Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge

Arnaud Arindra Adiyoso Setio[a,1], Alberto Traverso[g,h,a,1], Thomas de Bel[o], Moira S.N. Berens[n], Cas van den Bogaard[o], Piergiorgio Cerello[h], Hao Chen[m], Qi Dou[m], Maria Evelina Fantacci[k,i], Bram Geurts[b], Robbert van der Gugten[n], Pheng Ann Heng[m], Bart Jansen[e,f], Michael M.J. de Kaste[n], Valentin Kotov[o], Jack Yu-Hung Lin[j], Jeroen T.M.C. Manders[n], Alexander Sóñora-Mengana[d,e,1], Juan Carlos García-Naranjo[d], Evgenia Papavasileiou[e], Mathias Prokop[a], Marco Saletta[h], Cornelia M Schaefer-Prokop[a,c], Ernst T. Scholten[a], Luuk Scholten[o], Miranda M. Snoeren[b], Ernesto Lopez Torres[l], Jef Vandemeulebroucke[e,f], Nicole Walasek[o], Guido C.A. Zuidhof[n], Bram van Ginneken[a,p], Colin Jacobs[a]

[a] *Diagnostic Image Analysis Group, Radboud University Medical Center, Nijmegen, The Netherlands*
[b] *Department of Radiology and Nuclear Medicine, Radboud University Medical Center, Nijmegen, The Netherlands*
[c] *Department of Radiology, Meander Medisch Centrum, Amersfoort, The Netherlands*
[d] *Centro de Biofísica Médica, Universidad de Oriente, Santiago de Cuba, Cuba*
[e] *Department of Electronics and Informatics, Vrije Universiteit Brussel, Brussels, Belgium*
[f] *imec, Leuven, Belgium*
[g] *Department of Applied Science and Technology, Polytechnic University of Turin, Turin, Italy*
[h] *Turin Section of Istituto Nazionale di Fisica Nucleare, Turin, Italy*
[i] *Pisa Section of Istituto Nazionale di Fisica Nucleare, Pisa, Italy*
[j] *Yan'an Xi Lu 129, 9th floor, Shanghai, China*
[k] *Department of Physics, University of Pisa, Pisa, Italy*
[l] *Center of Applied Technologies and Nuclear Development, La Habana, Cuba*
[m] *Department of Computer Science and Engineering, The Chinese University of Hong Kong, China*
[n] *Radboud University, Nijmegen, The Netherlands*
[o] *Institute for Computing and Information Sciences, Radboud University Nijmegen, Nijmegen, The Netherlands*
[p] *Fraunhofer MEVIS, Bremen, Germany*

Abstract

Automatic detection of pulmonary nodules in thoracic computed tomography (CT) scans has been an active area of research for the last two decades. However, there have only been few studies that provide a comparative performance evaluation of different systems on a common database. We have therefore set up the LUNA16 challenge, an objective evaluation framework for automatic nodule detection algorithms using the largest publicly available reference database of chest CT scans, the LIDC-IDRI data set. In LUNA16, participants develop their algorithm and upload their predictions on 888 CT scans in one of the two tracks: 1) the complete nodule detection track where a complete CAD system should be developed, or 2) the false positive reduction track where a provided set of nodule candidates should be classified. This paper describes the setup of LUNA16 and presents the results of the challenge so far. Moreover, the impact of combining individual systems on the detection performance was also investigated. It was observed that the leading solutions employed convolutional networks and used the provided set of nodule candidates. The combination of these solutions achieved an excellent sensitivity of over 95% at fewer than 1.0 false positives per scan. This highlights the potential of combining algorithms to improve the detection performance. Our observer study with four expert readers has shown that the best system detects nodules that were missed by expert readers who originally annotated the LIDC-IDRI data. We released this set of additional nodules for further development of CAD systems.

Keywords: pulmonary nodules, computed tomography, computer-aided detection, medical image challenges, deep learning, convolutional networks

1. Introduction

Lung cancer is the deadliest cancer worldwide, accounting for approximately 27% of cancer-related deaths in the United States (American Cancer Society (2016)). The NLST trial showed that three annual screening rounds of high-risk subjects using low-dose computed tomography (CT) reduced lung cancer mortality after 7 years by 20% in comparison to screening with chest radiography (Aberle et al. (2011)). As a result of this trial and subsequent modeling studies, lung cancer screening programs using low-dose CT are currently being implemented in the U.S. and other countries will likely follow soon. One of the major challenges arising from the implementation of these screening programs is the enormous amount of CT images that must be analyzed by radiologists.

In the last two decades, researchers have been developing Computer-Aided-Detection (CAD) systems for automatic detection of pulmonary nodules. CAD systems are intended to make the interpretation of CT images faster and more accurate, hereby improving the cost-effectiveness of the screening program. The typical setup of a CAD system consists of: 1) preprocessing, 2) nodule candidate detection, and 3) false positive reduction. Preprocessing is typically used to standardize the data, restrict the search space for nodules to the lungs, and reduce noise and image artifacts. The candidate detection stage aims to detect nodule candidates at a very high sensitivity, which typically comes with many false positives. Subsequently, the false positive reduction stage reduces the number of false positives among the candidates and generates the final set of CAD marks.

Although a large number of CAD systems have been proposed (Bergtholdt et al. (2016); Torres et al. (2015); van Ginneken et al. (2015); Brown et al. (2014); Jacobs et al. (2014); Choi and Choi (2013); Tan et al. (2013); Teramoto and Fujita (2013); Cascio et al. (2012); Guo and Li (2012); Camarlinghi et al. (2011); Tan et al. (2011); Riccardi et al. (2011); Messay et al. (2010); Golosio et al. (2009); Murphy et al. (2009)), there have only been few studies providing an objective comparative evaluation framework using a common database. The reported performances of published CAD systems can vary substantially because different data sets were used for training and evaluation (Firmino et al. (2014); Jacobs et al. (2016)). Moreover, substantial variability among radiologists on what constitutes a nodule has been reported (Armato et al. (2009)). Consequently, it is difficult to directly and objectively compare different CAD systems. The evaluation of different systems using the same framework provides unique information that can be leveraged to further improve the existing systems and develop novel solutions.

ANODE09 was the first comparative study aimed towards evaluating nodule detection algorithms (van Ginneken et al. (2010)). This challenge has allowed groups to evaluate their algorithms on a shared set of scans obtained from a lung cancer screening trial. However, this study only included 50 scans from a single center, all acquired using one type of scanner and scan protocol. In addition, the ANODE09 set contained a limited number of larger nodules, which generally have a higher suspicion of malignancy. Evaluation on a larger and more diverse image database is therefore needed.

In this paper, we introduce a novel evaluation framework for automatic detection of nodules in CT images. A large data set, containing 888 CT scans with annotations from the publicly available LIDC-IDRI database (Armato et al. (2011)), is provided for both training and testing. A web framework has been developed to efficiently evaluate algorithms and compare the result with the other algorithms. The impact of combining multiple candidate detection approaches and false positive reduction stages was also evaluated.

The key contributions of this paper are as follows. (1) We describe and provide an objective web framework for evaluating nodule detection algorithms using the largest publicly available data set; (2) We report the performance of algorithms submitted to the framework and investigate the impact of combining individual algorithms on the detection performance. We show that the combination of classical candidate detectors and a combination of deep learning architectures processing these candidates generates excellent results, better than any individual system; (3) We update the LIDC-IDRI reference standard by identifying nodules that were missed in the original LIDC-IDRI annotation process.

2. Data

The data set was collected from the largest publicly available reference database for lung nodules: the LIDC-IDRI (Armato et al. (2011); Clark et al. (2013); Armato III et al. (2015)). This database is available from NCI's Cancer Imaging Archive[2] under a Creative Commons Attribution 3.0 Unsupported License.

[1] These authors contributed equally to this work

[2] https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI

The LIDC-IDRI database contains a total of 1018 CT scans. CT images come with associated XML files with annotations from four experienced radiologists. The database is very heterogeneous: It consists of clinical dose and low-dose CT scans collected from seven different participating academic institutions, and a wide range of scanner models and acquisition parameters.

As recommended by Naidich et al. (2013); Manos et al. (2014) and the American College of Radiology (Kazerooni et al. (2014)), thin-slice CT scans should be used for the management of pulmonary nodules. Therefore, we discarded scans with a slice thickness greater than 3 mm. On top of that, scans with inconsistent slice spacing or missing slices were also excluded. This led to the final list of 888 scans. These scans were provided as MetaImage (.mhd) images that can be accessed and downloaded from the LUNA16 website[3]. A more extensive data set description is provided in our previous study (Jacobs et al. (2016)).

Each LIDC-IDRI scan was annotated by experienced thoracic radiologists in a two-phase reading process. In the initial blinded reading phase, four radiologists independently annotated scans and marked all suspicious lesions as: nodule $\geq$ 3 mm; nodule < 3 mm; non-nodule (any other pulmonary abnormality). For lesions annotated as nodule $\geq$ 3 mm, diameter measurements were provided. In a subsequent unblinded reading phase, the anonymized blinded results of all other radiologists were revealed to each radiologist, who then independently reviewed all marks. No consensus was forced.

In the 888 scans, a total of 36,378 annotations were made by the radiologists. We only considered annotations categorized as nodules $\geq$ 3 mm as relevant lesions, as nodules < 3 mm, and non-nodule lesions are not considered relevant for lung cancer screening protocols (Aberle et al. (2011)). Nodules could be annotated by multiple radiologists; annotations from different readers that were located closer than the sum of their radii were merged. In this case, position and diameters of these merged annotations were averaged. This resulted in a set of 2,290, 1,602, 1,186, and 777 nodules annotated by at least 1, 2, 3, or 4 radiologists, respectively. We considered the 1,186 nodules annotated by the majority of the radiologists (at least 3 out of 4 radiologists) as the positive examples in our reference standard. These are the lesions that the algorithms should detect. Other findings (1,104 nodules annotated by less than 3 out of 4 radiologists, 11,509 “nodule < 3 mm” annotations, and 19,004 “non-nodule” annotations) were considered “irrelevant findings” and marks on these locations were not counted as false positives nor as true positives in the final analysis; the same approach was used by (van Ginneken et al. (2010); Jacobs et al. (2016)). Irrelevant findings were excluded in the evaluation because they constitute pulmonary abnormalities that could be important for different clinical diagnosis (Armato et al. (2011)). As such, a CAD mark on such a lesion is not a true false positive mark. It also alleviates the problem of disagreement as to what constitutes a nodule (Armato et al. (2009); van Ginneken et al. (2010)).

