Medical & Biological Engineering & Computing (2020) 58:1803–1815
https://doi.org/10.1007/s11517-020-02197-7

ORIGINAL ARTICLE

Automated lung cancer diagnosis using three-dimensional convolutional neural networks

Gustavo Perez[1] · Pablo Arbelaez[1]

Received: 27 September 2019 / Accepted: 20 May 2020 / Published online: 5 June 2020

Abstract

Lung cancer is the deadliest cancer worldwide. It has been shown that early detection using low-dose computer tomography (LDCT) scans can reduce deaths caused by this disease. We present a general framework for the detection of lung cancer in chest LDCT images. Our method consists of a nodule detector trained on the LIDC-IDRI dataset followed by a cancer predictor trained on the Kaggle DSB 2017 dataset and evaluated on the IEEE International Symposium on Biomedical Imaging (ISBI) 2018 Lung Nodule Malignancy Prediction test set. Our candidate extraction approach is effective to produce accurate candidates with a recall of 99.6%. In addition, our false positive reduction stage classifies successfully the candidates and increases precision by a factor of 2000. Our cancer predictor obtained a ROC AUC of 0.913 and was ranked 1st place at the ISBI 2018 Lung Nodule Malignancy Prediction challenge.

Keywords Computed tomography · Computer-aided diagnosis · Convolutional neural networks · Deep learning · Lung cancer

1 Introduction

Cancer is the main cause of death worldwide, accounting for 8.2 million deaths per year approximately. Lung cancer leads this list with 1.69 million deaths per year [50]. Early detection with the aid of low-dose computer tomography (LDCT) scans has shown to reduce lung cancer mortality by 16 to 20% compared with standard chest X-ray among adults [2]. Unlike conventional X-rays, LDCT scanning provides very detailed images of many types of tissue in three dimensions, which avoid the overlapping of several layers of different tissues in a single image. In January 2013, the American Cancer Society issued guidelines for early detection of lung cancer based on a systematic review of the evidence. These guidelines endorse a process of shared decision-making between clinicians who have access to high-volume lung cancer screening programs [2]. Lung cancer is a malignant lung tumor characterized by uncontrolled cell growth in tissues of the lung [27]. This growth can spread beyond the lung by the process of metastasis into nearby tissue or other parts of the body [11]. Although the *majority of lung nodules (*at least 60 percent of nodules overall) are not cancerous [23], lung cancer diagnosis requires the identification of non-lung tissues to be able to perform a biopsy on them and confirm that they are benign. Therefore, nodule detection is directly related to cancer diagnosis. However, the consensus in lung nodule detection by radiologists is less than 52% when detecting nodules of any size [3]. As shown in Fig. 1, the difficulty in the early diagnosis of lung cancer is due to the variability in shape and size of nodules, and the high unbalance between the nodules and other lung structures and tissues.

A great amount of research has been conducted over the past two decades in computer-aided detection (CAD) systems for lung cancer in LDCT scans [18, 19]. A large number of systems for cancer detection have been proposed in the literature [48]. However, low sensitivity and high false positive rates are still issues that prevent the use of these systems in the daily clinical practice.

Recently, significant research has been done with the use of deep learning techniques, following their recent success for detection, segmentation, and recognition on natural [12, 17, 33, 36, 39, 43] and medical [13, 21, 24, 25] images makes inescapable the application of these machine learning methods for lung cancer CAD systems. Due to the variability and the

✉ Gustavo Perez
ga.perezs@uniandes.edu.co

[1] Universidad de los Andes, Cra 1 N 18A-12, Bogota 111711, Colombia

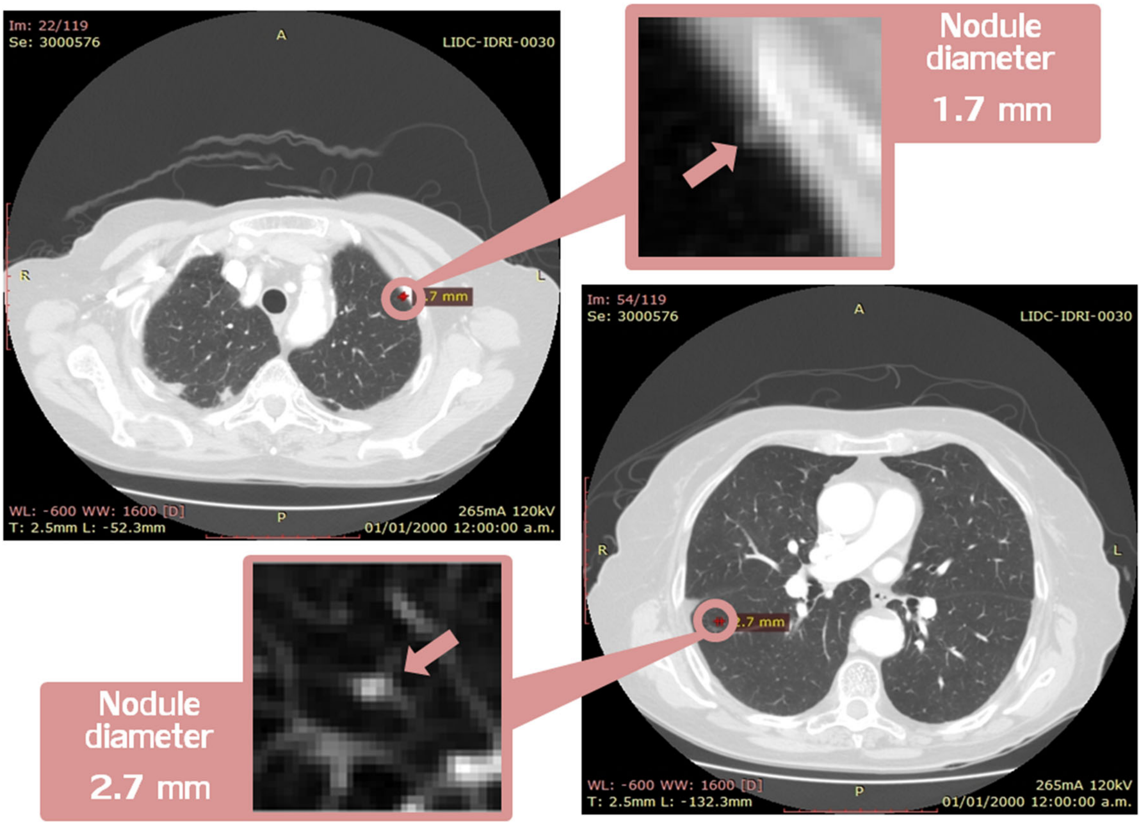


Fig. 1 Examples of annotated nodules < 3 mm on the LIDC-IDRI dataset. Left, juxtapleural nodule of diameter 1.7 mm. Right, parenchymal nodule of diameter 2.7 mm surrounded by vessels

high unbalance between nodules and other lung structures, handcrafted features for this task are difficult. In contrast, automatically learned features from a convolutional neural network yield conceptual abstractions by each layer in a hierarchical way and typically outperform handcrafted features.

