applied sciences

MDPI

Article

Lung Segmentation in CT Images: A Residual U-Net Approach on a Cross-Cohort Dataset

Joana Sousa [1,2]*, Tania Pereira [1], Francisco Silva [1,3], Miguel C. Silva [4], Ana T. Vilares [4], António Cunha [1,5] and Hélder P. Oliveira [1,3]

1 INESC TEC—Institute for Systems and Computer Engineering, Technology and Science, 4200-465 Porto, Portugal; tania.pereira@inesctec.pt (T.P.); francisco.c.silva@inesctec.pt (F.S.); acunha@utad.pt (A.C.); helder.f.oliveira@inesctec.pt (H.P.O.)
2 FEUP—Faculty of Engineering, University of Porto, 4200-465 Porto, Portugal
3 FCUP—Faculty of Science, University of Porto, 4200-465 Porto, Portugal
4 CHUSJ—Centro Hospitalar e Universitário de São João, 4200-319 Porto, Portugal; miguel.ncds@gmail.com (M.C.S.); ana_mvilares@hotmail.com (A.T.V.)
5 UTAD—Institute for Systems and Computer Engineering, University of Trás-os-Montes and Alto Douro, 5001-801 Vila Real, Portugal
* Correspondence: joana.v.sousa@inesctec.pt

Abstract: Lung cancer is one of the most common causes of cancer-related mortality, and since the majority of cases are diagnosed when the tumor is in an advanced stage, the 5-year survival rate is dismally low. Nevertheless, the chances of survival can increase if the tumor is identified early on, which can be achieved through screening with computed tomography (CT). The clinical evaluation of CT images is a very time-consuming task and computed-aided diagnosis systems can help reduce this burden. The segmentation of the lungs is usually the first step taken in image analysis automatic models of the thorax. However, this task is very challenging since the lungs present high variability in shape and size. Moreover, the co-occurrence of other respiratory comorbidities alongside lung cancer is frequent, and each pathology can present its own scope of CT imaging appearances. This work investigated the development of a deep learning model, whose architecture consists of the combination of two structures, a U-Net and a ResNet34. The proposed model was designed on a cross-cohort dataset and it achieved a mean dice similarity coefficient (*DSC*) higher than 0.93 for the 4 different cohorts tested. The segmentation masks were qualitatively evaluated by two experienced radiologists to identify the main limitations of the developed model, despite the good overall performance obtained. The performance per pathology was assessed, and the results confirmed a small degradation for consolidation and pneumocystis pneumonia cases, with a *DSC* of 0.9015 ± 0.2140 and 0.8750 ± 0.1290, respectively. This work represents a relevant assessment of the lung segmentation model, taking into consideration the pathological cases that can be found in the clinical routine, since a global assessment could not detail the fragilities of the model.

Keywords: lung segmentation; deep learning; CT images; cross-cohort; clinical assessment

Citation: Sousa, J.; Pereira, T.; Silva, F.; Silva, M.C.; Vilares, A.T.; Cunha, A.; Oliveira, H.P. Lung Segmentation in CT Images: A Residual U-Net Approach on a Cross-Cohort Dataset. *Appl. Sci.* **2022**, *12*, 1959. https://doi.org/10.3390/app12041959

Academic Editor: Jan Egger

Received: 20 January 2022
Accepted: 9 February 2022
Published: 13 February 2022

1. Introduction

Respiratory diseases are the leading causes of death worldwide, and among the most common causes are asthma, chronic obstructive pulmonary disease (COPD), acute respiratory infections, tuberculosis, and lung cancer, contributing to the global burden of respiratory diseases [1,2]. Lung cancer is one of the most common causes of cancer-related mortality. In 2020, approximately 2.2 million people were diagnosed with lung cancer and about 1.79 million individuals died from this condition [3]. The majority of patients are diagnosed in advanced stages, resulting in small chances of 5-year survival rate of around 3.9%. When lung cancer is identified in early stages, which can be achieved through screening, these probabilities increase up to approximately 54% [4]. Among the available

imaging modalities for screening, computed tomography (CT) has shown the highest reduction in cancer mortality [3].

Clinical assessment of CT images is a time-consuming task that is prone to discrepancies as a result of the subjective physicians interpretation, and computer-aided diagnosis (CAD) systems can help reduce this burden [4,5] . In automatic image analysis systems of the thorax, the segmentation of the lungs usually constitutes the first stage of processing, in order to reduce the computational cost [6,7] and to regularize the prediction task of the CAD by eliminating unnecessary information, or allow feature extraction from a region of interest for machine learning approaches [8–11]. This task is very challenging given that the lungs present high variability in shape, size, and volume. Moreover, the presence of abnormalities in the lung parenchyma, such as consolidations and cavities, makes segmentation even more difficult, leading to inaccurate delineations [6]. Current lung segmentation methodologies are able to segment lungs, exhibiting characteristics that are within a specific spectrum of patterns [6], but fail in the cases that present pathologies with complex patterns, for which they were not trained [6], as a consequence of a lack of diverse training data [7].

DL approaches, applied to medical imaging tasks, have been gaining popularity in recent years [5] and are preferred to other techniques, such as traditional imaging processing [12]. U-Net-based approaches have shown promising results on the segmentation tasks of medical images [13]. Skourt et al. [14] proposed a U-Net-based model composed of a contracting path similar to the one of U-Net (two convolutions followed by rectified linear unit (ReLU) activation and max-pooling, repeated four times) and an expansive path in which the upsampled layer is concatenated with a cropped corresponding feature map of the first path. Images from the Lung Image Database Consortium's Image Database Resource Initiative (LIDC-IDRI) dataset were manually segmented to generate the ground truth and later used to train and test the model. The average dice similarity coefficient (*DSC*) achieved was 0.9502. Shaziya et al. [15] also presented a U-Net model, with a contracting path, formed by three blocks of convolutional layers, ReLU and max-pooling, and an expansive path, formed by three blocks, two of them with two convolutions, concatenation with the corresponding feature map of the contracting path, and upsampling. The last one is similar to the previous blocks, except that it presents three convolutions, instead of two, and following the upsampling, there were two more convolutions and a dropout layer, followed by the output layer. The input images, with dimensions of 128×128, were resized to 32×32, in order to reduce computational time. Data augmentation was performed by rotation using the available training samples, to increase the number of images used to develop the model, and an accuracy of 0.9678 is achieved. Yoo et al. [16] presented 2D and 3D U-Net models for the segmentation of the lungs as one region and separately. The 2D model has an input dimension of $512 \times 512 \times 1$ and it is formed by 4 encoders and 4 decoders, in which bilinear interpolation is used for the upsampling step. On the other hand, the 3D model has an input dimension of $512 \times 512 \times 8$, 3 downsampling steps, and 3 upsampling steps with trilinear interpolation. For both models, softmax function is used in the output layer and cross entropy is used as a loss function. Two types of models were trained, one for the segmentation of the whole lung region and another for the separated segmentation of each lung. Concerning the latter, each ground truth was separated into two additional masks, each containing only one of the lungs. Afterwards, each of these masks was flipped horizontally and used as training for the segmentation of the opposite lung. The University Hospitals of Geneva's Interstitial Lung Disease (HUG-ILD) dataset was used as external validation. The 2D model presents a *DSC* of 0.9840 and 0.9840 for the whole segmentation and for the separated segmentation, respectively. Khanna et al. [17] introduced a residual U-Net for the lung segmentation in CT images. Initially, to improve the number of available training images, data augmentation was performed via flips, rotation, zooming, and shifting. The residual U-Net architecture is a combination of ResNet and U-Net architectures; hence, it presents an encoder path and a decoder path, each one with four stages.

