Re-Identification and Growth Detection of Pulmonary Nodules without Image Registration Using 3D Siamese Neural Networks

Xavier Rafael-Palou[a,b,*], Anton Aubanell[c], Ilaria Bonavita[a], Mario Ceresa[b], Gemma Piella[b], Vicent Ribas[a], Miguel A. González Ballester[b,d]

[a]*Eurecat Centre Tecnològic de Catalunya, eHealth Unit, Barcelona, Spain*
[b]*BCN MedTech, Dept. of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona, Spain*
[c]*Vall d'Hebron University Hospital, Barcelona, Spain*
[d]*ICREA, Barcelona, Spain*

Abstract

Lung cancer follow-up is a complex, error prone, and time consuming task for clinical radiologists. Several lung CT scan images taken at different time points of a given patient need to be individually inspected, looking for possible cancerogenous nodules. Radiologists mainly focus their attention in nodule size, density, and growth to assess the existence of malignancy. In this study, we present a novel method based on a 3D siamese neural network, for the re-identification of nodules in a pair of CT scans of the same patient without the need for image registration. The network was integrated into a two-stage automatic pipeline to detect, match, and predict nodule growth given pairs of CT scans. Results on an independent test set reported a nodule detection sensitivity of 94.7%, an accuracy for temporal nodule matching of 88.8%, and a sensitivity of 92.0% with a precision of 88.4% for nodule growth detection.

Keywords:
2000 MSC: 41A05, 41A10, 65D05, 65D17 Lung cancer, Nodule detection, Nodule growth, Transfer learning, Deep Learning

1. Introduction

Lung cancer is the leading cause of cancer death, regardless of gender or ethnicity. Only 19% of all people diagnosed with lung cancer will survive after 5 years, but this percentage improves dramatically when the disease is diagnosed at early stages (Noone et al., 2018).

Small lung nodules are the most common expression of early lung cancer. Their variability in size, texture, and morphology make it difficult to detect them even for clinical specialists. The use of thin-slice helical chest computed tomography (CT) together with the recommendations established by clinical guidelines, such as those of the Fleischner Society (MacMahon et al., 2017), have allowed improving nodule detection rates as well as better identifying the malignancy of incidental nodules. However, recommendations for borderline and complex cases are still vague and open to the judgment and experience of the clinicians.

Current clinical criteria for assessing pulmonary nodule changes are based on visual comparison and diameter measurements from the axial slices of the initial and follow-up CT images (Larici et al., 2017). Three-dimensional assessment provides more accurate and precise nodule measurements, especially for small nodules (Ko et al., 2012). However, it requires the segmentation of the nodule, which is a time-consuming process and highly subjected to intra- and inter- observer variability. This is

*Corresponding author
Email address: xavier.rafael@eurecat.org (Xavier Rafael-Palou)

why it is rarely used in a typical clinical workflow.

Computer-assisted diagnosis (CAD) systems are expected to assist in clinical decision by providing relevant information such as accurate growth rates, increase in solid component, or change in density of the nodules. This information could help specialists to reduce the number of studies for a problematic nodule, decreasing the diagnostic time and, hopefully, reducing the classification of the neoplasm, which should lead to a reduction in morbid mortality (American College of Radiology, 2014).

Recent advances in deep neural networks (Goodfellow et al., 2016) have allowed increasing substantially the performances reported by conventional image processing methods in nodule detection (Setio et al., 2017), segmentation (Messay et al., 2015), and malignancy classification (Ciompi et al., 2017). Some of the main advantages of using deep neural networks rely in their ability to learn and extract, in a very effective way, intricate patterns from the raw data without any previous feature engineering, reuse these patterns in different locations of the image, and even transfer them to different domains (Weiss et al., 2016). Despite the recent explosion of methods based on deep neural networks in the lung cancer domain, most of them are focused on the analysis of a single CT scan.

Few CAD systems (Ardila et al., 2019) have been proposed for the automatic support of lung cancer follow-up. Major developments in the field are mainly limited by the lack of open datasets with annotated series of CTs. To analyze series of CT scans, prior and follow-up lung exams have to be initially registered to facilitate, for instance, the correct re-identification of pulmonary nodules. Several factors compromise the effectiveness of the registration process, such as the variability in the image size and resolution originated by the use of different CT scans, and the variability in the position and breath cycle of the patients when performing the scanning.

Although current medical image registration methods (Song et al., 2017), especially non-linear (Rühaak et al., 2017), report accurate CT alignments, they are still slow and introduce some distortions in the intrinsic structure of the lung, hindering their wide clinical acceptance (Viergever et al., 2016). In addition, other complexities must be addressed, regardless of the quality of the image registration, to enable a proper nodule re-identification, such as the existence of several nodules close to each other, and/or the alteration in texture, size, and even location of the nodules due to disease progression. Therefore, more research is still needed to reliably include the nodule re-identification in different CT scans, in automated tools to support physicians in the analysis of longitudinal studies of lung cancer.

This work aims to take a step in this direction, and proposes a novel approach for the re-identification of pulmonary nodules. In particular, we propose a 3D Siamese neural network Koch (2015) to predict the most likely matching nodules from a series of lung CT scans of the same patient. This approach does not require prior registration of the CT scans, avoiding some of the shortcomings that it entails. In addition, to demonstrate the value of this approach, we integrate it into an automated pipeline aimed to detect the growth of pulmonary nodules over time.

The contributions of this paper with respect to previous works is two-fold. First, we investigate and provide several models for re-identifying lung nodules in CT scans series, relying directly on 3D volumetric data, transfer learning, and siamese neural networks. In this sense, to the best of our knowledge, this would be the first time that the problem of pulmonary nodule re-identification is addressed through deep learning techniques. Secondly, we build and evaluate an automatic pipeline that integrates the proposed models to predict nodule growth from logitudinal CTs.

2. Related work

2.1. Automated nodule re-identification

Lung nodule re-identification (i.e. matching) between current and former CT examinations is necessary for assessing nodule growth or shrinkage. While the majority of lung cancer CAD systems found in the literature focus on the nodule detection task (Loyman and Greenspan, 2019), relatively few automated nodule matching systems have been proposed (partly because of the limited availability of follow up datasets).

An early CAD system for nodule re-identification in series of lung CT scans was proposed in (Ko and Betke, 2001). They reported high performances (86% nodules re-identified) using 8 patients (310 nodules), although some parts of the system required manual intervention (lung apex identification) and no train/test split was reported. In Lee et al. (2007) a commercial CAD system

was evaluated for nodule re-identification for 30 patients (210 nodules) with lung metastasis, reaching a matching rate of 67%. In a cohort of 54 pairs of low-dose multi detector CT screening, a CAD system successfully matched 91.3% of nodules ≥4mm (Beigelman-Aubry et al., 2007). In another commercial CAD evaluation study (Tao et al., 2009) a matching rate of 92.7% was achieved in three serial CT scans from 40 subjects with 143 nodules from the NLST[1]. Another CAD system evaluation (Koo et al., 2012) for automated lung nodule matching using annotations from 4 experts in 57 patients reported between 79% and 92% of accuracy scores. Deep learning-based CAD systems for analysis of longitudinal lung cancer studies are practically nonexistent in literature. An exception is in (Ardila et al., 2019), where a CAD system for end-to-end lung cancer screening is proposed. However, nodule matching was not directly tackled in the study.

All these CAD systems rely on registration of the lungs in the different CT examinations. Performing an accurate registration of lung images is particularly challenging (Murphy et al., 2010) due to the high deformability of the lung tissue and the volume changes during the breathing cycle. Previous studies (Hong et al., 2008; Sun et al., 2007) evaluated methods for rigid and non-rigid registration for matching lung nodules on sequential chest CT scans. Murphy et al. (2011) presented the results of the EMPIRE10 pulmonary image registration challenge, which comprised a comprehensive evaluation and comparison of more than 20 algorithms on 30 thoracic CT pairs. Top-5 algorithms were using different non-rigid transformations. Although non-rigid registration is usually more accurate than rigid registration, rigid registration is substantially more computationally efficient, potentially making it more useful in a busy clinical setting in which real-time processing is necessary. A more recent and complete review of registration methods for medical image series analysis can be found in (Song et al., 2017). The choice of the right registration method and of the correct evaluation metric to assess its performance are of crucial importance as they can affect the results of the analysis.

[1]https://www.cancer.gov/types/lung/research/nlst

2.2. Siamese Neural Networks

The problem of nodule re-identification can be closely related to the one of recognizing the same object in different images. This type of problems has been successfully addressed by siamese neural networks (Bromley et al., 1994) (SNNs). They are designed as two sibling networks, connected by a distance layer at the top, trained to predict matching or mismatching between two input images. The original architecture, first introduced for the problem of signature verification, was later extended by Koch (2015) using convolutional layers and adjusting the optimization metric with a weighted L1 distance between the twin feature vectors of both networks.

