UIP-net: a Decoder-Encoder CNN for the Detection and Quantification of Usual Interstitial Pneumoniae Pattern in Lung CT Scan Images

Rossana Buongiorno[1], Danila Germanese[1], Chiara Romei[2], Laura Tavanti[3], Annalisa De Liperi[2], and Sara Colantonio[1]

[1] Institute of Information Science and Technologies (ISTI), National Research Council (CNR), Pisa, Italy
[2] 2nd Radiology Unit, Pisa University Hospital, Pisa, Italy
[3] Pulmonary Unit, Pisa University Hospital, Pisa, Italy

Abstract. A key step of the diagnosis of Idiopathic Pulmonary Fibrosis (IPF) is the examination of high-resolution computed tomography images (HRCT). IPF exhibits a typical radiological pattern, named Usual Interstitial Pneumoniae (UIP) pattern, which can be detected in non-invasive HRCT investigations, thus avoiding surgical lung biopsy. Unfortunately, the visual recognition and quantification of UIP pattern can be challenging even for experienced radiologists due to the poor inter and intra-reader agreement.

This study aimed to develop a tool for the semantic segmentation and the quantification of UIP pattern in patients with IPF using a deep-learning method based on a Convolutional Neural Network (CNN), called UIP-net. The proposed CNN, based on an encoder-decoder architecture, takes as input a thoracic HRCT image and outputs a binary mask for the automatic discrimination between UIP pattern and healthy lung parenchyma. To train and evaluate the CNN, a dataset of 5000 images, derived by 20 CT scans of different patients, was used. The network performance yielded 96.7% BF-score and 85.9% sensitivity. Once trained and tested, the UIP-net was used to obtain the segmentations of other 60 CT scans of different patients to estimate the volume of lungs affected by the UIP pattern. The measurements were compared with those obtained using the reference software for the automatic detection of UIP pattern, named Computer Aided Lungs Informatics for Pathology Evaluation and Rating (CALIPER), through the Bland-Altman plot. The network performance assessed in terms of both BF-score and sensitivity on the test-set and resulting from the comparison with CALIPER demonstrated that CNNs have the potential to reliably detect and quantify pulmonary disease in order to evaluate its progression and become a supportive tool for radiologists.

Keywords: Deep-learning · Convolutional Neural Network · Idiopatic Pulmonary Fibrosis.

1 Introduction

The term *Interstitial Lung Diseases* (ILDs) refers to a large group of lung disorders, most of which cause scars of the interstitium, usually referred to as pulmonary fibrosis. Fibrosis reduces the ability of the air sacs to capture and carry oxygen into the bloodstream, leading to a progressive loss of the ability to breathe. Although ILDs are rare if taken individually, together they represent the most frequent cause of non-obstructive chronic lung disease. The Idiopatic Pulmonary Fibrosis (IPF) is a chronic, progressive fibrosing interstitial pneumonia, which is classified among the ILDs with the poorest prognosis [1]. The high variability and unpredictability of IPF course have traditionally made its clinical management hard. The recent introduction of antifibrotic drugs has opened novel therapeutic options for mild to moderate IPF [2]. In this respect, treatment decisions highly rely on the assessment and quantification of IPF impact on the interstitium and its progression over time. High-Resolution Computed Tomography (HRCT) has demonstrated to have a key role in this frame, as it represents a non-invasive diagnostic modality to evaluate and quantify the extent of lung interstitium interested by IPF [3]. In fact, IPF shows a typical radiological pattern, called Usual Interstitial Pneumonia (UIP) pattern, whose presence is usually assessed by radiologists to diagnose IPF. The HRCT features that characterize the UIP pattern are the presence and positioning of specific lung parenchymal anomalies, known as *honeycombing*, *ground-glass opacification* and *fine reticulation* [4]. These anomalies appear in the HRCT scans with specific textural characteristics that are detected via a visual inspection of the imaging data. Assessing the diffusion of these anomalies is instrumental to understand the impact of IPF and to monitor its evolution over time. Quantitative and reliable approaches are in high demand in this respect, as the visual examination by radiologists suffers, by its nature, of poor reproducibility [5].

To overcome this issue, much research is being conducted to develop new techniques for automatic detection of lung diseases that may support radiologists during the diagnostic pathway, particularly in HRCT image analysis.

CALIPER (*Computer Aided Lung Informatics for Pathology Evaluation and Rating*) is a software tool developed by the Biomedical Imaging Resource Laboratory at the Mayo Clinic for the automatic detection and quantification of CT anomalies in HRCT images of ILDs [10]. CALIPER uses histogram signatures to characterize and quantify parenchymal disease on HRCT and it was developed using pathologically confirmed imaging data evaluated by expert radiologist consensus. It is currently considered as the most viable instrument by radiologists. Nevertheless, it is not an open-source tool and its performance varies based on the acquisition context, thus on CT scanners and protocols and on the spatial kernel used by the image reconstruction algorithm.