The *DSC* was used as a loss function. In order to improve the model accuracy, a connected component algorithm was applied to remove non-lung regions. The Lung Nodule Analysis 2016 (LUNA16) and the Vessel Segmentation in the Lung 2012 (VESSEL12) datasets were used for training, whereas the model evaluation was performed with the HUG-ILD dataset. Two different architectures, ResNet34 and ResNet50, were implemented alongside the U-Net. It was verified that the latter presents slightly better results, achieving an average *DSC* of 0.9868.

Lung segmentation can be a very difficult task due to the influence of other pathologies that produce imaging changes. The presence of other respiratory comorbidities alongside lung cancer is frequent. Lung cancer and COPD, both predominantly caused by cigarette smoking, are closely linked, and each condition presents its own range of characteristic imaging features [18]. For this reason, apart from lung-cancer-specific patterns, it is imperative that the segmentation models are likewise able to identify features particular to other pulmonary conditions. The main goal of this work is the development of a deep-learning-based model for lung segmentation in CT images that must be robust on a cross-cohort dataset and capable of coping with the numerous and variable imaging appearances, derived from different pathologies with physiological heterogeneities.

2. Material and Methods

This section presents the multiple datasets used in the current study—the data selected for both training and performance evaluation of the model—and gives a detailed description of each dataset. It also describes the preprocessing steps taken and the segmentation model developed.

2.1. Datasets

2.1.1. Lung CT Segmentation Challenge 2017

The Lung CT Segmentation Challenge (LCTSC) 2017 [19] dataset was part of a competition in which the goal was the development of algorithms for the segmentation of several organs at risk in CT images for radiation treatment planning. The data was collected from 3 different institutions, making a total of 60 CT scans. The dataset is divided into 2 subsets: 1 contains 36 scans that are intended to be used for training (36-LCTSC), and the other subset contains 24 scans intended to be used for the assessment of the developed models (24-LCTSC). The number of slices along the z-axis per scan varies between 103 and 279 and their axial resolution is 512×512. The slice spacing is of 1.02 ± 0.11 mm and the slice thickness is of 2.65 ± 0.38 mm. The ground truth of the original images contains the delineation of five anatomical structures: esophagus, heart, left and right lungs, and spinal cord. Given that the lungs are the only organs of interest for this work, a binary ground truth containing solely the pulmonary regions was generated for each slice, using the information regarding these organs extracted from the DICOM RSTRUCT file.

2.1.2. Lung Nodule Analysis 2016

The Lung Nodule Analysis 2016 (LUNA16) [20] dataset was also part of a competition and it was developed to provide a large set for the comparison and evaluation of CAD systems designed for the detection of pulmonary nodules. This database contains 888 CT scans with annotations from another public dataset, LIDC-IDRI, and lung masks are available for each one of them. The scans were divided in 10 folders, intended to be used into a 10-fold cross-validation manner. For this work, only two of them were used, being randomly chosen from all the available folders. The number of slices varies between 103 and 733. The slice spacing is of 0.69 ± 0.09 mm and the slice thickness is of 1.60 ± 0.74 mm.

2.1.3. University Hospitals of Geneva—Interstitial Lung Disease

Motivated by the low availability of public collections of Interstitial Lung Disease (ILD) cases, the University Hospitals of Geneva Interstitial Lung Disease (HUG-ILD) database [21] was created to provide a public platform of interstitial lung disease (ILD)

cases for the development and evaluation of CAD systems. This collection comprises the 13 most common histological diagnosis of ILD, including conditions such as ground-glass, emphysema, fibrosis, consolidations, reticulation, and micronodules. The HUG-ILD database provides 112 CT scans with their respective binary lung segmentation masks. The number of slices along the z-axis per scan varies between 14 and 60 and the spacing between slices is within the range of 10–15 mm. Each slice is a matrix with 512 $\times$ 512 dimension. The slice spacing is of 0.70 ± 0.10 mm and the slice thickness is of 1.00 ± 0.00 mm.

2.1.4. Vessel Segmentation in the Lung 2012

Similar to the LCTSC dataset, the Vessel Segmentation in the Lung 2012 (VESSEL12) [22] dataset was part of a competition and its goal is to serve as a mean of comparison for (semi) automatic models for the vessel segmentation in lung CT scans. This database contains 10 patients with CT scans and the correspondent binary lung masks and comprises cases of alveolar inflammation, diffuse interstitial lung disease, and emphysema. Three extra scans are available as well, but were not intended to be used as part of an evaluation set. All thirteen scans have corresponding binary lung masks. The number of slices along the z-axis , varies between 355 and 534 and their resolution is of 512 $\times$ 512. The slice spacing is of 0.74 ± 0.09 mm and the slice thickness is of 0.88 ± 0.15 mm.

2.1.5. University Hospital Center of São João

A private dataset of 141 patients with lung cancer was collected in the University Hospital Center of São João (CHUSJ) and contains severe cases of this pathology. Semantic features were annotated for each scan and lung binary masks were generated for 27 of the available scans. The number of slices of these scans varies between 61 and 281 and their resolution is 512 $\times$ 512. The slice spacing is of 0.71 ± 0.08 mm and the slice thickness is of 3.07 ± 0.38 mm.

2.1.6. Summary of Cross-Cohort Dataset

In total, five datasets were used. Table 1 describes the number of scans used from each dataset. These datasets allow the development and test of the segmentation model in cross-cohorts, ensuring the heterogeneity of the data fundamental for a good generalization of the learning model.

Table 1. Final number of patients with CT scans used from each dataset.

Dataset	# CT Scans
LCTSC [19]	60
LUNA16 [20]	176
HUG-ILD [21]	112
VESSEL12 [22]	10
CHUSJ	27

2.2. Pre-Processing

When using data obtained from multiple sources, whether they are different institutions, different CT equipment, patients, or acquisition protocols, there is a high probability that the CT scans present variations, such as in image resolution, the field of view, and slice spacing and thickness. Each pixel of the 2D slices is represented by a value corresponding to the X-ray attenuation and it is expressed in Hounsfield Units (HU). Thus, the images were submitted to a HU *min-max* normalization, in which values between -1000 HU (HU_{min}) and 400 HU (HU_{max}) are rescaled into a range of $[0, 1]$. Then, as a second step of the preprocessing phase, the input images were resized to a dimension of 128 $\times$ 128 via bilinear interpolation, in order to reduce the computational cost. Regarding voxel spacing, the images were not rescaled in the z-direction.

2.3. Learning Model

Based on the model proposed by Khanna et al. [17], a hybrid architecture structure, depicted in Figure 1, was designed, consisting of the combination of the U-Net and ResNet34 architectures. This structure demonstrated better performance when compared to other simpler DL structures, such as U-Net, and when using LUNA16 dataset [17]. Following the typical structure of a U-Net, the model is composed of an encoder path and a decoder path, each one formed by five stages. First, the input images were submitted to 7×7 convolution, batch normalization (BN), and max-pooling, similarly to the first block of the ResNet34; likewise, each of the remaining 4 blocks that follow comprises residual units. Each unit contains a convolutional layer, followed by BN and parametric rectified linear unit (PReLU) activation, and a second convolutional layer followed by BN. In the end, in each unit, the input, the designated shortcut connection (SC), is added to the output of the unit and then submitted to a PReLU activation to produce a final result. Because max-pooling is not performed, the first convolution of the first residual unit of each residual block (RB), from RB2 to RB4, is applied with stride two, in order to reduce the dimension. Moreover, the SC of this unit is also submitted to a convolutional layer, so that the dimensions are in conformity. The blocks from RB1 to RB4 contain 3, 4, 6, and 3 residual units, respectively. Concerning the decoder path, the first four stages comprise upsampling via 2D transpose convolutions, followed by concatenation with the output of the corresponding block of the encoder path. Thereafter, the set of operations, including convolution, BN, and rectified linear unit (ReLU) activation, is performed two times. At the final stage, the output of the previous block is submitted to upsampling, followed by convolution, BN, and ReLU. At last, 1×1 convolution and sigmoid activation function are performed, producing the final segmentation result, a probability mask.

