

Pulmonary nodule detection in CT images using multi-view convolutional networks

Arnaud A. A. Setio, Francesco Ciompi, Geert Litjens, Paul Gerke, Colin Jacobs, Clara I. Sánchez, Bram van Ginneken

Abstract—Automatic detection of pulmonary nodules in Computed Tomography (CT) has been an important research topic for over a decade. The majority of existing Computer-Aided Detection (CAD) systems require hand-crafted features and ad-hoc classification frameworks to achieve a high detection performance. We propose a novel CAD system for pulmonary nodules using multi-view convolutional networks (ConvNets), for which discriminative features are automatically learnt from the training data. The network is fed with nodule candidates obtained by combining three candidate detectors specifically designed for solid, subsolid, and large nodules. For each candidate, a set of 2-D patches from differently oriented planes is extracted. The proposed architecture comprises multiple streams of 2-D ConvNets, for which the outputs are combined using a dedicated fusion method to get the final classification. Data augmentation and dropout are applied to reduce overfitting. The proposed CAD system is evaluated on 888 scans from the publicly available LIDC-IDRI dataset. Compared with other methods applied on a large number of LIDC-IDRI scans, our method reaches higher detection sensitivity of 85.4% and 90.1% at 1 and 4 false positives per scan, respectively. We performed an additional evaluation on an independent dataset from the ANODE09 challenge and confirmed that our algorithm outperforms previously proposed methods that rely on manually defined features.

Index Terms—pulmonary nodule, computer-aided detection, lung cancer, deep learning, computed tomography, convolutional networks.

I. INTRODUCTION

LUNG cancer is the leading cause of cancer death worldwide [1]. The seminal National Lung Screening Trial [2] showed a reduction of 20% in lung cancer mortality in high-risk subjects scanned with low-dose Computed Tomography (CT), compared to the control group that received chest radiography. As a consequence of this result, lung cancer screening programs with low-dose CT imaging are being implemented in the US. Currently, only 15% of all diagnosed lung cancers are detected at an early stage, which causes a five-year survival rate of only 16%. The aim of screening is to detect cancers in an earlier stage when curative treatment options are better.

Arnaud A. A. Setio is with the Diagnostic Image Analysis Group at the Department of Radiology and Nuclear Medicine, Radboud University Medical Center, Nijmegen, the Netherlands, e-mail: Arnaud.ArindraAdiyoso@radboudumc.nl.

Francesco Ciompi, Geert Litjens, Paul Gerke, Colin Jacobs, and Clara I. Sánchez are with the Diagnostic Image Analysis Group at the Department of Radiology and Nuclear Medicine, Radboud University Medical Center, Nijmegen, the Netherlands.

Bram van Ginneken is with the Diagnostic Image Analysis Group at the Department of Radiology and Nuclear Medicine, Radboud University Medical Center, Nijmegen, the Netherlands, and Fraunhofer MEVIS, Bremen, Germany.

The implementation of screening would mean a significant increase of reading effort for radiologists. Computer-Aided Detection (CAD) systems have been developed to assist radiologists in the reading process and thereby potentially making lung cancer screening more cost-effective [3–5]. The architecture of a CAD system typically consists of two stages: 1) nodule candidates detection and 2) false positive reduction. The aim of the first step is to detect nodule candidates at a very high sensitivity, which typically implies the presence of many false positives. Techniques such as double thresholding and morphological operations are often used to detect a large set of candidates [4, 5]. False positives are subsequently reduced in a second stage. Typically, a large set of dedicated features is extracted and a supervised classification scheme is used [3–5].

Although it has been shown that CAD systems improve the reading efficiency of radiologists, a considerable number of nodules remains undetected at low false positive rates, prohibiting the use of CAD in clinical practice [6, 7]. Figure 1 illustrates that nodules come with a wide variation in shapes, sizes, and types (e.g. solid, subsolid, calcified, pleural, etc.). In addition, the number of nodules from different categories are highly imbalanced and many irregular lesions that are visible in CT are not nodules. As a consequence, extracting underlying characteristics of nodules is difficult and requires many heuristic steps. Techniques to detect lesions with a broad spectrum of appearances are needed to improve the performance of CAD systems.

In the last years, spurred by the large amount of available data and computational power of modern-day computers, convolutional networks (ConvNets) [8, 9] have been shown to outperform the state-of-the-art in several computer vision applications [10–13]. ConvNets have also been introduced in the field of medical image analysis [14–18]. Because ConvNets can be trained end-to-end in a supervised fashion while learning highly discriminative features, removing the need for handcrafting nodule descriptors, they are well suited to be used for the false positive reduction step of a pulmonary nodule CAD system. To the best of our knowledge, the work of Lo et al. [19] is the only study which used ConvNets specifically trained for pulmonary nodule detection, and was solely applied to chest radiography images.

Although ConvNets have been shown to outperform other supervised learning methods, only few studies extended the use of conventional 2-D ConvNets to the analysis of volumetric 3-D images [14, 17, 20]. In all these studies, volumetric candidates are firstly decomposed into fixed triplanar views (sagittal, coronal, and axial planes). Thereafter, each plane is

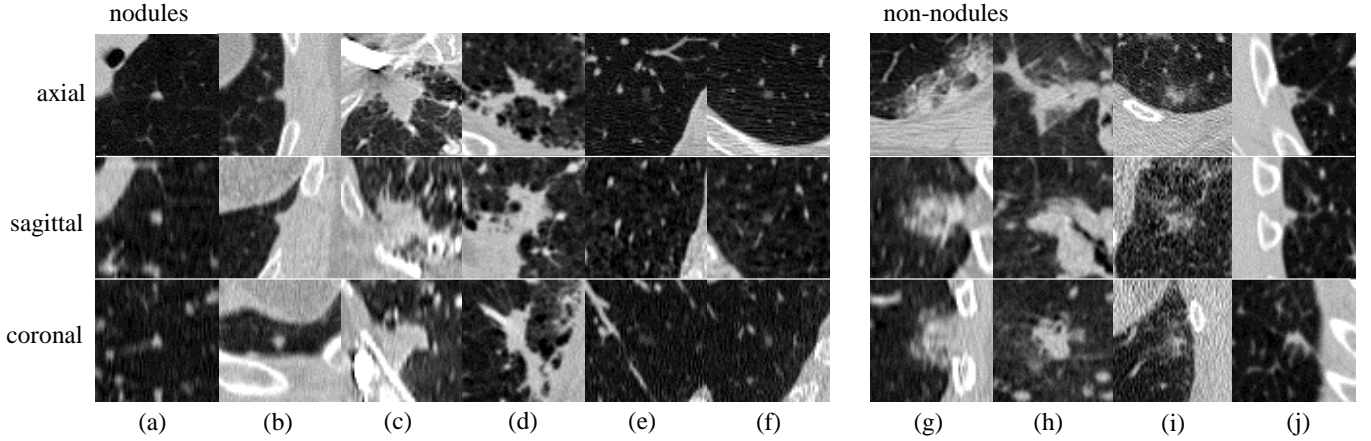


Fig. 1: Examples of lesions (nodules and non-nodules) in sagittal, coronal, and axial view. Lesions are located in the center of the box (50×50 mm). The left set of images are nodules with a wide range of morphological characteristic: (a) solid nodule, (b) pleural nodule, (c-d) large nodules with irregular shape, (e-f) subsolid nodules. The right set of images are irregular lesions that are not related with nodules or cancers. These examples illustrate that designing features for accurate detection and classification of nodules may not be trivial.

processed using a multi-view architecture, for which streams of 2-D ConvNets are applied to all patches and output units are combined using data fusion technique, such as late-fusion [14], committee-fusion [10, 20], or the combination of both fusion methods [17]. Although all of these fusion methods show promising performance gain, how different methods compare with each other remains an open question.

