3D DEEP LEARNING FOR DETECTING PULMONARY NODULES IN CT SCANS

Ross Gruetzemacher
Doctoral Student
Department of Systems & Technology
Raymond J. Harbert College of Business
Auburn University

Ashish Gupta, PhD*
Harbert College Advisory Council Faculty Fellow
Associate Professor of Analytics
Department of Systems & Technology
Raymond J. Harbert College of Business
Auburn University
***Corresponding Author: ashish.gupta@auburn.edu, ph: (334) 844-6456**

David Paradice, PhD
Harbert Eminent Scholar and Department Chair
Department of Systems & Technology
Raymond J. Harbert College of Business
Auburn University

NOTE: This is a preprint, electronic version of an article published in the Journal of the American Medical Informatics Association:

Gruetzemacher, R., Gupta, A. and Paradice, D., 2018. 3D deep learning for detecting pulmonary nodules in CT scans. *Journal of the American Medical Informatics Association*, *25*(10), pp.1301-1310.

ABSTRACT

Objective: To demonstrate and test the validity of a novel deep learning based system for the automated detection of pulmonary nodules.

Materials and Methods: The proposed system uses two 3D deep learning models, one for each of the essential tasks of computer-aided nodule detection: candidate generation and false positive reduction. 888 scans from the LIDC-IDRI dataset were used for training and evaluation.

Results: Results for candidate generation on the test data indicated a detection rate of 94.77% with 30.39 false positives per scan while the test results for false positive reduction exhibited a sensitivity of 94.21% with 1.789 false positives per scan. The overall system detection rate on the test data was 89.29% with 1.789 false positives per scan.

Discussion: An extensive and rigorous validation was conducted that soundly demonstrated the performance of the proposed system as among the best existing systems. The system demonstrated a novel combination of 3D deep neural network architectures and was the first system to use deep learning for both candidate generation and false positive reduction that has been evaluated using a substantial test dataset. The results strongly support the ability of deep learning pulmonary nodule detection systems to generalize to unseen data. The source code and trained model weights have been made available.

Conclusion: A novel deep neural network based pulmonary nodule detection system is demonstrated and validated. Overall, the results indicate that the system is comparable with the best existing systems.

BACKGROUND

Lung cancer is the leading cause of cancer mortality throughout the world [1]. Regular screening of high-risk individuals using low-dose computed tomography (CT) has been shown to reduce mortality in lung cancer patients [2]. Errors in cancer diagnosis are the most costly and detrimental type of diagnosis errors [3] with type II errors being particularly common in lung cancer diagnosis [4]. Subject to errors in both reading and interpretation, lung cancer diagnosis is highly error prone, however, many errors from diagnosis based on CT imagery can be reduced with a second reader [5]. Due to the large number of individuals at high-risk of lung cancer, regular screening with or without the assistance from a second reader could impose significant workflow and workload challenges for radiologists and clinical staff. Instead, computer-aided detection (CAD) systems have the potential to aid radiologists in lung cancer screening by reducing reading times or acting as a second reader.

CAD systems for pulmonary nodule detection consist of two essential tasks: candidate generation (CG) and false positive reduction (FPR) [6]. Various methods have been proposed for CG including thresholding-based methods [7-9], shape-based methods [10 11], rule-based methods [12] and filtering-based methods [13]. The best results for this task have been demonstrated using a 2D multiscale enhancement filter [14-16]. Methods for FPR are also varied, and include methods similar to those used for CG [17-19]. Neural networks were first used for FPR of nodule candidates over 20 years ago [20 21], and have been accepted as a leading method for FPR in radiographs for over a decade [22 23].

In recent years, deep neural networks (DNNs) have been responsible for significant improvements in a variety of complex tasks. The potential of DNNs for such improvements was first demonstrated at the 2012 ImageNet general image classification competition by Krizhevsky *et al*. [24]. Krizhevsky's team used the first DNN in the competition's history, winning by such a large margin that each of the top ten teams in the following year's competition were also using DNNs. In the years since, deep learning has been applied to a variety of medical imaging tasks [25 26], many of which involve segmentation [27-

30]. For some medical imaging tasks deep learning methods have been demonstrated performing at or above the level of licensed experts [31 32]. DNNs have also been applied for FPR [33-37], and just recently for pulmonary nodule segmentation [38 39].

