Lung nodule detection from CT scans using 3D convolutional neural networks without candidate selection

Natalia M. Jenuwine[a], Sunny N. Mahesh[b], Jacob D. Furst[c], Daniela S. Raicu[c]
[a]University of Michigan, 500 S. State St., Ann Arbor, MI, USA 48109; [b]Arizona State University, Tempe, AZ, USA 85281; [c]College of Computing and Digital Media, DePaul University, 243 S. Wabash Ave., Chicago, IL, USA 60604

ABSTRACT

Early detection of lung nodules from CT scans is key to improving lung cancer treatment, but poses a significant challenge for radiologists due to the high throughput required of them. Computer-Aided Detection (CADe) systems aim to automatically detect these nodules with computer algorithms, thus improving diagnosis. These systems typically use a candidate selection step, which identifies all objects that resemble nodules, followed by a machine learning classifier which separates true nodules from false positives. We create a CADe system that uses a 3D convolutional neural network (CNN) to detect nodules in CT scans without a candidate selection step. Using data from the LIDC database, we train a 3D CNN to analyze subvolumes from anywhere within a CT scan and output the probability that each subvolume contains a nodule. Once trained, we apply our CNN to detect nodules from entire scans, by systematically dividing the scan into overlapping subvolumes which we input into the CNN to obtain the corresponding probabilities. By enabling our network to process an entire scan, we expect to streamline the detection process while maintaining its effectiveness. Our results imply that with continued training using an iterative training scheme, the one-step approach has the potential to be highly effective.

Keywords: Convolutional Neural Networks, 3D Convolutional Neural Networks, Lung Nodules, Detection, CT Scans, Computer-Aided Diagnosis

1. INTRODUCTION

Lung cancer is the leading cancer in mortality rates, taking approximately 1.1 million lives each year worldwide[1]. Radiologists can detect lung cancer by analyzing computed tomography (CT) scans to find lung nodules, small round tumors that have the potential to become malignant. Early detection of nodules allows for early treatment, significantly reducing the risk of mortality. However, due to the limited number of radiologists and the increasing number of patients, radiologists often have minimal time to analyze each scan. This makes it difficult for them to detect every nodule, leading to many missed diagnoses[2].

Computer-Aided Detection (CADe) systems serve as a powerful aid to radiologists. They aim to help radiologists work efficiently by immediately directing their attention to suspicious areas and providing a second opinion in uncertain cases. Our purpose is to build a CADe system that is more efficient than current options. Specifically, most CADe systems for nodule detection divide the task into two steps: selecting potential nodule candidates and classifying whether the candidates are nodules or not. We propose to streamline the process by bypassing candidate selection and approaching the task in a single step. We hypothesize that a CADe system that uses a 3D CNN to identify lung nodules in CT scans without a candidate detection step has the potential to be just as effective as existing systems that do use a candidate selection step, while eliminating needless complexity.

2. BACKGROUND

The classical approach to Computer-Aided Detection of lung nodules in CT scans divides the problem into two steps. First, candidates that resemble nodules are extracted from the entire scan; and second, these candidates are classified into true nodules and false positives using a machine learning model. In the earliest attempts, classifiers were based on hand-crafted features, typically corresponding to a priori knowledge about the properties of nodules. Treating each candidate as a discrete object, these methods extract a wide range of discriminative features, describing properties including shape[3,4], texture[3], intensity[4] and curvature[5]. They often obtain decent results, detecting as many as 92% of nodules at 10 FPs per scan[3]. However, these methods depend on the painstaking development and selection of relevant features, and only give the machine learning model access to quantified descriptions of the images rather than the images themselves.

Following the successful use of 2D convolutional neural networks (CNNs) for detecting and classifying objects found in natural images, several CADe systems[6,7] aimed to incorporate this method in place of a feature-based classifier. As CNNs are designed to directly interpret pixel data and take advantage of the spatial structure of images, they appeared to be well-suited for the task of analyzing CT images. However, they achieved limited success, and have not, to the best of our knowledge, been able to surpass the performance of feature-based approaches. This is likely due to their inability to capture context outside of a single 2-dimensional image. Some attempts have made to overcome this by combining predictions on adjacent or orthogonal slices[6], leading to slightly higher performance.