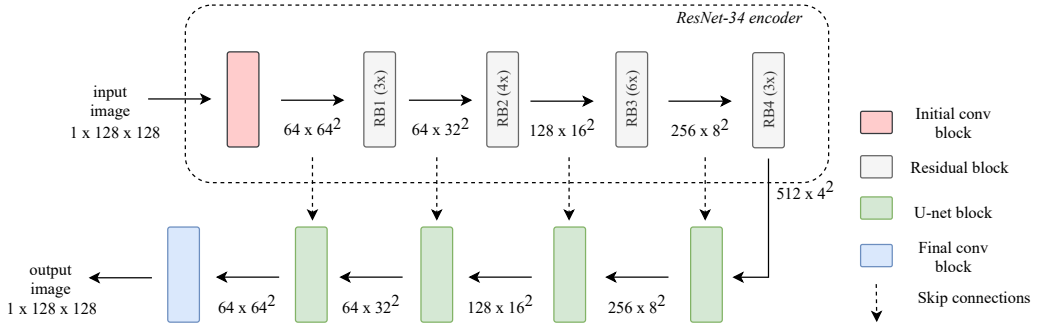


Figure 1. Hybrid structure that results from combining the ResNet-34 and U-Net architectures.

2.4. Training

The loss function used in the training process was based on the *DSC* , due to the fact that it has proven to be a useful measurement for this type of tasks [23]. The *DSC* is given by Equation (1), in which *DSC* is the dice similarity coefficient, X, corresponds to the ground truth mask, Y, is the predicted mask, $X \cap Y$ is the area of overlap of the two images, and $X + Y$ is the total number of pixels of the two images.

$$DSC = \frac{2(X \cap Y)}{X + Y} \tag{1}$$

This equation provides a measure of similarity between two images; thus, in order to measure the loss between the ground truth and the predicted mask, Equation (2) was used, in which DSC_{loss} represents the loss and *DSC* corresponds to the dice similarity coefficient.

$$DSC_{loss} = 1 - DSC \tag{2}$$

With the aim of finding the best hyper-parameters, experiments with different combinations of the optimizer, learning rate, and batch size were performed. Table 2 presents the values taken by each one of these hyper-parameters.

Table 2. Variable hyper-parameters and their values.

Hyper-Parameter	Value
Optimizer	Adam
Learning rate	0.00001, 0.0001, 0.001
Batch size	4, 8, 16, 32

The 36-LCTSC and the LUNA16 datasets were chosen to be used for training since these data are not intended to be used for evaluation (Table 3). As for the validation set, 30% of the training set was chosen randomly to be used as validation data and fixed with a seed to ensure that this distribution was the same across all experiments. At last, regarding the evaluation set, taking into account the importance of the capability of a model to cope with the different lung heterogeneities and its generalization ability, 4 distinct datasets: 24-LCTSC, HUG-ILD, VESSEL12, and CHUSJ, comprising a variety of disease patterns and differences in CT imaging protocols, were used. All tests were performed in each dataset separately. A summary of the distribution of the data per training and test sets is represented in Table 3.

Table 3. Distribution of the data per training, validation, and test set.

Task	Dataset	# CT Scans	# CT Images
Training	36-LCTSC + LUNA16	212	34,969
Validation			14,986
Test	24-LCTSC	24	3675
	HUG-ILD	112	2978
	VESSEL12	10	4279
	CHUSJ	27	3340

2.5. Evaluation

The evaluation metrics used were the *DSC*, given by Equation (1), Hausdorff distance (HD) and average symmetric surface distance (ASSD) [24]. The HD metric is given by Equation (3) in which $H(A, B)$ is the Hausdorff distance, A and B are two distinct objects, $h(A, B)$ is the maximum distance of any point of A to its nearest point in B, and vice versa for $h(B, A)$.

$$H(A, B) = max(h(A, B), h(B, A)) \tag{3}$$

The ASSD metric is given by Equation (4), in which $ASD(A, B)$ is the average of distances between the points of the borders of the ground truth, A, and the predicted mask, B, $S(A)$ is the set of border points belonging to A, $S(B)$ is the set of border points belonging to B, $\sum_{s_A \in S(A)}(ds_A, S(B))$ is the sum of distances of all border points of A to B—vice versa for $\sum_{s_B \in S(B)}(ds_B, S(A))$—and $S(A) + S(B)$ is the sum of all border points of A and B.

$$ASD(A, B) = \frac{\sum_{s_A \in S(A)}(ds_A, S(B)) + \sum_{s_B \in S(B)}(ds_B, S(A))}{\mid S(A) + S(B) \mid} \tag{4}$$

The HD and ASSD are metrics that take into account pixel spacing and each scan can present its own value for this image property. Therefore, the preliminary results of these two metrics obtained for each image were multiplied by the correspondent image pixel spacing, in order to produce normalized metrics. The three metrics were determined

between lung masks from the datasets (used as ground truth) and the segmentation mask produced by the model developed.

Additionally, a clinical assessment of the results was performed by two experienced radiologists. They performed a visual analysis of the 40 randomly selected cases and they evaluated and discussed the cases for the learning models that failed the segmentation.

3. Results and Discussion

This section includes the results obtained in the quantitative and qualitative evaluations for each test dataset, a clinical assessment performed by two radiologists, and the limitations found.