In this study, we present an automatic pulmonary nodules detection system using ConvNets specifically trained for the task at hand. The contributions of this paper are as follows: (1) We formulate the candidate detection as a combination of three detectors specifically designed to detect solid, subsolid and large solid nodules. In this way, we cover a large spectrum of nodule types that can be found in chest CT, thereby boosting the sensitivity of the candidate detection step. (2) We formulate a novel false positive reduction step using multi-view ConvNets for pulmonary nodule detection. Each candidate in a 3-D scan is described by multiple 2-D patches. Each 2-D patch is processed in a separate stream of ConvNets and is combined to obtain the final prediction. The hypothesis is that a combination of predictions from multiple views provides 3-D information of a nodule candidate in a CT scan. (3) We provide a performance benchmark on how different fusion methods influence the detection performance for the task at hand. We also evaluate the impact of adding more views on the performance of each architecture and include an external validation on a completely independent dataset.

II. MATERIALS

A. LIDC-IDRI

We trained and validated the proposed CAD system using the large publicly available dataset, Lung Image Database Consortium (LIDC-IDRI) [21]. LIDC-IDRI contains a heterogeneous sets of 1,018 cases from seven institutions. The slice thickness of CT images varies from 0.6 mm to 5.0 mm with a median of 2.0 mm. The reference standard is set by

manual annotations from four radiologists who reviewed each scan in two reading rounds. In the first blinded reading round, suspicious lesions were independently annotated and each of them was categorized as non-nodule, nodule < 3 mm, or nodule ≥ 3 mm. Manual 3-D segmentation was performed only for lesions categorized as nodules ≥ 3 mm. In the second reading round, annotations from all four radiologists were reviewed in an unblinded fashion and each radiologist decided to either accept or reject each annotation.

In our experiments, we excluded thick-slice scans (> 2.5 mm), as these are not recommended anymore [22, 23], and scans with inconsistent slice spacing, obtaining 888 scans. We considered only annotations categorized as nodule ≥ 3 mm. Nodules < 3 mm are not considered relevant according to current screening protocols [2, 24]. As nodules could be annotated by multiple readers, we merged annotations that are distant less than the sum of their radii. For these merged annotations, the diameters, volumes, and coordinates were averaged. We selected nodules ≥ 3 mm accepted by the majority of radiologists (3 or 4 out of 4 radiologists) as reference standard. For this study, this resulted in a set of 1,186 nodules. Any non-nodule, nodule < 3 mm, or nodules ≥ 3 mm accepted by the minority was not counted as false positive, because marks by a CAD system on such locations are not necessarily undesirable.

B. ANODE09

In order to further validate the performance of the proposed system on a dataset completely independent from the training set, we used data from the ANODE09 challenge [6]. The ANODE09 dataset consists of 55 CT scans. Each scan was annotated by two observers in a blinded fashion. Five scans were provided as training cases, while the remaining 50 cases were provided as testing cases. The reference standard for testing cases is not publicly available.