In recent years, 3D CNNs, which classify volumes of pixel data rather than 2-dimensional images, have seen extensive use for classification of nodule candidates in CT scans[8,9,10,11,12]. They have been experimentally shown to obtain superior results to comparable 2D CNNs[8], presumably due to their ability to fully analyze the 3D spatial structure present in CT scans. 3D CNN approaches have been highly successful compared to both 2D and feature-based attempts, obtaining sensitivities in the range of 95% at only 1 FP per scan[13]. However, these systems still typically make use of candidate selection, either by another neural network[9] or by a rules-based algorithm[8,10], in order to obtain inputs both for training and for applying their CNN classifier. As none of their candidate detection systems have a perfect sensitivity, the inputs to their classifiers are already missing some percentage of the nodules.

In this paper, we propose a new paradigm for detection, which aims to bypass candidate selection and approach the task in a single step using a 3D CNN. The CNN classifies subvolumes of a scan as either containing or not containing a nodule, without being trained on pre-selected candidates. It is then applied to detect nodules from entire CT scans by systematically sampling subvolumes from the scan and classifying them. To our knowledge, such an approach has been applied only once before with limited success, obtaining around 70% sensitivity at 10 FPs per scan[11]. We propose to incorporate a more robust training scheme, as well as recent advances in machine learning methods and processing power, to show that such a system is viable.

3. METHODS

An overview of our methods is depicted in Figure 1. First we optimize and train a 3D CNN on a large dataset of positive and negative example subvolumes, using an iterative training scheme (1). To evaluate a full CT scan, our system divides the scan systematically into overlapping subvolumes (2), which are input to the CNN (3) to obtain predictions representing the probability that each subvolume contains a nodule (4). Various probability thresholds can be applied to obtain sets binary predictions (5), which we compare to ground truth locations of nodules to evaluate the success of our system (6).

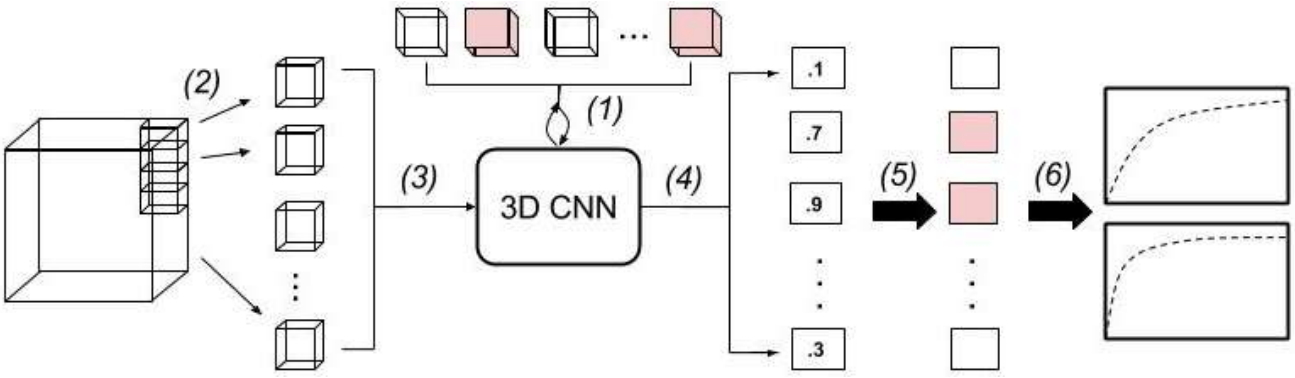

Figure 1. Overview of our pipeline for processing a CT scan.

3.1 Dataset

We trained and validated our system using data from the NIH/NCI Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) database, an open-source database intended as a benchmark for developing CAD systems[14,15,16]. It includes thoracic CT scans from 1018 patients, where each scan consists of 65 to 764 DICOM images.

Each LIDC scan has been annotated by up to four radiologists, each of whom marked the boundaries of any nodules found with a radius >3mm. The database provides coordinates of these boundaries, which we used to generate training data and to evaluate the accuracy of our system.