3. LUNA16 challenge

The proposed evaluation framework was coined the LUng Nodule Analysis 2016 (LUNA16) challenge. LUNA16 invites participants to develop a CAD system that automatically detects pulmonary nodules in CT scans. The challenge provides the data set and the reference annotations described in Section 2. This data set can be used for training of the systems and the evaluation of the algorithms is performed on the same data set. This makes LUNA16 a completely open challenge. To prevent biased results as a result of training and testing on the same data set, participants are instructed to perform cross-validation in the manner described in the following subsections. The LUNA16 website allows participants to submit the results. Submitted results are automatically evaluated and presented on the website.

3.1. Challenge tracks

The challenge consists of two separate tracks: (1) complete nodule detection and (2) false positive reduction.

The complete nodule detection track requires the participants to develop a complete CAD system, meaning that the only input into the system is a CT scan.

In the false positive reduction track, participants are only required to classify a number of locations in each scan as being a nodule or not. This is equivalent to the so-called false positive reduction step in many published CAD systems. For this track, a list of nodule candidates computed using existing nodule candidates detection algorithms is supplied to the participants (see Section 4.1). This can be seen as a typical machine learning task, where a two class classification (nodule/not-nodule) has to be performed. We included this track in the challenge to encourage the participation of teams with experience in image classification tasks but no particular background on the analysis of medical images. As further support, we included a tutorial on the LUNA16 website on how to extract cubes and patches around the nodule candidate locations in CT scans.

[3] https://luna16.grand-challenge.org/

3.2. *Cross-validation*

Participants are required to perform 10-fold cross-validation when they use the provided LIDC-IDRI data both as training and as test data. The data set has been randomly split into ten subsets of equal size on a patient level. The subsets can be directly downloaded from the LUNA16 website. The following steps describe how to perform 10-fold cross-validation (for fold N):

1. Split the data set into a test set and a training set (Subset N is used as test set and the remaining folds are used as the training set).
2. For the 'false positive reduction' track, test and training candidates should be extracted on the corresponding test and training set.
3. Train the algorithm on the training set.
4. Test the trained algorithm on the test set and generate the result file.
5. After iterating this process over all folds, merge the result files to get the result for all cases.

3.3. *Evaluation*

The results of the algorithms must be submitted online in the form of a comma separated value (csv) file. The csv file contains all marks produced by the CAD system. For each CAD mark, a position (image identifier, x, y, and z coordinates) and a score should be provided. The higher the score, the more likely the location is a true nodule.

A CAD mark is considered a true positive if it is located within a distance r from the center of any nodule included in the reference standard, where r is set to the radius of the reference nodule. When a nodule is detected by multiple CAD marks, the CAD mark with the highest score is selected. CAD marks that detect irrelevant findings are discarded from the analysis and are not considered as either false positive or true positive. CAD marks not falling into previous categories are marked as false positives.

Results are evaluated using the Free-Response Receiver Operating Characteristic (FROC) analysis (International Commission on Radiation Units and Measurements (2008)). The sensitivity is defined as the fraction of detected true positives (TPs) divided by the number of nodules in our reference standard. In the FROC curve, sensitivity is plotted as a function of the average number of false positives per scan (FPs/scan). For each scan, we take a maximum of 100 CAD marks that were given the highest scores. The 95% confidence interval of the FROC curve are computed using bootstrapping with 1,000 bootstraps, as detailed in Efron and Tibshirani (1994). In order to evaluate and compare different systems easily, we defined one overall output score. The overall score is defined as the average of the sensitivity at seven predefined false positive rates: 1/8, 1/4, 1/2, 1, 2, 4, and 8 FPs per scan. The performance metric was introduced in the ANODE09 challenge and is referred to as the Competition Performance Metric (CPM) in Niemeijer et al. (2011).

The evaluation script is publicly available on the LUNA16 website and can thus be viewed and used by all participants.

4. Methods

In this section we provide a brief description of the algorithms applied in the LUNA16 challenge. As of 31 October 2016, seven systems have been applied to the complete nodule detection track and five systems have been applied to the false positive reduction track. The candidate detection algorithms that were used to generate candidates for false positive reduction track are presented in Section 4.1; the systems submitted to the complete detection system track are described in Section 4.2; systems submitted to the false positive reduction track are detailed in Section 4.3.

4.1. *Candidate detection*

All candidate detection algorithms were developed as part of published CAD systems (Murphy et al. (2009); Jacobs et al. (2014); Setio et al. (2015); Tan et al. (2011); Torres et al. (2015)), some of which are included in the complete nodule detection track. As candidates from multiple algorithms are likely to be complementary, we merged all candidates using the procedure described in Section 4.1.6. The list of merged candidates can be downloaded from the LUNA16 website and can be used by teams that want to participate in the false positive reduction track.

4.1.1. *ISICAD*

The generic nodule candidate detection algorithm was developed by Murphy et al. (2009). First, the image is downsampled from 512 × 512 to 256 × 256 with the number of slices reduced to form isotropic resolution. Thereafter, Shape Index (SI), and curvedness (CV) are computed at every voxel in the lung volume as follows:

$$SI = \frac{2}{\pi}\arctan(\frac{k_1 + k_2}{k_1 - k_2})$$

$$CV = \sqrt[2]{k_1^2 + k_2^2}$$

where k_1 and k_2 are principal curvatures computed using first and second order derivatives of the image with a Gaussian blur of scale $\sigma = 1$ voxel. After SI and CV are computed, thresholding on these values is applied to obtain seed points for nodule candidates. These seed points represent voxels which may lie on a nodule surface. Seeds are expanded using broader thresholds to form voxel clusters. To reduce the number of the clusters, clusters within 3 voxels are merged recursively. The center of the mass of the cluster is considered to be the point of interest. The algorithm was developed using a data set from a large European lung cancer screening trial.

4.1.2. SubsolidCAD

The candidate detection algorithm was built to detect subsolid nodules, which are less common but more likely to be cancerous (Henschke et al. (2002)). The candidate detection algorithm by Jacobs et al. (2014) applies a double threshold density mask. The HU values commonly observed in subsolid nodules are used, ranging between -750 HU and -300 HU. Since partial volume effects may occur at the boundaries of the lungs, vessels, and airways, a morphological opening using spherical structuring element (3 voxels diameter) is applied to remove these structures. Next, connected component analysis is performed. Components with a volume smaller than 34 mm^3 are discarded from the list of candidates as subsolid nodules with a diameter smaller than 5 mm do not require follow-up CT. The centers of the candidate regions are used as nodule candidate locations. The algorithm was developed using a data set from a large European lung cancer screening trial.

4.1.3. LargeCAD

The algorithm has the function of detecting large nodules (Setio et al. (2015)). Large solid nodules (≥ 10 mm) have surface/shape index values or specific intensity range that is not captured by the two previously described nodule detection algorithms. An intensity threshold of -300 HU (usually corresponding to solid nodules) is applied in combination with multiple morphological operations. Thereafter, all connected voxels are clustered using connected component analysis; clusters with an equivalent diameter outside the range [8,40] mm are discarded. The algorithm was developed using the data set used by LUNA16.

4.1.4. ETROCAD

The applied method uses the detector system proposed by Tan et al. (2011). Isotropic re-sampling of the image to a voxel dimension of 1 mm^3 is applied in the preprocessing step. The nodule candidate algorithm consists of a nodule segmentation method based on nodule and vessel enhancement filters and a computed divergence feature to locate the centers of the nodule clusters. Three different set of filters (Li et al. (2003, 2004)) are applied to detect different types of nodules: isolated, juxtavascular, and juxtapleural nodules. To better estimate the location of the nodule centers and reduce the FP rate, the maxima of the divergence of the normalized gradient (DNG) of the image $k = \nabla(\overrightarrow{w})$ is used, where $\overrightarrow{w} = \frac{\overrightarrow{\Delta L}}{\|\overrightarrow{\Delta L}\|}$ and L is the image intensity. The enhancement filters and DNG are calculated at different scales in order to detect the seed points for different sizes of nodules.

Thresholding on the filtered image and DNG is applied to obtain the list of candidates. Different thresholds on the filtered image and the nodule-enhanced image are applied for isolated nodules, juxtavascular nodules, and juxtapleural nodules to get candidate locations. Finally, to ensure that a single nodule is represented by a single mark, cluster merging is performed. The algorithm was developed using a set of scans from LIDC-IDRI.

4.1.5. M5L

The candidate detection algorithm proposed by Torres et al. (2015) consists of two different algorithms: LungCAM and Voxel-Based Neural Approach (VBNA).

LungCAM is inspired based on the life-cycle of ants colonies (Cerello et al. (2010)). The lung internal structures are segmented by iteratively deploying ant colonies in voxels with intensity above a predefined thresholds. The ant colony moves to a specific destination and releases pheromones based on a set of rules (Chialvo and Millonas (1995)). Voxels visited by ant colonies are removed and new ant colonies are deployed in not-yet-visited voxels. Iterative thresholding of the pheromone maps is applied to obtain a list of candidates. The probability P_{ij} that a candidate destination is chosen is defined as:

$$P_{ij}(v_i \rightarrow v_j) = \frac{W(\sigma_j)}{\sum_{n=1,26} W(\sigma_n)}$$

where $W(\sigma_j)$ depends on the amount of pheromone in voxel v_j. The algorithm ends when all the ants in the colony have died.