All cases were collected from the University Medical Center

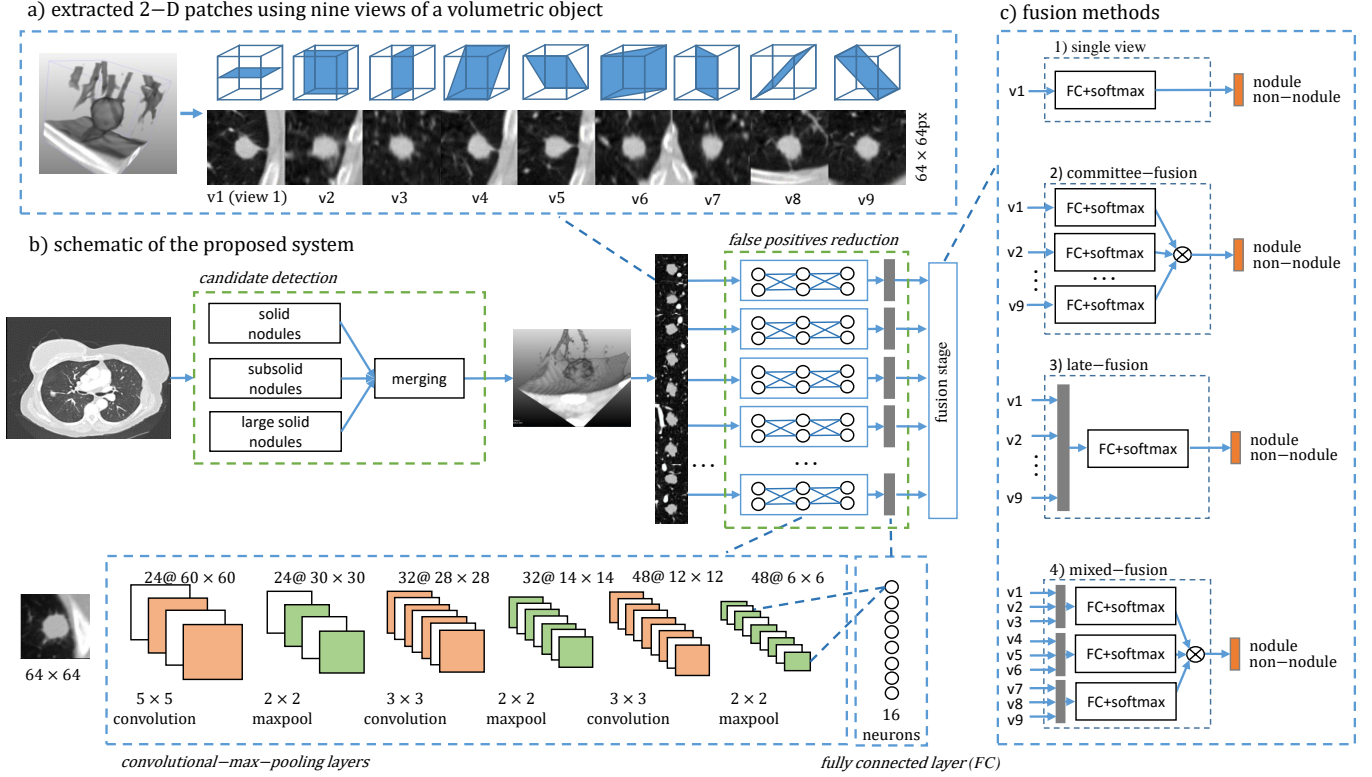


Fig. 2: An overview of the proposed CAD system. (a) An example of extracted 2-D patches from nine symmetrical planes of a cube. The candidate is located at the center of the patch with a bounding box of 50×50 mm and 64×64 px. (b) Candidates are detected by merging the outputs of detectors specifically designed for solid, subsolid and large nodules. The false positive reduction stage is implemented as a combination of multiple ConvNets. Each of the ConvNets stream processes 2-D patches extracted from a specific view. (c) Different methods for fusing the output of each ConvNet stream. Grey and orange boxes represent concatenated neurons from the first fully connected layers and the nodule classification output. Neurons are combined using fully connected layers with softmax or a fixed combiner (product-rule).

Utrecht and originated from the largest CT lung cancer screening trial in Europe, the NELSON study [24]. The images were reconstructed at 1.0 mm thickness. A web-based framework for objective evaluation of nodule detection algorithms is available¹, where the results of CAD systems can be uploaded for benchmarking.

III. METHODS

The architecture of the proposed CAD system is schematized in Figure 2. Two main stages are incorporated: 1) candidates detection and 2) false positives reduction. We applied three candidates detectors specifically designed for solid, subsolid, and large solid nodules. The combination of these detectors is applied to increase the detection sensitivity of nodules. Note that nodules have a large variations in both size and morphological characteristics. For each candidate, we extract multiple 2-D views in fixed planes. Each 2-D view is then processed by one ConvNets stream. The ConvNets features are then fused to compute a final score. In the next sections we describe the CAD system in details.

A. Candidates detection

Candidate detection algorithms play an important role on the performance of any CAD system, as it determines the maximum detection sensitivity of subsequent stages. Candidate detection algorithms should ideally detect all suspicious lesions. However, the morphological variation of nodules is often greater than what a single candidate detection algorithm can detect.

To detect a wider spectrum of nodules, we applied a combination of multiple candidate detection algorithms. Three existing CAD systems are used to detect nodule candidates [3, 5, 25]. Each algorithm aims at a specific type of nodules, namely solid nodules, subsolid nodules, and large solid nodules. For each candidate, the position $\vec{p} = (x, y, z)$ and the nodule probability are given. Three sets of nodule candidates are computed and are merged in order to maximize the sensitivity of the detector. The candidates located closer than 5 mm to each others are merged. For these combined candidates, the position \vec{p} and nodule probability are averaged [26].

The methods for candidate detection stage, for which the locations of volume of interest (VOI) are obtained, are described in the following paragraphs.

For *solid* nodules, we implemented the technique proposed by Murphy et al. [3]. For each voxel in the lungs, shape index

¹<http://anode09.grand-challenge.org/>

and curvedness are computed, and thresholding is applied on the two measures to define the seedpoints. An automatic segmentation method is executed at the seedpoints to obtain clusters of interest. Subsequently, clusters located close to each other are merged. Finally, we discard clusters with a volume $< 40 \text{ mm}^3$.

For *subsolid* nodules, we implemented the technique proposed by Jacobs et al. [5]. A double-threshold density mask (-750, -350 Hounsfield Unit (HU)) is first performed to obtain a mask with voxels of interest. Morphological opening is applied to remove connected clusters, followed by 3D connected component analysis. Clusters for which the centers of mass are within 5 mm are merged. Finally, an accurate segmentation of the candidates is obtained by using a previously published nodule segmentation algorithm [27].

Large solid nodules ($\geq 10 \text{ mm}$) have surface/shape index values that are locally different from smaller solid lesions and have a specific intensity range that is not captured by both solid and subsolid nodules detection algorithms [25]. Therefore, the two aforementioned algorithms do not perform well in detecting large solid nodules. In addition, large solid nodules attached to the pleural wall may be excluded by lung segmentation algorithms since the contrast with the pleura is low. For these reasons, we implemented a third detector that consists of three steps: (1) post-processing of lung segmentation by applying a rolling-ball algorithm to the segmentation mask, which includes large nodules attached to the pleura in the lung segmentation; (2) density thresholding (-300 HU), to obtain a mask with voxels of interest; (3) morphological opening in a multi-stage fashion to get candidate clusters, where we start with large structuring elements to extract larger nodules, and progressively continue with smaller structuring elements to extract smaller nodules.

One issue with training an algorithm using highly unbalanced data is that the learned parameters may be skewed toward characteristics of the most common candidates (e.g. vessels) while overlooking important characteristics of rarer nodules. To prevent overfitting on highly prevalent false positives, we discarded candidates with a low likelihood for being nodules. The probability threshold is empirically set to reduce a large number of false positives while maintaining high detection sensitivity.

B. Patches extraction

For each candidate, we extracted multiple 2-D patches of $50 \times 50 \text{ mm}$ centered on \vec{p} . The size of the patch was chosen in order to have all nodules ($\leq 30 \text{ mm}$) fully visible on the 2-D views and include sufficient context information to aid in the classification of the candidate. We resized each $50 \times 50 \text{ mm}$ patch to a size of $64 \times 64 \text{ px}$, working at the resolution of 0.78 mm , which corresponds to the typical resolution of thin slice CT data. The pixel intensity range is rescaled from (-1000, 400 HU) to (0, 1). Intensity outside the given range is clipped.

In order to extract patches, we first consider a cube of $50 \times 50 \times 50 \text{ mm}$, which encloses the candidate. Nine patches are extracted on planes corresponding to the plane of symmetry in a cube. Similar to [14, 17, 20], three planes of symmetry

that are parallel to a pair of faces of the cube are used. These planes are commonly known as sagittal, coronal, and axial planes. The other six planes are the planes of symmetry that cut two opposite faces of cubes in diagonals. Such a plane contains two opposite edges of the cube and four vertices. Examples of extracted patches are shown in Figure 2a.

C. 2-D ConvNets configuration

To obtain discriminative patterns from a patch, 2-D ConvNets are used. A pilot-study on smaller dataset was performed to determine the number of layers and the size of the kernels. Several configurations were pre-determined and the best configuration was selected. The used 2-D ConvNets consist of 3 consecutive convolutional layers and max-pooling layers (see Fig. 2b). The input of the network is a 64×64 patch. The first convolutional layer consists of 24 kernels of size $5 \times 5 \times 1$. The second convolutional layer consists of 32 kernels of size $3 \times 3 \times 24$. The third convolutional layer consists of 48 kernels of size $3 \times 3 \times 32$. The max-pooling layer is given by the maximum values in non-overlapping windows of size 2×2 (stride of 2). Rectified linear units (ReLU) [11] are used in the convolutional layers, where the activation a for a given input x is obtained as $a = \max(0, x)$. The last layer is a fully connected layer with 16 output units.