This study aims to provide a tool for UIP pattern recognition based on a low-cost and real-time Machine Learning (ML) method to obtain UIP-pattern volume measurements based on a different approach than CALIPER, in the attempt to eventually overcome the aforementioned limits. The method relies on a fully-convolutional neural network (CNN), called UIP-net, which takes as input

a lung HRCT image and returns the corresponding binary map discriminating disease and normal tissue. This preliminary work firstly investigates whether CALIPER might be reproduced, to open the way to further investigation on many different scenarios (e.g., using UIP-net on images acquired by different scanners and reconstructed with different spatial kernels), possibly leveraging an unsupervised approach towards more generalizable results.

The paper is organized as follows: Section 2 describes the state of the art in the field of ML techniques applied to the detect UIP pattern and IPF biomarkers from lung HRCT scans; in Section 3 the UIP-net, that is the CNN here proposed for the detection of UIP patterns, is presented; then, in Section 4 the experimental setup and results provided by the UIP-net are described. Finally, Section 5 concludes the paper.

2 State of the art

Modern CT scanners allow for assessing anatomical and physiological properties providing high-definition volumetric images with an excellent spatial and temporal resolution. Computerized algorithm for HRCT image analysis, namely quantitative CT (QCT), gives a non-invasive mean for direct visualization, characterization and quantification of anatomic structures in order to obtain rapid and reproducible digital IPF biomarkers [6]. Indeed, several studies showed that QCT may overcome the issue of the inter-observer variability and could provide more consistent prognostic indexes. Furthermore, QCT may allow to extract CT features that are not visually recognisable and to objectively keep track of the disease progression [7]. The most relevant QCT methods for the assessment of ILD in patients with IPF are based on densitometric and local histogram analysis and textural analysis.

CT histogram provides a distribution of X-ray attenuation allowing the calculation of mean value, skewness and kurtosis that may give a measure of the extent of fibrosis. For example, both kurtosis and skewness showed correlation with functional test such as Forced Vital Capacity (FVC)[8]. It was demonstrated also that mean value, skewness and kurtosis are correlated with survival in patient with ILDs [9]. However, this approach is not sufficient to quantify the extent of every single interstitial lung abnormalities in patient with IPF, therefore more sophisticated textural analysis have been implemented.

Texture analysis consists in the quantitative description of the structural arrangement of pixels of different intensities and their relationship to the surrounding environment. Given the heterogeneity of lung parenchyma both in healthy subjects and in the presence of IPF, a correct interpretation of HRTC images may rely on texture analysis.

CALIPER (*Computer-Aided Lung Informatics for Pathology Evaluation and Rating*) can be considered as the most performing method based on texture analysis for IPF pattern visualization. This tool integrates a texture matching method with the analysis of histogram features of voxels for the automated lung parenchymal characterization and quantification of pulmonary disease on

HRCT images. This process automatically labels each pixel as belonging to one of seven specific parenchymal patterns: normal, ground-glass opacity (GGO), reticular density, honeycombing, and mild, moderate, or severe low-attenuation areas (see Figure (1)). It has been demonstrated that compared to visual scoring, CALIPER results are strongly correlated with functional tests [11], overall survival and decline of pulmonary capacity [12, 13].

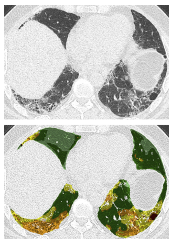

Fig. 1. HRCT image of UIP with CALIPER characterization. Top: Reticulation, groundglass opacity with a honeycomb cyst in the left lower lobe. Bottom: Color overlay image highlighting parenchymal patterns characterized by CALIPER: normal lung (light and dark green), ground-glass opacity (yellow), reticulation (orange), and honeycombing (red).

2.1 Deep Learning and Convolutional Neural Networks

The discussed methods, based on histogram and texture analysis, involve *handcrafted* features, that means manually engineered features, which are fed to machine learning classifiers to locally recognize patterns in lung tissue. More recently, advanced AI techniques, such as Deep Learning, outperformed such methods by adopting *learned features*, that are automatically obtained from the layers of the neural network thus overcoming the issues related to human bias.

Deep learning has achieved impressive results in several medical image classification tasks but only few methods have been proposed for IPF radiological pattern classification through Convolutional Neural Networks (CNNs).

