**2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference**

The Impact of Interstitial Diseases Patterns on Lung CT Segmentation

Francisco Silva, Tania Pereira, Joana Morgado, António Cunha and Hélder P. Oliveira *(Member, IEEE)*

***Abstract*—Lung segmentation represents a fundamental step in the development of computer-aided decision systems for the investigation of interstitial lung diseases. In a holistic lung analysis, eliminating background areas from Computed Tomography (CT) images is essential to avoid the inclusion of noise information and spend unnecessary computational resources on non-relevant data. However, the major challenge in this segmentation task relies on the ability of the models to deal with imaging manifestations associated with severe disease. Based on U-net, a general biomedical image segmentation architecture, we proposed a light-weight and faster architecture. In this 2D approach, experiments were conducted with a combination of two publicly available databases to improve the heterogeneity of the training data. Results showed that, when compared to the original U-net, the proposed architecture maintained performance levels, achieving 0.894 ± 0.060, 4.493 ± 0.633 and 4.457 ± 0.628 for DSC, HD and HD-95 metrics, respectively, when using all patients from the ILD database for testing only, while allowing a more efficient computational usage. Quantitative and qualitative evaluations on the ability to cope with high-density lung patterns associated with severe disease were conducted, supporting the idea that more representative and diverse data is necessary to build robust and reliable segmentation tools.**

***Index Terms*—Deep Learning, Lung Segmentation, CT Images, Interstitial Lung Diseases.**

I. Introduction

The lung is the most vulnerable internal organ, due to the constant exposition to the external environment [1]. Respiratory diseases comprise a large variety of pathologies, from lung cancer to chronic obstructive pulmonary disease (COPD), and they are among the most common causes of severe illness and death worldwide [1]. In general, the automatic methods developed to help with lung disease detection and diagnosis would need to segment this organ for further analysis of the internal structures. For this reason, a robust method for lung segmentation is one of the first requirements for computer-aided decision (CAD) in lung problems. From the literature, there are several works on the field; however, on severe pathological cases or abnormalities, they usually originate inaccurate segmentations [2]. Since those segmentation methods were normally trained with datasets that not cover all the pathological heterogeneities, they have difficulty in dealing with the extreme cases. The most recent proposed approaches are based on deep learning techniques and trained and tested with data from private institutions in order to build methods that can deal with the physiological changes and keep the performance of the segmentation [3]. However, the development of such models using only publicly available sources of data remains a challenge in this research field, given the lack of representative and heterogeneous public datasets.

This study presents an investigation on the impact of the imaging manifestations associated with severe interstitial diseases in the lung segmentation task, especially high-density abnormalities, such as fibrosis, pneumonia, consolidation, as well as pulmonary nodules. Using only public databases, a lighter U-net architecture is proposed, and the conducted experiments were designed to improve the training data diversity in order to increase the ability of the algorithms to cope with the presence of pathological regions.

II. Materials and Methods

A. Datasets

1) Lung CT Segmentation Challenge 2017: The Lung CT Segmentation Challenge (LCTSC) [4] is a data collection provided in association with a segmentation competition regarding thoracic organs at risk (OAR): esophagus, heart, lungs and spinal cord. This database comprises training (LCTSC-36) and evaluation (LCTSC-24) datasets for 60 cases from 3 different institutions with different clinical practices. Thus, CT slice thickness took values of 1 mm, 2.5 mm and 3 mm. In this database, the tumor regions are excluded for most of the data, as well as the trachea and main bronchus (secondary bronchi may be included or excluded) [4].

2) ILD Dataset: The ILD database [5] is a CT dataset collected at the University Hospitals of Geneva (HUG), which comprises CT scans for a cohort of 128 patients diagnosed with lung parenchyma diseases, with available binary lung masks for a total of 113 patients. Considering data acquisition protocol, CT scans present a slice thickness value ranging from 1mm to 2mm, space between slices of $10 - 15$ mm and pixel spacing in (x, y) directions ranges from $0.4 - 1$ mm [5]. Regarding contouring guidelines, the lung mask ground-truth includes all pathological regions, as well as the trachea and bronchi structures.

3) NSCLC-Radiogenomics: The NSCLC-Radiogenomics dataset [6] is a publicly available CT image collection covering a cohort of 211 patients with non-small cell lung cancer

F. Silva and T. Pereira are with the INESC TEC - Institute for Systems and Computer Engineering, Technology and Science, Portugal.