Our general framework is shown in Fig. 2. To achieve the goal of cancer diagnosis from a LDCT scan, we implement a stage for pre-processing using filtering and lung extraction from the entire volume for each subject. From the extracted lungs, we generate nodule candidates using morphological operations. We then use the extracted candidates to train a three-dimensional convolutional neural network for nodule classification and false positive reduction. With the top-scored detected nodules, we train a cancer predictor to produce a final malignancy score per subject. We conduct experiments on the largest publicly available database with individual nodule annotations, the Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) for our nodule detector and on the IEEE International Symposium on Biomedical Imaging (ISBI) 2018 Lung Nodule Malignancy Prediction dataset for our cancer predictor. Our computer-aided system for the detection of lung cancer was ranked 1st place at the ISBI 2018 Lung Nodule Malignancy Prediction challenge [22]. In order to ensure reproducibility of our results and to promote further research on automated lung cancer diagnosis, our source code and pre-trained models are publicly available at https://github.com/BCV-Uniandes/LungCancerDiagnosis-pytorch.

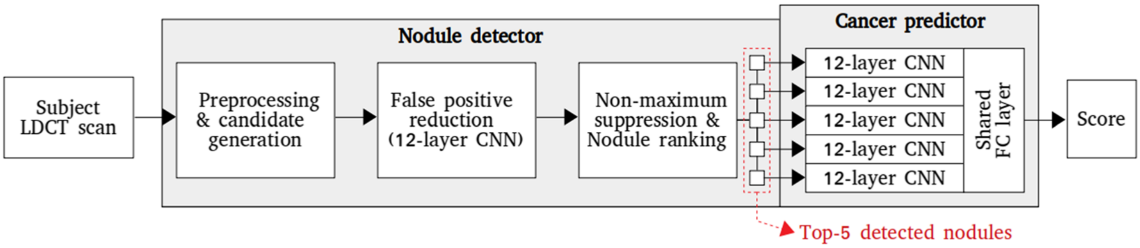


Fig. 2 Proposed method: pre-processing for noise reduction and lung extraction with a mask, candidate generation using morphological operations, nodule classification with a three-dimensional convolutional neural network to reduce false positives and increase precision, and a 5-way convolutional neural network to obtain a final cancer probability for each subject

Table 1 Annotation statistics over 1010 subjects from the LIDC-IDRI dataset. For this project, we use the included nodules by at least one radiologist (i.e., total included nodules). "Included by 4 annotators" refer to nodules included in the unblinded stage by all 4 radiologists

Total nodules	11,608	100%	-
Total included nodules	*6287*	*54.2%*	*100%*
Included by 4 annotators	3233	27.8%	51.4%

Italic indicates the best result

2 Related work

Computer-assisted lung cancer diagnosis is divided into two main problems, i.e., lung nodule detection and lung cancer prediction. Several algorithms have focused on nodule detection as a critical intermediate step for the prediction of lung cancer [9, 15, 29, 30, 41, 45]. Other approaches try to predict cancer nodules from nodule candidate patches, avoiding explicit detection of nodules [15, 26, 34, 44, 47]. Our framework learns an intermediate nodule detector whose detections are then used as input for cancer prediction.

Several algorithms rely on thresholding methods and morphological operations for nodule segmentation, followed by feature extraction and classification. In 2007, Dolejsi et al. [9] proposed an algorithm for segmentation of nodules in two separate ways, morphological closing and thresholding, to find juxtapleural nodules and 3D blob detector with multiscale filtration to locate non-pleural nodule candidates. For classification, linear and multi-threshold classifiers were used. In 2007, Osman et al. [30] proposed a CAD system using template matching over the 3D volume to generate candidates. The false positive reduction was made using connected components and the sum of differences of densities in the surrounding pixels. In 2012, Sudha et al. [41] proposed a global thresholding algorithm following an iterative approach for lung volume extraction. The nodule segmentation stage was made by thresholding and morphological reconstruction. Another method, using template matching for nodule segmentation, was proposed by Tartar et al. [45] in 2013. False positive reduction was conducted with decision trees. Other algorithms using morphological operations for candidate extraction and different types of classifiers for false positive reduction have also been proposed [15, 29].

In the case of deep learning strategies, few methods added convolutional neural networks (CNN) in addition to handcrafted features or use CNN to extract features and classify using a different methodology [5, 49]. Some research use deep learning but has focused on the classification of already detected nodule candidates from the LIDC-IDRI dataset [14, 40, 42], the LUNA16 challenge dataset (which is based on LIDC-IDRI), and the Multicentric Italian Lung Detection (MILD) [31] trial [7]. In these cases, the problem is addressed as classification of nodules from given candidate centroids that were detected with previously published CAD systems [15, 23, 34, 40, 47]. The number of false positives to be classified by these methods is almost 25 times less than our extracted candidates, but the highest sensitivity reached by these methods is around 87% (in the candidate generation stage) for all sized nodules. Since the nodule classification algorithms are evaluated over the total previously detected nodules and not over the total ground truth nodules of each subject, the classification of nodules from previously detected algorithms is a problem with a lower difficulty degree.

Table 2 Statistics from the Kaggle DSB 2017 dataset

Total subjects	1890	100%
Subjects with public labels	1384	73.2%
Subjects without public labels	506	26.7%

Table 3 Statistics from the ISBI 2018 dataset

Total subjects	100	100%
Subjects with public labels	30	30%
Subjects without public labels	70	70%

Nodule detection and/or classification starting from subject LDCT has also been studied using deep learning tools. One common strategy is to use 2D CNN with orthogonal or continuous nodule–centered patches [4, 5, 20, 32, 34, 46] because of its straightforward adaptation from standard CNN architectures designed for detection and/or classification of natural images [12, 17, 33, 36, 39, 43]. However, LDCT like many other biomedical image acquisition methods incorporates 3D information which natural images do not have and are not exploited with 2D CNN. A better strategy is to use CNN with 3D convolutions to take advantage of the 3D information. Several methods use 3D CNN for nodule detection, classification, and/or stratification [10, 14, 38].

Transfer learning or the use of learned features (pre-trained models) from different domains is a common strategy used in deep learning when the amount of data for training is not large enough. Pre-trained models have been used for lung nodule feature extraction [49] and nodule classification [5, 38].

Multi-path CNN architectures are used when the context information is considered important in the detection or classification of an object. Several methods have resorted to the use of such architectures for nodule detection [35] and nodule classification [6, 37].

What most methods lack is a unified strategy to diagnose lung cancer starting from a subject LDCT. Our proposed algorithm uses a single subject LDCT and outputs a probability of cancer/non-cancer. We take advantage of the 3D information throughout the entire method. Also, we use a multi-path architecture (i.e., for our cancer predictor), not as a multi-scale

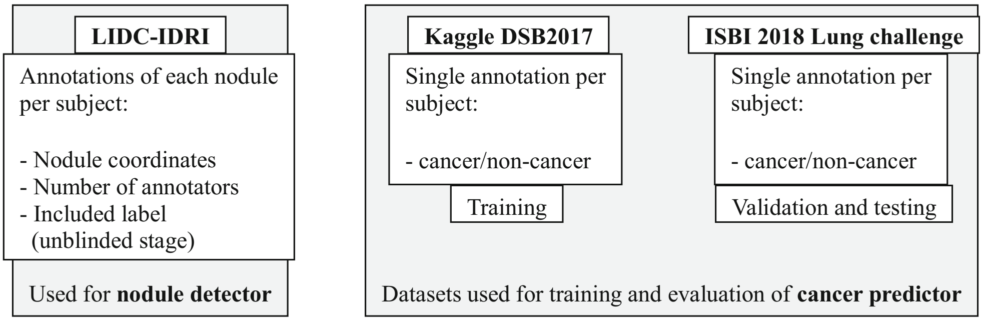

LIDC-IDRI
Annotations of each nodule per subject:
- Nodule coordinates
- Number of annotators
- Included label (unblinded stage)
Used for **nodule detector**
Kaggle DSB2017
Single annotation per subject:
- cancer/non-cancer
Training
ISBI 2018 Lung challenge
Single annotation per subject:
- cancer/non-cancer
Validation and testing
Datasets used for training and evaluation of **cancer predictor**

Fig. 3 Summary of the three datasets used. LIDC-IDRI dataset has independent nodule annotations, and it is used to train and validate the nodule detector. Kaggle DSB 2017 and ISBI 2018 lung challenge datasets have a single label of cancer/non-cancer per subject, and are used to train and evaluate respectively the cancer predictor

feature extractor–like [6, 35, 37], but to combine the information of the highest ranked nodules from our nodule detector to give a single diagnosis.