The slice thicknesses of the scans – that is, the width of the volume of tissue projected onto each CT image – vary widely, ranging from 0.5 to 5mm. This poses a significant challenge for 3D CNNs, as it is difficult to generalize patterns across scans that were generated differently. Since the American College of Radiology (ACR) recommends thin-slice CT scans for evaluating lung nodules[17], we excluded scans with slice thickness >2.5mm. Of the 666 remaining scans, 86 were set aside for the evaluation of our system.

Of the 2692 nodules for which boundary coordinates are provided, we excluded from our training data any nodule that is an outlier in its maximum x, y, or z dimension, reducing the dataset to 2302 nodules. We define outliers as values that fall more than 1.5 times the interquartile range above the third quartile, yielding upper limits of 31 pixels × 29 pixels × 13 slices. The equivalent lower limits were below zero and thus not applicable. Exclusion was based on pixel dimensions rather than true size dimensions because this allowed us to further narrow the size of our input subvolumes, allowing our CNN to focus on smaller, more typical nodules. We also excluded nodules found by only 1 or 2 radiologists, as there is too low a certainty that these are truly nodules. These steps left us with a dataset of 1106 nodules, of which 775 lie within our 666 training scans satisfying the (ACR) CT slice thickness criteria of greater than 2.5mm.

3.2 Input data

The inputs to our network must be sufficiently small to provide detections that are localized enough to be useful. At the same time, they must be large enough to a) fully contain the largest nodules, allowing the network to capture boundary and contextual information and b) analyze an entire scan, which consists of several hundred 512 × 512 images, within a reasonable timeframe. In accordance with the size of the largest nodules in our dataset, we chose an input size of 40 pixels × 40 pixels × 18 slices in order to allow the network to capture contextual information around these nodules.

We used the coordinates provided by the LIDC database to determine minimum, maximum, and central x, y, and z coordinates for each of our 775 nodules. We used these to construct subvolumes of the chosen input size centered on each nodule. We then augmented this data with randomly selected transformations of up to 10 pixels in the x and y planes and up to 4 slices in the z plane, as well as reflections and rotations of each nodule in all possible directions, increasing our dataset to 5425 subvolumes.

We generated examples of areas without nodules by randomly sampling subvolumes from our training scans until we had a balanced dataset. To ensure that our negative examples did not overlap any nodules, we excluded the areas containing all nodules, including outliers and those found by two or fewer radiologists, from the area that could be sampled. This method allowed us to quickly generate a huge number of negative training examples.

3.3 Convolutional neural network

We constructed a 3D CNN which takes in subvolumes of a set size and outputs the probability that each subvolume contains a nodule. Our network consists of convolutional, ReLU, and max-pooling layers which extract volumetric features, followed by fully connected layers and a softmax function which combine these features into a probabilistic classification. The kernels were initialized using Gaussian distributions scaled according to kernel size, as proposed by He et al[18]. We used categorical cross-entropy as our cost function, and the Adam optimization algorithm[19] to manage the adjustment of weights through gradient descent. The training of our network incorporated batch normalization[20], dropout of kernels and nodes[21], and early stopping of training[22] to avoid overfitting.

The architecture of our CNN was modeled on the VGG-16 network developed at the University of Oxford, winner of the 2014 ImageNet challenge[23]. In this architecture, the image is processed by two to three convolutional layers followed by a max pooling layer; this process is then repeated four additional times, with double the number of kernels each time. The output of the final max pooling layer is then processed by three fully connected layers to create the output. We created a scaled-down adaptation of this network, consisting of four convolutional layers alternated with two max pooling layers, followed by two fully connected layers (including the output layer), as displayed in Figure 2. Due to the small dimensions of our input subvolumes, we opted for convolutional kernels of size 3 × 3 × 3 with zero-padding, a max pooling field of 2 × 2 × 2, both applied with a stride of 1.