SNNs have been extensively used in computer vision matching problems such as tracking objects in videos (Tao et al., 2016), matching pedestrians across multiple camera views (Varior et al., 2016), and matching corresponding patches in satellite images (Hughes et al., 2018).

In the medical image domain, SNNs have been used primarily to extract a latent representation for content-based image retrieval. For instance, Chung and Weng (2017) proposed a SNN, pre-trained on the ImageNet dataset and using a contrastive loss function (Hadsell et al., 2006) to retrieve similar images to the query, using a publicly available dataset of diabetic retinopathy fundus images. Another example is the work by Cai et al. (2019), which applied SNNs to retrieve similar images from several medical image databases of lung, pancreas, and brain. As far as we know, SNNs have not yet been applied to re-identify nodules in a series of lung CT scans.

3. Method

3.1. Nodule re-identification

To solve the problem of nodule re-identification in a pair of CTs of the same patient taken at different time points, we propose building a SNN (Koch, 2015). An appealing characteristic of SNNs is that they rely on a distance metric computed on features extracted automatically by a deep learning network. This should allow greatly accelerating and simplifying the nodule re-identification process avoiding to introduce a registration technique as source of variability and error in the analysis.

Siamese neural networks are composed of a feature extraction component in which two subnetworks (with

shared architecture and weights) process a pair of images at a time to produce two embedding feature vectors directly from the images. A second component (i.e. the head of the network) aims to classify whether the two embedding feature arrays are similar or not. To assess this, the features are passed to a pairwise distance layer that computes a similarity score.

In a previous study (Bonavita et al., 2019), we trained a deep convolution neural network (CNN) for nodule classification able to effectively reduce the number of false positives in the nodule detection problem. In the present work, we have adjusted that network improving its final performance. In particular, we propose a 3D CNN based on a ResNet-34 architecture that expects nodule patches of 32x32x32. As described in the original paper, the patches are pre-processed crops done around the center of the annotated nodules of the lung CT. The nodule classification network was trained from scratch using a large amount of nodule candidates (> 750K) from the LUNA-16 challenge dataset (Setio et al., 2017). Further details on its architecture and performance are shown in the supplementary material (S1).

In the current study, we removed the fully connected layers of the nodule classification network to use it as the backbone of the sibling networks of the feature extraction component of the SNNs. Figure-1 shows the SNN architecture for the nodule re-identification problem. In this figure, we can observe the two components. First, the feature extraction component, which pre-processes the input nodule patches (i.e. taken at different time points, T1 and T2) and uses the sibling network to extract the corresponding feature maps. Second, the classification component composed of the head of the network that predicts if both feature maps are similar or not. These feature maps (solid arrows in Figure-1) come from different levels of the pre-trained sibling networks. Further details about the feature maps and the network heads are described in Subsection 3.1.2 and 3.1.3, respectively.

Different SNNs configurations were proposed (Table-1) to gain further insights into the best parameterizations. To allow a fair comparison of the configurations, we trained the SNNs with the same parameter values. Concisely, the number of epochs was set to 150, the learning rate to 1e-4, the batch size to 8, dropout to 0.3, the early stopping at 10 epochs without any significant improvement, and Adam (Kingma and Ba, 2014) was used for optimization. Finally, random rotation, flip, and zoom were applied for data augmentation.

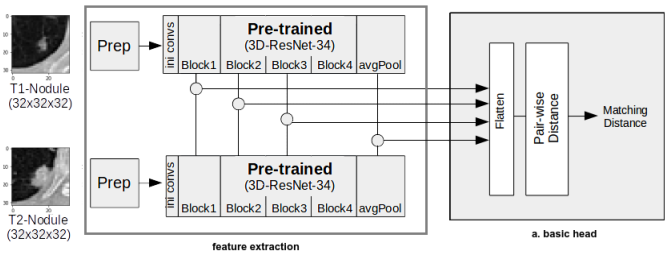

Figure 1: Siamese network proposed for lung nodule re-identification. The network is composed of a feature extraction and a basic head network to perform the prediction.

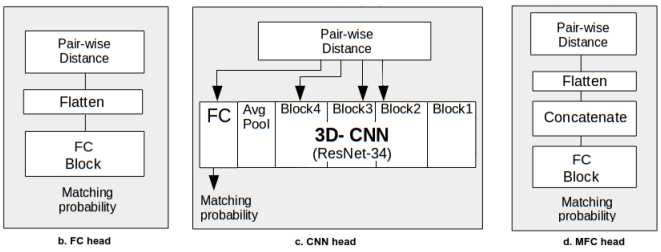

Figure 2: Alternative head networks to configure different siamese networks.

	Pre-trained	Feature maps	Head	Loss
FIBC	Frozen	Individual	Basic	Contrastive
UIBC	Unfrozen	Individual	Basic	Contrastive
FIFB	Frozen	Individual	FC	BCE
UIFB	Unfrozen	Individual	FC	BCE
FICB	Frozen	Individual	CNN	BCE
UICB	Unfrozen	Individual	CNN	BCE
FCMB	Frozen	Combined	MFC	BCE
UCMB	Unfrozen	Combined	MFC	BCE

Table 1: List of the different siamese network configurations. The index column contains the acronyms of the networks, resulting from joining the first letter of the options placed in the next 4 columns.

Below we describe in more detail the main configurations and parameters used in the experiments.

3.1.1. Pre-trained network weights

Two configuration values were proposed for this setting: frozen and unfrozen. Usually, the weights of the pre-

trained networks in a SNN remain frozen. In this study the pre-trained network had a related but slightly different learning goal than the target (siamese) network. Thus, we allowed also the option of unfreezing the weights of the pre-trained network and updating them during the back-propagation steps of the siamese network training process. To un/freeze the networks, we dis/abled the option to update all the weights and biases of the pre-trained layers during training.

3.1.2. Feature maps

We propose two options: using the feature maps individually and combining the feature maps together. Feature maps extracted from the first layers of a CNN refer to low-level and less domain-specific representations (e.g lines, circles, spikes), whereas features extracted from deeper layers are generally more high level and domain-related representation (e.g. morphology, texture). To analyze the potential of both general and more specific nodule features, we used features from different depths of the network (i.e., from the last layer of each of the 4 convolution blocks that holds the pre-trained Resnet-34 network). The resulting feature maps were obtained after a forward-passing through the network for each of the nodule images of the whole dataset. Table-2 shows the layer name, the number of filters per layer, the output dimension of each filter, and the total number of parameters for each of the selected feature maps.

Layer	Filters	Dimension	Total params
layer1	64	[16,8,8]	65536
layer2	128	[8,4,4]	16384
layer3	256	[4,2,2]	4096
avgpool	1	[1,1,512]	512

Table 2: Layers selected from the pre-trained part of the SNNs.

We designed experiments to evaluate each of the possible feature maps, i.e. 4 individual features maps - one per layer - and 11 feature maps resulting from combinations - (4 over 2) + (4 over 3) + (4 over 4).

3.1.3. Siamese heads

We proposed four different head networks, one meant to follow a more conventional siamese architecture and the others with more exploratory purposes, more precisely:

1. A basic head network (Figure-1) composed of a flatten (to homogenize all features to one dimension) and a pairwise distance (i.e. L1) layer, just after the feature extraction part of the network.
2. A fully connected (FC) head network (Figure-2b) composed of a pairwise distance, a flatten, and a FC block layer. The FC block comprises a FC layer (with 64 units), a batch norm, a ReLU, a dropout layer and a final FC layer (with one unit). This classifier head aims at finding non-linear patterns among the merged features (from both sibling networks).
3. A CNN head network (Figure-2c) composed of a pairwise distance layer and a clean (without pre-trained weights) ResNet-34 CNN. Several arrows connect the pairwise distance layer with this clean ResNet-34. There are as many arrows as pre-trained layers used to extract the features. The arrows redirect the features to a specific part of the clean ResNet-34. The redirection had to make compatible the dimensions of the output from the previous layer with the layers of the input. For instance, features extracted from last layer of block1 were linked to the initial layer of the block2, features from layer2 were linked to the initial layer of the block3 and so on. This head network aimed at exploring non-linear patterns between features but without loosing the space dimension (i.e. no flattening of the features was done between the pairwise layer and the clean ResNet-34).
4. A multi-features combined (MFC) head network (Figure-2d) composed of a pairwise distance layer, a flatten layer, a concatenation layer (to merge all features), and a FC (already described above). This head network aimed at exploring combination of features from different parts of the network.

It is important to note that in the basic head network, the pairwise distance layer not only computes the batch-wise L1-distance between each component of the previously flattened input vectors, but also it sums the components up to eventually generate an output of size bs $\times$ shape (where bs is the batch size). This is done to accommodate the expected inputs of the contrastive loss function with which the basic head network is configured. For

the rest of the head networks, the pairwise distance layer does not perform any reducing sum operation, leaving its input and output with the same size bs $\times$ 1 $\times$ z $\times$ y $\times$ x, and therefore, allowing its output to be exploited more deeply with additional layers (for example, convolutional or fully connected).