D. ConvNets fusion

Previous studies proposed various methods on how different output of ConvNets can be combined [10, 14, 17, 20, 28]. However, none of the studies compared different fusion methods and the optimal fusion method given a specific task remains a question. Three approaches for fusing multiple 2-D ConvNets are investigated:

1) *committee-fusion*: One of the most commonly used fusion method is by applying a committee-based combiner to the output predictions of several ConvNets [10, 20]. The motivation is to divide the detection task of 3-D object into several simpler 2-D detection tasks. We connected the output of the fully connected layer of each stream to a classification layer that consists of an additional fully connected layer with softmax activation function. The output of the softmax activation is the probability of the input image being a nodule. Each stream of ConvNets is trained separately using patches from a specific view and the output predictions are combined using a product-rule on the output probabilities [20], as shown in Figure 2c.

2) *late-fusion*: The late-fusion method [14, 28] concatenates the outputs of the first fully connected layers and connects the concatenated outputs directly to the classification layer (see Figure 2c). With such method, the classification layer can learn the 3-D characteristics by comparing the outputs of multiple ConvNets. In this configuration, the parameters of the convolutional layers for different streams are shared.

3) *mixed-fusion*: Mixed-fusion is a combination of the previous two approaches. Similar to Roth et al. [17], multiple late-fused ConvNets are implemented using a fixed number of orthogonal planes. Taking an advantage of having more

views, the prediction of the system is improved by combining multiple late-fused ConvNets in a committee. We divide nine patches into three independent sets; each set contains three different patches. Although other methods can be used to compose these sets of patches (e.g. random sets of triplanar patches), we attempted to compare all fusion methods fairly by keeping the same input information for each configuration.

E. Training

We performed evaluation in 5-fold cross-validation. For each training, we used 3 folds for training, 1 fold for validation, and 1 fold for testing. One of the challenges of using ConvNets is to efficiently optimize the weights of ConvNets given the training dataset. RMSProp [29], a learning algorithm that adaptively adjust each weight by considering a moving average of its squared gradient, is used to optimize the model. We set the learning rate $\eta = 0.001$, $\rho = 0.9$ and $\epsilon = e^{-6}$. The cost is measured by using cross-entropy error and the weights are updated using mini-batches of 128 examples. Dropout [30] with a probability of 0.5 is implemented on the output of the first fully connected layer as regularization. Training is stopped when the accuracy on the validation dataset does not improve after 3 epochs. We initialized the weights using normalized initialization proposed by Glorot and Bengio [31]. The biases were initialised with zero.

F. Data augmentation

Optimization of ConvNets using an imbalanced dataset can mislead the learning algorithm to local optima, where the predictions are biased toward the most frequent samples and overfitting occurs. Data augmentation is applied to prevent overfitting by adding invariances to the existing dataset.

1) *Training data augmentation*: As the number of nodules is much smaller than the number of non-nodules, augmentation is only performed on nodules. This process is applied for training and validation purposes. We translated the position of the candidates by 1 mm in each axis and scaled the patches to 40, 45, 50, and 55 mm. The translation is set to 1 mm in order to keep the nodules (>3 mm) to be captured properly in the patch. We further balanced the dataset by randomly upsampling the candidates from the nodule class.

2) *Test-data augmentation*: Data augmentation on the testing dataset has been shown to improve the performance of ConvNets [11, 12]. It may also improve the robustness of the system as candidates are evaluated on many possible conditions, such as analyzing the input image at several scales. Test-data augmentation (TDA) is performed by rescaling the patches to 40, 45, 50, and 55 mm, for which each of them is processed by ConvNets-CAD independently. We obtained the final prediction by averaging predictions computed from the augmented data.

G. Evaluation

Two performance metrics were measured: 1) area under the ROC curve (AUC) and 2) Competition Performance Metric (CPM) [26], which measures the average sensitivity at seven

TABLE I: Detection sensitivity of candidate detection algorithms

Total number of CT scans: 888				
Total number of nodules: 1,186				
Candidate detection	Detected nodules	Sensitivity (%)	False Positives (FPs)	FPs per scan
Solid	1,016	85.7	292,413	329.3
Subsolid	428	36.1	255,027	287.2
Large solid	377	31.8	41,816	47.1
Combined set	1,120	94.4	543,160	611.7
Reduced set	1,106	93.3	239,041	269.2

TABLE II: Performance benchmark of ConvNets configurations on LIDC-IDRI dataset. The best score for each performance metric is marked in bold.

Configuration	Number of views	AUC	CPM
single-view	1	0.969	0.481
committee-fusion	3	0.981	0.696
	9	0.987	0.780
late-fusion	3	0.987	0.742
	9	0.993	0.827
mixed-fusion	3*3	0.996	0.824

operating points of the FROC curve: 1/8, 1/4, 1/2, 1, 2, 4, and 8 FPs/scan. AUC shows the performance of ConvNets on classifying candidates as nodules or non-nodules while CPM shows the performance of CAD at operating points that are likely used in practice. It has to be noted that a system with higher AUC score may not necessarily result in higher CPM. We also computed the 95% confidence interval and the p-value using bootstrapping with 1,000 bootstraps [32].

IV. EXPERIMENTAL RESULTS

A. Candidates detection

The performance of individual candidate detection algorithms, as well as the combined algorithm is shown in Table I. When considered separately, the three approaches for solid, subsolid and large candidate detection give sensitivity of 85.7%, 36.1% and 31.8%, respectively. After the three candidate detection algorithms are combined, a sensitivity of 94.4% (1,120/1,186) is achieved. This shows that the three approaches are complementary and that combination is a better baseline for the false positive reduction. The reduced set indicates the set of candidates after removing those given a low likelihood for being nodules. The threshold is empirically set to 2.48×10^{-7} , which maintains 1,106 nodules (93.25%) with 239,041 FPs (269.2 FPs/scan).

B. False positives reduction

The performance benchmark of different ConvNets configurations tested on the LIDC-IDRI dataset is summarized in Table II. Given a set of candidates, applying ConvNets for nodules/non-nodules classification task yields an area under

the ROC (AUC) score up to 0.996. An average sensitivity of 0.828 at seven operating points is achieved using a late-fusion approach. We found that adding test-data augmentation does not significantly improve the detection performance (p-value = 0.46), as shown in Figure 3b. Combined with candidate detection algorithm that detect 93.1% nodules at 269.2 FPs/scan, our proposed method achieves a sensitivity of 85.4% and 90.1% at 1 and 4 FPs/scan, respectively.