3.1. Performance Results

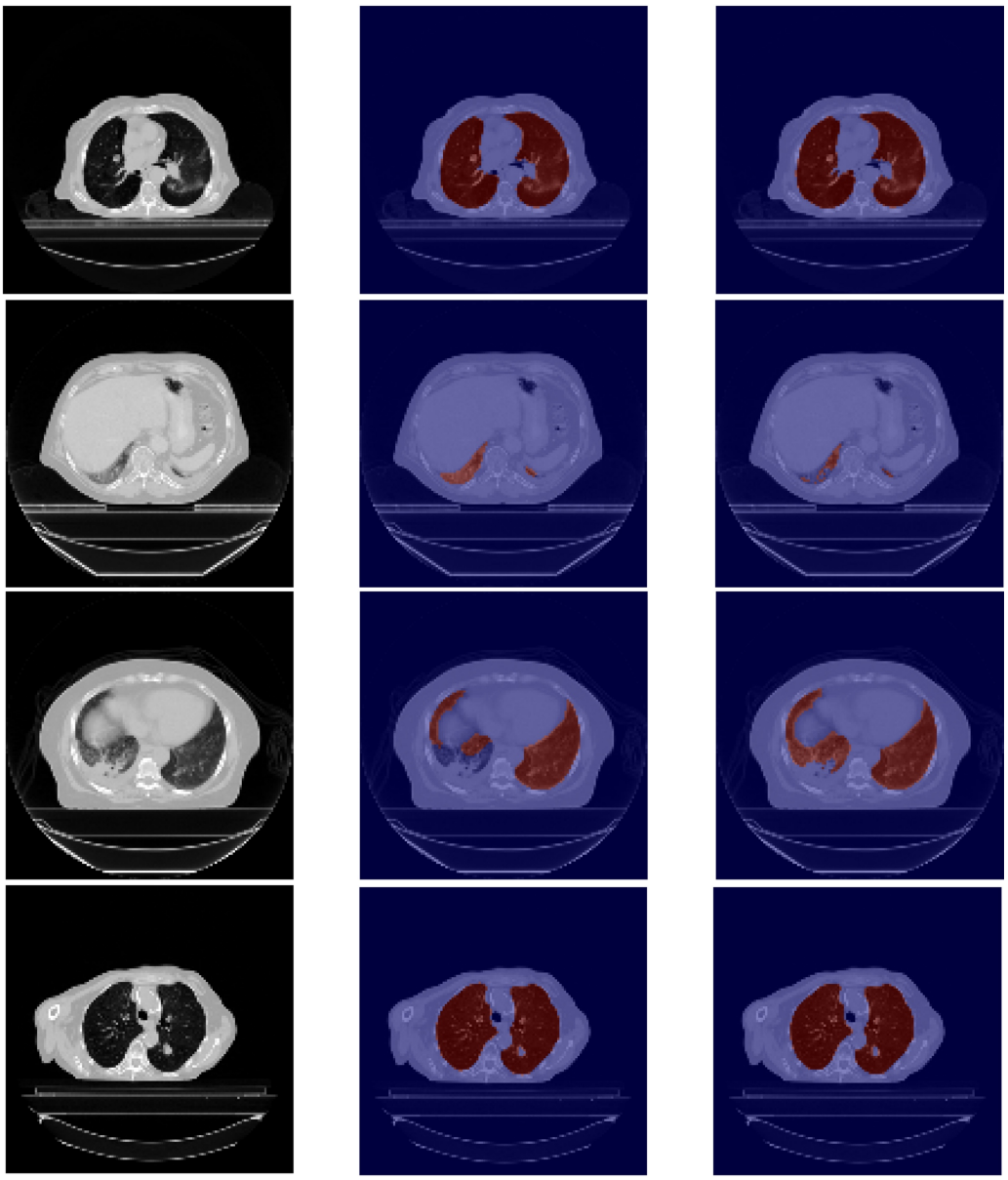
The set of hyper-parameters that lead to the best performance is as follows: Adam optimizer, the learning rate of 0.0001, and batch size of 8. The results obtained for each test dataset are presented in Table 4.

Table 4. Mean and standard deviation (std) results of the three metrics: dice similarity coefficient (*DSC*), Hausdorff distance (HD), and average symmetric surface distance (ASSD) for each dataset. The maximum value of the range for HD and ASSD metrics was obtained by considering that the maximum distance between 2 distinct objects in an image of 128 × 128 corresponded to the diagonal of that image.

Test Set	*DSC*	HD (mm)	ASSD (mm)
	Mean ± std	Mean ± std	Mean ± std
LCTSC	0.9472 ± 0.1752	3.7202 ± 5.7966	0.3359 ± 1.1534
HUG-ILD	0.9334 ± 0.1372	5.1783 ± 5.3090	0.4381 ± 1.4245
VESSEL12	0.9778 ± 0.2142	1.9395 ± 3.8952	0.1167 ± 0.5317
CHUSJ	0.9339 ± 0.1298	4.0943 ± 6.9651	0.4639 ± 1.5110
Range	[0–1]	[0–181.0193]	[0–181.0193]

The results from the three metrics in the analysis are consistent, showing better segmentations for the VESSEL12 dataset and worst results for HUG-ILD. However, the difference between the worst and the best results are very small, showing a good confidence that the segmentation model is robust to the great variability of the pathological cases used in the test sets. Overall, the model is able to generate good results and because the four test datasets present differences between them regarding acquisition protocols, pathologies included, and segmentation guidelines, the results for each one of them will be discussed individually.

With respect to the 24-LCTSC dataset, the model is generally able to correctly segment the pulmonary images (see the first row in Figure 2), but fails to identify their initial slices, which correspond to the base of the lung, leading to a decrease in the *DSC* (see the second row in Figure 2). Furthermore, for one of the patients, the masks produced by the model seem to be more accurate than the ground truth images, as the latter excludes part of the lung parenchyma. An example is shown in the third row in Figure 2. Therefore, even though this contributes to a lower *DSC* due to the discrepancy between them, the predicted mask is more precise. On the other hand, there are cases in which nodules are not included in the ground truth, and there are elements that the model does not include as well, giving rise to a higher *DSC*, although incorrectly classified. An example is depicted in Figure 2, fourth row.



(**a**) CT Image. (**b**) Ground Truth. (**c**) Predicted Mask.

Figure 2. Examples of LCTSC images, the ground truth and the predicted mask. From top to bottom the examples are, respectively: good segmentation example; an example in which the model fails to segment the base of the lung; an example of a ground truth image that excludes part of the parenchyma of the lung; an example that excludes a nodule in its ground truth, and which is also misclassified by the model.

Regarding the HUG-ILD dataset, the model successfully segments the majority of scans and in some cases, it does not exclude pulmonary regions presenting a higher density. The model was also assessed on this dataset on a pattern level, i.e., each pattern was evaluated individually, to gain a better understanding of its behavior, and the results are presented in Table 5.

Table 5. Results obtained for the patterns included in the HUG-ILD dataset. The row "Other" refers to scans that presented more than one pattern.

Pattern	Dice Similarity Coefficient	# CT Scans
	Mean ± Standard Deviation	
Micronodules	0.9545 ± 0.1372	22
Bronchioectasis	0.9377 ± 0.1845	2
Emphysema	0.9442 ± 0.1748	1
Fibrosis	0.9192 ± 0.1752	24
Macronodules	0.9281 ± 0.2142	3
Reticulation	0.9375 ± 0.1389	3
Consolidation	0.9015 ± 0.2140	2
Ground-glass	0.9436 ± 0.1298	15
Pneumocystis carinii pneumonia	0.8750 ± 0.1290	2
Other	0.9255 ± 0.1731	30

The results presented in Table 5 can be visually verified in Figure 3. For the specific cases of micronodules, bronchioectasis, emphysema, and some cases of fibrosis, the model is able to segment the entire lung area, as these are patterns which do not contain a higher contrast in tissues density (see respective examples on the first four rows of Figure 3). In general, those pathological cases showed a slightly better performance (Table 5).

In contrast, for cases of macronodules, reticulation, consolidation, ground-glass, and *pneumocystis carinii* pneumonia, the model presents a difficulty in performing such tasks in the regions of higher density (see last five examples (rows) of Figure 3). Besides that, the scans from this dataset include the trachea and other respiratory structures (apart from the lungs) in their ground truth, elements that are not identified by the model, and thus contributing to a lower metric (see rows two–five in Figure 3). Once again the model fails to identify the slices corresponding to the base of the lung (see the second row in Figure 6).

Concerning the VESSEL12 dataset, the model is able to produce good segmentation masks (see Figure 4), in general, for all ten scans, which were translated on a higher *DSC*, since this dataset does not contain intricate patterns. Moreover, the model does not erroneously classify other darker structures that are present in some slices as lungs.

As for the CHUSJ dataset, the model demonstrates, once again, a difficulty in the segmentation of the base of the lung (see the first row of Figure 5). Nonetheless, what contributes most to the decrease in the *DSC* is the large masses of higher density that are present in the majority of the scans and which the model does not identify. The model correctly segments the surrounding pulmonary tissue and these masses are the only structures that are not included in its predicted masks. Examples of these scans are depicted in the last two rows of Figure 5.