Walsh et al. used a pre-trained Neural Network (NN) for the discrimination of UIP and not-UIP patterns, with training data labelled by expert radiologists [14]. Anthimopoulos et al. designed and tested a CNN for the classification of 7 anomalies within 2-D patches of HRCT images: healthy, GGO, micronodules, consolidation, reticulation, honeycombing, combination of GGO and reticulation[15], as shown in Figure (2). The proposed CNN reached an accuracy of 85% showing the potential of CNNs in IPF pattern recognition. Also Kim et al. developed a CNN for the classification of lung tissue in 2-D images; in this specific case, CNN outperformed a Support Vector Machine classification algorithm[16].

Several studies focused on the development of deep neural networks also for segmentation tasks. Anthimopoulos et al. designed a network that outperformed the traditional classification methods with less computational power and few segmentation errors [14, 16, 17]. Agarwala et al. pre-trained a P-net using daily

photographs and after a fine-tuning, the parameters of the network have been modified in order to optimize the performance on thoracic HRCT images including only ILD manifestations[18]. The network had good capacity in detecting fibrosis and emphysema even if the number of labelled patches used as training data was exiguous. Although the reported deep neural networks provided good results, they were not able to overcome the issues related to texture recognition of UIP patterns in HRCT images. The work reported in this paper aimed to design a CNN for the detection of UIP patterns preserving texture details during image processing. The network, named UIP-net, exploits the descriptive capability of neural networks to improve the diagnostic accuracy compared to the existing methods for quantitative image analysis of IPF. The network has been trained and tested on a dataset of 5000 images. In addition, according to the opinion of an expert radiologist, it provided acceptable results compared to CALIPER.

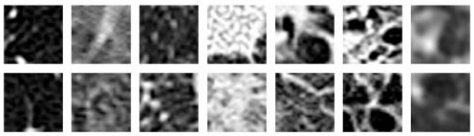

Fig. 2. Healthy tissue and typical ILD patterns from left to right: healthy, GGO, micronodules, consolidation, reticulation, honeycombing, combination of GGO and reticulation[15].

3 Data and Methods

3.1 Data

For the training and test of UIP-net, 20 HRCT volumetric scans of patients with IPF from the 2nd Radiology Unit database of Pisa University Hospital were used. Each scan had about 250 slices with 512×512 pixels per slice, thus the dataset had a total of about 5000 images. The scans were acquired using the same CT scan (Siemens Sensation 64) and acquisition protocol. Each slice had the same pixel spacing of 0.7 mm.

First of all, CT scans were processed by CALIPER in order to obtain the *ground truth*. CALIPER provided colour images with the segmentation of the areas corresponding to the parenchymal anomalies characteristic of the UIP pattern: yellow for Ground Glass Opacity (GGO), orange for reticulation and red for honeycombing, as shown in Figure (1). The colour images returned by CALIPER were imported in Matlab®(version R2019b) for obtaining the corresponding binary masks. These masks had pixels values equals to 0 or 1 if belonging to normal tissue or to UIP patterns, respectively. They represented the desired outputs of UIP-net: that means, UIP-net was trained to provide a binary mask for each gray-level input image with pixels equals to 0 or 1 if belonging to normal tissue or disease, respectively.

In order to test the usefulness of UIP net in quantifying the volume of the disease, another 60 CT scans of different patients and acquired with the same scanner were used. Also these scans had gray-scale slices with 512×512 pixels per slice.

3.2 Methods

Data Pre-processing In order to reduce the computational complexity of training and improve the speed of convergence of the model, the original images were pre-processed using Matlab®.

Nonetheless, no filtering was applied for preserving the intensity difference of adjacent pixels. On the other hand, to optimize the amount of data to be analyzed, the number of nonzero pixels was decreased through a Fuzzy c-means (FCM) algorithm. Two clusters were defined: the background (with the abdomen) and the foreground (i.e., the lungs). Pixels with a probability greater than 70% to belong to background were set to zero, those with a probability greater than 70% to belong to foreground were kept unchanged. After that, both the images and the ground truth were cropped for reducing the size of the Field of View (FOV). Thus, pre-processed images had 492×492 pixels with nonzero values only within the lungs (see Figure (3)).

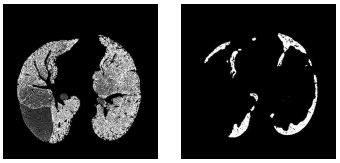

Fig. 3. On the left: an example of pre-processed cropped image with non-zeros pixels only within the lungs. On the right: an example of the ground truth obtained from CALIPER.

UIP-net architecture First, in order to design the optimal architecture of UIP-net, the problem was carefully analyzed and the requirements of the model were defined:

1. Since UIP patterns are characterized by typical textural features, the network should preserve the excellent image quality, in term of spatial resolution and bit depth, and be able to capture texture details;
2. The network should reduce data loss during training;
3. The network should be trainable and provide good results even:
   (a) with few examples because big datasets are not always available;
   (b) with few computational resources to make the network an accessible tool.