The impact of two important parameters is observed: 1) fusing model and 2) number of views. Table II shows that fully optimized fusing models (late-fusion) lead to a better detection performance with a CPM score of 0.828, in comparison with committee-fusion (CPM score of 0.780, p-value < 0.001) and mixed-fusion (CPM score of 0.823, p-value = 0.029). The performance of the proposed ConvNets-CAD system in terms of Free-response Receiver Operating Characteristic (FROC) curve is depicted in Figure 3a. We also show a consistent improvement of the performance of ConvNets when more views are considered in the architecture, as shown in Figure 3c.

C. Comparison with existing CAD

To put the proposed CAD in a broader context, we reported the performance of existing CAD systems that use the LIDC-IDRI dataset for development in Table IV. In comparison with CAD systems that have been evaluated with a large number of scans (≥ 360 scans), we show that ConvNets specifically trained for the task at hand outperforms existing CAD systems.

We also applied the proposed CAD system on scans from completely independent ANODE09 dataset. The predictions were submitted to the ANODE09 evaluation system and performance were evaluated. Two sets of candidates were used. The first set contains candidates obtained by combining the candidate detection approaches described in Section III-A. The second set contains candidates only from ISICAD [3], a subset of (solid) candidates that is used in the first set of candidates. The motivation is to evaluate if the ConvNets, which is the main contribution of our work, can achieve better performance in comparison with conventional feature extraction method, given the same candidates. We used the ConvNets with the late-fusion approach and the test-data augmentation to compute the nodule probability. Although we have shown that the usage of TDA does not significantly improve the performance in LIDC-IDRI dataset, we found that it substantially improves the detection performance when the ConvNets are applied to the independent dataset, ANODE09.

Table III shows the scores of the proposed ConvNets-CAD in comparison with other CAD systems in ANODE09. When only considering solid nodules candidates, the proposed ConvNets-CAD achieves the CPM score of 0.637, which outperforms other CAD systems. When TDA was not applied, a CPM of 0.492 was achieved using the same set of candidates. The scores of the other CAD systems are obtained either from ANODE09 website or from published articles if their scores are not available on the website [33]. The performance of ConvNets-CAD (solid) for different nodule categories in ANODE09 dataset is shown in Figure 4.

TABLE III: Performance benchmark of CAD systems on ANODE09 dataset. The performance of ConvNets-CAD using two different sets of candidates are included.

Method	Score (CPM)
ConvNets-CAD (solid candidates)	0.637
ISICAD [3]	0.632
M5L [33]	0.619
ConvNets-CAD (all candidates)	0.598
lungCAM [33]	0.564
FlyerScan [4]	0.552
Pisa team [34]	0.293
Philips [6]	0.231
FujitaLab [6]	0.212

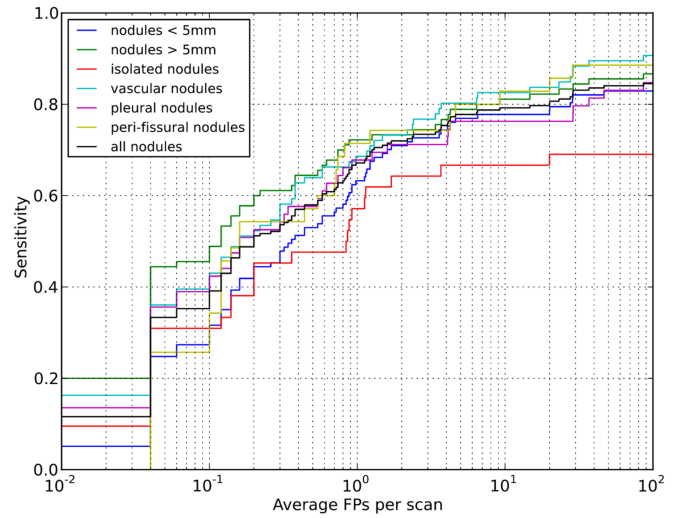


Fig. 4: FROC of proposed CAD for different nodule categories on ANODE09 dataset.

V. DISCUSSION

In this study, a novel pulmonary nodule detection CAD system using a multi-view convolutional network is proposed. Among CAD systems that are evaluated on a large part of the publicly available LIDC-IDRI dataset (≥ 360 scans), our proposed CAD system achieves the best performance, outperforming many conventional CAD systems (see Table IV). We also show that the proposed system is better than our previous CAD system that apply the off-the-shelf OverFeat network trained on million natural images of ImageNet dataset [20]. It suggests that training ConvNets specifically for the task at hand is crucial. The possibility of learning features from training dataset allows the network to learn classifying objects with a high degree of variation, which is suitable for the problem of pulmonary nodule detection.

We applied a combination of multiple candidate detection algorithms to localize suspicious lesions. Table I shows that combining multiple candidate detection algorithms boosts the detection sensitivity from 85.7% to 93.3% while maintaining a similar number of false positives. It is worth noting that subsolid and large nodules represent a small group of nodules.

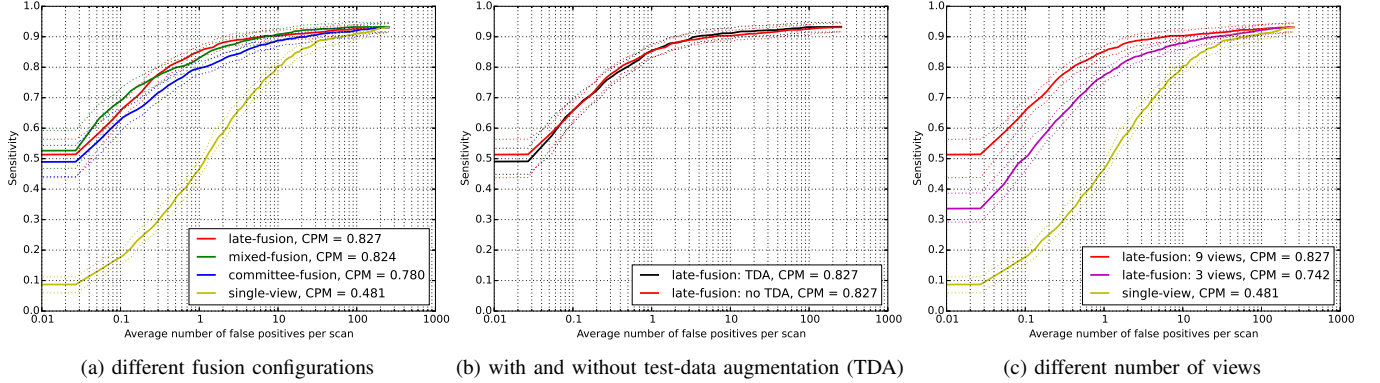


Fig. 3: FROC curves of ConvNets architectures with different configurations. Dashed curves show the 95% confidence interval estimated using bootstrapping.

However, they both add important subsets of nodules that are more likely to be cancerous. Although combining algorithms improves the sensitivity of the given candidate detectors, 6.7% of nodules remain undetected. A more general ConvNets framework that both localizes and classifies nodules directly from CT scans is planned as a future work.