VBNA uses two different procedures to detect nodules inside the lung parenchyma (Li et al. (2003); Retico et al. (2008)) and nodules attached to the pleura (Retico et al. (2009)). The nodules inside the lung parenchyma are detected using a dedicated dot-enhancement filter. Since nodules can manifest with a different size range, a multi-scale approach is followed (Li et al. (2003)). Nodule candidate locations are defined as the local maxima of the filtered image. The pleural nodules are detected by computing the surface normal at the lung wall. To build the normal, a marching cube algorithm is used. For each voxel inside the lung, the number of surface normals passing through are accumulated. Pleural candidates are defined as the local maxima of the accumulated scores. The algorithm was developed using a set of scans from LIDC-IDRI, ANODE09, and ITALUNG-CT.

4.1.6. Combining candidate detection algorithms

The combination of different CAD systems has been shown to improve the overall detection performance for nodule detection in chest CT (van Ginneken et al. (2010); Niemeijer et al. (2011)). The previously described candidate detection algorithms used different approaches to detect nodules and are therefore likely to detect different sets of nodules. Consequently, the combination of multiple algorithms may improve the detection sensitivity of nodules and would therefore be a better baseline for the false positive reduction systems.

To combine the results of multiple candidate detection algorithms, we concatenated the lists of candidates, where candidates located closer than 5 mm to each others were merged. The position of the merged candidates were averaged. Candidates located outside the lung region were discarded, as they were irrelevant for nodule detection. The lung region is determined based on the lung segmentation algorithm proposed by van Rikxoort et al. (2009). As the algorithm may exclude nodules attached to the lung wall, a slack border of 10 mm was applied.

4.2. Complete nodule detection system

The seven methods that were submitted to the complete nodule detection track are described in this section.

4.2.1. ZNET

ZNET uses ConvNets for both candidate detection and false positive reduction. As a preprocessing step, CT images are resampled to isotropic resolution of 0.5 mm. Candidate detection is extracted based on the probability map given by U-Net (Ronneberger et al. (2015)). U-net is applied on each axial slice. Before applying U-Net, the resampled input slice is cropped to 512 × 512. The candidates are extracted based on the slice-based probability map output of the U-net. A threholding is applied to obtain candidate masks. The threshold was determined on the validation subset, maximizing the number of detected nodules. Thereafter, a morphological erosion operation with a 4-neighborhood kernel is used to remove partial volume effects. The candidates are then grouped by performing connected component analysis. The center of mass of the components represent the coordinates of the candidates. The false positive reduction is described in Section 4.3.4. Both candidate detection and false positive reduction stages were trained in a cross-validation using the provided folds from LUNA16.

4.2.2. Aidence

Aidence is a company developing computer assisted diagnosis tools for radiologists based on deep learning (http://aidence.com/). The LUNAAidence algorithm uses end-to-end ConvNets trained on a subset of studies from the National Lung Screening Trial (NLST) with additional annotation provided by in-house radiologists. The LUNA16 data set was used for validation purposes only and was not used as training data. A detailed description is not available because of commercial confidentiality.

4.2.3. JianPeiCAD

JianPeiCAD is a system developed by Hangzhou Jianpei Technology Co. Ltd., a company based on Hangzhou, China (http://www.jianpeicn.com). The algorithm follows the common two stage work-flow of nodule detection: Candidate detection and false positive reduction. A multi-scale rule-based screening is applied to obtain nodule candidates. The false positive reduction uses 3D ConvNets with wide channels, which are trained using data augmentation to prevent overfitting. The system was developed using in-house resources (Chinese patient CT images and CT devices from local-vendors) and the LUNA16 data set was used as further validation for patients outside China. A detailed description is not available because of commercial confidentiality.

4.2.4. MOT_M5Lv1

The Multi Opening and Threshold CAD is a fully automatic CAD developed to be included into the M5L system (Torres et al. (2015)). The lung volume is obtained using 3D region growing, with trachea exclusion and lung separation procedures. The candidate detection algorithm was developed based on the method proposed by Messay et al. (2010). Multiple gray level-thresholding and morphological processing is used to detect and segment nodule candidates. Several modifications to the sequence of threshold and opening radius as well as the merging procedure are made. Subsequently, a dedicated nodule segmentation method (Kuhnigk et al. (2006)) is applied to separate nodules from vascular structures during the segmentation step. The false positive reduction computes 15 features, among which geometrical (e.g. radius, sphericity, skewness of distance from center) and intensity features (e.g. average, standard deviation, maximum, entropy). Classification is performed using feed-forward neural networks that consists of 1 input layer with 15 input units, 1 hidden layer with 31 units, and 1 output layer with 1 output unit. The algorithm was developed using the LUNA16 data set.

4.2.5. *VISIACTLung*

This submission contains the results of the commercially available Visia$^{\text{TM}}$ CT Lung CAD system, version 5.3 (MeVis Medical Solutions AG, Bremen, Germany). This is an FDA approved CAD system designed to assist radiologists in the detection of solid pulmonary nodules during review of multidetector CT scans of the chest. It is intended to be used as an adjunct, alerting the radiologist after his or her initial reading of the scan to regions of interest (ROIs) that may have been initially overlooked. A detailed description is not available because of commercial confidentiality.

4.2.6. *ETROCAD*

ETROCAD is a CAD system adapted from Tan et al. (2011). The candidate detection algorithm is described in Section 4.1.4. The false positive reduction stage uses a dedicated feature extraction and classification algorithm. For each candidate, a set of features is computed, including invariant features defined on a 3D gauge coordinates system (Salden et al. (1991)), shape features, and regional features. The classification is performed using a SVM classifier with a radial basis function. The algorithm was developed using a set of scans from LIDC-IDRI.

4.2.7. *M5LCAD*

M5LCAD is a CAD system developed by Torres et al. (2015), which consists of two algorithms: LungCAM and VBNA. This CAD system uses the candidate detector algorithms described in section 4.1.5. The false positive reduction stage of LungCAM computes a set of 13 features for nodule candidate analysis, including spatial, intensity, and shape features. The set of features is used to classify the candidates using a feed-forward artificial neural network (FFNN). The FFNN architecture consists of 13 input neurons, 1 hidden layer with 25 neurons, and 1 neuron as output layer. The false positive reduction of VBNA performs the classification using a standard three-layered FFNN using the raw voxels as the feature vector (Retico et al. (2008, 2009)). The algorithm was developed using a set of scans from LIDC-IDRI, ANODE09, and ITALUNG-CT.

4.3. *False positive reduction systems*

The five methods that were applied to the false positive reduction track are described in this section.

4.3.1. *CUMedVis*

CUMedVis uses multi-level contextual 3D ConvNets developed by Dou et al. (2016). To tackle challenges coming from variations of nodule sizes, types, and geometry characteristics, a system that consists of three different 3D ConvNets architectures (*Archi-a*, *Archi-b*, *Archi-c*) was presented. Each subsystem uses an input image with different receptive field so that multiple levels of contextual information surrounding the suspicious location could be incorporated.

Archi-a has a receptive field of 20×20×6. Three convolutional layers are used with 64 kernels of $5 \times 5 \times 3$, $5 \times 5 \times 3$, $5 \times 5 \times 1$, respectively. Thereafter, a fully-connected layer with 150 output units and a softmax layer are applied. *Archi-b* has a receptive field of $30 \times 30 \times 10$. The first convolutional layer with 64 kernels of $5 \times 5 \times 3$ is used followed by a max-pooling layer with kernel $2 \times 2 \times 1$. Two convolutional layers each with 64 kernels of $5 \times 5 \times 3$ are then added, finalized by a fully-connected layer with 250 output units and a softmax layer. *Archi-c* has the largest receptive field of $40 \times 40 \times 26$. After the first convolutional layer with 64 kernels of $5 \times 5 \times 3$, a max-pooling layer with kernel $2 \times 2 \times 2$ is used. Thereafter, two convolutional layers each with 64 kernels of $5 \times 5 \times 3$ are added. Finally, a fully-connected layer with 250 output units and a softmax layer are established. The prediction probabilities from the three ConvNets architectures are fused with weighted linear combination to produce the final prediction for a given candidate.

For pre-processing, voxel intensities are clipped into the interval from -1000 to 400 HU and normalized into the range of 0 to 1. To deal with the class imbalance between the false positives and nodules, translation (one voxel along each axis) and rotation (90^0, 180^0, 270^0 within the transverse plane) augmentations are performed on the nodules. The weights are initialized using a Gaussian distribution and are optimized using the standard back-propagation with momentum (Sutskever et al. (2013)). A dropout strategy (Hinton et al. (2012)) was applied during training. The system was implemented using Theano (Bastien et al. (2012)) and a GPU of NVIDIA TITAN Z was used for acceleration. The algorithm was developed using the cerebral microbleeds data set and was further optimized using LUNA16 data set.

4.3.2. *JackFPR*

The proposed method uses a similar multi-level contextual 3D ConvNet architecture as presented by Dou et al. (2016). It uses the three architectures (*Archi-a*, *Archi-b*, *Archi-c*) described in Section 4.3.1 with several modifications. Exponential activation units were used as the activation functions. So instead of combining the predictions of three ConvNets using linear combination, the fully-connected layers from three architectures were concatenated and connected to a fully-connected layer with 128 output units. The last fully-connected layer was then followed by a softmax layer to obtain the prediction.

The training was performed for 240 epochs with 1,024 iterations per epoch. Xavier initialization (Glorot and Bengio (2010)) was used as the weight initialization and Nesterov accelerated Stochastic Gradient Descent (SGD) was used. Cross-entropy loss, L2 regularization loss, and center loss were used as the cost function. Center loss penalizes the difference between a running average of learned features for each class and sample class features seen during the particular batch (Wen et al. (2016)). The learning rate was set to 0.005 for the first 5 epochs as a warming up. Thereafter, the learning rate was set to 0.01 and was reduced by 1/10 every 80 epochs. Data augmentation was performed and dropout was applied to combat over-fitting. The algorithm was developed using LUNA16 data set.