On the basis of these assumptions, UIP-net was inspired by [19] and designed with an Encoder-Decoder structure as in Figure (4). With respect to [19], the design of UIP-net architecture provides for:

1. the suppression of batch-normalization and pooling layers in order to prevent an excessive loss of information;
2. fewer layers to reduce the number of operations performed on the images;
3. the introduction of the tanh activation function for the last layer (instead of the softmax), in order to improve the speed of convergence of the model and make the network stable against sudden changes of the input.

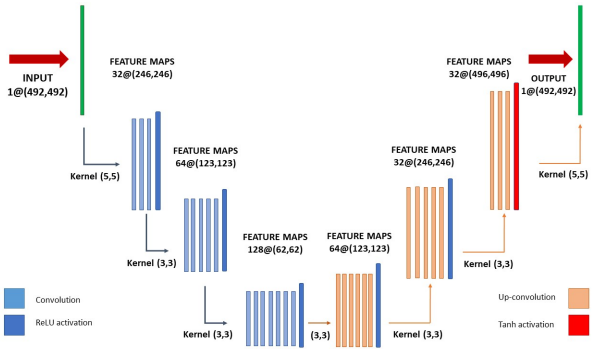

Fig. 4. UIP-net architecture.

The input layer of the UIP-net takes a 492×492 image and is followed by three convolutional layers and three de-convolutional layers. The size of receptive field was set to 3×3 for each layer, except for the first and the last one that have 5×5 kernel, in order to capture characteristic local structure of texture. Each convolutional layer doubles the number of features maps outputted. Thus, the first, the second and the third layers return 32, 64 and 128 features maps, respectively. On the other hand, two of the three de-convolutional layers produce 64 and 32 feature maps, while the third one keeps the same number of features maps, changing only the size of images. The last convolutional layer, finally, merges all the features maps into one with 492×492 pixels. Therefore, the output layer returns a binary mask with the segmentation of UIP patterns detected in the input image, keeping the same size. Each layer has a ReLU activation function, except the last one which, as mentioned above, has tanh activation.

Since no pooling was carried out between the convolutional layers to prevent loss of information, the stride was set to 2 in order to halve the size of the image after each layer. The padding was set to $\frac{k-1}{2}$ with k equal to the size of kernel (i.e 5×5 for the first and last layers, 3×3 for the others).

Training Method The UIP-net was trained by minimising the binary cross-entropy using Adam optimizer.

After some experiments, it was proved that the network works well with default values, namely the learning rate equal to 0.001, the exponential decay rates

for the moving average of the gradient equal to 0.9, and the squared gradient equal to 0.999.

Furthermore, Dice score monitored the network performances during training, comparing the segmentation made by UIP-net (S) with the ground truth (G), according to eq. 1.

$$D = \frac{2|S \cap G|}{|S| + |G|} \tag{1}$$

Finally, the weight updates were performed in mini-batches and the number of samples per batch was set to 10.

4 Experimental Setup and Results

4.1 Experimental Setup

The number of HRCT scans available for the training set was set to 13. Since each scan had about 215 slices, the training set consisted of about 3200 examples. Nevertheless, in order to avoid overfitting, it was necessary to establish how many images were strictly necessary.

Also the number of the epochs was set in order to stop training once the model performance stops improving on a validation set.

Thus, the validation set used for the fine tuning of the hyper-parameters was made up of 30% the examples, randomly extracted from the training set at each epoch.

A total of 20 trainings were carried-out:

1. 200, 400, 800, 1600, 3200 samples were fed to UIP-net for the same number of epochs;
2. UIP-net was trained for 50, 100, 150, 200 epochs keeping unchanged the number of examples.

Loss function and Dice score were monitored during training in order to choose both the correct number of examples and epochs.

As shown in Figure (5), the best model was the one trained by 800 samples for 50 epochs, with a binary cross-entropy on training and validation set of 0.14 and 0.097 respectively, and a Dice score of 0.81 and 0.78, respectively.

For this model, a 5-fold cross validation scheme was adopted to ensure the validity of the results. On average over all folds, the number of slices was 640 images for training and 160 for testing, while Dice score was 76.19% with a deviation standard of 3.54%.

The discussed method was implemented using Keras and TensorFlow framework and coded in Python 3.7.

