error! Not a valid bookmark self-reference.** are slightly different from those in Table 2. For predicting consensus TOT, our method was close to the first rating by observer Obs1 ($Obs1_{T1}$), but Obs2 was closer to the consensus than our method (**Error! Not a valid bookmark self-reference.**). For GG the model had a fair agreement with the consensus, while the observers had moderate agreement, and for RET the model's agreement was moderate, but moderate/substantial for observers.

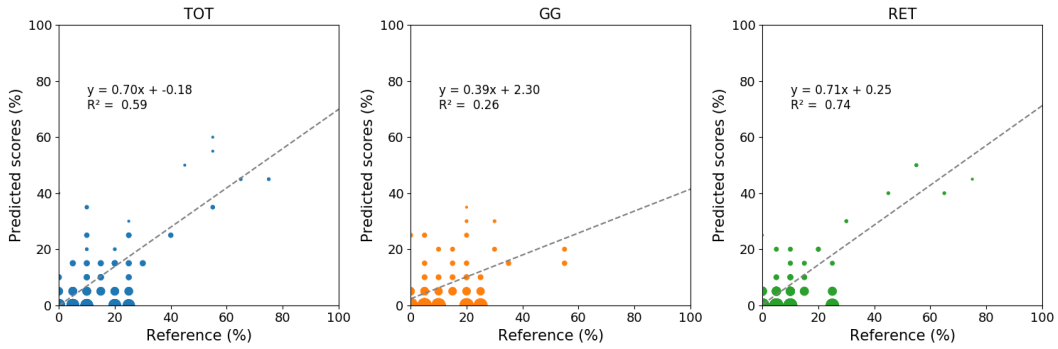
Table 3. Comparison between human performance and our method (with pre-trained weights, data synthesis and balanced sampling) in a subset of 16 patients. T1 and T2 denote the first and second observations.

	TOT			GG			RET		
	MAE[%]↓	κ↑	ICC↑	MAE[%]↓	κ↑	ICC↑	MAE[%]↓	κ↑	ICC↑
$Obs1_{T1}$	7.06 (7.97)	0.51	0.73	5.63 (6.14)	0.44	0.68	4.94 (7.60)	0.56	0.76
$Obs1_{T2}$	6.19 (6.63)	0.58	0.82	5.38 (7.53)	0.46	0.59	4.75 (7.24)	0.58	0.78
$Obs2_{T1}$	6.56 (7.61)	0.58	0.80	5.38 (7.53)	0.48	0.63	4.63 (6.16)	0.61	**0.84**
$Obs2_{T2}$	**4.94 (6.45)**	**0.67**	**0.86**	**4.94 (5.99)**	**0.55**	**0.75**	**4.19 (6.96)**	**0.63**	0.80
Our method	8.13 (8.15)	0.49	0.73	6.94 (8.46)	0.33	0.47	4.81 (6.35)	0.60	0.83

The scatter plots of the scoring results from our network and from the second observation from Obs2 ($Obs2_{T2}$) are shown in Figure 5. Because the predictions and scores coincided frequently as they are both defined with a precision of 5 percent, we used circles of different sizes to denote the numbers of points in the same positions. From the two figures we can find that our method had a systemic underestimation in the GG prediction. Additionally, our method performed slightly worse than the best human observation ($Obs2_{T2}$) on the TOT scoring. However, it had competitive performance with the best human expert on the RET scoring.



(A)



(B)

Figure 5. Reference score vs. the predicted score from (A) $Obs2_{T2}$ and (B) our method.

4. DISCUSSION AND CONCLUSION

We proposed a two-step framework to automatically predict Goh score for lung disease in systemic sclerosis. A simple data synthesis method was proposed in this work. Despite its simplicity, we found that data synthesis was surprisingly effective in improving the prediction of all the three patterns. This confirms the importance of data diversity and label balance. To the best of our knowledge, this is the first work to predict SSc scores automatically; therefore there is no existing similar literature to perform a fair comparison. We compared the performance of our method with two experienced human experts and showed that it performed slightly worse than the best human observation on TOT and GG prediction, but its performance was competitive for RET prediction.

Our framework provides a potential objective alternative for Goh scoring system of systemic sclerosis. In addition, with the help of our framework, Goh scoring could become suitable for daily practice because the scoring time per patient was decreased from dozens of minutes to a couple of seconds, which helps the Goh scoring system to be transferred from a research tool to a clinical tool.