4.3.3. *DIAG CONVNET*

This method uses multi-view ConvNets proposed by Setio et al. (2016). For each candidate, nine 65×65 patches of 50×50 mm from different views are extracted. Each view corresponds to a different plane of symmetry in a cube and is processed using a stream of 2D ConvNets. The ConvNets stream consists of 3 consecutive convolutional layers and max-pooling layers: The first is formed by 24 kernels of 5×5; the second by 32 kernels of 3×3; and the third by 48 kernels of 3×3. Weights are initialized randomly and updated during training. The max-pooling layer is used to reduce the size of patches by half. The last layer is a fully connected layer with 16 output units. Rectified linear units (ReLU) are used as the activation functions. The fusion of the different ConvNets is performed using the late fusion method (Prasoon et al. (2013); Karpathy et al. (2014)). Fully-connected layers from all streams are concatenated and are connected directly to a softmax layer. This approach allows the network to learn 3D characteristics by comparing outputs from multiple ConvNets streams. In this approach, all the parameters of the convolutional layers from different streams are shared.

Data augmentation was applied on candidates in the training set to increase the variance of presentable candidates. For each candidate, random zooming $[0.9, 1.1]$ and random rotation $[-20°, +20°]$ were performed. To prevent over-fitting during training, random positive and negative candidates with equal distribution were sampled in a batch of 64 samples. Validation was performed every 1,024 batches. Training was stopped when the area under the curve of the receiver operating characteristic on the validation data set does not improve after 3 epochs. Xavier initialization (Glorot and Bengio (2010)) was used as the weight initialization. The weights were optimized using RMSProp (Tieleman and Hinton (2012)), and evaluation was performed in a 10-fold cross validation. Compared to the original work (Setio et al. (2016)), the submitted system uses an ensemble of three multi-view ConvNets trained using different random seed-points, averaging out biases from training using random samples. The system was implemented using Theano (Bastien et al. (2012)); a NVIDIA TITAN X GPU was used for acceleration. Three different architectures were evaluated by Setio et al. (2016) using the same LUNA16 data set and the best performing architecture was selected. The algorithm was further optimized using the same LUNA16 data set.

4.3.4. *ZNET*

ZNET used the recently published wide residual networks (Zagoruyko and Komodakis (2016)). For each candidate, 64×64 patches from the axial, sagittal and coronal views were extracted. Each patch was processed separately by the wide residual networks. The predicted output values of the network for these three different patches were averaged to obtain the final prediction. The architecture used 4 sets of consecutive convolutional layers. The first set consisted of 1 convolutional layer with 16 kernels of 3×3. Sets two to four consisted of 10 convolutional layers with a stride of two, each with 96 kernels of 3×3, 192 kernels of 3×3, and 384 kernels of 3×3, respectively. Each set also had a $1 \times 1 \times N$ projection convolution in their skip connection, where N is the number of kernels in the corresponding set. The fourth layer was additionally connected to a global average pooling layer, resulting in a $1 \times 1 \times 384$ output image connected to the softmax layer.

Xavier initialization was used for weight initialization (Glorot and Bengio (2010)) and ADAM was used as the optimization method (Xu et al. (2015)). Leaky Rectified Linear Units were used as nonlinearities throughout the network. Data augmentation (flipping, rotation, zooming and translation) was applied not only to the training data set but also the test data set in order to improve the test set scores. The learning rate was reduced over time: Learning rate is decreased by 90% after epochs 80 and 125. All convolutional networks were implemented using the Lasagne and Theano libraries (Dieleman et al. (2015); Bastien et al. (2012)). The training was performed on a computer cluster using large range of CUDA enabled graphics cards including the Tesla K40M, Titan X, GTX 980, GTX 970, GTX 760 and the GTX 950M. The algorithm was developed using the LUNA16 data set.

4.3.5. CADIMI

This method used multi-slice ConvNets. For each axial, sagittal, and coronal view, three patches are extracted at three locations: The plane in the exact candidate location and the planes 2 mm in both directions of the remaining free axes (x, y, and z). The patches were concatenated as three-dimensional arrays, which resulted in patches of $52 \times 52 \times 3$ mm centered around the candidate location. The network consisted of 2D ConvNets with three consecutive convolutional layers and max-pooling: The first convolutional layer used 24 kernels of 5×5; the second used 32 kernels of 3×3; and the third 48 kernels of 3×3. The output of the last max-pooling was connected to fully-connected layer of 512 output units. ReLU was used as the activation function, and the last fully-connected layer was connected to a softmax layer.

Training was performed one time using patches from all three views for 80 epochs. For each epoch, all positive patches and 20,000 random negative patches were used. In order to tackle the problem of data imbalance, data augmentation (vertical/horizontal flip and random cropping) was applied. During testing, 5 patches (1 center patch and 4 patches with $[-4, +4]$ translation in two axes) were extracted from each view. These patches were processed using a single trained network and the predictions were averaged. Batch normalization was applied after each max-pooling layer to reduce overfitting. The weights were initialized using the uniform initialization (He et al. (2015b)). Nesterov accelerated SGD with a learning rate of 0.01, a decay of 0.001, and a momentum of 0.9 is used. The system was implemented using the Lasagne and Theano libraries (Dieleman et al. (2015); Bastien et al. (2012)). The algorithm was developed using the LUNA16 data set.

4.4. Combining false positive reduction systems

The combination of multiple classification methods, also known as an ensemble method, has been used in many machine-learning problems to improve the prediction performance (Dietterich (2000)). As systems applied in the false positive reduction track use the same set of candidates, the impact of combining multiple methods could be evaluated. In this study, we combined CAD results from the systems in the false positive reduction track. The combination is performed by simply averaging the probabilities given by the systems. Such is a common approach in optimizing the performance of deep learning architectures (Szegedy et al. (2014); He et al. (2015a)).

4.5. Observer study

To evaluate the potential of CAD systems to detect nodules missed by human readers, and to elucidate the nature of the false positives detected by the CAD systems, an observer study was performed. In the observer study, CAD marks from the combination of false positive reduction systems were assessed to identify if there were additional nodules detected. The reading process was performed by four expert readers independently.

We extracted all CAD marks at 0.25 FPs/scan that were categorized as false positives to be further analyzed by expert readers. To reduce the readers' workload, research scientists read and removed CAD marks that were obvious false positives (e.g. vessels, ribs, diaphragm) beforehand. Thereafter, CAD marks that were close to annotated lesions in LIDC-IDRI but were missed by our hit criteria (thus considered as false positives) were discarded. Most of these lesions were non-nodular and therefore were not well captured by the defined hit criteria (radius of the corresponding lesion). This operation resulted in a set of 127 marks that were potentially nodules. As a similar observer study was performed in our previous study (Jacobs et al. (2016)), marks which were already evaluated on this CT data by radiologists were not read again and the scores of the four radiologists from the previous study were used. Last, we asked the expert readers to review and annotate the remaining marks as: nodule ≥ 3 mm, nodule < 3 mm, or false positives. Measurement tools were made available to readers during the process in order to enable size evaluation.

5. Results

In this section, we present the results achieved by all individual systems described in Section 4. The results of combining multiple algorithms are provided.

5.1. Candidate detection

Table 1 summarizes the performance of individual candidate detection algorithms and their top performing combinations. The best detection sensitivity of 92.9% was achieved by ETROCAD. When multiple candidate detection algorithms were combined, the sensitivity substantially improved up to 98.3% (1,166/1,186 nodules), higher than the sensitivity of any individual system. This illustrates the potential of combining multiple candidate detection algorithms to improve the sensitivity of CAD systems.

5.2. Complete nodule detection track

The FROC curves of the systems on the complete nodule detection track are shown in Figure 1a. In this track, the best score was achieved by ZNET with a CPM of 0.811. Other systems show comparable performance. It was observed that the relatively large differences in terms of sensitivity at low FPs/scan substantially influences the overall scores of the systems.

5.3. False positive reduction track

The FROC curves of the systems on the false positive reduction track are shown in Figure 1b. The best average score was achieved by CuMedVis, with a CPM of 0.908. Table 2 shows all possible system combinations, where the sensitivities of the combined systems were higher than the sensitivity achieved by the best system. Although all false positive reduction systems are based on ConvNets, it is evident that combining ConvNets with different configurations further improves the overall sensitivity as shown in Table 2. The p-value was defined as the probability that a system's CPM is higher or lower than the reference system's CPM. A p-value below 0.002 is considered to be statistically significant after Bonferroni correction ($m = 30$).

5.4. Performance based on nodule type

To assess the performance of the algorithms on different types of nodules (non-solid, part-solid, and solid), additional analysis was performed. The nodule type was derived based on the morphological characteristic scored by the LIDC-IDRI radiologists. The nodule was labeled ”non-solid” if the majority of the radiologists gave a texture score of 1, ”solid” for a majority score of 5, and ”part-solid” if the two previous criterion did not hold. Using this labelling strategy, the LUNA16 data set consisted of 64 non-solid nodules, 189 part-solid nodules, and 933 solid nodules. The performance of the algorithms on different set of nodules are tabulated in Table 3.

5.5. Analysis of false positives: observer study

A summary of the observer study is shown in Table 4. Among 127 CAD marks, 108, 91, 69, and 41 CAD marks were accepted as nodules ≥ 3 mm by at least 1, 2, 3, or 4 readers, respectively; 6 out of 19 remaining CAD marks were considered as nodule < 3 mm. Examples of nodules found in this observer study are shown in Figure 2c. We shared the set of additional nodules on the LUNA16 website to be used for further development of CAD systems.

6. Discussion

In this study, we presented LUNA16: A novel evaluation framework for automatic nodule detection algorithms. The aim of the study was to supply the research community a framework to test and compare algorithms on a common large database with a standardized evaluation protocol. This allows the community to objectively evaluate different CAD systems and push forward the development of state of the art nodule detection algorithms. The submitted systems were described and the performance was evaluated. We showed that the combination of multiple false positive reduction algorithms applied on a combined set of candidates outperformed any individual system. This highlights the potential of combining algorithms to improve the detection performance.