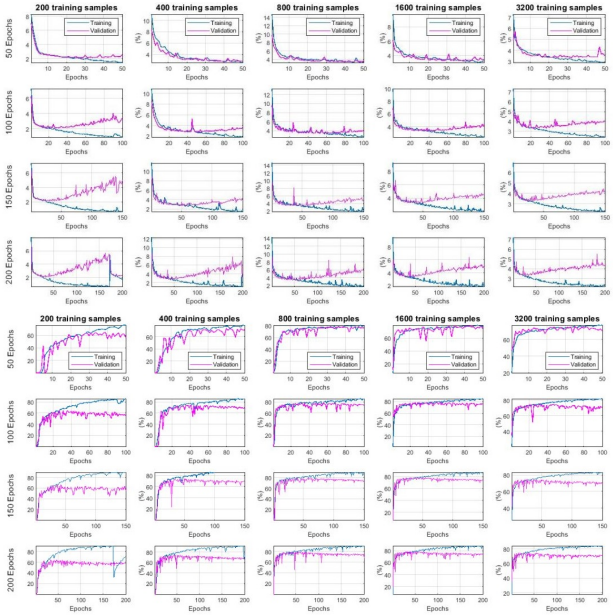

Fig. 5. Loss function (top) and Dice score (bottom) during training.

4.2 Results

The performances of the network were evaluated on the test set, consisting of 7 HRCT scans, with 1800 images in total. A quantitative performance analysis was performed, followed by qualitative assessment according to the opinion of an expert radiologist on the predicted segmentation (see Figure (6)).

Finally, 60 HRCT scans were used to compute the volume of UIP pattern detected by the UIP-net. The measurements were then compared with those obtained with CALIPER.

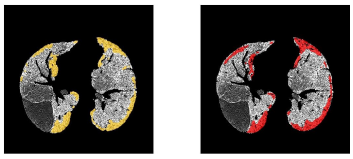

Fig. 6. On the left: original input image with ground truth overlapped (yellow). On the right: original input image with predicted segmentation of UIP-patterns overlapped (red). The predicted segmentations outputted by the network were assessed both quantitatively and qualitatively.

Quantitative Analysis The quantitative evaluation measures were: Dice and Boundary F1 (BF) contour matching score (BF-score), sensitivity and specificity.

BF score is a metric that tends to correlate better with human qualitative assessment than Dice score. It measures how close the predicted boundary of an object matches the ground truth boundary. The BF score is defined as the harmonic mean of the precision (P) and recall (R) values calculated within a distance error tolerance (typically 0.75% of the image diagonal [20]) to decide whether a point on the predicted boundary has a match on the ground truth boundary or not. BF-score can be defined through P and R according to the following eq. (2):

$$BF = \frac{2PR}{(P+R)} \tag{2}$$

Sensitivity and specificity could be defined on the basis of true/false positive and true/false negative, as shown in eq. (3) and (4).

$$Sensitivity = \frac{TP}{(TP+FN)} \tag{3}$$

$$Specificity = \frac{TN}{(TN+FP)} \tag{4}$$

In Table 1 evaluation measures on the test set are shown, with maximum values highlighted in green and minimum in red, while in Table 2 mean value and standard deviation of all HRCT scans of the test set are shown.

Table 1. Quantitative evaluation measures on test set.

	P1	P2	P3	P4	P5	P6	P7
Dice	74.94%	73.3%	61.58%	59.4%	58.4%	57%	56.65%
BF-score	78.96%	72.9%	81.16%	78.68%	81.32%	75%	83.73%
Sens	83.52%	83.51%	84.57%	74.46%	80.39%	74%	77.03%
Spec	98.75%	98.33%	97.98%	98.7%	99%	99%	98.17%

Table 2. Mean value and standard deviation of evaluation measures

	Dice	BF-score	Sensitivity	Specificity
Mean value	63.1%	78.8%	79.6%	98.5%
Standard deviation	7%	3%	4.4%	0.4%

Qualitative Assessment For a more comprehensive evaluation process, the performance of UIP-net was assessed through a qualitative visual analysis of the predicted segmentation performed by an expert radiologist. This showed that, compared to the ground truth (see Figure (7)):

1. UIP-net detected some patterns missed by CALIPER expecially where the amount of diseased tissue is high (see Fig 7);
2. UIP-net detected lung disease in the intestine and labelled vessels and airways as lung tissue. On the contrary, CALIPER manages to discriminate both vessels and airways.

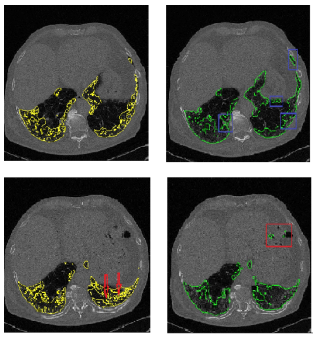

Fig. 7. On the left: the ground truth. On the right: UIP-net predicted segmentations. At the top: in the blue boxes, there are pixels detected by UIP-net but ignored by CALIPER. At thehttps://www.overleaf.com/project/5f686a151c2b180001f801de bottom: on the left, the arrows point to the vessels segmented by CALIPER, but ignored by UIP-net; on the right, the red box highlights the intestine mistakenly segmented by UIP-net.