3.1.4. Loss functions

We explored two options: a contrastive loss and a binary cross entropy (BCE) loss function. Traditionally, SNNs are trained using a contrastive loss (Hadsell et al., 2006) function. This function encodes both similarity and dissimilarity (between the feature maps) independently in a loss function. It ensures that semantically similar pairs are embedded close together while forcing the dissimilar pairs to be apart from each other. Another option to train these networks is through a prediction error-based approach. For our case we adopted the binary cross entropy loss. This implied to apply a sigmoid function on the outputs to transform them into probability values (between 0 and 1).

3.2. Nodule growth detection pipeline

A valuable application of nodule re-identification is to predict nodule growth between current and follow-up CT scans of a patient. This is a crucial, complex, and time-consuming task for lung cancer assessment since nodule growth has a clear predictive importance for benignity and malignancy (Gurney et al., 1993). Thus, further efforts are required to support clinicians to increase the precision and effectiveness of such endeavour.

To this end, we propose an end-to-end pipeline (Figure-3) comprised of two different components: 1) a nodule detector that, given a pair of CTs of the same patient but taken at different time points, generates a list of nodule candidates per each CT, and 2) a nodule matching component (embedding the siamese networks) that, given the list of nodule candidates of the CTs, matches the nodules and computes the difference in diameter between them.

3.2.1. Nodule detector

To build the nodule detector, we followed the work of Liao et al. (2019), with which they won the Data Science Bowl lung cancer challenge[2]. The authors proposed a 3D Faster-RCNN (Ren et al., 2015) scheme for nodule detection. The backbone of the network was similar to the U-net (Ronneberger et al., 2015) architecture, in which the information flows not only in a classical bottom-up way but also between the encoder and decoder parts of the network thanks to some symmetric links (or short-cuts) that bound both parts of the network. The output of this network were probability feature maps, useful for the lung cancer classification problem.

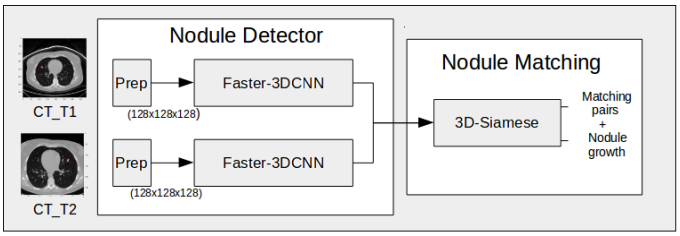

Figure 3: Nodule growth detection pipeline.

To the original network, we proposed attaching a double CNN head as in (Ren et al., 2015). One head was used for regression and the other for classification. The regression branch infers the center (x,y,z locations) and the diameter of the nodule, while the classification branch predicts the probability of being a nodule.

The input lung CT was pre-processed before entering the nodule detection network. The image was resampled to an isotropic resolution (1 $\times$ 1 $\times$ 1mm), pixel intensities clipped between [-1000, 600] HU and normalized between 0 and 1. The full lung image, without any previous lung segmentation, was then split in overlapping patches (due to memory constraints) of 128 $\times$ 128 $\times$ 128 with an overlap of 32 pixels per dimension. Since the location of the patch may influence the decision of whether it is a nodule and whether it is malignant, we also introduce the location information in the network as in (Liao et al., 2019). Thus, each patch was fed to the network together with its corresponding location crop of size 32 $\times$ 32 $\times$ 32 $\times$ 3, which contains the location of the patch image with respect to the whole lung image. The final network architecture used for nodule detection as well as the perfor-

[2]https://www.kaggle.com/c/data-science-bowl-2017

mance obtained in LUNA-16 (Setio et al., 2017) dataset can be found in the supplementary material (S2).

3.2.2. Nodule matching

This component performs the re-identification of the nodules among all CT pairs. To do this, for each pair of CTs, we took each candidate found at T1 and we paired with each of the candidates found at T2. The pairs were pre-processed following the specifications described in Section 4.1, and then they were fed to the SNN. The network, trained off-line, provided a matching probability for each pair of candidates. The pairs with the highest probability were selected as the matching ones.

To assess the performance of this process, we computed for each pair of CTs, whether the candidate at T2 predicted with highest probability by the SNN, matched with the annotated nodule at T2. Additionally, we computed the time required for finding the matching nodules. We repeated this process for each of the SNN configurations.

Once having predicted all matching nodules for each pair of CTs, the pipeline returns the nodule growth along with the location and diameter of the matching nodules. The nodule growth is calculated directly by the difference between the predicted nodule diameters at T1 and T2 for each pair of lung CTs.

To evaluate the nodule growth detection, we selected all the correctly matched CT pairs and compared whether the nodule growth difference was of the same sign in both ground truth and predicted. True positive (TP) and false negative (FN) cases were those that had (in both ground truth and predicted) positive and negative growth differences, respectively. A false positive (FP) case was considered when the predicted growth difference was positive and the ground truth one was negative; and a false negative (FN) was considered in the opposite case.

4. Experiments and results

4.1. Evaluation datasets

4.1.1. LUNA-16

In this work we used an updated version of the LIDC dataset (Armato III et al., 2015) provided in the LUNA-16 challenge (Setio et al., 2017), which includes only scans with at least one lesion of size $\geq$ 3 mm marked as a nodule by at least three of the four radiologists. The LUNA-16 dataset consists of 888 CT scans comprising a total of 1186 nodules. Annotations with coordinates of each nodule in the three spatial axes inferred from the original LIDC annotations are also provided.

4.1.2. VH-Lung

This dataset was designed specifically to identify and follow up suspicious lung nodules in time. Ethics approval was obtained from the Medication Research Ethics Committee of Vall d'Hebrón University Hospital (Barcelona) with reference number PR(AG)111/2019 presented on 01/03/2019.

Inclusion criteria were patients without a previous neoplasia, with a confirmed diagnosis, and with visible nodules ($\geq$ 5 mm) in at least two consecutive CT scans. The interval between current and previous CT examinations ranged from 32 to 2464 days. These nodules were located in the three spatial axes by two different specialists at each time point and quantified by another experienced radiologist. The size mean of annotated nodules was 11.08 $\pm$ 5.35 at T1 and 13.49 $\pm$ 5.18 at T2, and the growth size mean is 2.41 $\pm$ 4.38 mm.

The chest helical CT studies were performed using different scanners: Phillips (Brilliance 16/64, iCT 256), Siemens (SOMATOM Perspective/ Definition) and General Electrics (LightSpeed16). Acquisition and reconstruction protocols were set according to subject biometrics and clinical inquiry: 100–120 kV, 33-196 mAs and exposure time 439-1170 ms. Each image had 512 $\times$ 512 pixels with 16-bit gray resolution, spacing between slices 0.75-1.5 mm and slice thickness 1-5 mm.

In total the dataset contains 151 patients with two thoracic CT scans. For each patient, the clinicians annotated only one relevant nodule in both CT scans. We randomly divided the dataset into two subsets, one for training (113 patients) with 70 cancers and 43 benign cases, and other for testing (38 patients) with 25 cancers and 13 benign cases.

4.2. Nodule re-identification

In this paper we propose the use of SNNs for nodule re-identification. In order to train the SNNs, we first identified positive cases, i.e. pairs of the same nodule from the same patient taken at different time points (T1 and T2), as well as negative cases made up of pairs of mismatched nodules. In the VH-Lung dataset we had already annotated (N=151) positive cases. To create the negative cases

we used the nodule locations of the VH-Lung dataset at T1 together with a random nodule location of the annotated nodule locations at T2 (avoiding to select correct nodule location). In total, we build a balanced dataset (N=302) composed of 226 CT pairs in the training set and 76 CT pairs in the test set, thus respecting the initial training/test (75% / 25%) partition of the VH-Lung dataset.

We optimized the different SNNs (Table-1) with the training data using a stratified 10-fold cross-validation, and we tested them with the testing set. Results of the best SNNs configurations are shown in Table-3. Additional SNNs configuration results can be found in Table-S4 (supplementary material).

In addition, we investigated the nodule re-identification performance in terms of nodule growth. In total we found 14 cases (CT pairs) with an increase in nodule diameter > 9 mm (aprox. Mean + 1.5*std), and 4 cases with a decrease in nodule diameter > 4 mm (aprox. Mean – 1.5*std). We labeled these cases as large growth changes (Other similar studies (Koo et al., 2012) defined large nodules as > 10 mm). We also found 50 cases with a nodule change ± 1 mm, labeling them as small growth changes, and the remaining 87 cases were labeled as medium growth changes. The results for our best method (FIFB) can be found in the Table-S5 of the supplementary material.