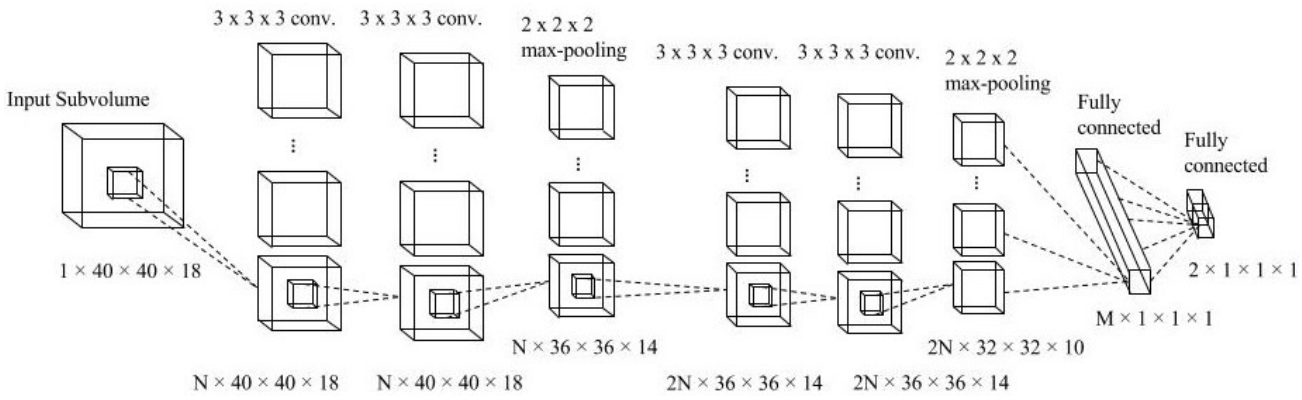

Figure 2. Architecture of our CNN. N is a hyperparameter representing the number of kernels in each convolutional layer; likewise, M represents the number of nodes in the first fully connected layer.

To select the remaining hyperparameters of our model - learning rate, dropout rate, batch size, number of kernels, and number of nodes in the fully connected layer - we performed a hyperparameter search. We used the Random Search optimization algorithm, which has been demonstrated to be more efficient than manual and grid-based search algorithms[24]. Our search consisted of 30 iterations, which according to binomial probability theory guarantees us a 95% chance of finding a result within a 10% interval of the true optimum. The performance of each set of hyperparameters was evaluated using 3-fold cross-validation on a subset of 1318 subvolumes from our training data.

3.4 Preprocessing experiments

We experimented with two different methods of preprocessing the pixel data to improve the performance of our CNN. The first of these is applying adaptive histogram equalization, a technique designed to enhance local contrast, to each subvolume. The second is intensity clipping, in which pixel values below or above a certain range are replaced with the minimum and maximum of that range. To select thresholds for clipping, we analyzed the frequencies of pixel values throughout the entire LIDC database to determine the upper and lower bounds for outliers. Based on this, we clipped to the range [-1434, 2446].

The two preprocessing methods yielded a total of four options - no preprocessing, histogram equalization, intensity clipping, and clipping followed by equalization. For each of these four options, we located the optimal hyperparameters as described above, trained the network, and then evaluated its performance on a validation dataset to determine the optimal method.

3.5 Network training

After determining the configuration of our network, we trained it on 85% of the balanced dataset of subvolumes described above, monitoring loss and accuracy on the remaining 15%. We ensured that the multiple augmented versions of a single nodule were not distributed between the testing and training sets, as this could artificially inflate the observed performance of the network. The network trained for 50 epochs, saving the weights at the points which obtained the lowest loss on the testing data.

To improve our network's performance when applied to full scans, we implemented an iterative training scheme. We segmented our training scans into overlapping subvolumes (as described below) and used the LIDC coordinates to determine which were positives, defined as any subvolume containing the center of a nodule, and which were negatives. We used all of the negatives from one scan, together with the complete set of positives, to further train the existing network, using class weights to compensate for our unbalanced dataset. We then applied this network to all of the negatives from the next two training scans, selected the false positives, and used those as the set of negative examples for the next iteration of training, allowing us to train selectively based on what was difficult for the network. At each iteration, 85% of the data was used for training and 15% for testing.

3.6 Application system

To apply our system to detect nodules from full scans, we use a "sliding box" approach, segmenting the scan into overlapping subvolumes and inputting each subvolume into our CNN. We chose a stride of 20 in the x and y directions and 9 in the z direction, as this is half the size of our inputs. This increases the likelihood that for any object within the scan, at least one subvolume will be roughly centered on it, and thus will be able to capture its entire shape as well as contextual information. For each subvolume it processes, the CNN outputs a probability. Various thresholds can be applied to these probabilities to obtain a set of detections.