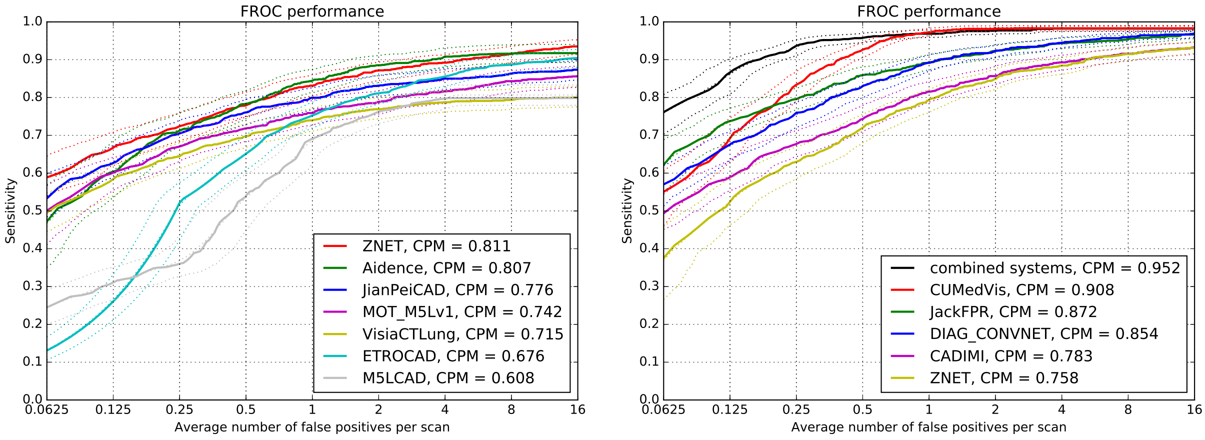
Candidate detection plays an important role of determining the maximum attainable detection sensitivity of a CAD system. The algorithms should ideally detect all nodules with an acceptable amount of false positives. Table 1 shows that the individual candidate detection algorithms achieve a detection sensitivity between 31.8% and 92.9%; combining different candidate detection algorithms improved the sensitivity up to 98.3%. While a smaller set of candidates (a combined set of only ISICAD, SubsolidCAD, and LargeCAD candidates) were also provided in the earlier phase of LUNA16, we here only reported the results of the systems that use the latest set of candidates that has a much higher sensitivity. The results of other systems that use the smaller set of candidates, resulting in lower scores, are available on the LUNA16 website. It is worth noting that the candidate detection systems used in this study do not employ

System name	Combination	Sensitivity	Best single sensitivity	Difference sensitivity	Total number of candidates	Average number of candidates / scan
ISICAD	■□□□□	0.856			298 256	335.9
SubsolidCAD	□■□□□	0.361			258 075	290.6
LargeCAD	□□■□□	0.318			42 281	47.6
M5L	□□□■□	0.768			19 687	22.2
ETROCAD	□□□□■	0.929			295 686	333.0
	■■□□□	0.918	0.857	0.062	520 319	585.9
	■□■□□	0.898	0.857	0.041	328 742	370.2
	■□□■□	0.917	0.857	0.061	308 047	346.9
	■□□□■	0.959	0.929	0.030	524 108	590.2
	□■■□□	0.523	0.361	0.162	295 476	332.7
	□■□■□	0.869	0.768	0.101	274 900	309.6
	□■□□■	0.954	0.929	0.024	518 058	583.4
	□□■■□	0.834	0.768	0.066	59 359	66.8
	□□■□■	0.945	0.929	0.016	319 405	359.7
	□□□■■	0.942	0.929	0.013	297 030	334.5
	■■■□□	0.944	0.857	0.088	551 065	620.6
	■■□■□	0.954	0.857	0.098	530 942	597.9
	■■□□■	0.977	0.929	0.048	728 162	820.0
	■□■■□	0.934	0.857	0.078	339 229	382.0
	■□■□■	0.964	0.929	0.035	548 523	617.7
	■□□■■	0.967	0.929	0.038	529 404	596.2
	□■■■□	0.900	0.768	0.132	310 323	349.5
	□■■□■	0.964	0.929	0.035	545 204	614.0
	□■□■■	0.965	0.929	0.035	524 726	590.9
	□□■■■	0.954	0.929	0.024	326 274	367.4
	■■■■□	0.980	0.929	0.051	750 838	845.5
	■■■□■	0.983	0.929	0.054	732 901	825.3
	■■□■■	0.970	0.929	0.040	553 327	623.1
	■□■■■	0.965	0.857	0.108	559 543	630.1
	□■■■■	0.970	0.929	0.040	551 227	620.8
	■■■■■	0.983	0.929	0.054	754 975	850.2

Table 1: The results of five candidate detection systems and all possible combinations. The filled squares indicate which systems were included in the combination. CPM: Competition Performance Metric.

System name	Combination	0.125	0.25	0.5	1	2	4	8	CPM	P-value	Difference CPM	Best single CPM
CUMedVis	■□□□□	0.677	0.834	0.927	0.972	0.981	0.983	0.983	0.908	Reference		
JackFPR	□■□□□	0.734	0.796	0.859	0.892	0.923	0.944	0.954	0.872	0.002		
DIAG CONVNET	□□■□□	0.669	0.760	0.831	0.892	0.923	0.945	0.960	0.854	<0.001		
CADIMI	□□□■□	0.583	0.677	0.743	0.815	0.857	0.893	0.916	0.783	<0.001		
ZNET	□□□□■	0.511	0.630	0.720	0.793	0.850	0.884	0.915	0.758	<0.001		
	■■□□□	0.809	0.901	0.962	0.976	0.981	0.981	0.982	0.942	<0.001	0.908	0.034
	■□■□□	0.831	0.917	0.965	0.979	0.981	0.981	0.981	0.948	<0.001	0.908	0.040
	■□□■□	0.802	0.903	0.948	0.976	0.979	0.979	0.980	0.938	<0.001	0.908	0.030
	■□□□■	0.831	0.927	0.968	0.976	0.979	0.981	0.981	0.949	<0.001	0.908	0.041
	□■■□□	0.745	0.826	0.864	0.906	0.948	0.958	0.969	0.888	0.042	0.872	0.016
	□■□■□	0.717	0.797	0.858	0.895	0.932	0.947	0.959	0.872	<0.001	0.872	0.000
	□■□□■	0.728	0.828	0.879	0.917	0.938	0.954	0.963	0.887	0.038	0.872	0.015
	□□■■□	0.550	0.680	0.796	0.869	0.912	0.938	0.959	0.815	<0.001	0.854	-0.040
	□□■□■	0.616	0.737	0.831	0.888	0.931	0.953	0.964	0.845	<0.001	0.854	-0.009
	□□□■■	0.602	0.732	0.812	0.852	0.884	0.913	0.946	0.820	<0.001	0.783	0.037
	■■■□□	0.821	0.898	0.954	0.975	0.981	0.982	0.982	0.942	<0.001	0.908	0.034
	■■□■□	0.816	0.897	0.945	0.970	0.980	0.980	0.980	0.938	<0.001	0.908	0.030
	■■□□■	0.843	0.911	0.957	0.978	0.981	0.981	0.981	0.947	<0.001	0.908	0.039
	■□■■□	0.817	0.912	0.954	0.968	0.975	0.979	0.982	0.941	<0.001	0.908	0.033
	■□■□■	0.859	0.937	0.958	0.969	0.976	0.982	0.982	0.952	<0.001	0.908	0.044
	■□□■■	0.820	0.907	0.946	0.968	0.976	0.981	0.981	0.940	<0.001	0.908	0.032
	□■■■□	0.720	0.802	0.864	0.916	0.941	0.960	0.970	0.882	0.010	0.872	0.010
	□■■□■	0.736	0.835	0.891	0.924	0.945	0.969	0.973	0.896	0.222	0.872	0.024
	□■□■■	0.741	0.815	0.874	0.918	0.938	0.954	0.965	0.887	0.024	0.872	0.015
	□□■■■	0.635	0.777	0.839	0.888	0.929	0.954	0.965	0.855	<0.001	0.854	0.001
	■■■■□	0.823	0.896	0.939	0.968	0.977	0.980	0.981	0.938	<0.001	0.908	0.030
	■■■□■	0.846	0.912	0.949	0.971	0.977	0.981	0.982	0.946	<0.001	0.908	0.037
	■■□■■	0.821	0.892	0.944	0.970	0.978	0.981	0.981	0.938	<0.001	0.908	0.030
	■□■■■	0.830	0.912	0.947	0.964	0.973	0.979	0.981	0.941	<0.001	0.908	0.033
	□■■■■	0.745	0.823	0.884	0.925	0.946	0.961	0.973	0.894	0.102	0.872	0.022
	■■■■■	0.836	0.896	0.940	0.965	0.976	0.981	0.982	0.939	<0.001	0.908	0.031

Table 2: The results of five false positive reduction systems and all possible combinations. The filled squares indicate which systems were included in the combination. CPM: Competition Performance Metric.



(a)

(b)

Figure 1: FROC curves of the systems in (a) the nodule detection track and (b) the false positive reduction track. Dashed curves represent the 95% confidence intervals estimated using bootstrapping.