Some limitations will be addressed in our future work. The performance of our current method is limited by the quality of synthetic images which include some artifacts. More advanced GAN-based data synthesis methods[13–15] will be introduced to generate more realistic images in the future. We will also introduce transfer learning from pre-trained medical imaging related task [16,17] instead of ImageNet. This may improve the performance further.

ACKNOWLEDGEMENT

Computing resources from the Academic Leiden Interdisciplinary Cluster Environment (ALICE) was provided by Leiden University. This work is supported by the China Scholarship Council No.202007720110.

REFERENCES

[1] Denton, C. P. and Khanna, D., "Systemic sclerosis," Lancet **390**(10103), 1685–1699 (2017).

[2] Goh, N. S. L., Desai, S. R., Veeraraghavan, S., Hansell, D. M., Copley, S. J., Maher, T. M., Corte, T. J., Sander, C. R., Ratoff, J., Devaraj, A., Bozovic, G., Denton, C. P., Black, C. M., Du Bois, R. M. and Wells, A. U., "Interstitial lung disease in systemic sclerosis: a simple staging system," Am. J. Respir. Crit. Care Med. **177**(11), 57–59 (2008).

[3] Williamson, L., "New reference atlas for pulmonary fibrosis severity score in systemic sclerosis," Lancet Respir. Med. **9**(2), 130–131 (2021).

[4] Desai, S. R., Veeraraghavan, S., Hansell, D. M., Nikolakopolou, A., Goh, N. S. L., Nicholson, A. G., Colby, T. V., Denton, C. P., Black, C. M., Du Bois, R. M. and Wells, A. U., "CT features of lung disease in patients with systemic scerosis: Comparison with idiopathic pulmonary fibrosis and nonspecific interstitial pneumonia," Radiology **232**(2), 560–567 (2004).

[5] Shadmi, R., Mazo, V., Bregman-Amitai, O. and Elnekave, E., "Fully-convolutional deep-learning based system for coronary calcium score prediction from non-contrast chest CT," Proc. - Int. Symp. Biomed. Imaging **2018-April**, 24–28 (2018).

[6] Wang, Y., Qiu, Y., Thai, T., Moore, K., Liu, H. and Zheng, B., "A two-step convolutional neural network based computer-aided detection scheme for automatically segmenting adipose tissue volume depicting on CT images," Comput. Methods Programs Biomed. **144**, 97–104 (2017).

[7] Simonyan, K. and Zisserman, A., "Very deep convolutional networks for large-scale image recognition," 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc. (2015).

[8] Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., Van Der, L. and Facebook, M., "Exploring the Limits of Weakly Supervised Pretraining" (2018).

[9] Acharya, S., Shukla, S., Mahajan, S. N., Banode, P., Mahure, C. and Mathew, L., "Interstitial lung disease in systemic sclerosis," J. Datta Meghe Inst. Med. Sci. Univ. **8**(1), 57–59 (2013).

[10] Li, W., Nie, S. D. and Cheng, J. J., "A fast automatic method of lung segmentation in CT images using mathematical morphology," IFMBE Proc. **14**(1), 2419–2422, Springer Verlag (2007).

[11] Pedregosa, F., Michel, V., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D.,

Pedregosa, F., Varoquaux, G., Gramfort, A., Thirion, B., Grisel, O., Dubourg, V., Passos, A., Brucher, M., Perrot andÉdouardand, M., Duchesnay, A. and Duchesnay, Fré., "Scikit-learn: Machine Learning in Python," J. Mach. Learn. Res. **12**, 2825–2830 (2011).

[12] Vallat, R., "Pingouin: statistics in Python," J. Open Source Softw. **3**(31), 1026 (2018).

[13] Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J. and Catanzaro, B., "High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs" (2018).

[14] Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J. and Greenspan, H., "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification," Neurocomputing **321**, 321–331 (2018).

[15] Uzunova, H., Ehrhardt, J. and Handels, H., "Generation of Annotated Brain Tumor MRIs with Tumor-induced Tissue Deformations for Training and Assessment of Neural Networks," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) **12264 LNCS**, 501–511 (2020).

[16] Chen, S., Ma, K. and Zheng, Y., "Med3D: Transfer Learning for 3D Medical Image Analysis" (2019).

[17] Zhou, Z., Sodha, V., Pang, J., Gotway, M. B. and Liang, J., "Models Genesis," Med. Image Anal. **67**, 101840 (2021).