A. Cunha is with the INESC TEC and UTAD, Portugal.

J. Morgado and H. P. Oliveira are with the INESC TEC and FCUP, Portugal.

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project UIDB/50014/2020.

The databases used in the experiments of this work ensure, on the correspondent published description article, that the necessary ethical approvals regarding data access were obtained.

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

(NSCLC); however, binary tumor masks are only provided for a total of 144 patients. Considering CT acquisition protocol, slice thickness ranges from 0.625 to 3 mm (median: 1.5 mm) [6].

B. Preprocessing

Since we addressed this task in a 2D perspective, and given the high spacing between slices from the ILD database CT scans $(10 - 15)$ mm [5], resampling was employed to standardize image representations only over the axial (x, y) plane. Thus, the pixel spacing was set to $[1.0, 1.0]$ mm. Additionally, pixel intensity values, measured in the Hounsfield Units (HU) scale, were normalized using the *min-max* normalization method and the HU window of -1000 to 400 HU. Lung binary masks also went through the same resampling operation to match the correspondent CT dimensions. Images were cropped by body region and resized for 256×256 pixels. Fig. 1 shows an example of a slice from the LCTSC [4] (first row) and the ILD [5] (second row) databases, along with the correspondent ground-truth mask and the result of the overlay of both images.

C. Deep Learning Model Architecture

The implemented architecture was based on U-Net [7], a general encoder-decoder based neural network specially designed for biomedical image segmentation tasks. U-Net comprises a contraction, bottleneck and expansion phases. Differing from the original model, the number of channels of each convolutional block was set to 16, 32, 64, 128 and 256. This means that, when compared with the original U-net, the proposed network was reduced in 4 times regarding the depth of each convolutional block, resulting in a difference of almost 16 times in the number of trainable parameters. A batch normalization layer was also added after each convolution, and the upsampling operation was employed by means of a sub-pixel convolutional layer, which learns an array of upsampling filters to upscale the low resolution feature maps into the higher resolution image [8].

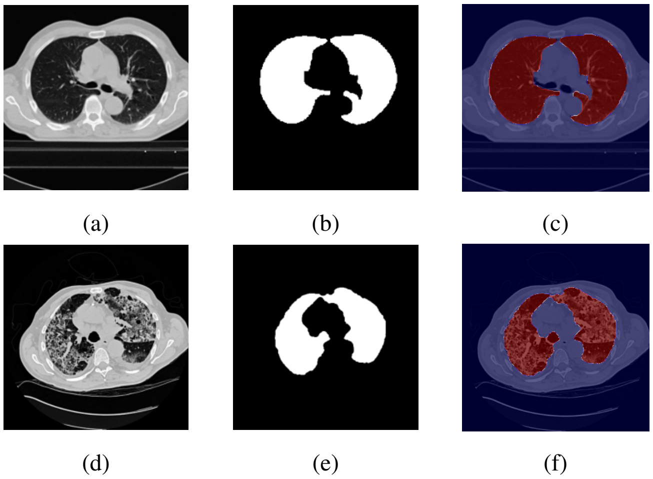

Fig. 1: Representation of one slice from the LCTSC and ILD databases. Figures (a) and (d) represent the raw CT slice, (b) and (e) the lung mask ground-truths and (c) and (f) the overlay between both, for the LCTSC and the ILD databases.

D. Experiment Design

The challenge training set (LCTSC-36) was considered the main data source for model training. Given the fact that tumor regions are excluded in the majority of cases in this database, along with the lack of pathological lung regions, the ILD database was used to improve the training data heterogeneity. To do this, a portion of ILD patients was selected to take part in the training set. This portion was iteratively increased, the performance metrics were assessed and a visual inspection of the ability to include severe pathological regions in the lung mask predictions was also employed. Only slices containing lung sections were included in these experiments. In testing phase, no post-processing operations were employed in order to reliably assess the capacity of the proposed approaches, without specific operations to overcome prediction mistakes. Moreover, the ability to include the tumor region was assessed using 144 CT images from the NSCLC-Radiogenomics database. For each CT scan, the slice with largest tumor section was segmented to compute the nodule overlap percentage.