In order to evaluate how the discussed issues affect the quantitative measures of the performance of UIP-net, a post-processing step was carried-out.

Firstly, false detections in the intestine were removed using the method proposed by Ross et al.[21] that allows to obtain masks containing only lungs. Briefly, this involves initial gray level thresholding using Otsu's method followed by morphological closing to fill in high attenuating areas within the lung field. In order to properly label airways outside the lung field, component region growing was applied. Once the region of the trachea is determined, an initial threshold and seed location are selected to initialize the region growing algorithm to extract the airway tree. The obtained masks contain lungs without airways but with vessels. Thus, in order to extract vessels from lungs, the method proposed by Sato et al.[22] was used. This involves 3-D line enhancement filtering with which accomplishes the following:

1. Recovery of line structures of various width, especially thin structures;
2. Removal of the effects of non-linear structures and of noise and artifacts.

Both methods were implemented using 3-D Slicer software and Chest Imaging Platform (CIP) framework. Once binary masks of the whole lungs and vessels were obtained, they were combined in order to get another one with only lungs which allowed to keep only segmented pixels belonging to lungs, thus removing those belonging to vessels and intestine (see Figure (8)).

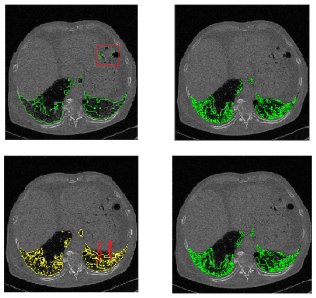

Fig. 8. On the top: predicted segmentations of UIP-net before (left) and after (right) post-processing. On the bottom: ground truth (left) compared to predicted segmentation of UIP-net after post-processing (right).

The same quantitative evaluation measures described in section 4.2 were then computed (see Table 3 with maximum values highlighted in green and minimum in red).

Table 3. Quantitative evaluation measures after post-processing.

	P1	P2	P3	P4	P5	P6	P7
Dice	84.43%	77.4%	65.77%	62.28%	75.43%	60.1%	64.58%
BF-score	96.07%	96.1%	96.71%	91%	93.6%	89.2%	94.14%
Sens	75%	84%	85.87%	76.3%	80.7%	75%	78.1%
Spec	100%	98.33%	98.66%	99.2%	99.3%	99%	98.87%

In Table 4 mean values with standard deviations of evaluation measures before post-processing and after post-processing are compared.

Table 4. Mean value and standard deviation of evaluation measures before and after post-processing

	Dice	BF-score	Sensitivity	Specificity
Before post-processing	63.10%±7%	78.8%±3%	79%±4.4%	98.5%±0.4%
After post-processing	70%±9%	93.8±2.8%	79.2±4.4%	99±0.5%

Volume estimation and comparison with CALIPER In order to evaluate the reliability of UIP-net in quantifying the volume of diseased tissue in the lungs, 60 HRCT scans were used. The scans were provided as input to the network to segment the UIP pattern. The segmentations were then imported in 3D-Slicer and, through CIP framework, the measures of volume of diseased tissue

in each scan (in cm^3) were computed. The comparison with those estimated by CALIPER was done through the Bland-Altman plot (see Table 5 and Figure 9) since it provides a visual representation of the agreement between two different methods.

The difference between UIP-net and CALIPER were acceptable: in Fig. 9 can be seen that most measures fall in the range between the lower and the upper Limit Of Agreement (LOA).

Table 5. Mean value (bias) and standard deviation of raw differences calculated between the measurements of volume obtained with UIP-net and CALIPER. Lower and Upper Limit of Agreement (LOA) were mean±1.96×standard deviation. Bias, lower LOA and upper LOA are shown on Bland-Altman plot as dashed lines.

	Raw Differences (cm^3)	Lower LOA	Upper LOA
UIP pattern volume	69.18041 ± 584.6433	-1076.72	1215.081

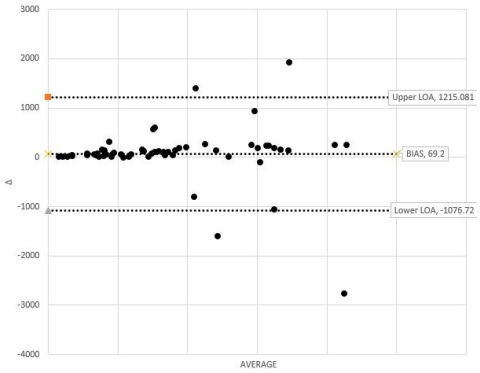

Fig. 9. Bland-Altman plot. Average and raw differences between UIP-net and CALIPER measurements (x-axis and y-axis respectively). The dashed lines indicate the lower and the upper LOA.