4.3. Nodule growth detection pipeline

For the evaluation of the initial stage of the pipeline described in Section 3.2, we first computed the performance of the pipeline to detect the annotated nodules (one per CT). To do this, we proposed different thresholds (1, 4, 8, 16, 32, and 64) or number of nodule candidates, and we computed per each CT whether the annotated nodule was in each subset of predicted nodule candidates (ranked by probability). To have a better estimation of the nodule detection performance, we repeated this process on 10 random train-test partitions (respecting the proposed size of the initial partitions of the dataset) of the VH-Lung. Results are plotted in Figure-4. This FROC curve (Setio et al., 2017), shows the sensitivity, in average, of finding the (only) annotated nodule per scan at different nodule candidate rates. As we can observe, in training the detector reaches a sensitivity of 0.951 with 32 nodule candidates (missing 10.5 ± 1.02 annotated nodules in 226 different CTs), and in test set a sensitivity of 0.973 with the same threshold (missing 2.5 ± 1.02 nodules in 76 CTs).

We therefore configured the nodule detection component of the pipeline with a threshold of 32 candidates per CT, since it empirically showed a good balance between sensitivity (real nodules detected) and precision (number of nodule candidates not really targeted by the clinicians) both in training and test.

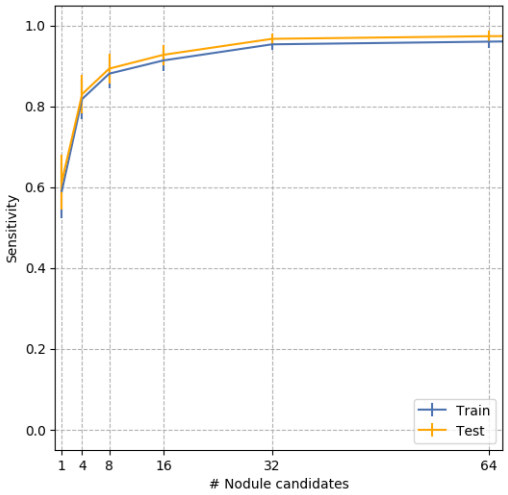

Figure 4: FROC-curve of the malignant nodule detection algorithm for training and test partition.

To gain insight into the complexity of the re-identification problem, we computed how many candidates were located within a chosen Euclidean distance from the nodule ground truth position (Figure-5). We defined 5 different distance thresholds: radius squared Euclidean distance (as used in the LUNA-16 challenge to accept a nodule detection as correct) and 4 fixed Euclidean distances (30, 20, 15, and 10 mm). For every distance, we computed the number of CTs in which 0, 1, 2, 5 or more than 10 candidates fell within the distance. Moreover, we computed an accuracy of detection for every distance choice by dividing the number of CTs for which only one candidate is within the distance by the total number of CTs. Results are shown in Table-4.

Next, we evaluated the performance of the best SNN (Table-3) for nodule re-identification using the location of the nodule candidates provided by the nodule detector. The best results were achieved by the FIFB network with

Configuration	Layer	tr_acc	val_acc	test_acc	test_prec	test_rec
FIBC	layer2	0.790 ± 0.013	0.775 ± 0.051	0.709 ± 0.003	0.806 ± 0.002	0.550 ± 0.007
UIBC	layer3	0.891 ± 0.009	0.864 ± 0.044	0.798 ± 0.018	0.765 ± 0.024	0.863 ± 0.036
FIFB	layer1	0.939 ± 0.025	0.899 ± 0.039	0.921 ± 0.036	0.905 ± 0.054	0.944 ± 0.038
UIFB	layer2	0.918 ± 0.037	0.890 ± 0.039	0.896 ± 0.028	0.871 ± 0.050	0.934 ± 0.017
FICB	layer1	0.867 ± 0.039	0.857 ± 0.060	0.831 ± 0.041	0.824 ± 0.075	0.860 ± 0.061
UICB	layer1	0.868 ± 0.063	0.888 ± 0.049	0.859 ± 0.070	0.842 ± 0.093	0.900 ± 0.046
FCMB	layer1, layer2	0.938 ± 0.034	0.882 ± 0.037	0.918 ± 0.017	0.907 ± 0.029	0.934 ± 0.035
UCMB	layer1, layer2, avgpool	0.954 ± 0.023	0.897 ± 0.045	0.925 ± 0.025	0.904 ± 0.040	0.952 ± 0.032

Table 3: Performance results (accuracy (acc), precision (prec) and recall (rec)) obtained on training (tr), validation (val) and test for the different SNN configurations. The meaning of the configured methods is detailed in Table1.

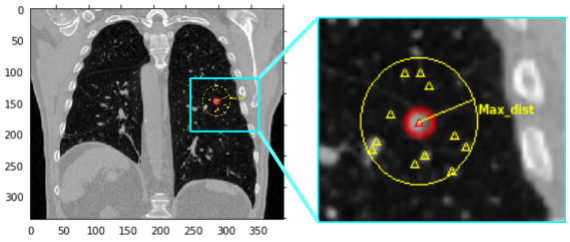

Figure 5: Candidates predicted (yellow marks) at a maximum distance from the ground truth centroid (red circle).

Distance	N=0	N=1	N=2	N=5	N>=10	Acc
radius^2	0	18	6	2	3	0.500
30 mm	1	22	7	1	0	0.611
20 mm	1	26	6	0	0	0.722
15 mm	1	32	3	0	0	0.888
10 mm	1	34	1	0	0	0.944
5 mm	3	33	0	0	0	0.916
3 mm	5	31	0	0	0	0.861
1.5 mm	18	18	0	0	0	0.500

Table 4: Number of CTs (in T2) containing N candidates located within a chosen euclidean distance from the actual nodule centroid. The accuracy score represents the number of CTs at N=1 respect the total of CTs.

only 4 CT-pairs incorrectly matched and an accuracy of 0.888. All results are presented in Table-5.

As in the standalone evaluation of our method, we also conducted some experiments with the best pipeline (FIFB) to investigate nodule re-identification performance in terms of nodule growth. Results are shown in Table-S6 of the supplementary material.

Configuration	Correct	Incorrect	Accuracy	Time(s)
FIBC	25	11	0.694	18.71
UIBC	27	9	0.750	36.01
FIFB	32	4	0.888	9.36
UIFB	30	6	0.834	12.73
FICB	30	6	0.834	20.12
UICB	28	8	0.777	20.16
FCMB	31	5	0.861	12.41
UCMB	31	5	0.861	19.10

Table 5: Results of the different nodule re-identification pipelines. The meaning of the configured methods is detailed in Table1.

Then, we evaluated the performance of the best pipeline (i.e. the pipeline configured with the FIFB network) for the nodule growth detection task. As explained in Section 4.2.2, a correct prediction was achieved when the difference on diameters between predicted and ground truth nodules had both the same sign. In this way, having 32 correctly identified cases (out of 36), we obtained a 0.92 of recall, a 0.88 of precision and a 0.90 of F1-score. The confusion matrix is shown in Figure-6.

Additionally, we assessed the precision in the measurement of the nodule growth prediction. Agreement between the predicted and ground-truth nodule growth vectors was assessed with a Bland-Altman (Altman and Bland, 1983; jaketmp, 2018) plot (Figure-7). The mean difference between the two measurements was 0.17 mm with a 95% confidence interval (from -3.35 to 3.70 mm). Predicted and ground-truth nodule growth vectors were not found statistically different on the basis of a 1-sample t-test (p-value = 0.99). Also, we computed the mean ab-

solute error of the predicted nodule growths (1.38 ± 1.17 mm), their mean squared error (3.26 ± 5.30 mm) and its coefficient of determination (r^2=0.71). Finally, Figure-8 shows the predicted and real difference of diameters for all CT pairs of the test dataset. To support the interpretation of this figure, we have included the axial slice with major diameter taken at time points T1 and T2 of an illustrative subset of nodules.

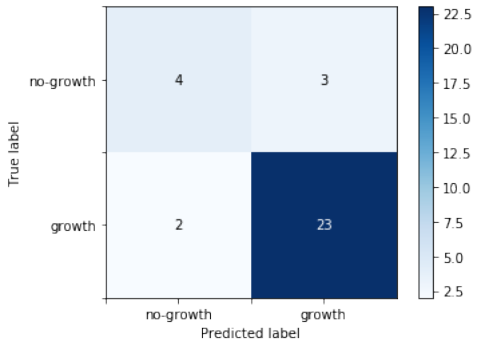

Figure 6: Confusion matrix for nodule growth prediction.

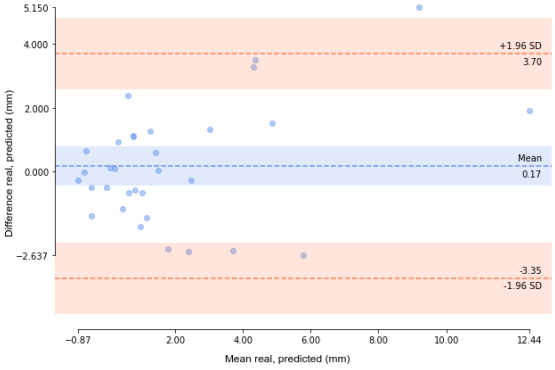

Figure 7: Bland–Altman plot for agreement between ground truth and predicted nodule growth.