E. Training Parameters

The Dice Loss [9] was the loss function used for optimization, based on the Dice score coefficient (DSC) performance metric. Models were trained with the Adam optimizer, using mini-batches of 8 images and a learning rate value of 0.001, with 10% decay every 5 epochs. Data augmentation techniques were employed to improve the model generalisation ability with random operations, including horizontal and vertical flips, image rotation and gaussian noise ($\mu = 0$, $\sigma = 0.1$).

F. Performance Metrics

1) Dice score coefficient: The DSC is a performance measure of relative overlap, ranging from 0 (no overlap) to 1 (perfect overlap), measuring the similarity between the predicted and the ground-truth masks.

2) Hausdorff distance: Given the predicted and ground-truth contours, the Hausdorff distance (HD) represents the maximum distance measured from a point in one set to its closest point in the other set. The 95th percentile HD (HD-95) measures the distance that is greater or equal than exactly 95% of all other distances, ensuring a more reliable evaluation in the presence of outliers.

III. Results

Performance results are depicted in Table I for the DSC, HD and HD-95 evaluation metrics, with values presented as mean ± standard deviation. Although the models were trained on a slice-level, the evaluation scores were computed in a scan-level. Training sets that contained patients from the

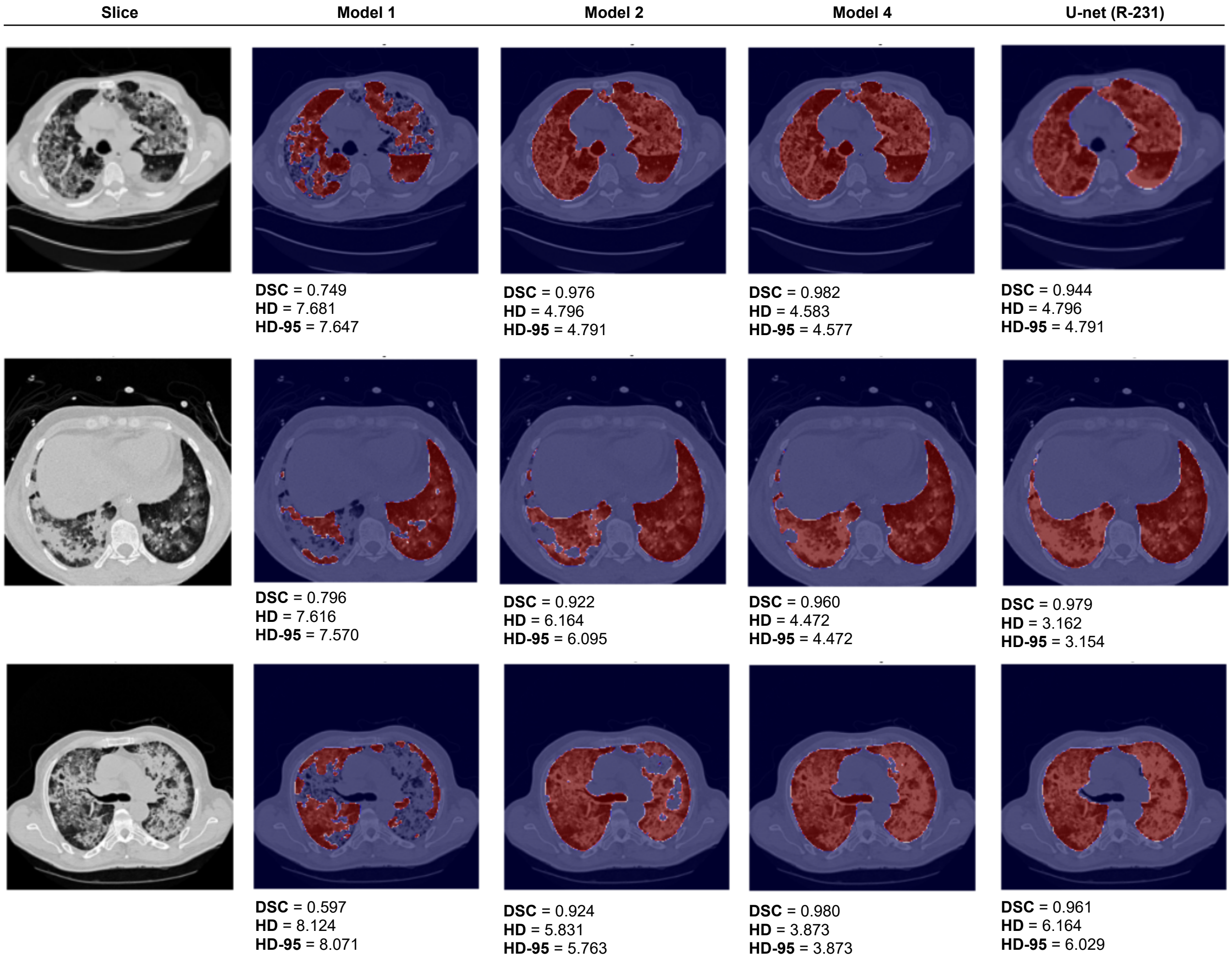

Fig. 2: Performance evaluation using the developed models (Table I), as well as the available model from [3] for inference (U-net (R-231)). Three distinct ILD patients were selected based on lung disease diagnosis: fibrosis, consolidation and pneumonia from the top to bottom rows.

ILD database allowed to increase the model ability of recognizing severe abnormalities as lung regions, improving scores in all evaluation metrics. Considering the proposed modifications regarding the depth of each convolutional block, results show that the proposed light-weight architecture does not cause a performance decrease in this segmentation task, as can be seen in Table I. Moreover, a random CT with $360 \times 480 \times 480$ dimensions was used to assess the inference times of both architectures, which showed that it takes 6.749 s to obtain the full volume lung mask prediction with the light-weight model, against the 30.294 s (almost 5 times more) necessary using the original U-net.