4.3 Discussion

Quantitative analysis of predicted segmentation of UIP-net and volume estimates followed by comparisons with CALIPER highlighted some aspects.

First of all, the lower values of Dice score, which was less correlated with human visual opinion than the others evaluation measures, could be due to the inhomogeneities of ILD. Indeed, Dice score works best with the segmentation of compact diseases like nodules, but ILD is mostly uneven within the lungs (Fig. 10).

On the other hand, BF-score was always consistent with visual assessment and took higher values than Dice score. At last, although specificity had always

highest values, it is not clear if such measure is accurate. As mentioned above, IPF is heterogeneously distributed within the lungs and tends to be present mostly in the middle and lower lung fields, so the many null pixels of apical slices of the same scan, might unbalance the measures (Figure (10)). On the contrary, sensitivity can be taken as a reference measure and it had similar values to BF-score.

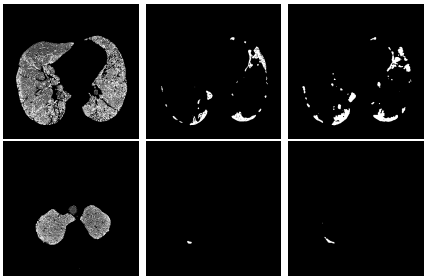

Fig. 10. On the top: an example of sparsity of the segmentation. On the bottom: an example of images with null pixels which unbalance the value of specificity. Original image (left), ground truth (center), predicted segmentation (right).

Overall, the UIP-net demonstrated good performance metrics and these results are encouraging. BF-score after post-processing reached a maximum of 83.73%. The post-processing results demonstrated that the misclassification of vessels and intenstine mostly affected the performance. In fact, all the computed indices increase: Dice score increases of 7%, while BF-score increases of 15% reaching a maximum of 96.71%. Thus, finer pre-processing can solve the issue.

Another consequence of a wrong pre-processing can be seen in Figure 11: some pixels belonging to the abdomen were mistaken for belonging to the lungs and appear on the final image given as input to UIP-net. Consequenlty, UIP-net analyze them and find the disease. This fact is reflected on the outliers in red of Figure 9 which represent the increased amount of diseased tissue detected.

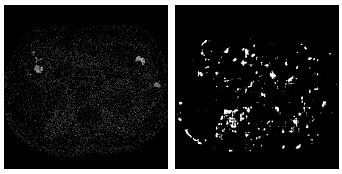

Fig. 11. Original image after an incorrect pre-processing (left): the abdomen are mistaken for lungs by the clustering algorithm. Segmentation mask outputted by UIP-net on the right.

5 Conclusion

In this work, a CNN named UIP-net was proposed for the detection of UIP patterns in HRCT images. A novel architecture was designed in order to preserve

fine details of the texture, thus taking advantage of the excellent quality of the images, both in term of spatial resolution and bit depth.

Future works will consist in:

1. Validating the current version of the UIP-net on additional data.
2. Modifying the network in order to:
   - take into account the 3-D nature of the UIP pattern;
   - provide a differential characterization of UIP patterns (e.g to discriminate between honeycombing and GGO).
3. Improving the generalizability and reliability of the CNN testing it on images belonging to different acquisition contexts. In this work, only images acquired with the same scanner and reconstructed with B60 kernel were involved.
4. Investigating unsupervised and label-independent learning, to boost UIP-net's performance and generalization ability on data not human-annotated.
5. Using the UIP-net to detect also HRCT manifestations of other diseases, first of all those produced by Covid-19.

Acknowledgment

The authors express all their gratitude to Brian J. Bartholomai from the Division of Radiology and Ronald Karwoski from Biomedical Imaging Resource of Mayo Clinic, MN, USA. Thanks to their useful support, it was possible to obtain the labelled images outputted by CALIPER used as ground truth. Without CALIPER, it would not have been possible to train the novel network proposed in this work.

References

1. S.L.F Walsh, Imaging biomarkers and staging in IPF, *Curr Opin Pulm Med*, vol. 24, pp. 445–452, 2018
2. H.J. Kim, D. Perlman, R. Tomic, Natural history of idiopathic pulmonary fibrosis, *Respir. Med.*, vol. 109, pp. 661–670, 2015
3. D.M. Hansell, J.G. Goldin, T.E. King et al., CT staging and monitoring of fibrotic interstitial lung diseases in clinical practice and treatment trials: a position paper from the Fleischner Society, *Lancet Respir. Med.*, vol. 3, pp. 483-496, 2015
4. G. Ragu et al., Diagnosis of Idiopathic Pulmonary Fibrosis. An Official ATS/ERS/JRS/ALAT Clinical Practice Guideline, *Am J Respir Crit Care Med*, vol. 198, pp. e44–e68, 2018
5. S.L. Walsh, L. Calandriello, N. Sverzellati et al., Interobserver agreement for the ATS/ERS/JRS/ALAT criteria for a UIP pattern on CT, *Thorax*, vol. 71, pp. 45-51, 2016
6. Goldin JG., Computed tomography as a biomarker in clinical trials imaging., *J Thorac Imaging*, vol. 28, pp. 291-297, 2013
7. Jacob J., Bartholmai BJ, Rajagopalan S, Kokosi M, Nair A, Karwoski R, et al., Mortality prediction in idiopathic pulmonary fibrosis: evaluation of computerbased CT analysis with conventional severity measures., *Eur Respir J*, vol.49, 2017.