Table 3: Performance benchmark of algorithms on different sets of nodules. Nodules are categorized based on the nodule type, which was derived based the morphological characteristic scored by the LIDC-IDRI radiologists (5-point scale: 1=non-solid, 3=part-solid, 5=solid). The nodule was labeled ”non-solid” if the majority of the radiologists gave a texture score of 1, ”solid” for a majority score of 5, and ”part-solid” if the two previous criterion did not hold. CPM score was used as the performance metric

System name	all (1,186)	non-solid (64)	part-solid (189)	solid (933)
Nodule detection track				
ZNET	0.811	0.663	0.735	0.836
Aidence	0.807	0.730	0.776	0.819
JianPeiCAD	0.776	0.248	0.725	0.825
MOT_M5v1	0.742	0.217	0.696	0.787
VisiaCTLung	0.715	0.033	0.652	0.775
ETROCAD	0.676	0.290	0.547	0.728
M5LCAD	0.608	0.156	0.549	0.650
False positive reduction track				
CUMedVis	0.908	0.908	0.912	0.907
JakFPR	0.872	0.636	0.845	0.893
DIAG_CONVNET	0.854	0.688	0.819	0.873
CADIMI	0.783	0.496	0.751	0.810
ZNET	0.758	0.498	0.685	0.790

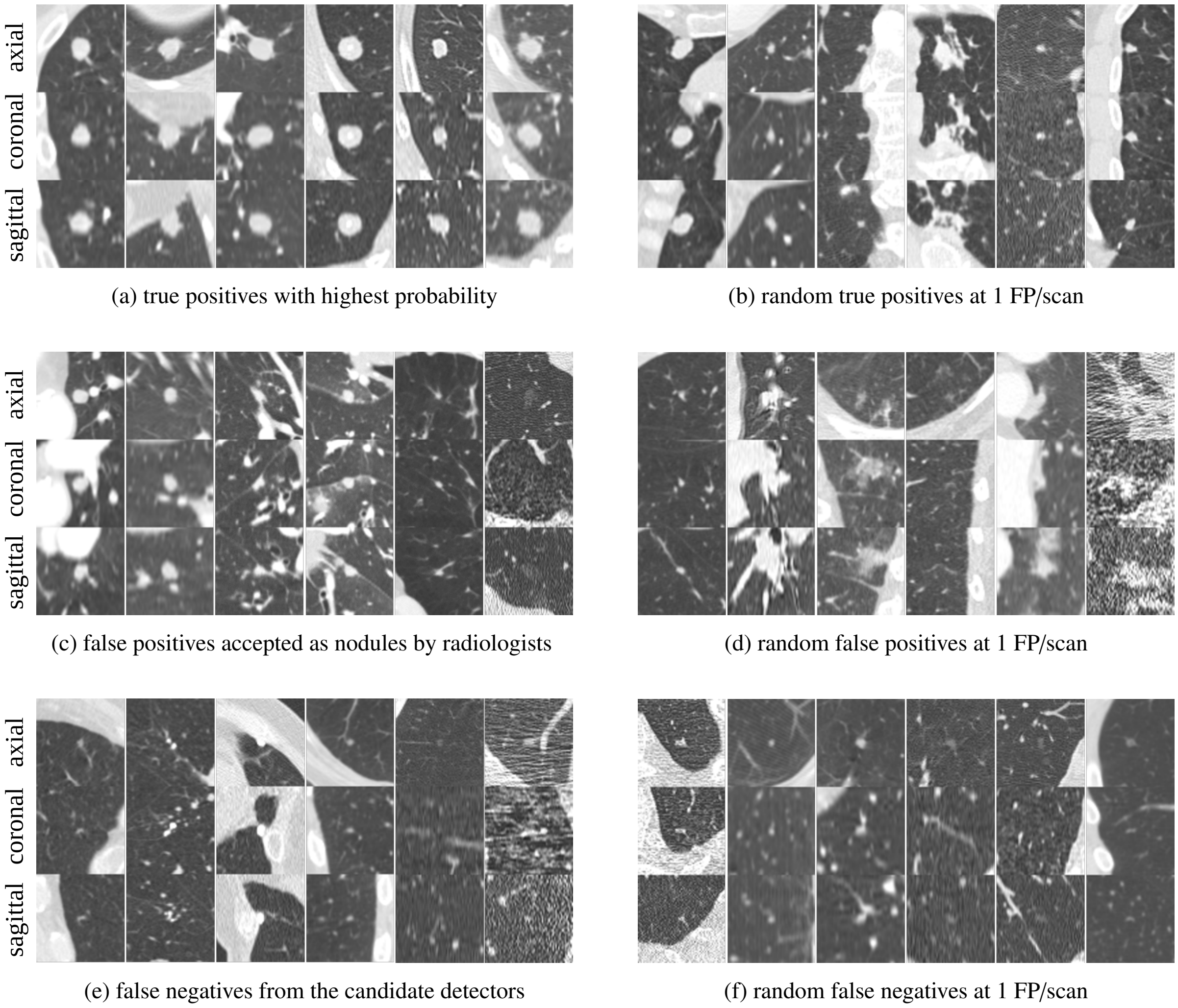
Table 4: An overview of the observer study on 222 false positives at 0.25 FPs/scan. The table shows the number of false positives that were accepted by the expert readers as nodule $\geq$3 mm at different agreement levels. The number of false positives that were not accepted as nodule $\geq$3 mm but were accepted as nodule<3 mm is also included.

Category	Number
nodule $\geq$3 mm - at least 1	108
nodule $\geq$3 mm - at least 2	91
nodule $\geq$3 mm - at least 3	69
nodule $\geq$3 mm - at least 4	41
nodule <3 mm	6

deep learning, while systems in the complete nodule detection track, e.g. ZNET, do employ ConvNets to detect nodule candidates.

In the complete nodule detection track, a total of seven systems were evaluated. Diverse methods were applied and different sets of data were used for training. When evaluated using the same data set, the detection sensitivity ranged between 69.1% and 91.5% at 1 and 8 FPs/scan, as shown in Figure 1a. Notably, the top three systems make use of ConvNets for their detection algorithms. While the variability of the performance is determined by the underlying methods, it is also affected by the training data used to develop the system (see also Table 5). This suggests the need of a standardized training data set for appropriate comparison of algorithms.

In the false positive reduction track, different systems for false positive reduction were evaluated given a common set of candidates and training data. A total of five systems were evaluated. ConvNets were used as the prediction model for all systems, which is in line with the recent trend of adopting deep learning in the medical image analysis domain. As shown in Figure 1b, all sys-



(a) true positives with highest probability

(b) random true positives at 1 FP/scan

(c) false positives accepted as nodules by radiologists

(d) random false positives at 1 FP/scan

(e) false negatives from the candidate detectors

(f) random false negatives at 1 FP/scan

Figure 2: Examples of true positives, false positives, and false negatives from the combined system. Each lesion is located at the center of the 50×50 mm patch in axial, coronal, and sagittal views.

tems achieve detection sensitivity between 79.3% and 98.3% at 1 and 8 FPs/scan. As the underlying method is similar, one could hypothesize that there could be little to no benefit when these systems are combined. Nevertheless, combining multiple ConvNets systems did substantially improve the detection performance (black line on Figure 1b); a detection sensitivity of over 95.0% was achieved at fewer than 1 FP/scan. Despite all methods being based on ConvNets, the differences in network parameters, such as selected architectures, random initialization methods, and input patches, apparently make these systems somewhat complementary for prediction, and this is leveraged by the (simple averaging) combination.

To provide a broader context to the results reported in this paper, we listed the performance of other published CAD systems that use LIDC-IDRI data in Table 5. For each CAD system, we listed the number of scans used in the validation data set, nodule inclusion criteria, the number of included nodules, and the reported CAD performance. Note that different subsets of LIDC-IDRI database were used; LUNA16 aims to make the CAD performance comparison more easy and more fair by using exactly the same data and evaluation protocol for each system. The CAD systems presented in Jacobs et al. (2016) are not listed in this table as these CAD systems also participated in the LUNA16 challenge and hence are already described in this paper.

The observer study showed that some false positives detected by the CAD systems are nodules that were missed during the manual annotations of LIDC-IDRI. The majority of these nodules were overlooked because they were small or less visible (e.g. ground-glass/non-solid nodules). Other nodules may be missed because there were multiple nodules in the corresponding scans, or because the nodules were part of a complex abnormality (e.g. an area of consolidation). While these nodules may be found during follow-up, detecting them early could provide essential clinical information (e.g. growth rate).

Examples of lesions detected or missed by the combined CAD system are shown in Figure 2. Nodules with a wide range of morphological characteristics are detected at 1 FP/scan, showing that ConvNets are capable of capturing morphological variation of nodules in the network. Larger nodules are unlikely to be missed, which is just as well as there is a strong positive correlation between size and malignancy risk. Most false positives are large vessels, scar tissue, spinal abnormalities, and other mediastinal structures. These false positives are a challenge. In scans from subjects with interstitial lung disease, there are, even in mild cases, regions with irregular opacities that can lead to a large number of erroneous nodule CAD marks. Other false positives are caused by motion artifacts and extreme noise. The false negatives were small and/or had irregular shapes. Improving the robustness of the candidate detection algorithms to detect small nodules should further improve the performance.

This study has several limitations. As the LIDC-IDRI is a web-accessible database for development and evaluation of CAD systems, all nodule annotations are publicly available. The usage of a completely open database is not a common setup for challenges. Typically, an independent test set is provided, for which the reference annotations are not made public. As a consequence, in LUNA16 teams could tune the parameters of their algorithm to show good performance on this particular data set, although the fact that LIDC/IDRI is a large set of scans from many different sources somewhat mitigates this risk. In order to allow performance comparison among different systems, we instructed participants that did not have their own training data to train their system in a particular cross-validation approach. Although this prevents some of the positive bias, positive bias may still remain if the design and architecture of a system are selected or optimized based on the full challenge data set. Previously published algorithms have also used LIDC-IDRI data to optimize the design of their systems. This may have influenced the design of algorithms used in LUNA16 as well. Moreover, a cross-validation approach introduces some risks as people may make a mistake that goes unnoticed while carrying out a cross-validation experiment. In fact, one team that originally participated in the challenge and reported excellent results had to withdraw because of a bug in the reinitialization of the network weights when starting training for the next fold in cross-validation. Unforeseen errors aside, allowing a cross-validation training procedure means that the presented systems are evaluated on test data while having been trained with data from the same sources (institutions, scanners, protocols). This may incur a positive bias in the reported results. This potential for bias is however also present in most, if not all, studies on the topic that have been previously published. Ruling out any possible bias is still an open problem for many machine learning competitions, even when the test data is not publicly available (Blum and Hardt (2015)). On this note, it is important to further validate the performance and the generalizability of the systems using a completely independent validation data set.

Possible approaches for system improvement are as follows. Most of the non-nodular abnormalities, especially in lungs with many irregularities, were still detected. Although they may be clinically relevant, the algorithm should be able to differentiate these abnormalities. Classifying abnormalities into a subset of taxonomy, similar to Esteva et al. (2017), may be more useful for clinical usage. A more direct application would be to assign malignancy scores for all nodules detected by the system. This would further optimize the lung cancer screening workflow, for which less suspicious nodules would not have to undergo extensive screening protocol.

A future challenge could incorporate an even larger data set split into a training data set with annotations and a dedicated test data set for evaluation in which the reference standard is kept hidden. This still introduces a risk that teams can visually inspect the test data and the output of their system, notice false positives and false negatives and use that information to improve their performance. This could be circumvented by letting teams upload their algorithms, e.g. as machine executables or software containers, and evaluate these on test data that is not released publicly.