The tuning of hyper-parameters that define the network architecture and the optimization procedure is crucial for optimal performance of ConvNets. However, the optimization of a large number of tunable parameters using a big dataset is impractical and remains a challenging research topic [35]. We conducted a pilot-experiment using a smaller dataset to optimize several hyper-parameters (i.e. number of layers, kernel size, learning rate, number of views, fusion method). Among these hyper-parameters, we identified two most critical parameters to tune, namely (1) the number of views and (2) the fusion method. These two parameters were further analyzed in experiments on the full LIDC-IDRI dataset. Other parameters were set to the best configuration found in pilot-experiments.

Figure 3c shows that incorporating more views in the architecture allows the network to achieve better performance. When all nine views are used, the FROC curve approaches the plateau at above 4 FPs/scan. This suggests that combining multiple views can be an effective approach for classifying 3-D objects, since simpler filters and fewer voxels are used compared to the isotropic 3-D volume ($64 \times 64 \times 64$ voxels). Following this trend, we expect that adding more views may slightly improve performance further. Experiments on different methods for fusing multiple 2-D ConvNets streams show that optimizing the combiner together with other parts of the network gives the best performance. This strategy allows the network to better learn the morphology of candidates from different perspectives, reducing errors caused by ambiguous information. As an example, vessels may be classified as nodules when the CAD system only processes one of its views. As a consequence, committee-fusion, which is commonly used in other works [10, 20], is sub-optimal for our architecture.

The evaluation on the ANODE09 dataset confirms that the proposed CAD system generalizes well on unseen data and performs accurately compared to other existing systems that

use hand-crafted features. When ConvNets are applied to a similar set of candidates as detected by the solid nodule detection algorithm ISICAD [3], a CPM score of 0.637 is achieved and is ranked first in ANODE09, outperforming ISICAD with a CPM score of 0.632. However, when candidates from the combined algorithm are used, the proposed system only achieves a CPM score of 0.598, outperformed by two systems: ISICAD [3] and M5L [33]. The reason for the deteriorated performance is the fact that the population of nodules on ANODE09 and that on LIDC-IDRI are different. ANODE09 dataset was randomly selected from the NELSON screening program to represent a screening scenario [6] while the LIDC-IDRI dataset was selected to capture the full spectrum of scans and nodules [21]. As a consequence, ANODE09 contains very few subsolid nodules and large nodules and additional candidates only contribute to more false positives. It is also worth noting that ISICAD [3], M5L [33], and lungCAM [33] were trained using a data set containing scans from the same data source of the ANODE09 study.

Examples of detected nodules, false positives, and false negatives are shown in Figure 5. Note that the system is able to detect nodules with a large variety of morphological characteristics. Figure 5b shows examples of false positives. We observed that a substantial number of false positives detected at 1 FP/scan are actually nodules (first and second column) that were missed by all four radiologists. This is a problem as all nodules are required to be detected for follow-up in screening scenario. Adding CAD systems in reading processes is expected to improve the annotation of lung nodules. Osteophytes (third column), which are important for quantification of spinal abnormalities, are also found as false positives. Other typical false positives include nodular-like structures, large vessels, mediastinal structures, and scarring. At 4 FPs/scan, most of undetected nodules are subsolid nodules or nodules with irregular shape, which are underrepresented in the training set. Further data balancing on nodule categories is expected to significantly improve the performance.

The ConvNets framework is implemented using Theano [42]. The computation time of ConvNets for a scan with on

TABLE IV: Summary of published CAD systems using LIDC-IDRI as dataset. Number of scans, reference standard criteria, and number of nodules used for validation are listed. Note that the LIDC-IDRI dataset has changed over-time, which partly explains why groups have used different subsets for their experiments. The reported performance at one or two operating points is provided.

CAD systems	Year	# scans	slice thickness	nodules size (mm)	agreement levels	# nodules	sensitivity (%) / FPs/scan	
Proposed system	-	888	≤ 2.5	≥ 3	at least 3	1,186	90.1 / 4.0	85.4 / 1.0
LungCAM (Torres <i>et al.</i> [33])	2015	949	NA	≥ 3	at least 2	1,749	80.0 / 8.0	
van Ginneken <i>et al.</i> [20])	2015	865	≤ 2.5	≥ 3	at least 3	1,147	76.0 / 4.0	73.0 / 1.0
Brown <i>et al.</i> [36]	2014	108	0.5-3	≥ 4	at least 3	68	75.0 / 2.0	
Choi and Choi[37]	2013	58	0.5-3	3-30	at least 1	151	95.3 / 2.3	
Tan <i>et al.</i> [38]	2013	360	NA	≥ 3	at least 4	NA	83.0 / 4.0	
Teramoto and Fujita [39]	2013	84	0.5-3	5-20	at least 1	103	80.0 / 4.2	
Cascio <i>et al.</i> [40]	2012	84	1.25-3	≥ 3	at least 1	148	97.0 / 6.1	88.0 / 2.5
Guo and Li [41]	2012	85	1.25-3	≥ 3	at least 3	111	80.0 / 7.4	75.0 / 2.8