As expected, the increase on training data diversity showed to improve the ability to deal with severe pathological regions (Fig. 2). In a quantitative analysis of the major improvement, the bottom slice showed a DSC increase from 0.597 to 0.980, and although performances of the U-net (R-231) [3] are usually lower than the ones obtained with Model 4, this was caused by differences in segmentation guidelines, especially the exclusion of trachea and bronchi regions in the development of the first. For this evaluation reported in Fig. 2, the three patients were selected from test sets, ensuring that these examples were not seen before when training.

The ability to include tumor regions in lung mask predictions was also assessed using the detailed lung cancer database, with results depicted in Table II. As expected, the lower the tumor section, the easier it was included in the lung mask prediction. Other tumor features as location in the lung, shape and texture also played an important role, which can be seen in the high standard deviation values presented.

IV. Discussion and Conclusions

In this work, we investigated the impact of the presence of severe pathological regions in an automatic lung segmentation task. Based on the idea that architectural innovations, by itself, have not increased performances over well-designed baseline models [3], [10], such as U-net [7], we proposed a simpler and lighter U-net to investigate the capacity of an even faster and simpler architecture. Multiple dataset combinations were employed in order to increase the representativeness and diversity of the training data.

TABLE I: Performance results considering the trained models. For each model, the Train and Test Datasets are presented, alongside with scores for DSC, HD and HD-95 evaluation metrics as mean ± standard deviation. For comparison, first row presents reference values obtained using the original U-net, and then Models 1–4 refer to the light-weight architecture.

			Performance Metrics (mean ± standard deviation)		
Model	**Train Dataset**	**Test Dataset**	**DSC**	**HD (mm)**	**HD-95 (mm)**
U-net [7]	LCTSC-36	LCTSC-24	0.954 ± 0.024	3.179 ± 0.432	3.162 ± 0.429
		ILD-all	0.891 ± 0.057	4.597 ± 0.651	4.561 ± 0.636
1	LCTSC-36	LCTSC-24	0.953 ± 0.025	3.194 ± 0.493	3.177 ± 0.490
		ILD-all	0.894 ± 0.060	4.493 ± 0.633	4.457 ± 0.628
2	LCTSC-36 + 10% ILD patients	LCTSC-24	0.950 ± 0.028	3.206 ± 0.452	3.189 ± 0.448
		ILD-test	0.946 ± 0.031	3.518 ± 0.463	3.496 ± 0.458
3	LCTSC-36 + 20% ILD patients	LCTSC-24	0.951 ± 0.025	3.285 ± 0.423	3.268 ± 0.420
		ILD-test	0.954 ± 0.029	3.309 ± 0.377	3.290 ± 0.374
4	LCTSC-36 + 40% ILD patients	LCTSC-24	0.954 ± 0.022	3.203 ± 0.426	3.187 ± 0.423
		ILD-test	0.962 ± 0.028	3.172 ± 0.413	3.156 ± 0.409

TABLE II: Evaluation of tumor inclusion ability using the developed models. Overlap scores were computed using only one slice per CT scan (the one with largest tumor section). Size was measured using the diagonal of the correspondent bounding-box.