7. Conclusions

We have presented a web-based framework for a fair and automated evaluation of nodule detection algorithms, using the largest publicly available data set of chest CT scans in which nodules were annotated by

Table 5: The performance summary of published CAD systems evaluated using LIDC-IDRI data sets. Note that different subsets of scans were used by different research groups.

CAD systems	Year	# scans	slice thickness	nodules size (mm)	agreement levels	# nodules	sensitivity (%) / FPs/scan	
Combined LUNA16	-	888	≤2.5	≥3	at least 3	1,186	98.2 / 4.0	96.9 / 1.0
Dou et al. (2016)	2016	888	≤2.5	≥3	at least 3	1,186	90.7 / 4.0	84.8 / 1.0
Setio et al. (2016)	2016	888	≤2.5	≥3	at least 3	1,186	90.1 / 4.0	85.4 / 1.0
Bergtholdt et al. (2016)	2016	243	-	≥3	at least 1	690	85.9 / 2.5	-
Torres et al. (2015)	2015	949	-	≥3	at least 2	1,749	80.0 / 8.0	-
van Ginneken et al. (2015)	2015	865	≤2.5	≥3	at least 3	1,147	76.0 / 4.0	73.0 / 1.0
Brown et al. (2014)	2014	108	0.5-3	≥4	at least 3	68	75.0 / 2.0	-
Choi and Choi (2013)	2013	58	0.5-3	3-30	at least 1	151	95.3 / 2.3	-
Tan et al. (2013)	2013	360	-	≥3	at least 4	-	83.0 / 4.0	-
Teramoto and Fujita (2013)	2013	84	0.5-3	5-20	at least 1	103	80.0 / 4.2	-
Cascio et al. (2012)	2012	84	1.25-3	≥3	at least 1	148	97.0 / 6.1	88.0 / 2.5
Guo and Li (2012)	2012	85	1.25-3	≥3	at least 3	111	80.0 / 7.4	75.0 / 2.8
Camarlinghi et al. (2011)	2011	69	0.5-2	>3	at least 2	114	80.0 / 3.0	-
Riccardi et al. (2011)	2011	154	0.5-3	≥3	at least 4	117	71.0 / 6.5	60.0 / 2.5
Tan et al. (2011)	2011	125	0.75-3	≥3	at least 4	80	87.5 / 4.0	-
Messay et al. (2010)	2010	84	1.3-3	≥3	at least 1	143	82.7 / 3.0	-

multiple exert human readers. We have shown that combining classical candidate detection algorithms and analyzing these candidates with convolutional networks yields excellent results. Additionally, we have provided an update to the LIDC-IDRI reference standard which includes additional nodules found by CAD. The LUNA16 challenge will remain open for new submissions and can therefore be used as a benchmarking framework for future CT nodule CAD development.

Acknowledgements

The authors acknowledge the National Cancer Institute and the Foundation for the National Institutes of Health and their critical role in the creation of the free publicly available LIDC-IDRI Database used in this study. This work was supported by a research grant from the Netherlands Organization for Scientific Research [project number 639.023.207]. A. Sóñora was supported by the Belgian Development Cooperation through the VLIR-UOS program with Universidad de Oriente in Santiago de Cuba. E.Papavasileiou was supported by a PhD grant of the Research Foundation Flanders (FWO).

References

Aberle, D.R., Adams, A.M., Berg, C.D., Black, W.C., Clapp, J.D., Fagerstrom, R.M., Gareen, I.F., Gatsonis, C., Marcus, P.M., Sicks, J.D., 2011. Reduced lung-cancer mortality with low-dose computed tomographic screening. New England Journal of Medicine 365, 395–409. doi:10.1056/NEJMoa1102873.

American Cancer Society, 2016. Cancer facts and figures 2016.

Armato, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., Kazerooni, E.A., MacMahon, H., Beek, E.J.R.V., Yankelevitz, D., Biancardi, A.M., Bland, P.H., Brown, M.S., Engelmann, R.M., Laderach, G.E., Max, D., Pais, R.C., Qing, D.P.Y., Roberts, R.Y., Smith, A.R., Starkey, A., Batrah, P., Caligiuri, P., Farooqi, A., Gladish, G.W., Jude, C.M., Munden, R.F., Petkovska, I., Quint, L.E., Schwartz, L.H., Sundaram, B., Dodd, L.E., Fenimore, C., Gur, D., Petrick, N., Freymann, J., Kirby, J., Hughes, B., Casteele, A.V., Gupte, S., Sallamm, M., Heath, M.D., Kuhn, M.H., Dharaiya, E., Burns, R., Fryd, D.S., Salganicoff, M., Anand, V., Shreter, U., Vastagh, S., Croft, B.Y., 2011. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. Medical Physics 38, 915–931. doi:10.1118/1.3528204.

Armato, S.G., Roberts, R.Y., Kocherginsky, M., Aberle, D.R., Kazerooni, E.A., Macmahon, H., van Beek, E.J.R., Yankelevitz, D., McLennan, G., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., P.Caligiuri, Quint, L.E., Sundaram, B., Croft, B.Y., Clarke, L.P., 2009. Assessment of radiologist performance in the detection of lung nodules: dependence on the definition of "truth". Academic Radiology 16, 28–38.

Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., Kazerooni, E.A., MacMahon, H., van Beek, E.J., Yankelevitz, D., Biancardi, A.M., Bland, P.H., Brown, M.S., Engelmann, R.M., Laderach, G.E., Max, D., Pais, R.C., Qing, D.P., Roberts, R.Y., Smith, A.R., Starkey, A., Batra, P., Caligiuri, P., Farooqi, A., Gladish, G.W., Jude, C.M., Munden, R.F., Petkovska, I., Quint, L.E., Schwartz, L.H., Sundaram, B., Dodd, L.E., Fenimore, C., Gur, D., Petrick, N., Freymann, J., Kirby, J., Hughes, B., Casteele, A.V., Gupte, S., Sallam, M., Heath, M.D., Kuhn, M.H., Dharaiya, E., Burns, R., Fryd, D.S., Salganicoff, M., Anand, V., Shreter, U., Vastagh, S., Croft, B.Y., Clarke, L.P., 2015. Data from LIDC-IDRI. doi:10.7937/K9/TCIA.2015.LO9QL9SX.

Bastien, F., Lamblin, P., Pascanu, R., Bergstra, J., Goodfellow, I., Bergeron, A., Bouchard, N., Warde-Farley, D., Bengio, Y., 2012. Theano: new features and speed improvements. arXiv preprint .

Bergtholdt, M., Wiemker, R., Klinder, T., 2016. Pulmonary nodule detection using a cascaded SVM classifier, in: Medical Imaging, International Society for Optics and Photonics. pp. 978513–978513. doi:10.1117/12.2216747.

Blum, A., Hardt, M., 2015. The ladder: A reliable leaderboard for machine learning competitions. arXiv preprint arXiv:1502.04585 .

Brown, M.S., Lo, P., Goldin, J.G., Barnoy, E., Kim, G.H.J., McNitt-Gray, M.F., Aberle, D.R., 2014. Toward clinically usable CAD for lung cancer screening with computed tomography. European Radiology doi:10.1007/s00330-014-3329-0.

Camarlinghi, N., Gori, I., Retico, A., Bellotti, R., Bosco, P., Cerello, P., Gargano, G., Torres, E.L., Megna, R., Peccarisi, M., Fantacci, M.E., 2011. Combination of computer-aided detection algorithms for automatic lung nodule identification. International Journal of Computer Assisted Radiology and Surgery doi:10.1007/s11548-011-0637-6.

Cascio, D., Magro, R., Fauci, F., Iacomi, M., Raso, G., 2012. Automatic detection of lung nodules in CT datasets based on stable 3D mass-spring models. Computers in Biology and Medicine doi:10.1016/j.compbiomed.2012.09.002.

Cerello, P., Cheran, S.C., Bagnasco, S., Bellotti, R., Bolanos, L., Catanzariti, E., De Nunzio, G., Fantacci, M.E., Fiorina, E., Gargano, G., et al., 2010. 3-d object segmentation using ant colonies. Pattern Recognition 43, 1476–1490. doi:10.1016/j.patcog.2009.10.007.

Chialvo, D.R., Millonas, M.M., 1995. How swarms build cognitive maps, in: The biology and technology of intelligent autonomous agents. Springer, pp. 439–450. doi:10.1.1.44.3015.

Choi, W.J., Choi, T.S., 2013. Automated pulmonary nodule detection system in computed tomography images: A hierarchical block classification approach. Entropy 15, 507–523. doi:10.3390/e15020507.

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., Tarbox, L., Prior, F., 2013. The cancer imaging archive (TCIA): Maintaining and operating a public information repository. Journal of Digital Imaging 26, 1045–1057. doi:10.1007/s10278-013-9622-7.

Dieleman, S., Schlüter, J., Raffel, C., Olson, E., Sønderby, S.K., Nouri, D., Maturana, D., Thoma, M., Battenberg, E., Kelly, J., et al., 2015. Lasagne: First release. Zenodo: Geneva, Switzerland doi:10.5281/zenodo.27878.

Dietterich, T.G., 2000. Ensemble methods in machine learning, in: Multiple Classifier Systems. Springer Science Heidelberg, pp. 1–15. doi:10.1007/3-540-45014-9_1.

Dou, Q., Chen, H., Yu, L., Qin, J., Heng, P.A., 2016. Multi-level contextual 3D CNNs for false positive reduction in pulmonary nodule detection. IEEE Transactions on Biomedical Engineering , 1–1doi:10.1109/tbme.2016.2613502.

Efron, B., Tibshirani, R.J., 1994. An introduction to the bootstrap. volume 57. CRC press.

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542, 115–118.