We evaluated the success of our system by analyzing the sensitivities and false positive (FP) rates at various thresholds. To calculate sensitivity (the percent of nodules that were detected), we considered all nodules found by 3 or more radiologists, including outliers. If at least one subvolume overlapping a nodule received a probability above the chosen threshold, the nodule was considered detected. Detections corresponding to nodules indicated by one or two radiologists were considered neither true nor false positives. Any subvolume scoring above the threshold which does not overlap a nodule was counted as a false positive. Contrary to most existing systems, FPs were counted here in terms of subvolumes, not in terms of misclassified "objects." Due to the overlap in subvolumes, a single object, depending on its size, could be evaluated 8 to 64 times, leading to an inflated FP rate. For ease of comparison, we created an adjusted FP metric by using the average number of subvolumes which detected each true nodule as an estimate for the number of FP subvolumes per FP "object." For example, if each true nodule had been detected, on average, in 4 different subvolumes; and our network had detected a total of 100 FPs; then we would estimate that these 100 subvolumes correspond to 25 unique objects. Using the above calculations, we constructed Free Response Operating Characteristic (FROC) curves comparing sensitivities to both original and adjusted FPs rates per scan.

4. RESULTS

4.1 Preprocessing experiments

When we tested our four preprocessing methods, as described above, the resulting comparison indicated intensity clipping to be the optimal method, obtaining an AUC score of .722. Meanwhile, histogram equalization substantially diminished the performance of the CNN. The comparison of the four methods is displayed in Figure 3.

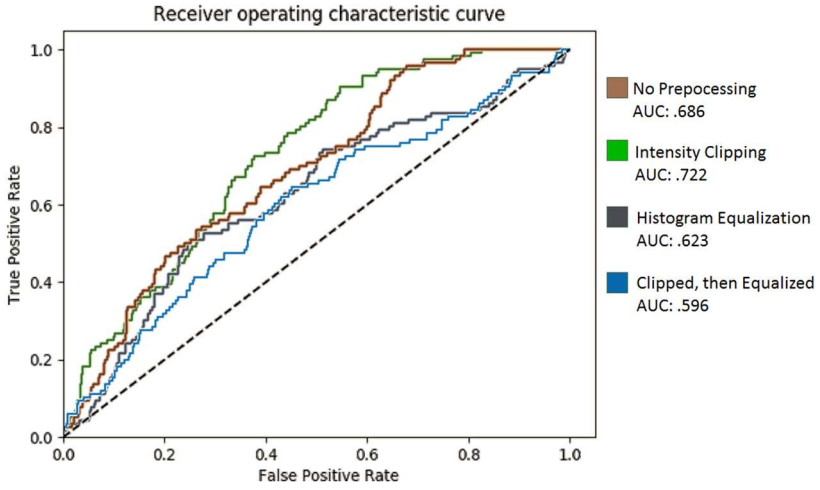

Figure 3. ROC curves representing performance on a balanced validation dataset of 232 subvolumes, and corresponding Area Under the Curve (AUC) scores, for each of four possible preprocessing methods.

Having selected intensity clipping as our method of preprocessing, and having obtained a high-performing set of hyperparameters for this method, we performed a second Random Search over a narrower parameter space in order to fine-tune the hyperparameters. This resulted in a learning rate of .00024, batch size of 60, dropout rates of .09 in the convolutional layers and .56 in the fully connected layer, 32 kernels in the first two convolutional layers, 64 kernels in the next two convolutional layers, and 64 nodes in the fully connected layer.

4.2 Network performance

Initial training yielded high sensitivities on our validation scans only at very high FP rates. However, the iterative training scheme created a significant improvement over the first few iterations. In the first iteration the training data included all FPs from one full scan, in the second it incorporated another two scans, and in the third it contained all FPs from an additional two scans, with all five of these scans selected randomly from those available for training. Figure 4 shows the network's performance over the course of this process, in terms of both original and adjusted FP rates.