	Tumor overlap (mean ± standard deviation)	
Tumor section size (mm)	**Model 1**	**Model 4**
$]\,0, 20\,[$	0.612 ± 0.323	0.739 ± 0.297
$[\,20, 40\,[$	0.560 ± 0.336	0.607 ± 0.337
$[\,40, 60\,[$	0.377 ± 0.227	0.441 ± 0.239
$[\,60, 80\,[$	0.323 ± 0.321	0.350 ± 0.331

Considering previous works, a large variety of deep learning-based approaches has been proposed for this task. In the majority of these works, privately collected data was used as primary training data source, which makes it more difficult to obtain a fair and reliable direct comparison. As a cause of using a diverse and heterogeneous training data, those models were always able to obtain the most successful results when comparing to the ones developed using only public data [3], [11]. The differences in contouring guidelines between databases must be considered and a basic quantitative evaluation might not be sufficient due to the impact of the inclusion or exclusion of some regions in performance metrics. The conducted qualitative evaluation regarding the ability to obtain reliable lung masks with severe disease is a useful representation of this idea. With Model 4, we were able to deal with the majority of segmentation mistakes, although some still persisted, which can be confirmed in the depicted examples. However, as a consequence of this mixed training data, trachea and bronchi regions were included in the predicted masks, which did not happen when using the selected inference model (U-net (R-231) [3]). This makes reproducibility and comparison between investigations impossible in some cases, and increases the urgency for the public access of diverse and representative datasets to develop universal tools for lung clinical research.

Lung segmentation is a critical processing task for several holistic lung analyses. By successfully dealing with the presence of tissue abnormalities, it results in the most efficient lung representation with the necessary information to be further investigated. Thus, the importance of the tumor information is indisputable so it must be included in the lung mask predictions. The tumor overlap reported results showed that the implemented solutions have not yet overcome this problem. With larger and more challenging tumor regions, the robustness of the models must be improved for more reliable and adequate mask predictions.

References

[1] Forum of International Respiratory Societies., *The Global Impact of Respiratory Disease- 2^a ed*, 2017.

[2] A. Mansoor, U. Bagci, B. Foster, Z. Xu, G. Z. Papadakis, L. R. Folio, J. K. Udupa, and D. J. Mollura, "Segmentation and image analysis of abnormal lungs at CT: Current approaches, challenges, and future trends," *Radiographics*, 2015.

[3] J. Hofmanninger, F. Prayer, J. Pan, S. Röhrich, H. Prosch, and G. Langs, "Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem," *European Radiology Experimental*, 2020.

[4] M. Yang, Jinzhong; Sharp, Greg; Veeraraghavan, Harini ; van Elmpt, Wouter ; Dekker, Andre; Lustberg, Tim; Gooding, "Data from Lung CT Segmentation Challenge. The Cancer Imaging Archive." 2017.

[5] A. Depeursinge, A. Vargas, A. Platon, A. Geissbuhler, P. A. Poletti, and H. Müller, "Building a reference multimedia database for interstitial lung diseases," *Comput. Med. Imaging Graph.*, vol. 36, no. 3, pp. 227–238, apr 2012.

[6] S. Bakr, O. Gevaert, S. Echegaray, K. Ayers, M. Zhou, M. Shafiq, H. Zheng, J. A. Benson, W. Zhang, A. N. Leung, M. Kadoch, C. D. Hoang, J. Shrager, A. Quon, D. L. Rubin, S. K. Plevritis, and S. Napel, "Data descriptor: A radiogenomic dataset of non-small cell lung cancer," *Sci. Data*, vol. 5, 2018.

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9351. Springer Verlag, may 2015, pp. 234–241.

[8] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 1874–1883, 2016.

[9] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations," in *Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support*. Cham: Springer International Publishing, 2017, pp. 240–248.

[10] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, p. 203–211, Dec 2020.

[11] A. Khanna, N. D. Londhe, S. Gupta, and A. Semwal, "A deep Residual U-Net convolutional neural network for automated lung segmentation in computed tomography images," *Biocybern. Biomed. Eng.*, vol. 40, no. 3, pp. 1314–1327, jul 2020.