3 Materials

3.1 LIDC-IDRI dataset

The LIDC-IDRI dataset[1] is produced by the LIDC and the IDRI [3] with a total of 1010 subjects. It is publicly available in DICOM format and the radiologists' annotations in XML markup. The annotations consist of the coordinates and the number of radiologists that annotated each nodule (i.e., each object in the lung region of the LDCT considered as a nodule by a radiologist). Also, for most nodules, it included information based on the subjective assessments of multiple experienced radiologists (e.g., lesion category, nodule outlines, and subtlety ratings) [3]. Each annotation was made by 4 radiologists in two stages, i.e., a blind stage and a second unblinded stage where each radiologist was presented with the marks placed by all radiologists in the blind stage. The total number of nodules of the LIDC-IDRI dataset is 11,608. Around half (54.2%) of the total nodules were included after the unblinded annotation stage (i.e., 6287 nodules). For this project, we consider only the included nodules (i.e., lesions labeled as nodules for at least one specialist and included after the unblinded second stage).

Table 1 shows the consensus in lung nodule detection by the four radiologists. As we can see, only 51.4% of the nodules that were included after the unblinded second stage are detected by the 4 specialists and only 27.8% of the total nodules from the blind initial stage. The consensus of 51.4%, which we will use as human performance for nodule detection task, shows the great difficulty of detecting early lung nodules, even for trained specialists.

We divide the dataset for the nodule detector randomly into 3 fixed sub-sets, i.e., 25% of subjects for training, 25% for validation of hyperparameters, and the remaining 50% for final testing. We use the average precision (AP) to evaluate the performance of our detector.

3.2 Kaggle DSB 2017 dataset [16]

The Kaggle DSB 2017 contains thousands of high-resolution lung scans provided by the National Cancer Institute with annotations of cancer/non-cancer for each subject. Of the total 1890 subjects, 1384 have public labels. Of the 1384 subjects with public labels, around 25% of the subjects are labeled with cancer. The scans are provided in DICOM format. Table 2 shows the statistics of the Kaggle DSB 2017 dataset. We use the 1384 labeled subjects for training of the multi-pathway cancer predictor.

3.3 ISBI 2018 lung cancer dataset

The ISBI 2018 Lung Nodule Malignancy Prediction challenge [22] use a set of 100 subjects with sequential LDCT (a total of 200 scans; one scan was taken in year 1999 and the second scan in year 2000 for each subject), including equal number of cancer and non-cancer cases. This dataset uses a subset of data from the National Lung Screening Trials (NLST)[2] [1]. The dataset provides annotations of cancer/non-cancer for each subject, and segmentation annotations of the index nodule (most critical nodule chosen by annotators of each LDCT) in Neuroimaging Informatics Technology Initiative (NIfTI) format. The dataset is divided by the challenge organizers in 30 subjects for training and 70 subjects for testing. The labels of the 70 test subjects are not public and evaluation is performed on the challenge server. The scans are provided in DICOM format. Table 3 shows statistics of the ISBI 2018 dataset.

We use the 30 subjects with public labels as validation set for our cancer predictor. The final score is evaluated on the entire test set in the ISBI 2018 challenge server. In our experiments, we use only the year 2000 scans of all subjects. We use the area under curve of the receiver operating characteristic (AUC ROC) to evaluate the performance of our predictor.

[1] LIDC-IDRI can be found at The Cancer Imaging Archive (TCIA): https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI

[2] NLST can be found at The Cancer Imaging Archive (TCIA): https://wiki.cancerimagingarchive.net/display/NLST

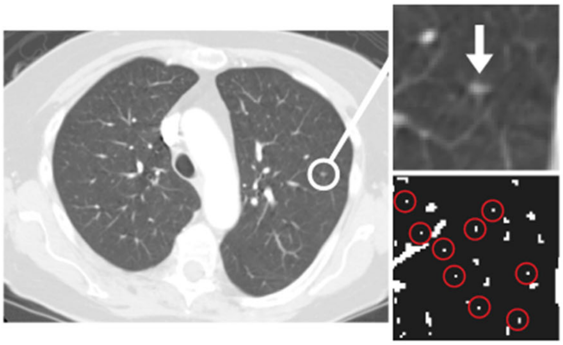


Fig. 4 Volume filtering to reduce noise from the original subject's LDCT scan. Filtering reduces the number of false positives (circled in red). Left, original volume and extracted objects without filtering in a zoomed patch. In red, we show the objects that are removed with filtering. Right, filtered volume using a 3D median filter

AUC ROC is the official metric from the ISBI 2018 Lung Nodule Malignancy Prediction challenge.

The three datasets used for our model are summarized in Fig. 3.

4 Proposed method

4.1 Nodule detector

We first present our lung nodule detector, which was originally introduced in [32]. It takes as input a subject's lung LDCT scan and gives as output the detected nodules with high probability of being malignant.

4.1.1 Input

We generate nodule candidates for each subject over the entire lung volume to benefit from the three-dimensional information provided by the LDCT scans. We interpolate the original LDCT volume into an isotropic volume in order to work with the same voxel size for all subjects.

4.1.2 Lung volume filtering and masking

We filter the volume using a 3D median filter for noise reduction. An example of a filtered volume is shown in Fig. 4. After filtering, we extract the lung volume with a calculated mask to avoid unnecessary information processing, which may lead to an increased number of false positives. This mask is produced