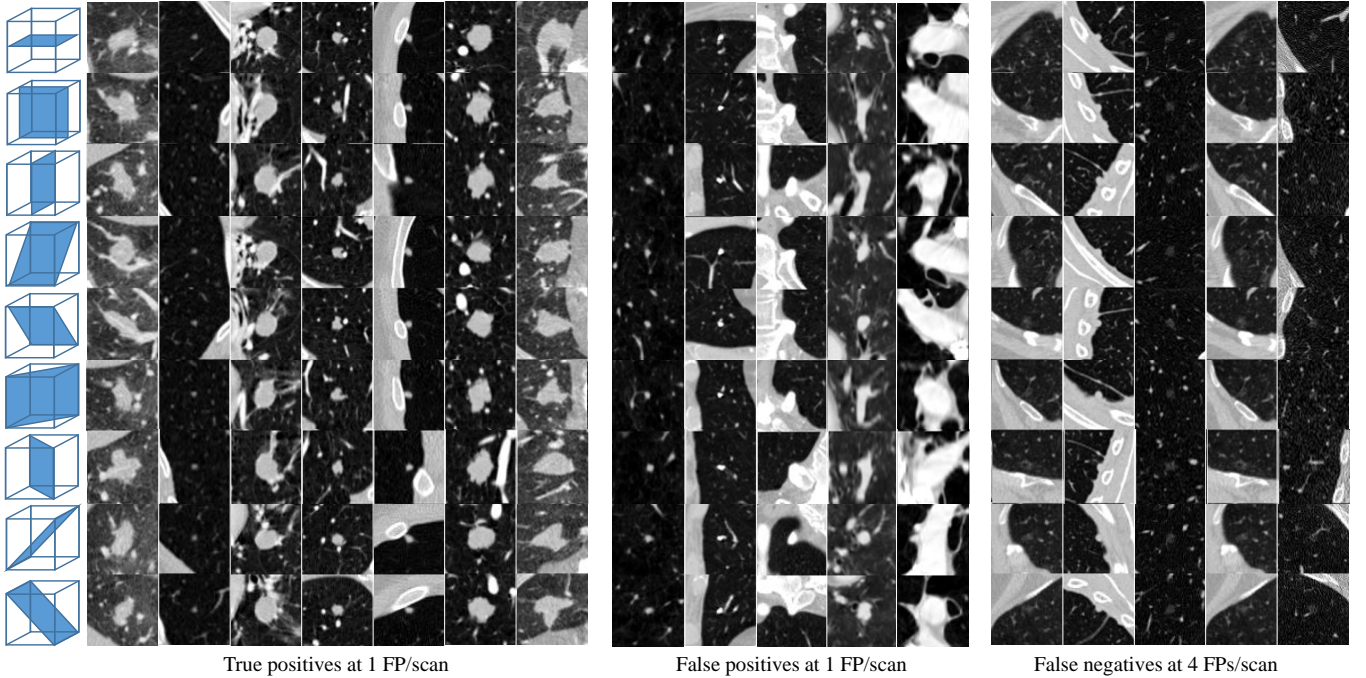


Fig. 5: Examples of lesions detected or missed by CAD system. Each column shows one lesion represented in patches viewed from different angles. The left set of lesions are nodules detected at 1 FP/scan. The middle set of lesions are false positives detected at 1 FP/scan. The right set of lesions are nodules missed at 4 FPs/scan. Most of the missed nodules are underrepresented in current dataset.

average 300 candidates per scan is 1 second on a standard PC with a GPU GeForce GTX570.

In the context of using the CAD system for lung cancer screening, the performance in terms of sensitivity should be improved. Several suggestions are proposed for future works. Information from 3-D input data could be exploited to train the ConvNets, even though this would increase the network complexity. Another interesting direction that might also improve performance is by adding features that could not be extracted from patches (e.g. context features).

VI. CONCLUSION

We have presented a CAD system for pulmonary nodule detection in CT scans based on multi-view convolutional networks. We have shown that the proposed ConvNets-CAD achieves good results for the nodule detection task and outperforms existing state-of-the-art CAD systems. Our findings suggest that designing manual features may no longer be compulsory to obtain an accurate pulmonary nodule CAD system. The promising results and the low computation time make the ConvNets-CAD highly suited to be used as a decision aid in a lung cancer screening scenario.

ACKNOWLEDGEMENT

This project was funded by a research grant from the Netherlands Organization for Scientific Research, project number 639.023.207.