4.4. Automatic lung CTs registration

We also computed lung nodule re-identification using conventional image registration methods. To do this, we aligned the CT pairs of the VH-Lung dataset and we computed how far apart were the nodule centroids, annotated by the radiologists, at T2 with the warped locations obtained after applying the transformation-fitted function on the nodule centroids at T1. To do this, we used two well-established methods for image alignment, one for rigid and other for non-rigid registration. Rather than exploring and fine-tunning new registration setups, we leveraged the Elastix (Klein et al., 2009) database[3] of published registration configurations. This is a publicly-available repository of configurations aimed at promoting research reproducibility. Therefore, for the rigid approach we selected a recent configuration already applied for CT images on (Al-Dhamari et al., 2017), and for the non-rigid approach we used an affine registration (Qiao et al., 2015) previously applied for lung CTs.

Table-6 shows the nodule re-identification performances obtained for the two registration methods on the train, test and the whole dataset. Correct cases were those in which the Euclidean distances between the location of the centroids at T2 and the warped locations of the centroids at T1 were less than the nodule' radius squared (same threshold as proposed in LUNA-16 challenge). Accuracy was obtained after summing all correct alignments divided by the total of CT pairs in the dataset. We also computed mean absolute errors (MAE) between the ground truth and the warped centroids and the average time required for performing the alignments.

5. Discussion

In this article, we provide a novel way to address the nodule re-identification problem. In particular, we propose a deep SNN that can directly re-identify nodules located in a series of pairs of CT scans without the need for any image registration.

The SNN allows matching pulmonary nodules in different CTs in a single stage by outputting a similarity score (i.e. the probability of being the same nodule). In contrast, standard techniques require at least two stages: first registering the image and then identifying matching nodules with some distance function. Moreover, with the proposed solution, no additional deformations/perturbations of the lung scan are performed, so that nodule measurements can be done directly from the image itself. Another

[3] http://elastix.bigr.nl/wiki

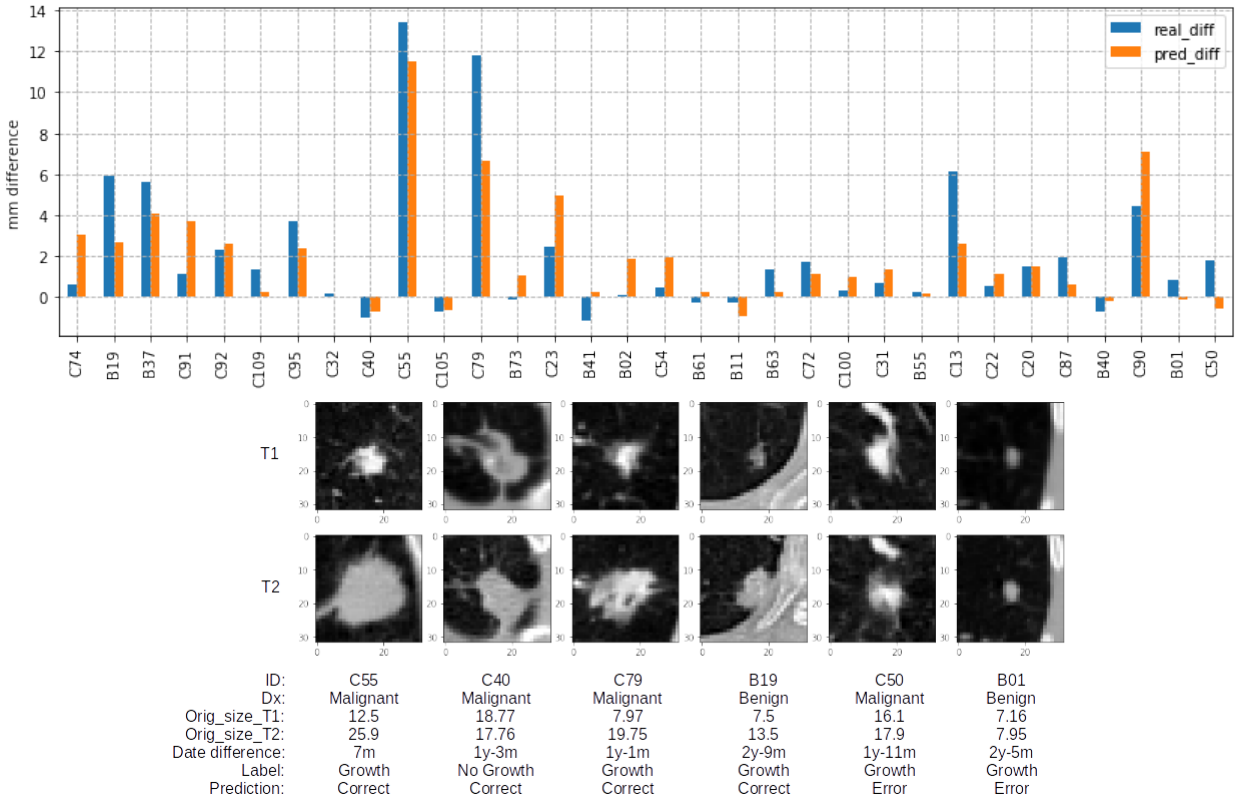

Figure 8: Comparison between real and predicted cases. Upper panel: diameter differences for all test set. Lower panel: axial slices at two time points of different nodules.

	Rigid			Non-Rigid		
	Accuracy	MAE (mm)	Time (s)	Accuracy	MAE (mm)	Time
Train (113 CT pairs)	0.672	30.8±44.2	52.6±10.0	0.761	23.8±39.7	82.2±12.5
Test (38 CT pairs)	0.684	29.6±38.7	52.9±7.7	0.605	30.2±44.3	82.8±9.5
All (151 pairs)	0.675	29.5±43.0	52.7±9.4	0.721	25.4±41.0	82.3±11.8

Table 6: Results after applying automatic registration using rigid and non-rigid approaches.

advantage is that the re-identification process is fast since all weights of the network have already been calculated during the training phase.

We designed and tested several SNN architectures in order to fully understand the complexities of the problem and find the best network configuration. To this end, we collected a longitudinal cohort of two CT scans per patient taken at different time-points. In each of the CT scans of the patients, the most suspicious nodule was annotated according to two different radiologists. Despite the richness of the cohort in terms of heterogeneity in the parameters that affect the image acquisition (e.g. scanners, protocols and setups), in the selected nodules (e.g. size, growth, malignancy), and in the temporal differences between CT studies, the total number of cases to test our approach was limited (38 patients, 25% of the total). Thus, the test set

may not be representative enough of the whole nodule spectrum. To mitigate this issue, despite having presented two different evaluation scenarios, more and diverse number of pulmonary nodules (with different morphologies, locations, sizes, growth rates, or degrees of malignancy) are recommended to collect for a more exhaustive validation of the present work.

As previously mentioned, we have provided two different evaluation scenarios with the intention of showing reliability and usefulness of our approach. In the first evaluation scenario, we trained the models with previous localized fixed image patches from 226 CTs pairs (doubling the original training partition with random negative cases) and we evaluated them using 10-fold cross validation as well as with image patches from 76 CT pairs from the independent test partition (doubling the original test partition with random negative cases). Results (Table-3) showed that, in general (7 out of 8 experiments), the networks obtained high accuracy scores, above 85% in validation and 80% in test. Indeed, several of the SNN configurations (e.g. FIFB, UCMB) achieved accuracy scores in test above 92%. Also, as shown in Table–3 there is no relevant performance gap between training, validation and test sets, which suggests that there is no overfitting.

Regarding the ability to re-identify matching nodules according to their growth (Table-S5 supplementary material), the best SNN (FIFB) obtained a high accuracy score both in training (99.1%) and in test (97.3%), and no significant differences in performance were found despite their nodule growth rates. However, the performance for identifying non-matching nodules was lower than that of the matching cases. In particular, the performance in training was 96.4% and in test 86.8%. This slight drop in test performance was due to errors for predicting non-matching nodules with moderate (2 out of 6 errors) to large change in size (3 out of 6 errors) between time-points. Beyond growth factor, other visual aspects, such as the density and size of the nodules at T0, were not relevant as they were equally distributed among the 6 mismatched pulmonary nodules in the test set. However, 4 of them were found in the left lung and 2 of them in the superior lobe. Also, 3 of these nodules were attached to blood vessels, 2 were close to or attached to the lung wall, and 1 of them was difficult to distinguish from the surrounding lung tissue at T0, whereas at T1 it was clearly visible. The usual appearance of the edges of these nodules was irregular (4 out of 6).