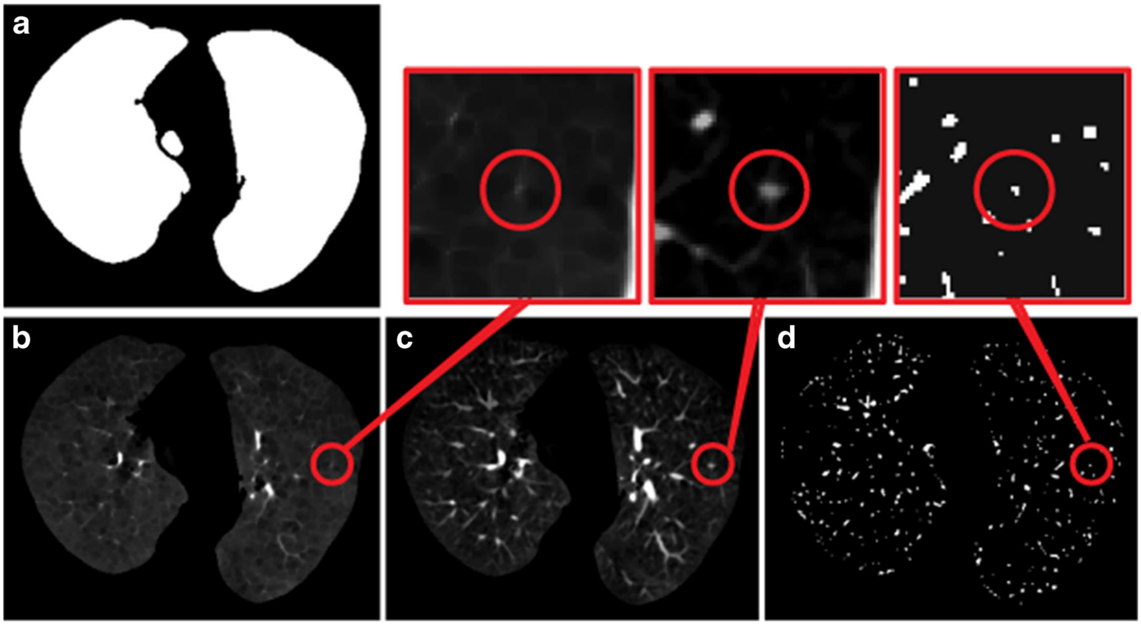


Fig. 5 Masking and candidate extraction process. **a** Calculated lung mask with thresholding and morphological operations. **b** Lungs after erosion. **c** Lungs after opening by reconstruction. **d** Regional maxima calculation

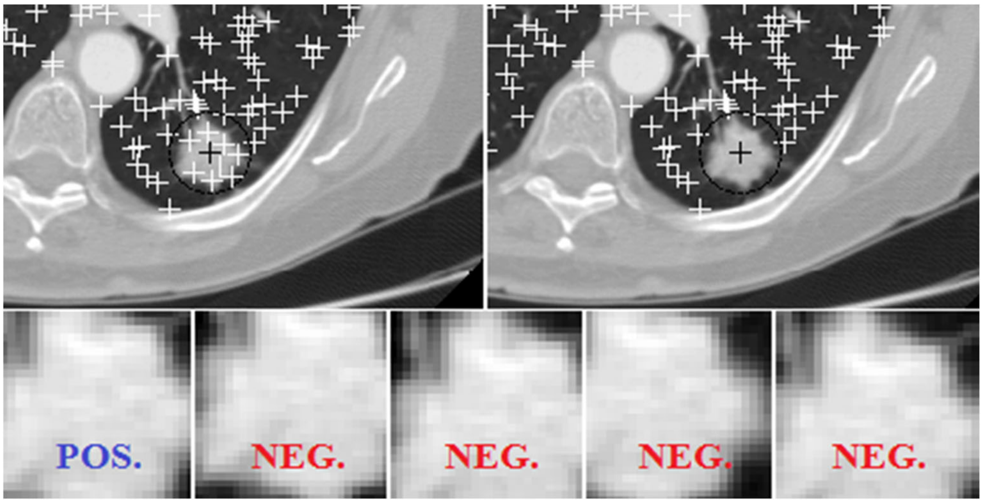


Fig. 6 Cleaning of negatives around the included nodules. Top-left, computed centroids without cleaning. Top-right, computed centroids with cleaning. Bottom, extracted candidates and corresponding labels without cleaning

for each subject with a thresholding operation. Given that LDCT scans in the dataset were produced by different machines, a fixed threshold pixel value does not give good results. Thus, we use a linear combination of the mean and standard deviation of each scan independently to get this value. Following thresholding, we use morphological closing to fill borders and holes, and to remove small objects and structures connected to the image border. An example of the resulting binary lung mask is shown in Fig. 5a .

4.1.3 Candidate generation

For candidate generation, we perform an opening by reconstruction over the extracted lung volume. We use this morphological operation given that the candidates are light regions in the scan. We use a marker volume created by eroding the 3D volume with an ellipsoid. The radius of the ellipsoid is 1 pixel and the height is the separation between two slices in each subject's volume. We carry out morphological reconstruction using the marker described above and the filtered image as mask. In addition, we calculate the regional maxima. Figure 5b to d show the result after erosion, opening, and regional maxima, respectively. The objective of the candidate generation is to extract all light components (higher density tissue) inside the lungs.

4.1.4 Nodule classification

From the regional maxima, we compute connected components per subject and their centroids. As shown in Fig. 6, we perform a cleaning stage of negatives around the included nodules with an experimentally estimated radius for the training dataset (non-maximum suppression for validation and test datasets).

We design and train a three-dimensional convolutional neural network (3D CNN) for false positive reduction with

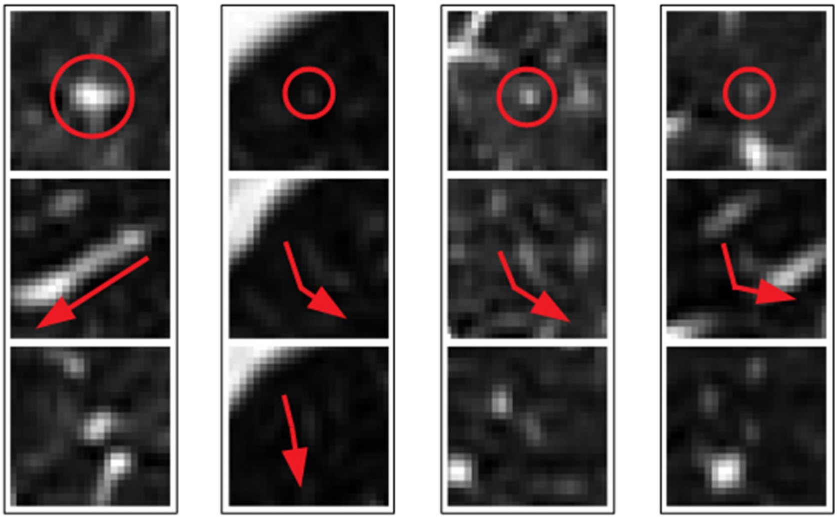


Fig. 7 Axial (top row), sagittal (middle row), and coronal (bottom row) planes of 4 sample nodule candidates (non-nodule in these cases). Top row, axial plane shows a round structure with higher density in the center of the image. Middle and bottom row, coronal and/or sagittal planes show a vessel-like structure

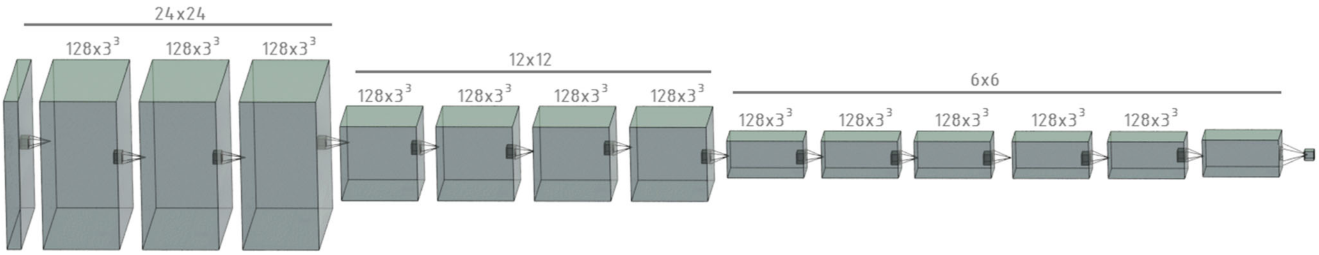


Fig. 8 Neural network architecture of twelve 3D convolutional layers with best results from our modular design of 3 × 3 × 3 filters. Sizes 24 × 24, 12 × 12, and 6 × 6 refer to the size of the feature responses: 24 × 24 as the input size, 12 × 12 after the first max. pooling, and 6 × 6 after the second max. pooling

3D candidates (volumes centered at the calculated centroid) as input. In contrast of 2D CNNs when using 3D convolutions, we analyze one additional spatial dimension which is important to differentiate nodules from other structures such as vessels that may look similar in one slice independently. As shown in Fig. 7, the axial plane of the objects (top row) shows a centered round structure with higher density. But in sagittal (middle row) and/or coronal (bottom row) planes, the object is a vessel-like structure.