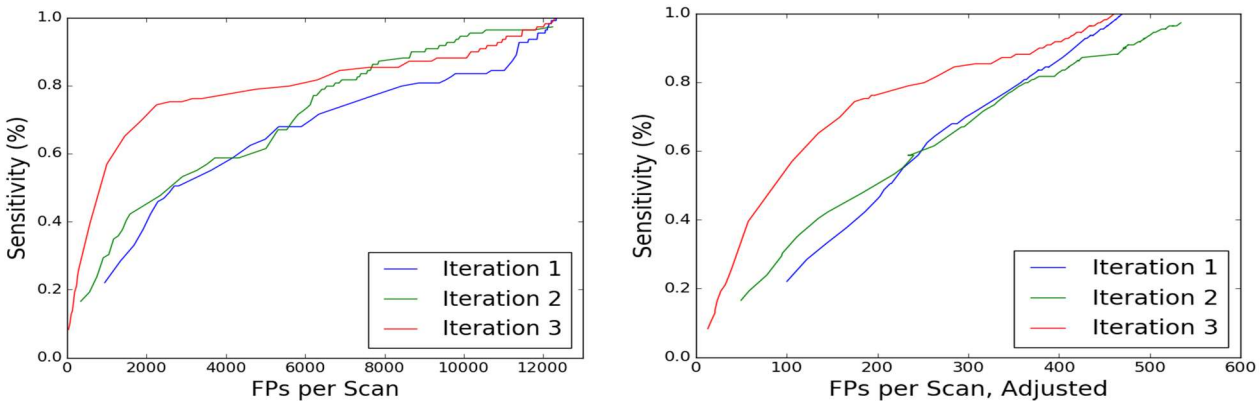

Figure 4. FROC curves representing performance of CNN on 86 validation CT scans over first 3 iterations of training

With each additional iteration, there is a visible increase in the area beneath the curve, indicating that the CNN overall obtains higher sensitivities at each FP rate. The improvement is especially noticeable between the second and third iterations. This only incorporates 5 training scans from among the hundreds available to us, serving as a proof of concept that continued iterative training could create much larger gains in performance.

5. CONCLUSION

We explored the possibility of a nodule detection system that does not require candidate selection for training or application. Our results indicate that with continued iterative training, a system that analyzes a CT scan in a single step using a 3D CNN has the potential to be highly effective. While initial performance was not accurate enough for the system to be usable, each iteration of training showed a clear improvement on the previous one, implying the possibility for much greater progress. As of now, the one-step approach to detection presents the possibility of being highly effective while bypassing unnecessary complexity, and warrants further investigation of its potential accuracy.

The performance of our system is not currently directly comparable to existing CADe systems presented in previous works, as disparities in network parameters, quantity and quality of training data, and other surrounding methods render direct comparisons across studies unreliable. Future work should seek to create equivalent systems, one with and one without candidate detection, allowing for a rigorous experimental comparison of the two approaches in terms of both complexity and accuracy.

Here we have presented a proof of concept showing that lung nodule detection from CT scans is possible without a candidate selection step. Further investigation is called for to expand upon this basis and work towards a reliable and accurate one-step detection system. Advanced training and evaluation techniques, such as dense sampling[8] and ensembling the predictions of several networks with variously sized inputs[12], have proved useful in constructing candidate-based systems. Incorporating such techniques into a one-step system could yield similar improvements in performance. Additionally, different strategies for interpreting, presenting, and combining the overlapping predictions that result from sliding-box evaluation should be explored. For example, converting a trained 3D CNN into a Fully Convolutional Network[9] has the potential to further streamline the detection process, while implementing a voting grid of adjacent inputs[11] may make the system's predictions more robust to error.

ACKNOWLEDGEMENTS

This work was possible in part by the National Science Foundation under the Research Experience for Undergraduates grant number 1659836. Results presented in this paper were obtained using the Chameleon testbed supported by the National Science Foundation grant number 1419141. The authors also acknowledge the NIH National Cancer Institute for its critical role in the creation of the free publicly available LIDC/IDRI Database used in this study.

REFERENCES

[1] International Agency for Research on Cancer, "Lung Cancer Estimated Incidence, Mortality and Prevalence Worldwide in 2012," GLOBOCAN 2012, 2015, http://globocan.iarc.fr/Pages/fact_sheets_cancer.aspx (1 August 2017).

[2] Fardanesh, M. and White, C., "Missed Lung Cancer on Chest Radiography and Computed Tomography," Seminars in Ultrasound, CT and MRI 33(4), 280-287 (2012).