One of the main factors contributing to the good performance is the use of transfer learning, namely initializing the backbone of the different SNNs with the weights of a previously trained 3D network. This can be noted by the fact that the simplest network configuration (FIBC), which it mainly performs a direct forward-pass mechanism of the input through the network, initialized with the weights of the transferred network, reaches, in our opinion, a considerable performance of 77.5% in validation and 71% in tests.

Regarding the loss functions configured in the different experiments, the methods using the BCE loss (which are based on probabilities) slightly outperformed the ones using the contrastive loss (which is based on distances). This can be seen in the difference in accuracy (3.5% in validation and 12% in test) obtained by the best network configured with probability-based loss function (FIFB) compared to the best network configured with loss function based on distance (UIBC).

Another finding was that unfreezing the weights of the pre-trained networks usually allowed for better performances. This is particularly evident in the UIBC case, which exceeded of almost 10% in validation and testing the corresponding frozen configuration (FIBC). Somehow, this finding was expected as weights were transferred from networks trained in a different, although closely related, domain.

With respect to the features used by the networks, we can observe (Table-3) that, in almost all the methods, the best performance was achieved by using features extracted by layer1 and/or layer2, while only for two methods it was achieved using features from layer3 and avgpool (i.e. the global average pooling). This may suggest that features encoding simple patterns (from earlier layers) are preferred for this problem, whereas layers that contains more specific features (from the last layers) are less useful. It is also worth noticing that networks combining features from different layers did not clearly outperform networks using features from a single layer. This is the case of UCMB in which the reported validation performances are just a bit lower (0.2%) than the performances reported by the FIFB configuration, although in test, UCMB outperformed by 0.4% the performance of FIFB.

Concerning the type of heads with which the networks

were configured, the best option was using fully connected layers (FC head). Surprisingly, networks with extra convolution layers before the fully connected layers (CNN head) achieved worse performances (1% and 6% less in validation and test, respectively) than networks with FC heads. This might suggest that adding extra convolution layers to find patterns between locally connected features increases the complexity of the model, leading to more weights to adjust but with the same amount of training data.

In the second evaluation scenario, more ambitious and practical, we integrated the SSNs into automatic pipelines intended first for the detection and re-identification of nodules, and then for the prediction of nodule growth given series of CTs of the same patient. This evaluation was done for both training (113 CT pairs) and testing (38 CT pairs) random partitions of the VH-Lung dataset.

The nodule detector component of the pipeline was configured to provide only the top-32 scored nodule candidates per CT. This threshold was empirically set based on the good balance between precision and recall in terms of nodule detection obtained in both training and test partitions of the VH-Lung dataset. In test, this component reported a nodule detection sensitivity of 97% in 32 nodule candidates (FP) per CT in average. This performance is far from 81.7% sensitivity in 0.125 FP per CT scan in (Huang et al., 2019) and from the results we obtained when training the nodule detector standalone (0.84 sensitivity with 1 FP, in average) in the LUNA-16 dataset. However the comparison is not fair since the nodule detector was not trained to find the most questionable nodule per patient according to radiologist but for detecting any nodule in the lungs, that is why more nodule candidates were needed to find the annotated nodules in the VH-Lung dataset.

Regarding the nodule re-identification step of the pipeline, the performances obtained by the different SSNs networks (Table-5) were lower than when evaluating the models standalone. This was expected since, as opposed to in training, where the patched images were cropped around the ground truth centroid of the nodules, in the pipeline the patches were cropped around the position predicted by the nodule detector, making its correct matching more difficult if the centroid position was not as precise. However, 5 out of 8 networks reached a nodule matching accuracy score above 80%, and the best network (FIFB) reached an accuracy of 88.8%.

In Table-S6 (supplementary material), we reported the performance of the different sub-processes of the best pipeline (FIFB) according to the growth of the nodules. Looking at the results, we can highlight that nodule detection and re-identification steps had high performances both in training (>92%, >85%) and testing (>94%, >88%). However, the training performance for growth detection in small nodules dropped down to 47%. This was not the case for moderate and large nodule changes in neither training nor testing. Different interrelated factors may explain this limitation. One reasonable factor could be the different data proportions between training and test set for this type of nodules. A second factor could be the errors in the ground truth annotations. Another factor could be the limitations from the nodule detector when out-coming the diameter for these nodules. More experiments and tests are required to improve this particular case.

Independently of the growth of the nodules, some common visual appearances were found along with the nodules incorrectly re-identified by the pipeline. In particular, from the 2 non-detected nodules at T0, we would highlight that both were solid and difficult to distinguish from the lung parenchyma (< 9 mm of diameter). From the 4 non-re-identified pair of nodules, 3 of them were malignant and greater than 10 mm at T0. Also, they were located on the right lung and close to or attached to the wall of the lung with irregular edges. Among the 5 pairs of nodules with incorrect growth classification, all of them were solid, 4 of them were malignant and 3 had subcentimeter diameters at T0. Moreover, 3 of them were in the lower right lobe of the lungs, whereas the others were in the upper left lobe. Furthermore, 3 of them were close to the lung wall, 2 had an attached vessel whereas another was close to the mediastinum. Regarding the characteristics of its edges, 2 were irregular and the other 3 smooth.

In terms of computational time, our approach achieved satisfactory performances being able to re-identify the nodules of the complete test set in times ranging from half a minute (in the worst case, UIBC) to less than 10 seconds (for the best configuration, FIFB), as can be seen in Table-5. This is a particularly appealing feature of our method, since even the most recent techniques for registration of lung CT images, necessary by any standard pipeline for nodule re-identification, require significantly more time,

for instance 5 minutes according to Rühaak et al. (2017) or approximately 1 minute by Zikri et al. (2019) per case. These processing times fluctuate substantially depending on the technique and the quality of the image registration.

To have a better intuition of the performances obtained using the proposed pipelines for the automatic nodule re-identification problem, we compared them with two conventional methods for lung image registration (Table-6). Both registration mechanisms were slower and did not outperform the performances reported by any of the configured pipelines. The accuracy differences using the worst (FIBC) and best (FIFB) pipelines compared with the rigid alignment were between 1% and 20.4%, and with the non-rigid alignment between 8.9% and 28.3%. Despite these differences in performance, more advanced registration techniques and further fine-tunning of its parameters would lead to greater re-identification performances. For example, in (Gu et al., 2011), the authors compared rigid and non-rigid registration methods for matching 60 diverse nodules in 60 lung CT pairs obtaining average registration errors (Euclidean distances between baseline and follow-up after alignment) between 9.5 and 10 mm. Also, in Jo et al. (2014), the authors using a rigid registration along with a rib based adjustment mechanism reported a registration errors of 17 ± 7mm for 69 lung nodules in 50 subjects with series of two CTs.

Compared to the latest CAD systems providing nodule re-identification (Tao et al., 2009; Koo et al., 2012), our method reports similar performances (92% accuracy) when evaluated standalone, but slightly below when integrated in pipelines. A number of factors may explain this difference. First, our approach is fully automated, whereas in those systems the position of the reference nodule, to match with, was given by the radiologists. Second, in those systems the data they used for evaluation was from lung cancer screening population, which makes the underlying lung tissue structure more consistent when compared to patients with lung metastases or from incidental cases like ours. Third, in our study, the total number of patients was more than double the number of patients used in these studies (40 and 53), which makes re-identification more difficult since the similarity of the lung structures between nodules is less plausible. In another related study (Jo et al., 2014) for lung nodule re-identification, they reported rates from 29% to 100% in 69 nodules from 50 different patients. However, in their experimental dataset, no severe lesions were reported (e.g. 14 nodules had no changes in diameter between corresponding nodules), and their method was evaluated using the entire cohort, making it difficult to know their ability to generalize to new cases.

Although the focus of the paper is the nodule re-identification, we also quantified and assessed nodule growth. To do this, we selected the best network for nodule re-identification (FIFB) and integrated it in the nodule-growth pipeline. In total, nodule growth was correctly detected in 27 cases and erroneously in 5 cases. However, only 2 of these errors were false negatives (that is, the pipeline failed to predict growth); one of them was on a benign nodule (B01) with growth difference of less than 1 mm, whereas the other was on a malignant nodule (C50) with growth difference of 1.8 mm. As shown in Figure-7, there is an agreement when comparing predicted and real nodule growths as most of the measures fall between the two standard deviations of the mean, there is a non-significant difference between them (p=0.99), and they show a good correlation score (r^2=0.71). Despite this positive results, the values obtained for the 95% limits of agreement (> 3mm) are still high. This was somehow expected as quantifying lung nodules is complex and subject to multiple variability factors (Li et al., 2015) (e.g slice thickness, reconstruction kernel algorithms, attachment of vessels, patient inspiration depth). An example of this was shown in a previous study (de Hoop et al., 2009), in which up to six different open software packages measured the volumetry of solid lung nodules, and reported large nodule inter-variabilities (from 16.4% to 22.3%) on repeated CTs of the same patient in a cohort of 20 patients.