We define a modular network for systematic exploration of CNN architectures. It consists of groups of convolutional layers with filters of a fixed size 3 by 3, batch normalization, and rectified linear unit (ReLU) activations [28]. ReLU is used to create non-linearities that reduce overfitting and regularize training. In the validation experiments, we change the number of convolutional layers before each pooling layer, the number of filters, and the value of hyperparameters such as batch size and learning rate. Also, the number of max. pooling layers is changed depending on the input size of the network, resulting in feature maps in the last convolutional layer with size from 2 × 2 to 12 × 12. Sizes 24 × 24, 12 × 12, and 6 × 6 refer to the size of the feature responses, i.e., 24 × 24 as the input size, 12 × 12 after the first max. pooling, and 6 × 6 after the second max. pooling. The architecture of the network with best results is shown in Fig. 8.

4.2 Cancer prediction

4.2.1 Predictor input

For a subject to have lung cancer, one malignant nodule is enough. However, in practice, the malignancy of a nodule is determined by a biopsy of a sample of the nodule tissue. Our proposed method relies only on the visual information of the LDCT. Therefore, we want our predictor model to use as much relevant information (i.e., detected nodules) as we can provide from our nodule detector. We feed our predictor model with the five top-scored nodules from our nodule detector. The decision of the number of inputs was given following this train of thought:

1. Because our nodule detector is not perfect, we should avoid feeding our predictor model with only the top-scored nodule. If the top-scored nodule is a bad detection, our complete method will fail to predict lung cancer. Also, we are assuming that the analysis of multiple detected nodules by our predictor (in contrast to using only the top-scored detected nodule) may help in the decision making of whether a subject has cancer or not.
2. Too many inputs (top-scored nodules) to our predictor model may be unnecessary due to all irrelevant detected objects that are non-nodules.
3. The upper boundary for the number of inputs to our predictor model is limited in any case but the amount of memory we can allocate in a single GPU.
4. Because the five top-scored detected nodules of every subject in the dataset received a score of at least 0.90 by our nodule detector, a smaller number of inputs were not considered.

4.2.2 Multi-path convolutional neural network

We train a multi-path network of 5 paths, all with the same 3D CNN architecture of the nodule classification network, as shown in Fig. 8. This cancer predictor is trained on the Kaggle Data Science Bowl 2017 challenge dataset (around 1400 subjects with labels of cancer/non-cancer for each

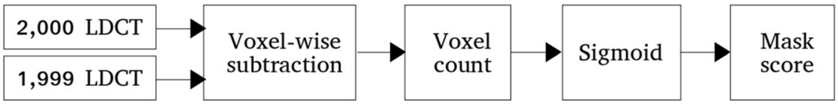


Fig. 9 Mask subtractor voxel by voxel between year 2000 LDCT and year 1999 CT. A subtraction voxel by voxel is done between the nodule of the year 2000 LDCT and 1999 LDCT of each subject voxel-wise subtraction. The number of voxels Voxel count resulting from the subtraction is then passed through a sigmoid (Sigmoid) to produce a probability between 0 and 1 (mask score)

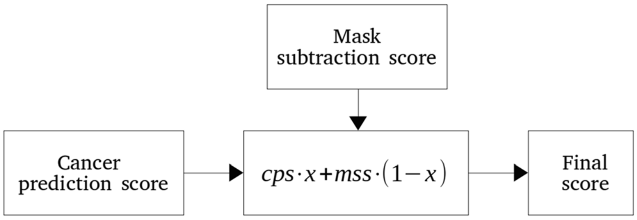


Fig. 10 Linear combination used to produce final malignancy score

subject CT) and validated on the data from the ISBI 2018 Lung Nodule Malignancy Prediction challenge (30 subjects). As the input of each of the 5 paths of the multi-path network, we use the 5 top detected nodules in the previous stage. For training, we perform data augmentation by a factor of 20 taking 5 random nodules from the 10 top scoring nodules of each subject. In the test set, only the 5 top-scored nodules are used. Figure 2 (right) shows our cancer predictor architecture.

4.3 Post-processing

An additional post-processing stage is applied using the released test segmentations by the ISBI lung nodule malignancy prediction challenge of the index nodule of each subject. A subtraction voxel by voxel is done between the nodule of the year 2000 LDCT and 1999 LDCT of each subject. The result of the subtraction is then passed through a sigmoid to produce a probability between 0 and 1. Figure 9 shows the mask subtraction of the nodule masks. Then, a linear combination (Fig. 10) with the result of the trained cancer predictor and the result from the subtraction gives us the final malignancy probability

Fig. 11 Candidate modalities. **a** 32 × 32 × 3 candidate with axial, sagittal, and coronal planes. **b** 32 × 32 × 9 candidate using 9 consecutive *z*-planes. **c** 24 × 24 × 24 candidate (best result)

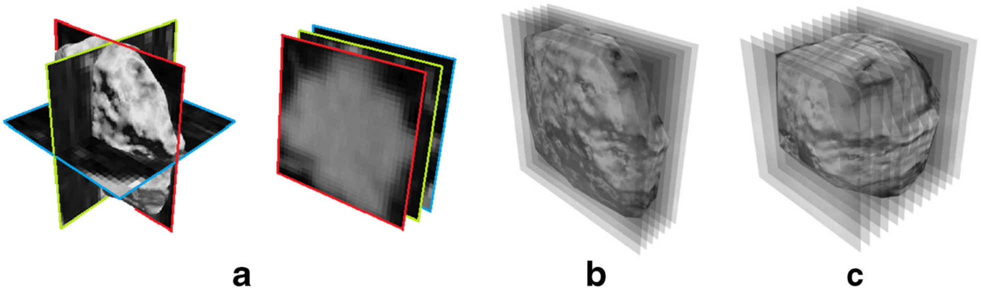


Fig. 12 Comparison between false positive reduction methods. HOG + SVM and best networks using convolutions in 2D, 2.5D (2D convolutions over 3D candidates), and 3D