3.2. Clinical Assessment

From the visual inspection, radiologists concluded that the model has a good overall performance, especially for healthy cases for which the model always set a correct segmentation. They tried to identify the physiological reasons that made the model fail and that can be taken into consideration in future work. The radiologists identified that the model tends to fail in areas of the lung that have different densities than expected.

In areas of higher density ("whiter" on the CT image) than the surrounding lung, resulting from involvement by interstitial lung disease or inflammatory/infectious pathology, they are not recognized by the model as lung tissue (see first three rows in Figure 6), although they corresponded to areas of "diseased" lung, involved by interstitial pathology.

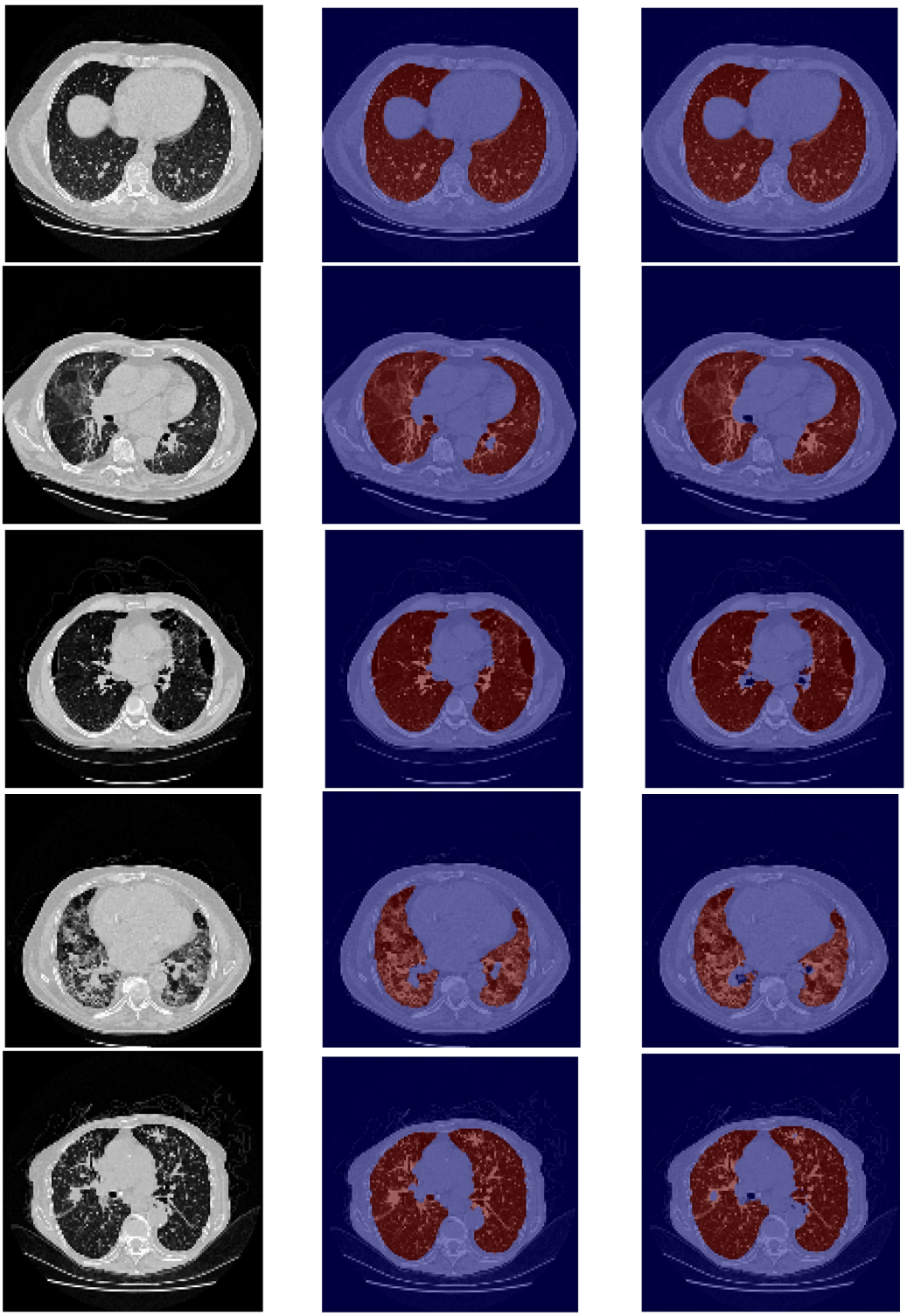
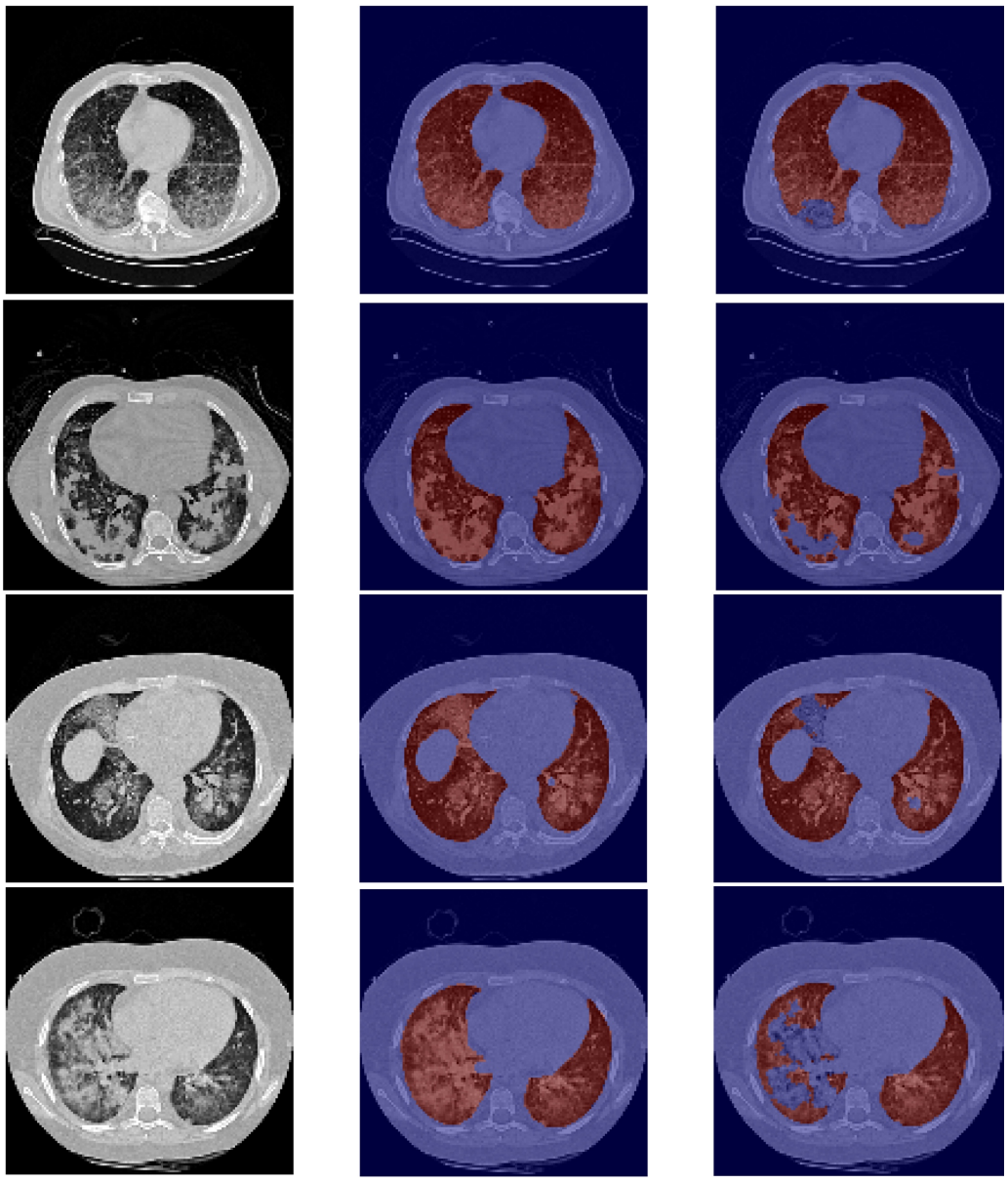
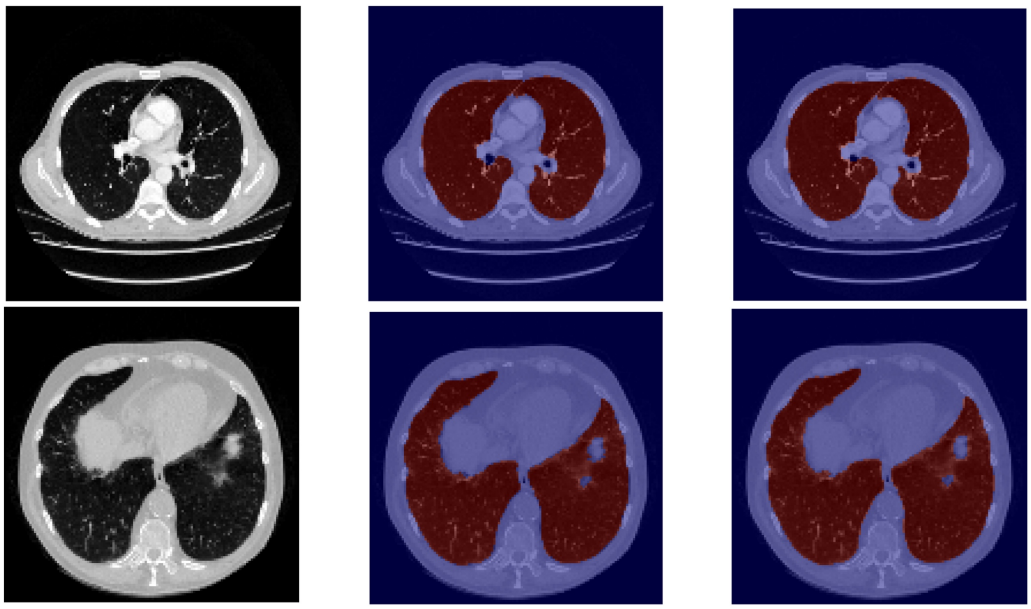