In our case, as we can see in the BA plot (Figure-7), the cases that experiment higher disagreements are those nodules with larger mean nodule growth (i.e. observations located in the right part). A reason that could explain it is that the nodule detector (which reports the nodule diameter) was trained in a database (LUNA-16) with a smaller nodule size distribution (8.30 ± 4.75 mm) than the one used for the evaluation of the pipeline (VH-Lung dataset with 12.45 ± 4.32 mm). Alternatives to address this issue could range from gathering more annotated data, increasing the distribution of large nodules by applying further data augmentation, implementing more sophisticated mechanisms (e.g. attention networks (Schlemper et al., 2019)) in the nodule detector, or instead of using the pre-

dicted diameter and centroid of the nodule detector, implementing deep nodule segmentation networks.

From a clinical point of view, the majority of the nodule differences were correctly classified (growth, no-growth) as shown in Figure-6. Indeed, we reported a mean absolute error of 1.38 ± 1.17mm in diameter with respect to the ground truth, which is slightly less than the 1.73 and 2.2mm of the variability error reported in different retrospective analysis (Revel et al., 2004; Kim et al., 2016) measuring changes in solid and subsolid nodules (<2cm) using only their diameter.

This study, however, is subject to several limitations. First, the limited number of cases to build our models. In the medical domain, longitudinal data is scarce, and much more complex to collect and manage than single time-point studies. Specially for lung cancer assessment, gathering large quantities of samples is even more difficult for different reasons. First, the disease in the early stages is asymptomatic and very aggressive, so when patients are explored, their pulmonary nodules often have clear signs of malignity, and radiologists do not require further studies for its diagnosis. Second, data is usually incomplete or missing, which suppose a real challenge in evolutionary studies. Although there are different initiatives that aim to screen large populations at risk (e.g. NLST), the access to these assets is not publicly open. Thus, having an insufficiently large dataset can negatively impact the performance of deep learning-based models. This is even more concerning for re-identification of lung nodules, since for each patient, twice as many images and annotations are needed. Another main limitation of the study is that the only expert annotation provided for nodule quantification was the major axial diameter. Although the diameter is the most common radiological measure used in practice for nodule growth assessment, using 3D measurements could lead to a more accurate quantification. In addition, if we would have had nodule measurements from more experts, we could have better explained the clinical variability, reporting more accurately the performance of our pipeline with respect to nodule growth prediction. Another limitation of our method could be on re-identifying structures with strong size variations. Some actions may be done to amend this aspect. First, retraining the model with larger input patch sizes. Second, making further data augmentation especially on image pairs with large size variation or collecting more cases of this typology. However, according to radiologists' recommendations, clinical guidelines (American College of Radiology, 2014), and literature (Larici et al., 2017), the challenge is to provide automatic support for growth detection at small/medium nodule change sizes, since larger nodules are easier to identify and substantial differences in growth ratio indicate a clear symptom of either malignancy (Siegelman et al., 1986) or benignity (Gurney et al., 1993). Finally, in this work, we focus on training and evaluating several SNNs to explore different configurations. Finer tuning of hyperparameters (e.g. the learning rates, batch sizes or dropout values) may lead to improved results.

Nevertheless, the automated re-identification of regions of interest in medical images over time, without the need to warp the inherent image structure, could be an appealing application beyond lung cancer assessment such as therapy follow-up as well as for different diseases located at different organs (e.g. prostate, breast cancer) in the body.

Several future works have been described in the paper, and some others are envisaged to extend the research presented in this paper. For example, it would be interesting to longitudinally evaluate the pipeline for more than one nodule per patient, or exploring the nodule spatial localization for the re-identification problem. Also, applying different feature fusion techniques, introducing different manners to weigh the feature maps, applying new techniques to reduce the dimensionality of the problem, as well as the use of segmentation could be some other research lines that would be worth exploring beyond this paper.

6. Conclusions

In this paper, we address the problem of automatic re-identification of pulmonary nodules in lung cancer follow-up studies, using siamese neural networks (SNNs) to rank similarity between nodules, which overpasses the need of image registration. This change of paradigm avoids possible image disturbances and provides computationally faster results. Different configurations of the conventional SNN were examined, ranging from the application of transfer learning, using different loss functions, to the combination of several feature maps of different network levels. The best results during the off-line

training of the SNNs reached accuracies (0.89 in cross-validation and 0.92 in test) similar to those reported by state of the art registration mechanisms. Finally, we embedded the best SNN into a two-stage nodule growth detection pipeline. Nodule re-identification results reported by the pipeline in an independent test set were fast (<10 seconds, matching 38 pairs of CTs) and precise (0.88 accuracy score). Nodule growth predictions were also accurate (0.92 sensitivity score), and both the predicted and the ground truth measurements were not significantly different (p=0.99).

Acknowledgments

This work was partially funded by the Industrial Doctorates Program (AGAUR) grant number DI087, and the Spanish Ministry of Economy and Competitiveness (Project INSPIRE FIS2017-89535-C2-2-R, Maria de Maeztu Units of Excellence Program MDM-2015-0502).

References

Al-Dhamari, I., Bauer, S., Paulus, D., Lissek, F., Jacob, R., 2017. Acir: automatic cochlea image registration, in: Medical Imaging 2017: Image Processing, International Society for Optics and Photonics. p. 1013310.

Altman, D.G., Bland, J.M., 1983. Measurement in medicine: the analysis of method comparison studies. Journal of the Royal Statistical Society: Series D (The Statistician) 32, 307–317.

Ardila, D., Kiraly, A.P., Bharadwaj, S., Choi, B., Reicher, J.J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G., et al., 2019. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nature Medicine 25, 954.

Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, Anthony P, .C.L.P., 2015. Data from LIDC-IDRI. the Cancer Imaging Archive `http://doi.org/10.7937/K9/TCIA.2015.LO9QL9SX`.

Beigelman-Aubry, C., Raffy, P., Yang, W., Castellino, R.A., Grenier, P.A., 2007. Computer-aided detection of solid lung nodules on follow-up MDCT screening: evaluation of detection, tracking, and reading time. American Journal of Roentgenology 189, 948–955.

Bonavita, I., Rafael-Palou, X., Ceresa, M., Piella, G., Ribas, V., González Ballester, M.A., 2019. Integration of convolutional neural networks for pulmonary nodule malignancy assessment in a lung cancer classification pipeline. Computer Methods and Programs in Biomedicine 185, 1–9.

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R., 1994. Signature verification using a "siamese" time delay neural network, in: Advances in Neural Information Processing Systems, pp. 737–744.

Cai, Y., Li, Y., Qiu, C., Ma, J., Gao, X., 2019. Medical image retrieval based on convolutional neural network and supervised hashing. IEEE Access 7, 51877–51885.

Chung, Y.A., Weng, W.H., 2017. Learning deep representations of medical images using siamese CNNs with application to content-based image retrieval. Advances in Neural Information Processing Systems. Workshop on Machine Learning for Health (ML4H) .

Ciompi, F., Chung, K., Van Riel, S.J., Setio, A.A.A., Gerke, P.K., Jacobs, C., Scholten, E.T., Schaefer-Prokop, C., Wille, M.M., Marchiano, A., et al., 2017. Towards automatic pulmonary nodule management in lung cancer screening with deep learning. Scientific Reports 7, 46479.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep learning. Cambridge: MIT press.

Gu, S., Wilson, D., Tan, J., Pu, J., 2011. Pulmonary nodule registration: Rigid or nonrigid? Medical Physics 38, 4406–4414.

Gurney, J.W., Lyddon, D.M., McKay, J.A., 1993. Determining the likelihood of malignancy in solitary pulmonary nodules with bayesian analysis. part ii. application. Radiology 186, 415–422.

Hadsell, R., Chopra, S., LeCun, Y., 2006. Dimensionality reduction by learning an invariant mapping, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), IEEE. pp. 1735–1742.

Hong, H., Lee, J., Yim, Y., 2008. Automatic lung nodule matching on sequential ct images. Computers in Biology and Medicine 38, 623 – 634.

de Hoop, B., Gietema, H., van Ginneken, B., Zanen, P., Groenewegen, G., Prokop, M., 2009. A comparison of six software packages for evaluation of solid lung nodules using semi-automated volumetry: what is the minimum increase in size to detect growth in repeated ct examinations. European radiology 19, 800–808.

Huang, W., Xue, Y., Wu, Y., 2019. A cad system for pulmonary nodule prediction based on deep three-dimensional convolutional neural networks and ensemble learning. PLOS ONE 14, 1–17. URL: https://doi.org/10.1371/journal.pone.0219369.

Hughes, L.H., Schmitt, M., Mou, L., Wang, Y., Zhu, X.X., 2018. Identifying corresponding patches in sar and optical images with a pseudo-siamese CNN. IEEE Geoscience and Remote Sensing Letters 15, 784–788.

jaketmp, 2018. jaketmp/pycompare: Looks both ways. URL: https://doi.org/10.5281/zenodo.1256204, doi:10.5281/zenodo.1256204.