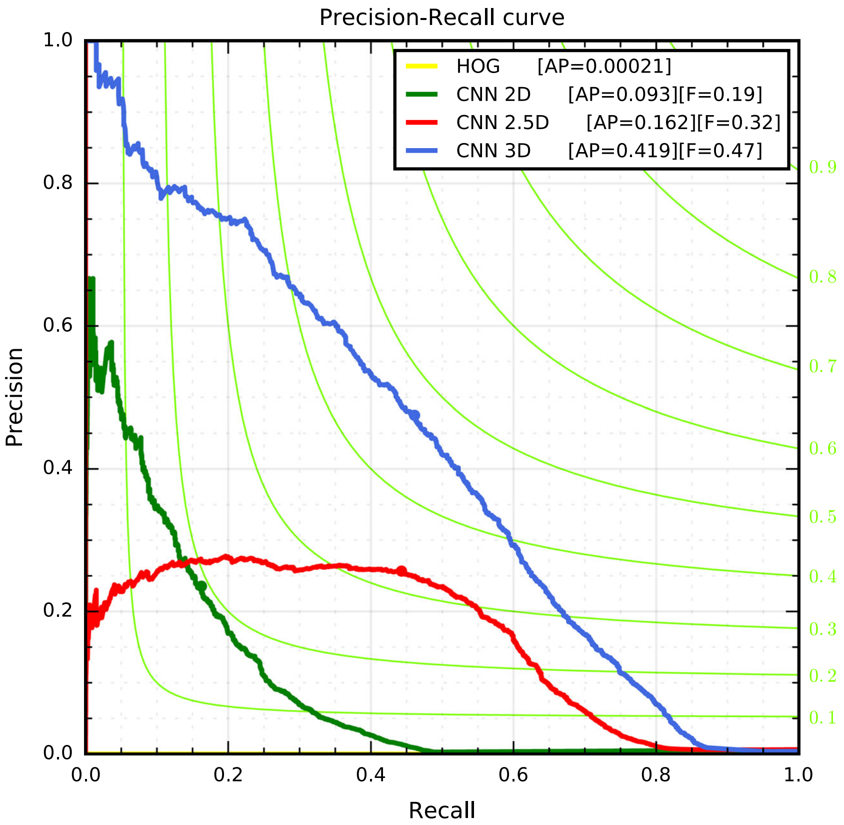


Table 4 Average precision obtained with different false positive reduction methods of the nodule detector

Method	AP (%)
No FP reduction method	0.006
HOG + SVM	0.021
CNN 2D	9.3
CNN 2.5D	16.2
CNN 3D	*41.9*

Italic indicates the best result

of each subject. The linear combination parameter x is adjusted empirically on the training set of the ISBI 2018 Lung Nodule Malignancy Prediction challenge dataset. When sequential LDCTs are not available, the output of our method is the cancer predictor score without the post-processing stage.

5 Experiments

5.1 Candidate generation

For candidate generation, we test different configurations of thresholding equations, several values for the erosion ellipsoid's radius and height, and different input connectivities. The best recall we obtained for this stage is 99.6% with 3154 included nodules (out of 3167) from a total of 25,221,581 generated candidates from the training/validation set. That gives a total of 25,218,427 false positives.

As stated before, the total amount of candidates for the training/validation set is around 25 million with approximately 3150 included nodules, which is extremely unbalanced. As a consequence, we perform data augmentation for training with image translations and horizontal reflections for each candidate. From the 3150 included nodules, we augment (by a factor of 216) to around 700,000 to have a representative number of positive nodules for training the CNN. We select randomly the same number of negatives (non-nodules) after augmentation to balance the training dataset. Therefore, our training/validation set is composed of around 1.4 million candidates.

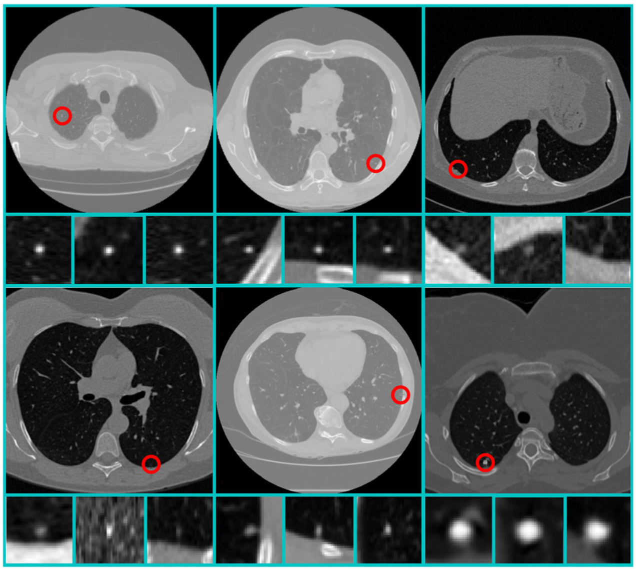


Fig. 13 Qualitative results of high scored nodule detections

Table 5 Results using different candidate size as input for 3D CNNs

Input size	# voxels	AP (%)
32 × 32 × 9	9216	13.2
16 × 16 × 16 (isotropic)	4096	32.0
24 × 24 × 24 (isotropic)	13,824	*41.9*

Italic indicates the best result

5.2 Nodule classification

Due to the variability of intensities in a LDCT scan and the grayscale nature of the images, we consider as a baseline an histogram of oriented gradients (HOG) for feature extraction of each candidate and train a support vector machine (SVM) for false positive reduction, because of its proven good performance as shape feature discriminator [8]. Although precision improves (from 0.0062 to 0.021), the number of false positives remains high.

Therefore, we decide to train a convolutional neural network to increase precision. We test 2D and 3D convolutions. For the 2D convolutions, we use 32 × 32 × 3 candidates as input for the network, using the axial, sagittal, and coronal planes centered on the calculated (Section 4.1.2) centroid (see Fig. 11a). We take this first approach for its simplicity and low GPU memory usage. For the second case, we use 2D convolutions on 32 × 32 × 9 candidates, using the 9 consecutive z-planes from the centroid (see Fig. 11b). In this case, we want to include more spatial information of each candidate and increase the precision. We also use 3D convolutions on 32 × 32 × 9 and other different candidate input sizes.

We use AP, which is the area under the precision-recall curve, to evaluate the performance of our nodule detector. We get best results using 3D convolutions with filters of size 3 × 3 × 3 and input volume size of 24 × 24 × 24. As shown in Fig. 12, we increase precision for all recall values with the 3D CNN approach.

In Table 4, we report results using different false positive reduction methods: HOG + SVM, bi-dimensional input nodule candidates with bi-dimensional convolutions (CNN 2D), three-dimensional input nodule candidates with bi-dimensional convolutions (CNN 2.5D), and three-dimensional input nodule candidates with three-dimensional convolutions (CNN 3D). The average precision obtained with deep learning methods (i.e., CNN 2D, 2.5D, and 3D) is

Table 6 AUC of the ROC curve obtained on the validation and test set with and without post-processing. Validation set corresponds to the ISBI 2018 Lung Nodule Malignancy Prediction challenge training set with public annotations. Test set corresponds to the ISBI 2018 Lung Nodule Malignancy Prediction challenge test set with private annotations. The test set results are evaluated on the challenge server

Method	AUC ROC (%)
Validation set	
5-way multi-path cancer predictor	88.7
5-way multi-path cancer predictor + 3D mask subtraction	93.7
Test set (private annotations)	
5-way multi-path cancer predictor + 3D mask subtraction	_91.3*_
Mehrtash et al. (2nd place)	89.7**

Italic indicates the best result

[*] Our method was ranked 1st place at the ISBI 2018 Lung Cancer challenge

[**] Method ranked 2nd place

increased to a greater extent than using HOG and SVM. CNN 3D gives the best results.

Figure 12 shows precision-recall curves of the results evaluated in the test set of the different methods used. We can see that the precision is greatly increased for all recall values using convolutional neural networks. As shown in the qualitative results in Fig. 13, our nodule detector is able to detect small parenchymal and juxtapleural nodules of clean and noisy LDCT scans.

5.2.1 Nodule classification CNN design experiments

Table 5 shows results with different input size for the 3D CNN approach. The first approach for 3D convolutions is using candidates of size 32 × 32 × 9 which is the input size that gives best results using 2D convolutions. The AP that we obtain with this candidate size is less than using 2D convolutions (13.2% with 3D convolutions and 16.2% with 2D convolutions). Then, we use smaller candidate size (16 × 16 × 16 instead of 32 × 32 × 9), but with more z-planes (16 planes instead of 9). The result is better than using 32 × 32 × 9 candidates even with less than half number of voxels. Finally, we increase the candidate size to 24 × 24 × 24 with which we obtain the best results.