Since the 2012 ImageNet competition, substantial progress has been made developing increasingly complex DNNs that have surpassed human-level performance in general image classification tasks [40]. In 2014, Szegedy *et al*. demonstrated an accuracy of 93.33% using an extensive DNN comprised of 22 layers [41]. A key feature of this network, dubbed GoogLeNet, was the inclusion of new components called inception modules which enabled the use of multiple convolutions in each layer of the network. In 2015, He *et al.* demonstrated an architecture for deep residual learning using a DNN comprised of 152 layers to achieve an accuracy of 96.43% [42]. This architecture, dubbed a residual network, employed a new feature for learning residual functions with reference to layer inputs.

Recent studies on nodule detection that utilize machine learning approaches, such as neural networks [6 34 39] and support vector machines [12 15 43 44], have several deficiencies. For example, several prior studies [37 39 43 44] exclude high resolution scans, several studies [37 39] use 20 year old deep learning architectures [45], and several [34 37 39] exclude small nodules and non-nodules from system evaluation. To address these deficiencies, we present a complete CAD system that uses novel, 3D adaptations of recent deep learning research [46-48], achieves performance comparable to the best existing systems, and exceeds the performance of all previous deep learning systems.

MATERIALS AND METHODS

Dataset

The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI) compiled a repository of CT scans specifically for the development of CAD systems for lung cancer [49 50]. This data is made available by the National Cancer Institute's Cancer Imaging Archive under the Creative

Commons Attribution 3.0 Unported License [50]. The LIDC-IDRI repository is comprised of 1018 openly available CT scans collected from five participating institutions with each of the scans including annotations by four radiologists. Lesions identified by radiologists were labeled as either large nodule (≥ 3mm), small nodule (< 3mm) or non-nodule. 2669 lesions were marked as a large nodule by at least one radiologist with 928 lesions being marked as large nodule by all four radiologists.

Although a large number of CAD systems have been demonstrated [12 34 43 44 51-53], comparing their relative performance has been challenging due to the lack of an objective evaluation framework. The LUNA16 dataset [6] was created in part to address this issue. It is a collection of 888 thin slice CT scans (*i.e.* slice thickness ≤ 3mm) of consistent slice spacing from the LIDC-IDRI dataset [54]. These scans were randomly divided into 10 bins for cross-validation purposes, and these bins were used in this study to enable reproducibility and aid in objective comparative evaluation. All nodules annotated as large nodules by three or four radiologists were considered as positive examples, resulting in 1186 nodules. Nodules labeled as large nodules by either one or two radiologists were considered inconclusive and were not considered as false positives in the evaluation. 11,509 small nodules and 19,004 non-nodules were considered inconclusive and not evaluated as false positives. This is consistent with the LUNA16 treatment of large nodules, but in this study these 30,513 other items were retained in evaluation and considered as false positives. Such items are irrelevant in the context of detecting and diagnosing lung cancer.

Method

As depicted in Figure 1, pulmonary nodules can vary greatly in density, shape, size, etc. A typical pulmonary nodule exists as an isolated lesion[1] in the parenchyma of the lung. Pulmonary nodules are frequently found attached to the pleural wall or vascular structures. Such juxtapleural and juxtavascular

[1] Lesions in the lungs less than 3cm are nodules.

nodules, depicted in the first and second rows of Figure 1, respectively, can complicate automated nodule detection. Other nodules that can be challenging for CAD systems to detect are low-density nodules, as seen in the third row of Figure 1. Due to the variety and unique challenges, CAD systems have had difficulty achieving high performance for all varieties.

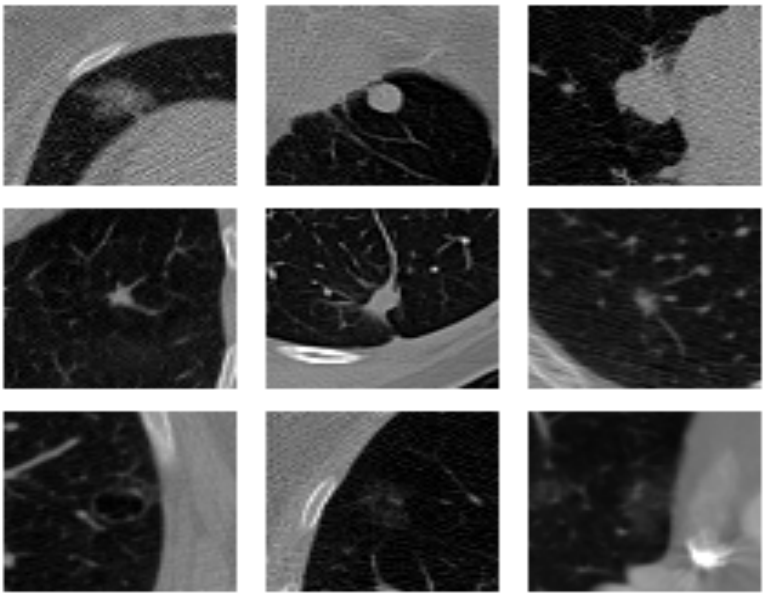

Figure 1. Different types of pulmonary nodules: juxtapleural (top row), juxtavascular (middle row) and low-density (bottom row).