[3] Talebpour, A.R., Hemmati, H. R. and Zarif Hosseinian, M., "Automatic lung nodules detection in computed tomography images using nodule filtering and neural networks," IEEE 22nd Iranian Conference on Electrical Engineering (ICEE), 1883-1887 (2014).

[4] Zhai, Z., Shi, D., Cheng, Y. and Guo, H., "Computer-Aided Detection of Lung Nodules with Fuzzy Min-Max Neural Network for False Positive Reduction," IEEE Sixth International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC) 66-69 (2014).

[5] Penedo, M.G., Carreira, M.J., Mosquera, A. and Cabello, D., "Computer-aided diagnosis: a neural-network-based approach to lung nodule detection," IEEE Transactions on Medical Imaging 17(6), 872-880 (1998).

[6] van Ginneken, B., Setio, A.A.A., Jacobs, C. and Ciompi, F., "Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans," IEEE 12th International Symposium on Biomedical Imaging (ISBI), 286-289 (2015).

[7] Jin, X., Zhang, Y. and Jin, Q., "Pulmonary Nodule Detection Based on CT Images Using CNN," IEEE 9th International Symposium Computational Intelligence and Design (ISCID), (2016).

[8] Huang X., Shan, J. and Vaidya, V., "Lung nodule detection in CT using 3D convolutional neural networks," IEEE 14th International Symposium on Biomedical Imaging (ISBI), 379-383 (2017).

[9] Hamidian, S., Sahiner, B., Petrick, N. and Pezeshk, A., "3D convolutional neural network for automatic detection of lung nodules in chest CT," Proceedings of SPIE Medical Imaging, 1013409 (2017).

[10] Anirudh, R., Thiagarajan, J.J., Bremer, T. and Kim, H., "Lung Nodule Detection Using 3D CNNs trained on weakly labeled data," Proc. SPIE 9785, 978532 (2016).

[11] Golan, R., Jacob, C. and Denzinger, J., "Lung nodule detection in CT images using deep convolutional neural networks," Proc. IEEE International Joint Conference on Neural Networks (IJCNN), 243-250 (2016).

[12] Dou, Q., Chen, H., Yu, L., Qin, J. and Heng, P., "Multilevel Contextual 3-D CNNs for False Positive Reduction in Pulmonary Nodule Detection," IEEE Transactions on Biomedical Engineering, 64(7), 1558 - 1567 (2016).

[13] Consortium for Open Medical Image Computing, "Results," Lung Nodule Analysis 2016, 14 August 2017, https://luna16.grand-challenge.org/results/ (15 August 2017).

[14] Armato, S.G. III, McLennan, G., Bidaut, L., McNitt-Gray, M.F., et al., "Data From LIDC-IDRI. The Cancer Imaging Archive," http://doi.org/10.7937/K9/TCIA.2015.LO9QL9SX (2015).

[15] Armato, S.G. III, McLennan, G., Bidaut, L., McNitt-Gray, M.F., et al., "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 (2011).

[16] Clark, K., Vendt, B., Smith, K., Freymann, J., et al., "The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository," Journal of Digital Imaging 26(6), 1045-1057 (2013).

[17] Setio, A.A.A., Traverso, A., de Bel, T., Berens, M.S.N., et al., "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge," Medical Image Analysis 42, 1-13 (2017).

[18] He, K., Zhang, X., Ren, S. and Sun, J., "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," http://arxiv.org/abs/1502.01852 (2015).

[19] Kingma, D. and Ba, J., "Adam: A Method for Stochastic Optimization," CoRR https://arxiv.org/abs/1412.6980 (2015).

[20] Ioffe, S. and Szegedy, C., "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," https://arxiv.org/abs/1502.03167 (2015).

[21] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R., "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," Journal of Machine Learning Research 15, 1929-1958 (2014).
[22] Prechelt, L., "Automatic early stopping using cross validation: quantifying the criteria," Neural Networks 11(4), 761-767 (1998).
[23] Simonyan, K. and Zisserman, A., "Very Deep Convolutional Networks for Large-Scale Image Recognition," CoRR https://arxiv.org/pdf/1409.1556.pdf (2015).
[24] Bergstra, J. and Bengio, Y., "Random Search for Hyper-Parameter Optimization," Journal of Machine Learning Research 13, 281-305 (2012).