Regarding the batch size, the trend is that the performance is better when it is smaller; the final batch size chosen is 16. As for the number of filters, we obtain the best results when the number (128 filters) is held constant in all layers. As for batch normalization, the performance of the network increases considerably after using it.

We obtain the best results by employing a neural network consisting of twelve 3D convolutional layers and one fully connected layer before the softmax that produces a probability of nodule/non-nodule. We also use max pooling after the 3rd and 7th layers. We use batch normalization for all convolutional layers, and the activation function for each one is a ReLU. The network architecture is shown in Fig. 8.

We train the network from scratch for 15 epochs using stochastic gradient descent and backpropagation with a fixed learning rate of 1e-4. The evaluation is performed on the test set of the LIDC-IDRI with an AP of 41.9%.

5.3 Cancer prediction

We use the same network architecture proposed as nodule detector for each of the paths of the cancer predictor in order to take advantage of the trained nodule feature extractors. We perform different tests changing the number of pathways (3, 5, and 10 lanes for the 3, 5, and 10 top-scored nodules of each subject), and using different amounts of data augmentation of the Kaggle DSB 2017 dataset. The best result on the validation dataset (30 training subjects of the ISBI 2018 dataset) is

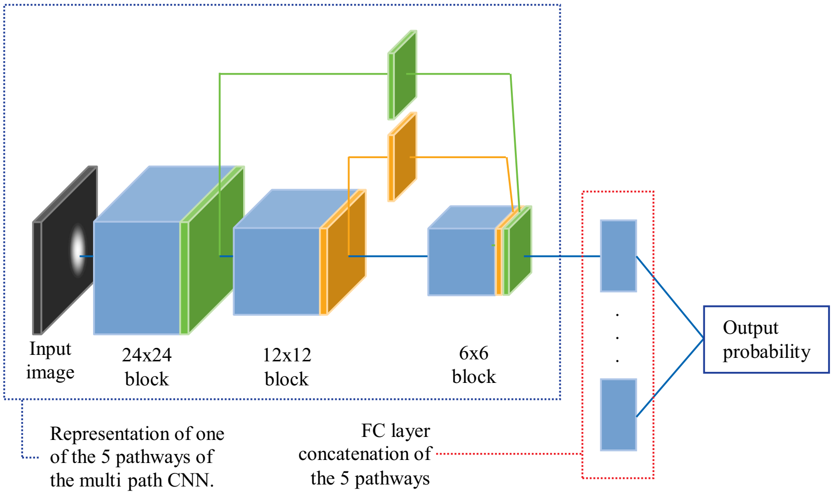

Input image
24x24 block
12x12 block
6x6 block
Output probability
Representation of one of the 5 pathways of the multi path CNN.
FC layer concatenation of the 5 pathways

Fig. 14 Modified predictor model using extracted features from previous layers. We use the extracted features from the last layer of each network block, apply convolutions to each one, and concatenate with the last convolutional layer. This modification is applied to all 5 pathways of our multi-path predictor model

Table 7 Results using the extracted features of previous layers

Method	AUC ROC (%)
5-way multi-path cancer predictor + 3D mask subtraction	93.7*
5-way multi-path cancer predictor + extracted features from previous layers + 3D mask subtraction	*97.3*

Italic indicates the best result

* Method ranked 1st place at the ISBI 2018 lung challenge

obtained using the 5 top-scored nodules from our nodule detector as input of a 5-way 3D CNN, data augmentation by a factor of 20 of the Kaggle DSB 2017 dataset for the training of the multi-path network, and a 3D mask subtraction post-processing. We train our multi-path predictor model for 5 epochs using as initialization parameters the trained weights from our detector. Best results are obtained using an increased learning rate of 1e-3, a batch size of 16.

In practice, it is easier to get a single LDCT of a subject than having sequential LDCTs for multiple years. Our method can predict cancer also with a single LDCT by removing the mask subtraction. Table 6 shows the final results of our method on our validation set with and without mask subtraction and on the ISBI 2018 lung challenge test set. Performance by method ranked 2nd place at the challenge and is included in Table 6.

Our model is able to successfully extract features from radiologist annotations (nodule level annotation of the LIDC-IDRI) and use them in conjunction with pathology annotations (DSB 2017 annotations). We show that using nodule detection reduces the difficulty of cancer diagnosis from a subject LDCT. The method ranked 2nd place[3] at the ISBI 2018 Lung Cancer challenge and also uses LIDC-IDRI nodule level information in conjunction with DSB 2017 patient level annotations.

Additional experiments are conducted after the ISBI 2018 Lung Cancer challenge using extracted features from previous layers in the multi-path predictor CNN. As shown in Fig. 14, we use the extracted features from the last layer of each network block (each network block as described in Fig. 8). Then, we apply a convolution layer with a stride value of 4 to the last layer of the 24 × 24 block, and convolution with a stride value of 2 to the last layer of the 12 × 12 block last layer. The two resulting layers (of size 6 × 6 each) are then concatenated with the last layer of the 6 × 6 block. This modification is applied to all 5 pathways of our multi-path predictor model. Using extracted features from previous layers exploits the global context information of the image improving the representational power of the model. With this modification, our model is able to increase its performance by 4%. These results are evaluated over our validation set (ISBI 2018 Lung Cancer challenge test labels are not publicly available). Results using extracted features from previous layers are presented in Table 7.

6 Conclusion

Although the problem of nodule detection is extremely unbalanced with high intra-class variance, our approach is able to detect lung nodules and predict cancer effectively. We design a candidate proposal method with almost perfect recall. In addition, we train a three-dimensional convolutional neural network that successfully classifies nodules from non-nodules and increases the precision by a factor of 2000 (compared with the HOG + SVM baseline) achieving a close to human performance in this challenging task. As for cancer diagnosis, we train a 5-way cancer predictor which was ranked 1st place at the ISBI 2018 Lung Nodule Malignancy Prediction challenge[4]. In order to ensure reproducibility of our results and to promote further research on automated lung cancer diagnosis, our source code and pre-trained models are publicly available at https://github.com/BCV-Uniandes/LungCancerDiagnosis-pytorch[5].

References

1. Aberle D et al (2011) Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 365:395–409. https://doi.org/10.1056/NEJMoa1102873
2. American Cancer Society (2015) American Cancer Society. Cancer Facts and Figures 2015. American Cancer Society, Atlanta
3. Armato S III et al (2011) The lung image database consortium LIDC and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. Med Phys 38:915–931
4. Chen S, et al., 2016. Automatic scoring of multiple semantic attributes with multi-task feature leverage: a study on pulmonary nodules in CT images. IEEE Trans Med Imaging, in press.
5. Ciompi F et al (2015) Automatic classification of pulmonary perifissural nodules in computed tomography using an ensemble of 2D views and a convolutional neural network out-of-the-box. Med Image Anal 26:195–202
6. Ciompi F, et al., 2016. Towards automatic pulmonary nodule management in lung cancer screening with deep learning. arXiv: 1610.09157.
7. Ciompi F, et al., 2017. Towards automatic pulmonary nodule management in lung cancer screening with deep learning.