Firmino, M., Morais, A.H., Mendona, R.M., Dantas, M.R., Hekis, H.R., Valentim, R.A., 2014. Computer-aided detection system for lung cancer in computed tomography scans: Review and future prospects. Biomedical Engineering Online 13, 41. doi:10.1186/1475-925X-13-41.

van Ginneken, B., Armato, S.G., de Hoop, B., van de Vorst, S., Duindam, T., Niemeijer, M., Murphy, K., Schilham, A.M.R., Retico, A., Fantacci, M.E., Camarlinghi, N., Bagagli, F., Gori, I., Hara, T., Fujita, H., Gargano, G., Belloti, R., Carlo, F.D., Megna, R., Tangaro, S., Bolanos, L., Cerello, P., Cheran, S.C., Torres, E.L., Prokop, M., 2010. Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: the ANODE09 study. Medical Image Analysis 14, 707–722. doi:10.1016/j.media.2010.05.005.

van Ginneken, B., Setio, A.A.A., Jacobs, C., Ciompi, F., 2015. Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans, in: IEEE International Symposium on Biomedical Imaging, pp. 286–289. doi:10.1109/ISBI.2015.7163869.

Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks., in: Aistats, pp. 249–256. doi:10.1.1.207.2059.

Golosio, B., Masala, G.L., Piccioli, A., Oliva, P., Carpinelli, M., Cataldo, R., Cerello, P., De Carlo, F., Falaschi, F., Fantacci, M.E., Gargano, G., Kasae, P., Torsello, M., 2009. A novel multithreshold method for nodule detection in lung CT. Medical Physics 36, 3607–3618. doi:10.1118/1.3160107.

Guo, W., Li, Q., 2012. High performance lung nodule detection schemes in CT using local and global information. Medical Physics 39, 5157–5168. doi:10.1118/1.4737109.

He, K., Zhang, X., Ren, S., Sun, J., 2015a. Deep residual learning for image recognition. arXiv:1512.03385 .

He, K., Zhang, X., Ren, S., Sun, J., 2015b. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 1026–1034. doi:10.1.1.725.4861.

Henschke, C.I., Yankelevitz, D.F., Mirtcheva, R., McGuinness, G., McCauley, D., Miettinen, O.S., 2002. CT screening for lung cancer: Frequency and significance of part-solid and nonsolid nodules. American Journal of Roentgenology 178, 1053–1057. doi:10.2214/ajr.178.5.1781053.

Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R., 2012. Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint .

International Commission on Radiation Units and Measurements, 2008. Receiver operating characteristic analysis in medical imaging. Journal of the ICRU 8, 1–62.

Jacobs, C., van Rikxoort, E.M., Twellmann, T., Scholten, E.T., de Jong, P.A., Kuhnigk, J.M., Oudkerk, M., de Koning, H.J., Prokop, M., Schaefer-Prokop, C., van Ginneken, B., 2014. Automatic detection of subsolid pulmonary nodules in thoracic computed tomography images. Medical Image Analysis 18, 374–384. doi:10.1016/j.media.2013.12.001.

Jacobs, C., van Rikxoort, E.M., Murphy, K., Prokop, M., Schaefer-Prokop, C.M., van Ginneken, B., 2016. Computer-aided detection of pulmonary nodules: a comparative study using the public LIDC/IDRI database. European Radiology doi:10.1007/s00330-015-4030-7.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L., 2014. Large-scale video classification with convolutional neural networks, in: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, IEEE. pp. 1725–1732. doi:10.1109/CVPR.2014.223.

Kazerooni, E.A., Austin, J.H., Black, W.C., Dyer, D.S., Hazelton, T.R., Leung, A.N., McNitt-Gray, M.F., Munden, R.F., Pipavath, S., 2014. ACR–STR practice parameter for the performance and reporting of lung cancer screening thoracic computed tomography (CT). Journal of Thoracic Imaging 29, 310–316. doi:10.1097/rti.0000000000000097.

Kuhnigk, J.M., Dicken, V., Bornemann, L., Bakai, A., Wormanns, D., Krass, S., Peitgen, H.O., 2006. Morphological segmentation and partial volume analysis for volumetry of solid pulmonary lesions in thoracic CT scans. IEEE Transactions on Medical Imaging 25, 417–434.

Li, Q., Arimura, H., Doi, K., 2004. Selective enhancement filters for lung nodules, intracranial aneurysms, and breast microcalcifications, in: International Congress Series, Elsevier. pp. 929–934.

doi:10.1016/j.ics.2004.03.372.

Li, Q., Sone, S., Doi, K., 2003. Selective enhancement filters for nodules, vessels, and airway walls in two- and three-dimensional CT scans. Medical Physics 30, 2040–2051. doi:10.1118/1.1581411.

Manos, D., Seely, J.M., Taylor, J., Borgaonkar, J., Roberts, H.C., Mayo, J.R., 2014. The lung reporting and data system (lu-rads): a proposal for computed tomography screening. Canadian Association of Radiologists Journal 65, 121–134. doi:10.1016/j.carj.2014.03.004.

Messay, T., Hardie, R.C., Rogers, S.K., 2010. A new computationally efficient CAD system for pulmonary nodule detection in CT imagery. Medical Image Analysis 14, 390–406. doi:10.1016/j.media.2010.02.004.

Murphy, K., van Ginneken, B., Schilham, A.M.R., de Hoop, B.J., Gietema, H.A., Prokop, M., 2009. A large scale evaluation of automatic pulmonary nodule detection in chest CT using local image features and k-nearest-neighbour classification. Medical Image Analysis 13, 757–770. doi:10.1016/j.media.2009.07.001.

Naidich, D.P., Bankier, A.A., MacMahon, H., Schaefer-Prokop, C.M., Pistolesi, M., Goo, J.M., Macchiarini, P., Crapo, J.D., Herold, C.J., Austin, J.H., Travis, W.D., 2013. Recommendations for the management of subsolid pulmonary nodules detected at CT: a statement from the fleischner society. Radiology 266, 304–317. doi:10.1148/radiol.12120628.

Niemeijer, M., Loog, M., Abràmoff, M.D., Viergever, M.A., Prokop, M., van Ginneken, B., 2011. On combining computer-aided detection systems. IEEE Transactions on Medical Imaging 30, 215–223. doi:10.1109/TMI.2010.2072789.

Prasoon, A., Petersen, K., Igel, C., Lauze, F., Dam, E., Nielsen, M., 2013. Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013, Springer. pp. 246–253. doi:10.1007/978-3-642-40763-5_31.

Retico, A., Delogu, P., Fantacci, M.E., Gori, I., Martinez, A.P., 2008. Lung nodule detection in low-dose and thin-slice computed tomography. Computers in Biology and Medicine 38, 525–534. doi:10.1016/j.compbiomed.2008.02.001.

Retico, A., Fantacci, M.E., Gori, I., Kasae, P., Golosio, B., Piccioli, A., Cerello, P., De Nunzio, G., Tangaro, S., 2009. Pleural nodule identification in low-dose and thin-slice lung computed tomography. Computers in Biology and Medicine 39, 1137–1144. doi:10.1016/j.compbiomed.2009.10.005.

Riccardi, A., Petkov, T.S., Ferri, G., Masotti, M., Campanini, R., 2011. Computer-aided detection of lung nodules via 3D fast radial transform, scale space representation, and Zernike MIP classification. Medical Physics 38, 1962–1971. doi:10.1118/1.3560427.

van Rikxoort, E.M., de Hoop, B., Viergever, M.A., Prokop, M., van Ginneken, B., 2009. Automatic lung segmentation from thoracic computed tomography scans using a hybrid approach with error detection. Medical Physics 36, 2934–2947. doi:10.1118/1.3147146.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. arXiv:150504597 doi:10.1007/978-3-319-24574-4_28.

Salden, A.H., Florack, L., ter Haar Romeny, B., 1991. Differential geometric description of 3d scalar images. 3D Computer Vision, Utrecht, The Netherlands, Technical Report , 91–23.

Setio, A.A.A., Ciompi, F., Litjens, G., Gerke, P., Jacobs, C., van Riel, S., Wille, M.W., Naqibullah, M., Sanchez, C., van Ginneken, B., 2016. Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks. IEEE Transactions on Medical Imaging doi:10.1109/TMI.2016.2536809.

Setio, A.A.A., Jacobs, C., Gelderblom, J., van Ginneken, B., 2015. Automatic detection of large pulmonary solid nodules in thoracic CT images. Medical Physics 42, 5642–5653. doi:10.1118/1.4929562.

Sutskever, I., Martens, J., Dahl, G., Hinton, G., 2013. On the importance of initialization and momentum in deep learning, in: Proceedings of the 30th international conference on machine learning (ICML-13), pp. 1139–1147.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2014. Going deeper with convolutions. arXiv:14094842v1 .

Tan, M., Deklerck, R., Cornelis, J., Jansen, B., 2013. Phased searching with neat in a time-scaled framework: Experiments on a computer-aided detection system for lung nodules. Artificial Intelligence in Medicine doi:10.1016/j.artmed.2013.07.002.

Tan, M., Deklerck, R., Jansen, B., Bister, M., Cornelis, J., 2011. A novel computer-aided lung nodule detection system for CT images. Medical Physics 38, 5630–5645. doi:10.1118/1.3633941.

Teramoto, A., Fujita, H., 2013. Fast lung nodule detection in chest CT images using cylindrical nodule-enhancement filter. International Journal of Computer Assisted Radiology and Surgery , 1–13doi:10.1007/s11548-012-0767-5.

Tieleman, T., Hinton, G., 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning 4.

Torres, E.L., Fiorina, E., Pennazio, F., Peroni, C., Saletta, M., Camarlinghi, N., Fantacci, M.E., Cerello, P., 2015. Large scale validation of the M5L lung CAD on heterogeneous CT datasets. Medical Physics 42, 1477–1489. doi:10.1118/1.4907970.

Wen, Y., Zhang, K., Li, Z., Qiao, Y., 2016. A discriminative feature learning approach for deep face recognition, in: European Conference on Computer Vision, Springer. pp. 499–515.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R.S., Bengio, Y., 2015. Show, attend and tell: Neural image caption generation with visual attention. arXiv preprint 2, 5.

Zagoruyko, S., Komodakis, N., 2016. Wide residual networks. arXiv preprint arXiv:arXiv:1605.07146.