REFERENCES

- [1] American Cancer Society. (2014) Cancer facts and figures 2014. [Online]. Available: <http://www.cancer.org/acs/groups/cid/documents/webcontent/003115-pdf.pdf>
- [2] D. R. Aberle, C. D. Berg, W. C. Black, T. R. Church, R. M. Fagerstrom, B. Galen, I. F. Gareen, C. Gatsonis, J. Goldin, J. K. Gohagan, B. Hillman, C. Jaffe, B. S. Kramer, D. Lynch, P. M. Marcus, M. Schnall, D. C. Sullivan, D. Sullivan, and C. J. Zylak, "The National Lung Screening Trial: overview and study design," *Radiology*, vol. 258, pp. 243–253, 2011.
- [3] K. Murphy, B. van Ginneken, A. M. R. Schilham, B. J. de Hoop, H. A. Gietema, and M. Prokop, "A large scale evaluation of automatic pulmonary nodule detection in chest CT using local image features and k-nearest-neighbour classification," *Medical Image Analysis*, vol. 13, pp. 757–770, 2009.
- [4] T. Messay, R. C. Hardie, and S. K. Rogers, "A new computationally efficient CAD system for pulmonary nodule detection in CT imagery," *Medical Image Analysis*, vol. 14, pp. 390–406, 2010.
- [5] C. Jacobs, E. M. van Rikxoort, T. Twellmann, E. T. Scholten, P. A. de Jong, J. M. Kuhnigk, M. Oudkerk, H. J. de Koning, M. Prokop, C. Schaefer-Prokop, and B. van Ginneken, "Automatic detection of subsolid pulmonary nodules in thoracic computed tomography images," *Medical Image Analysis*, vol. 18, pp. 374–384, 2014.
- [6] B. van Ginneken, S. G. Armato, B. de Hoop, S. van de Vorst, T. Duindam, M. Niemeijer, K. Murphy, A. M. R. Schilham, A. Retico, M. E. Fantacci, N. Camarlinghi, F. Bagagli, I. Gori, T. Hara, H. Fujita, G. Gargano, R. Bellotti, F. D. Carlo, R. Megna, S. Tangaro, L. Bolanos, P. Cerello, S. C. Cheran, E. L. Torres, and M. Prokop, "Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: the ANODE09 study," *Medical Image Analysis*, vol. 14, pp. 707–722, 2010.
- [7] M. Firmino, A. H. Morais, R. M. Mendona, M. R. Dantas, H. R. Hekis, and R. A. Valentim, "Computer-aided detection system for lung cancer in computed tomography scans: Review and future prospects," *Biomedical Engineering Online*, vol. 13, p. 41, 2014.
- [8] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, pp. 2278–2324, 1998.
- [9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.
- [10] D. Cireřan, U. Meier, J. Masci, and J. Schmidhuber, "Multi-column deep neural network for traffic sign classification," *Neural Networks*, vol. 32, pp. 333–338, 2012.
- [11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.
- [12] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *British Machine Vision Conference*, 2014.
- [13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:14091556*, 2014.
- [14] A. Prasoon, K. Petersen, C. Igel, F. Lauze, E. Dam, and M. Nielsen, "Deep feature learning for knee cartilage segmentation using a tri-planar convolutional neural network," in *Medical Image Computing and Computer-Assisted Intervention*, ser. Lecture Notes in Computer Science, vol. 8150, 2013, pp. 246–253.
- [15] A. A. Cruz-Roa, J. E. Arevalo Ovalle, A. Madabhushi, and F. A. González Osorio, "A deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection," in *Medical Image Computing and Computer-Assisted Intervention*, ser. Lecture Notes in Computer Science, vol. 8150, 2013, pp. 403–410.
- [16] R. Li, W. Zhang, H.-I. Suk, L. Wang, J. Li, D. Shen, and S. Ji, "Deep learning based imaging data completion for improved brain disease diagnosis," in *Medical Image Computing and Computer-Assisted Intervention*, ser. Lecture Notes in Computer Science, vol. 8675, 2014, pp. 305–312.
- [17] H. R. Roth, L. Lu, A. Seff, K. M. Cherry, J. Hoffman, S. Wang, J. Liu, E. Turkbey, and R. M. Summers, "A new 2.5D representation for lymph node detection using random sets of deep convolutional neural network observations," in *Medical Image Computing and Computer-Assisted Intervention*, ser. Lecture Notes in Computer Science, vol. 8673, 2014, pp. 520–527.
- [18] T. Brosch, Y. Yoo, D. K. B. Li, A. Traboulsee, and R. Tam, "Modeling the variability in brain morphology and lesion distribution in multiple sclerosis by deep learning," in *Medical Image Computing and Computer-Assisted Intervention*, ser. Lecture Notes in Computer Science, vol. 8674, 2014, pp. 462–469.
- [19] S.-C. Lo, H.-P. Chan, J.-S. Lin, H. Li, M. T. Freedman, and S. K. Mun, "Artificial convolution neural network for medical image pattern recognition," *Neural Networks*, vol. 8, pp. 1201–1214, 1995.
- [20] B. van Ginneken, A. A. A. Setio, C. Jacobs, and F. Ciampi, "Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans," in *IEEE International Symposium on Biomedical Imaging*, 2015, pp. 286–289.
- [21] S. G. Armato, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, E. A. Kazerooni, H. MacMahon, E. J. R. V. Beek, D. Yankelevitz, A. M. Biancardi, P. H. Bland, M. S. Brown, R. M. Engelmann, G. E. Laderach, D. Max, R. C. Pais, D. P. Y. Qing, R. Y. Roberts, A. R. Smith, A. Starkey, P. Batrah, P. Caligiuri, A. Farooqi, G. W. Gladish, C. M. Jude, R. F. Munden, I. Petkovska, L. E. Quint, L. H. Schwartz, B. Sundaram, L. E. Dodd, C. Fenimore, D. Gur, N. Petrick, J. Freymann, J. Kirby, B. Hughes, A. V. Castele, S. Gupte, M. Sallamm, M. D. Heath, M. H. Kuhn, E. Dharaiya, R. Burns, D. S. Fryd, M. Salganicoff, V. Anand, U. Shreter, S. Vastagh, and B. Y. Croft, "The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans," *Medical Physics*, vol. 38, pp. 915–931, 2011.
- [22] D. P. Naidich, A. A. Bankier, H. MacMahon, C. M. Schaefer-Prokop, M. Pistolesi, J. M. Goo, P. Macchiarini, J. D. Crapo, C. J. Herold, J. H. Austin, and W. D. Travis, "Recommendations for the management of subsolid pulmonary nodules detected at CT: a statement from the Fleischner society," *Radiology*, vol. 266, pp. 304–317, 2013.
- [23] D. Manos, J. M. Seely, J. Taylor, J. Borgaonkar, H. C. Roberts, and J. R. Mayo, "The lung reporting and data system (LU-RADS): a proposal for computed tomography screening," *Canadian Association of Radiologists Journal*, vol. 65, pp. 121–134, 2014.
- [24] D. M. Xu, H. Gietema, H. de Koning, R. Vernhout, K. Nackaerts, M. Prokop, C. Weenink, J. Lammers, H. Groen, M. Oudkerk, and R. van Klaveren, "Nodule management protocol of the NELSON randomised lung cancer screening trial," *Lung Cancer*, vol. 54, pp. 177–184, 2006.
- [25] A. A. A. Setio, C. Jacobs, J. Gelderblom, and B. van Ginneken, "Automatic detection of large pulmonary solid nodules in thoracic CT images," *Medical Physics*, vol. 42, no. 10, pp. 5642–5653, 2015.
- [26] M. Niemeijer, M. Loog, M. D. Abramoff, M. A. Viergever, M. Prokop, and B. van Ginneken, "On combining computer-aided detection systems," *IEEE Transactions on Medical Imaging*, vol. 30, pp. 215–223, 2011.
- [27] J. M. Kuhnigk, V. Dicken, L. Bornemann, A. Bakai, D. Wormanns, S. Krass, and H. O. Peitgen, "Morphological segmentation and partial volume analysis for volumetry of solid pulmonary lesions in thoracic CT scans," *IEEE Transactions on Medical Imaging*, vol. 25, pp. 417–434, 2006.
- [28] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Computer Vision and Pattern Recognition*. IEEE, 2014, pp. 1725–1732.
- [29] Y. N. Dauphin, H. de Vries, J. Chung, and Y. Bengio, "RMSProp and equilibrated adaptive learning rates for non-convex optimization," *CoRR*, vol. abs/1502.04390, 2015.
- [30] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.
- [31] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *International Conference on Artificial Intelligence and Statistics*, 2010, pp. 249–256.
- [32] B. Efron and R. J. Tibshirani, *An introduction to the bootstrap*. CRC press, 1994, vol. 57.
- [33] E. L. Torres, E. Fiorina, F. Pennazio, C. Peroni, M. Saletta, N. Camarlinghi, M. E. Fantacci, and P. Cerello, "Large scale validation of the M5L lung CAD on heterogeneous CT datasets," *Medical Physics*, vol. 42, pp. 1477–1489, 2015.
- [34] A. Retico, F. Bagagli, N. Camarlinghi, C. Carpentieri, M. E. Fantacci, and I. Gori, "A voxel-based neural approach (VBNA) to identify lung nodules in the ANODE09 study," in *Proceedings of the SPIE*, vol. 7260, 2014.

- 2009, pp. 72 601S–1–72 601S–8.
- [35] J. Bergstra and Y. Bengio, “Random search for hyper-parameter optimization,” *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 281–305, 2012.
 - [36] M. S. Brown, P. Lo, J. G. Goldin, E. Barnoy, G. H. J. Kim, M. F. McNitt-Gray, and D. R. Aberle, “Toward clinically usable CAD for lung cancer screening with computed tomography,” *European Radiology*, 2014.
 - [37] W.-J. Choi and T.-S. Choi, “Automated pulmonary nodule detection system in computed tomography images: A hierarchical block classification approach,” *Entropy*, vol. 15, pp. 507–523, 2013.
 - [38] M. Tan, R. Deklerck, J. Cornelis, and B. Jansen, “Phased searching with neat in a time-scaled framework: Experiments on a computer-aided detection system for lung nodules,” *Artificial Intelligence in Medicine*, 2013.
 - [39] A. Teramoto and H. Fujita, “Fast lung nodule detection in chest CT images using cylindrical nodule-enhancement filter,” *International Journal of Computer Assisted Radiology and Surgery*, pp. 1–13, 2013.
 - [40] D. Cascio, R. Magro, F. Fauci, M. Iacomi, and G. Raso, “Automatic detection of lung nodules in CT datasets based on stable 3D mass-spring models,” *Computers in Biology and Medicine*, 2012.
 - [41] W. Guo and Q. Li, “High performance lung nodule detection schemes in CT using local and global information,” *Medical Physics*, vol. 39, pp. 5157–5168, 2012.
 - [42] F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. J. Goodfellow, A. Bergeron, N. Bouchard, and Y. Bengio, “Theano: new features and speed improvements,” Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop, 2012.