[3] Method by Mehrtash et al. Currently not published. Information can be found at: https://www.rsipvision.com/ComputerVisionNews-2018May/28/

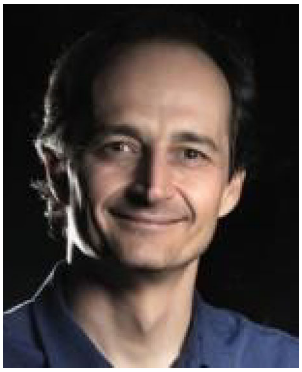
[4] ISBI 2018 lung cancer challenge results can be found at: https://bit.ly/2JPNnGS

[5] https://github.com/BCV-Uniandes/LungCancerDiagnosis-pytorch

8. Dalal N, Triggs B, 2005. Histograms of oriented gradients for human detection, in: In CVPR, pp. 886–893.
9. Dolejsi M, Kybic J, 2007. Automatic two-step detection of pulmonary nodules. Proceedings of SPIE 6514, 3j–1–3j–12.
10. Dou Q, et al., 2016. Multi-level contextual 3D CNNs for false positive reduction in pulmonary nodule detection, in press.
11. Falk S, Williams C, 2010. "Chapter 1". Lung cancer—the facts (3rd ed.). Oxford University Press. pp.3–4.
12. He K, et al., 2015. Deep residual learning for image recognition. CoRR abs/1512.03385. ArXiv:1512.03385.
13. Hu Z et al (2018) Deep learning for image-based cancer detection and diagnosis–a survey. Pattern Recognition Volume 83:134–149
14. Hussein S, et al., 2017. Risk stratification of lung nodules using 3D CNN-based multi-task learning. IPMI .
15. Jacobs C et al (2014a) Automatic detection of subsolid pulmonary nodules in thoracic computed tomography images. Med Image Anal 18:374–384
16. Kaggle Data Science Bowl, 2017. Data Science Bowl 2017. https://www.kaggle.com/c/data-science-bowl-2017.
17. Krizhevsky A, et al., 2012. Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems 25. Curran Associates, Inc., pp. 1097–1105.
18. Lee S et al (2012) Automated detection of ling nodules in computed tomography images: a review. Mach Vis Appl 23:151–163
19. Li Q (2007) Recent progress in computer-aided diagnosis of lung nodules on thin-section CT. Comput Med Imaging Graph 31:248–257
20. Li W, et al., 2016. Pulmonary nodule classification with deep convolutional neural networks on computed tomography images. Computational and Mathematical Methods in Medicine, 6215085.
21. Litjents G et al (2017) A survey on deep learning in medical image analysis. Med Image Anal 42:60–88
22. Lung Nodule Malignancy Prediction Challenge, 2018. ISBI 2018 - lung nodule malignancy prediction, based on sequential CT scans. http://isbichallenges.cloudapp.net/competitions.
23. Massion PP, Walker RC (2014) Indeterminate pulmonary nodules: risk for having or for developing lung cancer? *Cancer Prev Res (Phila)* 7(12):1173–1178. https://doi.org/10.1158/1940-6207.CAPR-14-0364
24. Meyer P et al (2018 Jul 1) 2017. Survey on deep learning for radiotherapy. Comput Biol Med 98:126–146. https://doi.org/10.1016/j.compbio-med.2018.05.018
25. Moria S et al (2018) Blood vessel segmentation algorithms — review of methods, datasets and evaluation metrics. Comput Methods Prog Biomed. https://doi.org/10.1016/j.cmpb.2018.02.001
26. Murphy K et al (2009) A large scale evaluation of automatic pulmonary nodule detection in chest CT using local image features and k-nearest-neighbour classification. Med Image Anal 13:757–770
27. NCI. 12 May 2015. Non-small cell lung cancer treatment –patient version. Archived from the original on 29 February 2016. Retrieved 5 March 2016.
28. Nair V, Hinton G, 2010. Rectified linear units improve restricted Boltzmann machines, in: Proceedings of the 27th International Conference on International Conference on Machine Learning, Omnipress, USA. pp. 807–814.
29. Oseas A et al (2014) Automatic detection of solitary lung nodules using quality threshold clustering, genetic algorithm and diversity index. Artif Intell Med 60:165–177
30. Osman S, Ucan O (2007) Lung nodule diagnosis using 3d template matching. Comput Biol Med 37:1167–1172
31. Pastorino U, et al., 2012. Annual or biennal CT screening versus observation in heavy smokers: 5-year results of the MILD trial. European Journal of Cancer Prevention.
32. Perez G, Arbelaez P, 2017. Automated detection of lung nodules with three-dimensional convolutional neural networks. Doi:https://doi.org/10.1117/12.2285954.
33. Ren S, et al., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in Neural Information Processing Systems 28. Curran Associates, Inc., pp. 91–99.
34. Setio A et al (2015) Automatic detection of large pulmonary solid nodules in thoracic CT images. Med Phys 42:5642–5653
35. Setio A et al (2016) Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks. IEEE Trans Med Imaging 35(5):1160–1169
36. Shelhamer E et al (2017) Fully convolutional networks for semantic segmentation. IEEE Trans Pattern Anal Mach Intell 39:640–651. https://doi.org/10.1109/TPAMI.2016.2572683
37. Shen W, et al., 2015. Multi-scale convolutional neural networks for lung nodule classification. In: Inf Process Med Imaging. Vol. 9123 of Lect Notes Comput Sci. pp. 588–599.
38. Shen W, et al., 2016. Learning from experts: developing transferable deep features for patient-level lung cancer prediction. In: Med Image Comput Comput Assist Interv. Vol. 9901 of Lect Notes Comput Sci. pp. 124–131.
39. Simonyan K, Zisserman A, 2014. Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556. ArXiv:1409.1556.
40. Song K, et al., 2015. Using deep learning for classification of lung nodules on computed tomography images.
41. Sudha V, Jayashree P (2012) Lung nodule detection in CT images using thresholding and morphological operations. International Journal on Emerging Science and Engineering (IJESE) 1:17–21
42. Sun W, et al., 2016. Computer aided lung cancer diagnosis with deep learning algorithms. Research gate.
43. Szegedy C, et al., 2015. Going deeper with convolutions, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
44. Tan M et al (2011) A novel computer-aided lung nodule detection system for CT images. Med Phys 38:5630–5645
45. Tartar A, Akan A, 2013. A new method for pulmonary nodule detection using decision trees. 35th Annual International Conference of the IEEE EMBS .
46. Teramoto A et al (2016) Automated detection of pulmonary nodules in PET/CT images: ensemble false-positive reduction using a convolutional neural network technique. Med Phys 43:2821–2827
47. Torres E et al (2015) Large scale validation of the M5L lung CAD on heterogeneous CT datasets. Med Phys 42:1477–1489
48. van Ginneken B et al (2010) Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: the ANODE09 study. Med Image Anal 14:707–722
49. van Ginneken B, et al., 2015. Off-the shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans. In: IEEE Int Symp Biomedical Imaging. pp. 286–289.
50. World Health Organization (2017) World Health Organization. Media Centre, Cancer http://www.who.int/mediacentre/factsheets/fs297/en/

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Gustavo Pérez received a M.Sc. in Biomedical Engineering at Universidad de los Andes where he spent three years as a graduate research assistant under the supervision of Pablo Arbeláez. He is currently a second year PhD student and a graduate research assistant in Computer Science at the University of Massachusetts Amherst.



Pablo Arbeláez received a PhD with honors in Applied Mathematics from the Universite Paris-Dauphine in 2005. He was a research scientist with the Computer Vision group at UC Berkeley from 2007 to 2014. He currently holds a faculty position at the Universidad de los Andes in Colombia.