Figure 3. *Cont.*



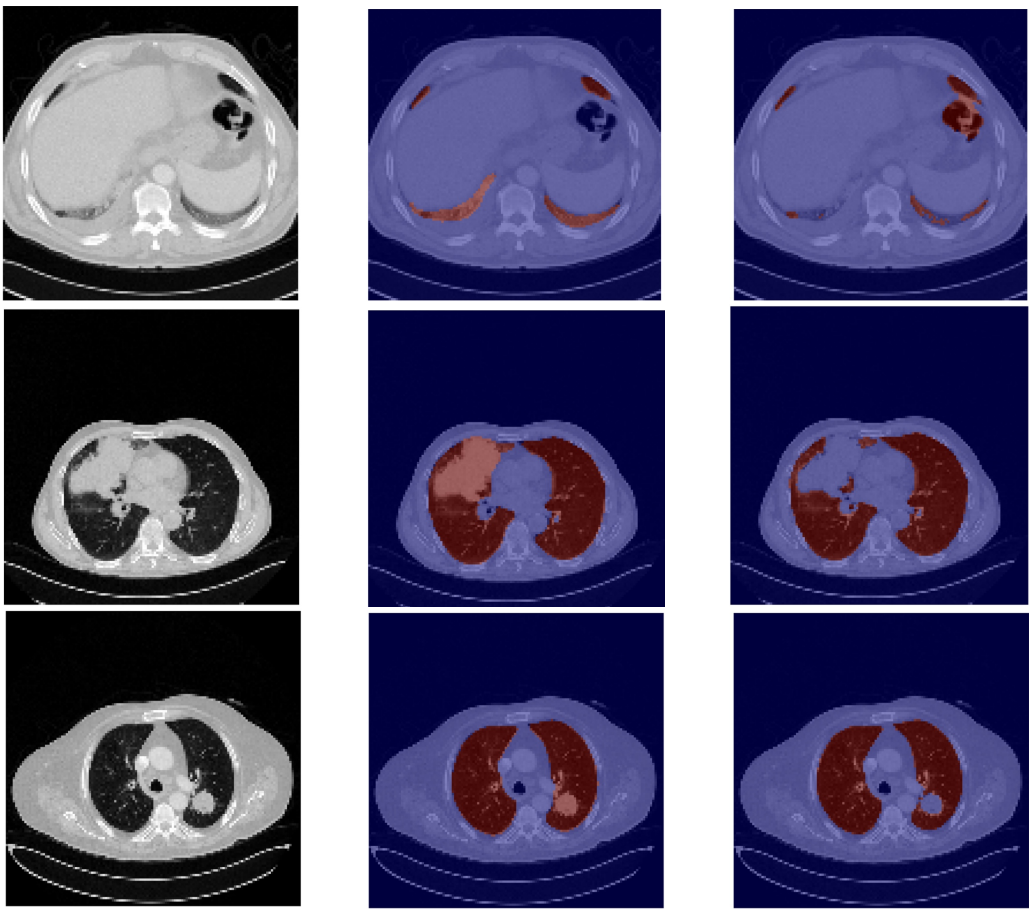
(**a**) CT Image. (**b**) Ground Truth. (**c**) Predicted Mask.

Figure 3. Examples of HUG-ILD images for pathological cases. The patterns of these images from top to bottom are, respectively, micronodules, bronchioectasis, emphysema, fibrosis, macronodules, reticulation with ground-glass, consolidation, ground-glass, and *pneumocystis carinii* pneumonia. For the last five examples, the model failed to segment part of the lung due to the pathological changes present in the image.



(**a**) CT Image. (**b**) Ground Truth. (**c**) Predicted Mask.

Figure 4. Examples of VESSEL12 images for which the model produces good segmentation masks.