Jo, H.H., Hong, H., Goo], J.M., 2014. Pulmonary nodule registration in serial ct scans using global rib matching and nodule template matching. Computers in Biology and Medicine 45, 87 – 97.

Kim, H., Park, C.M., Song, Y.S., Sunwoo, L., Choi, Y.R., Im Kim, J., Kim, J.H., Bae, J.S., Lee, J.H., Goo, J.M., 2016. Measurement variability of persistent pulmonary subsolid nodules on same-day repeat CT: what is the threshold to determine true nodule growth during follow-up? PLoS One 11, e0148853.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization, in: Proceedings of the 3rd International Conference on Learning Representations.

Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P., 2009. Elastix: a toolbox for intensity-based medical image registration. IEEE transactions on medical imaging 29, 196–205.

Ko, J., Betke, M., 2001. Chest ct: automated nodule detection and assessment of change over time–preliminary experience. Radiology 218, 267—273. URL: https://doi.org/10.1148/radiology.218.1.r01ja39267, doi:10.1148/radiology.218.1.r01ja39267.

Ko, J.P., Berman, E.J., Kaur, M., Babb, J.S., Bomsztyk, E., Greenberg, A.K., Naidich, D.P., Rusinek, H., 2012. Pulmonary nodules: growth rate assessment in patients by using serial CT and three-dimensional volumetry. Radiology 262, 662–671.

Koch, G., 2015. Siamese neural networks for one-shot image recognition, in: International Conference on Machine Learning. Workshop on Deep Learning, vol. 2.

Koo, C.W., Anand, V., Girvin, F., Wickstrom, M.L., Fantauzzi, J.P., Bogoni, L., Babb, J.S., Ko, J.P., 2012. Improved efficiency of CT interpretation using an automated lung nodule matching program. American Journal of Roentgenology 199, 91–95.

Larici, A.R., Farchione, A., Franchi, P., Ciliberto, M., Cicchetti, G., Calandriello, L., del Ciello, A., Bonomo, L., 2017. Lung nodules: size still matters. European Respiratory Review 26, 170025.

Lee, K.W., Kim, M., Gierada, D.S., Bae, K.T., 2007. Performance of a computer-aided program for automated matching of metastatic pulmonary nodules detected on follow-up chest CT. American Journal of Roentgenology 189, 1077–1081.

Li, Q., Gavrielides, M.A., Sahiner, B., Myers, K.J., Zeng, R., Petrick, N., 2015. Statistical analysis of lung nodule volume measurements with ct in a large-scale phantom study. Medical physics 42, 3932–3947.

Liao, F., Liang, M., Li, Z., Hu, X., Song, S., 2019. Evaluate the malignancy of pulmonary nodules using the 3-D deep leaky noisy-or network. IEEE Transactions on Neural Networks and Learning Systems 30, 3484–3495.

Loyman, M., Greenspan, H., 2019. Lung nodule retrieval using semantic similarity estimates, in: Medical Imaging 2019: Computer-Aided Diagnosis, International Society for Optics and Photonics. p. 109503P.

MacMahon, H., Naidich, D.P., Goo, J.M., Lee, K.S., Leung, A.N., Mayo, J.R., Mehta, A.C., Ohno, Y., Powell, C.A., Prokop, M., et al., 2017. Guidelines for management of incidental pulmonary nodules detected on CT images: from the fleischner society 2017. Radiology 284, 228–243.

Messay, T., Hardie, R.C., Tuinstra, T.R., 2015. Segmentation of pulmonary nodules in computed tomography using a regression neural network approach and its application to the lung image database consortium and image database resource initiative dataset. Medical Image Analysis 22, 48–62.

Murphy, K., Van Ginneken, B., Reinhardt, J., Kabus, S., Ding, K., Deng, X., Pluim, J., 2010. Evaluation of methods for pulmonary image registration: The empire10 study. Grand Challenges in Medical Image Analysis 2010, 11–22.

Murphy, K., van Ginneken, B., Reinhardt, J.M., Kabus, S., Ding, K., Deng, X., Cao, K., Du, K., Christensen, G.E., Garcia, V., Vercauteren, T., Ayache, N., Commowick, O., Malandain, G., Glocker, B., Paragios, N., Navab, N., Gorbunova, V., Sporring, J., de Bruijne, M., Han, X., Heinrich, M.P., Schnabel, J.A., Jenkinson, M., Lorenz, C., Modat, M., McClelland, J.R., Ourselin, S., Muenzing, S.E.A., Viergever, M.A., De Nigris, D., Collins, D.L., Arbel, T., Peroni, M., Li, R., Sharp, G.C., Schmidt-Richberg, A., Ehrhardt, J., Werner, R., Smeets, D., Loeckx, D., Song, G., Tustison, N., Avants, B., Gee, J.C., Staring, M., Klein, S., Stoel, B.C., Urschler, M., Werlberger, M., Vandemeulebroucke, J., Rit, S., Sarrut, D., Pluim, J.P.W., 2011. Evaluation of registration methods on thoracic ct: The empire10 challenge. IEEE Transactions on Medical Imaging 30, 1901–1920. doi:10.1109/TMI.2011.2158349.

Noone, A., Howlader, N., Krapcho, M., Miller, D., Brest, A., Yu, M., Ruhl, J., Tatalovich, Z., Mariotto, A., Lewis, D., et al., 2018. Seer cancer statistics review, 1975-2015. Bethesda, MD: National Cancer Institute .

Qiao, Y., van Lew, B., Lelieveldt, B.P., Staring, M., 2015. Fast automatic step size estimation for gradient descent optimization of image registration. IEEE transactions on medical imaging 35, 391–403.

American College of Radiology, 2014. Lung CT screening reporting and data system (lung-RADS). Reston, VA: American College of Radiology .

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster R-CNN: Towards real-time object detection with region proposal networks, in: Advances in Neural Information Processing Systems, pp. 91–99.

Revel, M.P., Bissery, A., Bienvenu, M., Aycard, L., Lefort, C., Frija, G., 2004. Are two-dimensional CT measurements of small noncalcified pulmonary nodules reliable? Radiology 231, 453–458.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 234–241.

Rühaak, J., Polzin, T., Heldmann, S., Simpson, I.J., Handels, H., Modersitzki, J., Heinrich, M.P., 2017. Estimation of large motion in lung CT by integrating regularized keypoint correspondences into dense deformable registration. IEEE Transactions on Medical Imaging 36, 1746–1757.

Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., Rueckert, D., 2019. Attention gated networks: Learning to leverage salient regions in medical images. Medical Image Analysis 53, 197 – 207. URL: http://www.sciencedirect.com/science/article/pii/S1361841518306133, doi:https://doi.org/10.1016/j.media.2019.01.012.

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., 2017. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. Medical Image Analysis 42, 1–13.

Siegelman, S.S., Khouri, N.F., Leo, F., Fishman, E.K., Braverman, R., Zerhouni, E., 1986. Solitary pulmonary nodules: Ct assessment. Radiology 160, 307–312.

Song, G., Han, J., Zhao, Y., Wang, Z., Du, H., 2017. A review on medical image registration as an optimization problem. Current Medical Imaging Reviews 13, 274–283.

Sun, S., Rubin, G.D., Paik, D., Steiner, R.M., Zhuge, F., Napel, S., 2007. Registration of lung nodules using a semi-rigid model: Method and preliminary results. Medical Physics 34, 613–626.

Tao, C., Gierada, D.S., Zhu, F., Pilgram, T.K., Wang, J.H., Bae, K.T., 2009. Automated matching of pulmonary nodules: evaluation in serial screening chest CT. American Journal of Roentgenology 192, 624–628.

Tao, R., Gavves, E., Smeulders, A.W., 2016. Siamese instance search for tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1420–1429.

Varior, R.R., Haloi, M., Wang, G., 2016. Gated siamese convolutional neural network architecture for human re-identification, in: European Conference on Computer Vision, Springer. pp. 791–808.

Viergever, M.A., Maintz, J.A., Klein, S., Murphy, K., Staring, M., Pluim, J.P., 2016. A survey of medical image registration – under review. Medical Image Analysis 33, 140 – 144. URL: http://www.sciencedirect.com/science/article/pii/S1361841516301074, doi:https://doi.org/10.1016/j.media.2016.06.030. 20th anniversary of the Medical Image Analysis journal (MedIA).

Weiss, K., Khoshgoftaar, T.M., Wang, D., 2016. A survey of transfer learning. Journal of Big data 3, 9.

Zikri, Y.K.B., Helguera, M., Cahill, N.D., Shrier, D., Linte, C.A., 2019. Toward an affine feature-based registration method for ground glass lung nodule tracking, in: ECCOMAS Thematic Conference on Computational Vision and Medical Image Processing, Springer. pp. 247–256.