8. HJ Kim e al., Comparison of the quantitative CT imaging biomarkers of idiopathic pulmonary Fibrosis at baseline and early change with an interval of 7 months., *Acad Radiol*, vol. 22, pp. 70–80, 2015
9. Alan C. Best et al., Idiopathic Pulmonary Fibrosis: Physiologic Tests, quantitative CT Indexes, and CT Visual Scores as Predictors of Mortality., *Radiology*, vol. 246, pp 935-940, 2008
10. Wu Xiaoping et al., Computed Tomographic Biomarkers in Idiopathic Pulmonary Fibrosis The Future of Quantitative Analysis., *American Journal of Respiratory and Critical Care Medicine*, vol. 199, pp. 12–21, 2018
11. Jacob J., Brian J. Bartholmai, Srinivasan Rajagopalan, Maria Kokosi, Arjun Nair, Ronald Karwoski, Simon L.F. Walsh, Athol U. Wells, David M. Hansell, Mortality prediction in idiopathic pulmonary Fibrosis: evaluation of computerbased CT analysis with conventional severity measures., *Eur Respir*, vol. 49, 2017
12. Teng Moua, Sushravya Raghunath, Srinivasan Rajagopalan, Ronald Karwoski, Brian Bartholmai, Jay Ryu, Richard Robb, Fabien Maldonado, Can progression of fibrosis as assessed by computer-aided lung informatics for pathology evaluation and rating (CALIPER) predict outcomes in patients with idiopathic pulmonary Fibrosis., *Chest*, vol. 140, 2011
13. Jacob J., Bartholmai BJ, Rajagopalan S, et al., Serial automated quantitative CT analysis in idiopathic pulmonary fibrosis: functional correlations and comparison with changes in visual CT scores., *Eur Radiol.*, vol. 28, pp. 1318-1327, 2018
14. Simon L F Walsh, Lucio Calandriello, Mario Silva, Nicola Sverzellati, Deep learning for classifying Fibrotic lung disease on high-resolution computed tomography: a case-cohort study., *Lancet Respir Med*, vol. 6, pp. 837–45, 2018
15. M. Anthimopoulos, Stergios Christodoulidis, Lukas Ebner, Andreas Christe, and Stavroula Mougiakakou, Lung Pattern Classification for Interstitial Lung Diseases Using a Deep Convolutional Neural Network., *IEEE Transactions on Medical Imaging*, vol. 35, 2016.
16. Kim GB, Jung KH, Lee Y, et al., Comparison of Shallow and Deep Learning Methods on Classifying the Regional Pattern of Diffuse Lung Disease., *J Digit Imaging*, vol. 31, pp. 415-424, 2018.
17. Olaf Ronneberger, Philipp Fischer, Thomas Brox, Convolutional networks for biomedical image segmentation, *Medical Image Computing and Computer-Assisted Inter vention*, 2015
18. Agarwala S, Kale M, Kumar D, et al., Deep learning for screening of interstitial lung disease patterns in high-resolution CT images, *Clinical Radiology*, vol. 75, 2020.
19. Liu F, Zhou Z, Jang H, Samsonov A, Zhao G, Kijowski R. Deep convolutional neural network and 3D deformable approach for tissue segmentation in musculoskeletal magnetic resonance imaging., *Magn Reson Med.*, vol. 79, pp. 2379-2391, 2018.
20. Csurka, Gabriela and Larlus, Diane, What is a good evaluation measure for semantic segmentation?, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, 2013.
21. Ross JC, Estépar RS, Díaz A, et al. Lung extraction, lobe segmentation and hierarchical region assessment for quantitative analysis on high resolution computed tomography images., *Med Image Comput Comput Assist Interv.*, vol. 12, pp. 690-98, 2009.
22. Sato Y, Nakajima S, Shiraga N, et al., Three-dimensional multi-scale line filter for segmentation and visualization of curvilinear structures in medical images., *Med Image Anal.*, vol. 2, pp. 143-168, 1998.