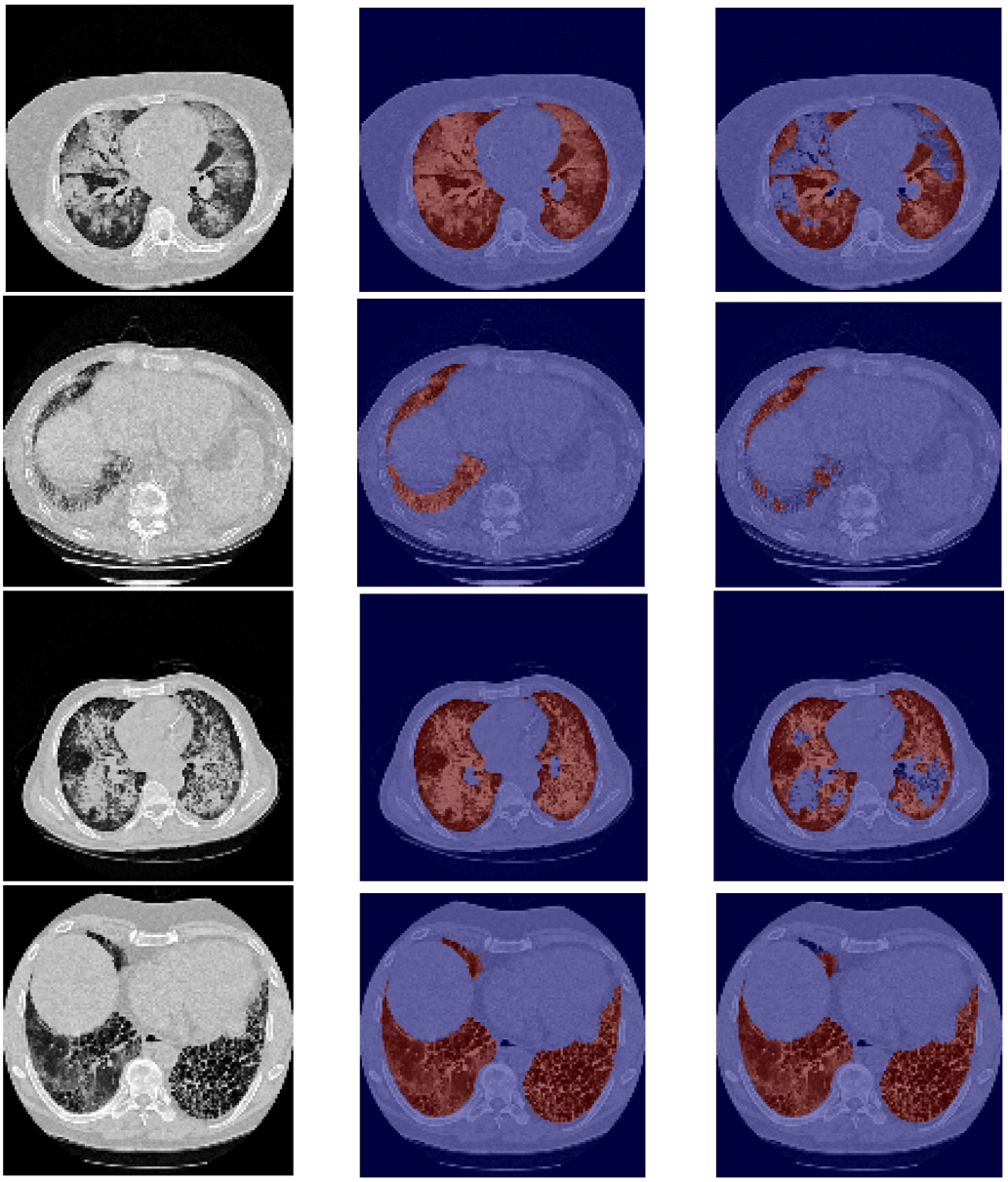
(**a**) CT Image. (**b**) Ground Truth. (**c**) Predicted Mask.

Figure 5. Examples of CHUSJ images, the ground truth and predicted mask. The first row is an example in which the model fails to correctly segment the pulmonary base; and second and third rows are examples in which the model does not segment the high density tumor masses.

In the case of lung neoplasms, something similar happens: the whole healthy lung is properly recognized; however, only the area of the lung mass, which is denser ("whiter") than the remaining lung, is not properly recognized (see last two rows in Figure 5).

A similar situation occurs in areas of the lung that are even less dense than usual ("blacker" on the CT image) which may also correspond to areas of "diseased" lung, in this case, areas of "air-trapping", emphysema, etc. (see the last row in Figure 6).

Lung pathologies present specific patterns, which correspond to changes in the density of the radiological image. Density increases for pulmonary consolidation present in pneumonia or other inflammatory/infectious processes; areas of opacification in ground-glass present innumerable causes, such as COVID, neoplasm, lung masses, and nodules. On the other hand, the density decreases with air-trapping, which is often associated with small airway disease and emphysema.



(**a**) CT Image. (**b**) Ground Truth. (**c**) Predicted Mask.

Figure 6. Examples from HUG-ILD dataset. First three rows show cases with higher density that the model does not classify as being lung tissue. Last row shows a darker pulmonary region misclassified by the model.

3.3. Limitations

The main motivation of this work was the combination of multiple cohorts of patients to train the learning model with a large spectrum of the heterogeneities that can be found in the population and the assessment of the segmentation performance on the most frequent lung diseases. The merge of cohorts covers the great majority of the pathophysiological patterns that can be found in lung diseases, and for this reason, the learning model fails only in the most extreme cases. Despite the overall good results, a large number of extreme pathological cases must be used in the training set to allow an even better generalization of the segmentation model in the future.

Another limitation comes from the annotations of the datasets. Those annotations come from different projects and followed different segmentation guidelines. There is no consensus on the inclusion/exclusion of some structures, such as airways or tumor masses. Examples of airways inclusion are shown in the second, third, fourth, and fifth rows in Figure 3 that belong to the HUG-ILD dataset, and examples of airways exclusion are shown in the last row in Figure 2, the first row in Figure 4, and the last row in Figure 5, that belong to the LCTSC, VESSEL12, and CHUSJ datasets, respectively. Examples of tumor masses inclusion are shown in last two rows in Figure 5 that belong to the CHUSJ dataset, and examples of exclusion are shown in the last row in Figure 2 that belong to the LCSTC dataset. Ideal, a very objective protocol of segmentation should be followed for the entire dataset annotation (training and test set) in order to not create label noise, which is responsible for overall quantitative performance degradation.

The need for massive and well-annotated datasets in the medical field is still one of the biggest limitations for the broad and impactful use of AI as support decision systems in the clinical routine. Unsupervised and semi-supervised approaches could be strategies to be used to overcome the lack of labeled medical data [25].

3.4. Methods Discussion

Considering the results obtained with the proposed approach and analyzing the different challenges ahead to overcome, some methodology improvements should be explored in the future, aiming to enhance the segmentation robustness of such models. Performance effects caused by increasing network complexity have been discussed in related application scenarios [7], being suggested that it will often be insufficient to overcome some specific problems. In this direction, the idea is raised that the path to more robust lung segmentation models would more likely comprise the development of models capable of being invariant in the presence of specific factors (e.g., severe pathological lung regions) that have caused significant performance drops. A robust segmentation model must not be influenced by the lung status itself, as diseased regions should be included in the predicted masks for further classification pipelines. In a different perspective, diving beyond traditional data augmentation techniques would also be an interesting idea to explore; the possibility of imitating some high-level properties of challenging tissue regions would enable us to reproduce the intended characteristic in any training example, which could result in more heterogeneous training data.

4. Conclusions

The proposed model was able to produce good results for the 24-LCTSC, the VESSEL12, and the HUG-ILD datasets. Nevertheless, it also generated poorer segmentation masks for the CHUSJ data, which mostly contains images with big tumor masses of higher density, and for some cases of the HUG-ILD data that contained complex patterns with high contrast tissues—all features that are not present in the training data. Thus, this work demonstrated that having a representative training database is crucial to build a robust segmentation model that is able to cope with complex patterns.

Subsequently, taking into account the wide variety of these elements and the limitations mentioned above, in particular the lack of well-annotated datasets, future models could be developed by making use of data augmentation techniques that would mimic the

missing imaging features. In addition, one could also explore the field of continuous learning that would possibly allow a model to continuously learn and improve its performance from the given data.

Author Contributions: J.S., T.P., F.S. and H.P.O. conceived the scientific idea, M.C.S. and A.T.V. provided the clinical insights. J.S. developed the software. All authors contributed to the critical discussion. J.S. and T.P. drafted the manuscript. All authors provided critical feedback and contributed to the final manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work is financed by National Funds through the Portuguese funding agency, FCT-Foundation for Science and Technology Portugal, within project LA/P/0063/2020, and a PhD Grant Number: 2021.05767.BD.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Ethics Committee of Centro Hospitalar de São João (290/18).

Informed Consent Statement: Patient consent was waived because it is a retrospective study.

Acknowledgments: We thank Inês Neves for performing the lung segmentation of the dataset from the University Hospital Center of São João.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Forum of International Respiratory Societies. *The Global Impact of Respiratory Disease*, 2nd ed; Technical Report; European Respiratory Society: Lausanne, Switzerland, 2017; ISBN 9781849840873. Available online: https://www.who.int/gard/publications/The_Global_Impact_of_Respiratory_Disease.pdf (accessed on 20 December 2021) .
2. GBD Chronic Respiratory Disease Collaborators. Prevalence and attributable health burden of chronic respiratory diseases, 1990–2017: A systematic analysis for the Global Burden of Disease Study 2017. *Lancet Respir. Med.* **2020**, *8*, 585–596. [CrossRef]
3. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.* **2021**, *71*, 209–249. [CrossRef]
4. Firmino, M.; Angelo, G.; Morais, H.; Dantas, M.; Valentim, R. Computer-aided detection (CADe) and diagnosis (CADx) system for lung cancer with likelihood of malignancy. *BioMed. Eng. OnLine* **2016**, *15*, 2. [CrossRef] [PubMed]
5. Zemouri, R.; Zerhouni, N.; Racoceanu, D. Deep Learning in the Biomedical Applications: Recent and Future Status. *Appl. Sci.* **2019**, *9*, 1526. [CrossRef]
6. Mansoor, A.; Bagci, U.; Foster, B.; Xu, Z.; Papadakis, G.Z.; Folio, L.R.; Udupa, J.K.; Mollura, D.J. Segmentation and Image Analysis of Abnormal Lungs at CT: Current Approaches, Challenges, and Future Trends. *RadioGraphics* **2015**, *35*, 1056–1076. [CrossRef] [PubMed]
7. Hofmanninger, J.; Prayer, F.; Pan, J.; Röhrich, S.; Prosch, H.; Langs, G. Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. *Eur. Radiol. Exp.* **2020**, *4*, 50. [CrossRef] [PubMed]
8. Chaturvedi, P.; Jhamb, A.; Vanani, M.; Nemade, V. Prediction and Classification of Lung Cancer Using Machine Learning Techniques. *IOP Conf. Ser. Mater. Sci. Eng.* **2021**, *1099*, 012059. [CrossRef]
9. Dandil, E. A Computer-Aided Pipeline for Automatic Lung Cancer Classification on Computed Tomography Scans. *J. Healthc. Eng.* **2018**, *2018*, 9409267. [CrossRef] [PubMed]
10. Morgado, J.; Pereira, T.; Silva, F.; Freitas, C.; Negrão, E.; de Lima, B.F.; da Silva, M.C.; Madureira, A.J.; Ramos, I.; Hespanhol, V.; et al. Machine Learning and Feature Selection Methods for EGFR Mutation Status Prediction in Lung Cancer. *Appl. Sci.* **2021**, *11*, 3273. [CrossRef]
11. Zhang, G.; Jiang, S.; Yang, Z.; Gong, L.; Ma, X.; Zhou, Z.; Bao, C.; Liu, Q. Automatic nodule detection for lung cancer in CT images: A review. *Comput. Biol. Med.* **2018**, *103*, 287–300. [CrossRef] [PubMed]
12. Hesamian, M.H.; Jia, W.; He, X. Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges. *J. Digit. Imaging* **2019**, *32*, 582–596. [CrossRef] [PubMed]
13. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Cham, Switzerland, 2015. [CrossRef]
14. Ait Skourt, B.; El Hassani, A.; Majda, A. Lung CT Image Segmentation Using Deep Neural Networks. *Procedia Comput. Sci.* **2018**, *127*, 109–113. [CrossRef]
15. Shaziya, H.; Shyamala, K.; Zaheer, R. Automatic Lung Segmentation on Thoracic CT Scans Using U-Net Convolutional Network. In Proceedings of the International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 3–5 April 2018; pp. 643–647. [CrossRef]

16. Yoo, S.; Yoon, S.; Lee, J.; Kim, K.; Choi, H.; Park, S.; Goo, J.M. Automated Lung Segmentation on Chest Computed Tomography Images with Extensive Lung Parenchymal Abnormalities Using a Deep Neural Network. *Korean J. Radiol.* **2020**, *22*, 476. [CrossRef] [PubMed]
17. Khanna, A.; Londhe, N.D.; Gupta, S.; Semwal, A. A deep Residual U-Net convolutional neural network for automated lung segmentation in computed tomography images. *Biocybern. Biomed. Eng.* **2020**, *40*, 1314–1327. [CrossRef]
18. Durham, A.L.; Adcock, I.M. The relationship between COPD and lung cancer. *Lung Cancer* **2015**, *90*, 121–127. [CrossRef] [PubMed]
19. Yang, J.; Sharp, G.; Veeraraghavan, H.; van Elmpt, W.; Dekker, A.; Lustberg, T.; Gooding, M. Data from Lung CT Segmentation Challenge. In *The Cancer Imaging Archive*; 2017. Available online: https://wiki.cancerimagingarchive.net/display/Public/Lung+CT+Segmentation+Challenge+2017#242845390e69ea3a95bd45b5b9ac731fb837aa14 (accessed on 20 December 2021). [CrossRef]
20. Setio, A.A.A.; Traverso, A.; de Bel, T.; Berens, M.S.; Van Den Bogaard, C.; Cerello, P.; Chen, H.; Dou, Q.; Fantacci, M.E.; Geurts, B.; et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge. *Med. Image Anal.* **2017**, *42*, 1–13. [CrossRef]
21. Depeursinge, A.; Vargas, A.; Platon, A.; Geissbuhler, A.; Poletti, P.A.; Müller, H. Building a Reference Multimedia Database for Interstitial Lung Diseases. *Comput. Med. Imaging Graph.* **2012**, *36*, 227–238. [CrossRef]
22. Rudyanto, R.D.; Kerkstra, S.; van Rikxoort, E.M.; Fetita, C.; Brillet, P.Y.; Lefevre, C.; Xue, W.; Zhu, X.; Liang, J.; Öksüz, İ.; et al. Comparing algorithms for automated vessel segmentation in computed tomography scans of the lung: The VESSEL12 study. *Med. Image Anal.* **2014**, *18*, 1217–1232. [CrossRef]
23. Zou, K.H.; Warfield, S.K.; Bharatha, A.; Tempany, C.M.; Kaus, M.R.; Haker, S.J.; Wells, W.M.r.; Jolesz, F.A.; Kikinis, R. Statistical validation of image segmentation quality based on a spatial overlap index. *Acad. Radiol.* **2004**, *11*, 178–189. [CrossRef]
24. Yeghiazaryan, V.; Voiculescu, I. Family of boundary overlap metrics for the evaluation of medical image segmentation. *J. Med. Imaging* **2018**, *5*, 015006. [CrossRef]
25. Zemouri, R.A.; Racoceanu, D. Innovative Deep Learning Approach for Biomedical Data Instantiation and Visualization. In *Deep Learning for Biomedical Data Analysis*; Springer: Cham, Switzerland, 2021.