Deep learning offers unique advantages over traditional methods for nodule detection. Rather than relying on mathematical techniques or hand-crafted features, deep learning enables learning internal feature representations directly from the input data [55 56]. The multiple processing layers in DNNs allow for learning these internal representations of the input data with multiple levels of abstraction [57]. These characteristics enable DNNs to learn feature representations for a large variety of features hierarchically, *i.e.* at different scales, and combine them for learning more complex features (*e.g.*

nodules). This capability is well suited for the unique challenges posed by nodule variety in pulmonary nodule detection.

We use DNNs for each of the two essential tasks of computer-aided pulmonary nodule detection. The first DNN is used for volume-to-volume prediction of nodules, *i.e.* segmentation, to identify potential pulmonary nodules within the input CT scan. These potential nodules that are generated are considered candidate nodules. This segmentation generates a large number of candidate nodules, many of which are false positives. The second DNN is used for reducing false positives among these nodule candidates through a binary classification of nodules and non-nodules. Figure 2 describes the system including the two primary tasks and the associated subtasks for each DNN.

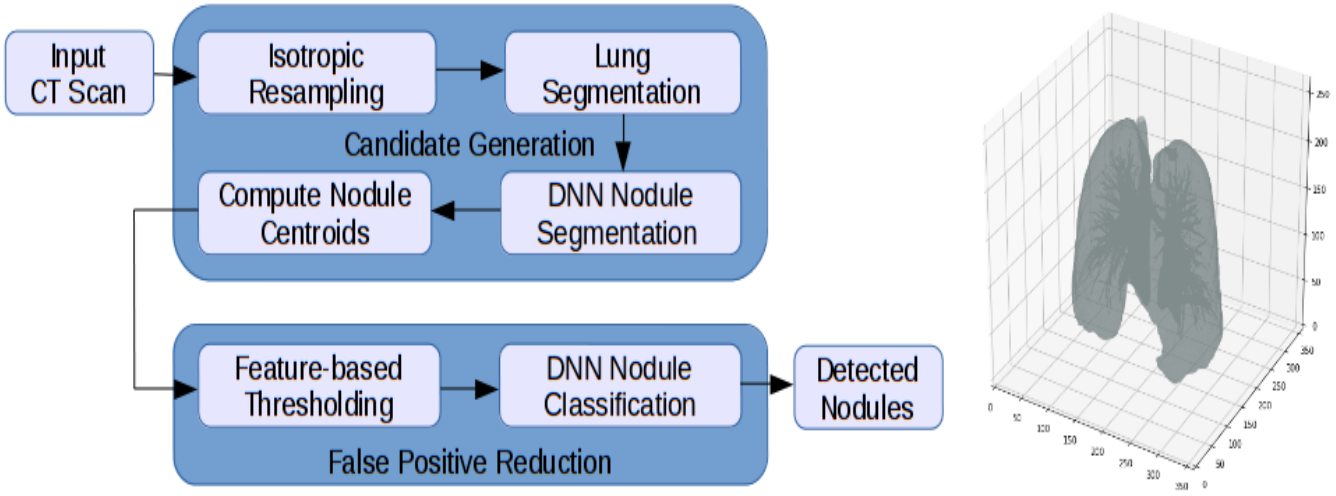

Figure 2. a) A process diagram of the proposed detection system. b) A segmented volume of the lungs.

The CT scans from the LUNA16 dataset range from 0.461mm to 0.977mm in pixel size and from 0.6mm to 3.0mm in slice thickness. Thus, the features depicted in the raw CT scans represent a scaled version of the actual features corresponding to the specifications of the imaging equipment used. To ensure uniformity of the data processed by the system, we begin with an isotropic resampling of the images,

using 3rd order spline interpolation to resize all voxels to a uniform size of $1mm^3$. As lung cancer screening is concerned solely with the lungs, a large portion of thoracic CT scans is irrelevant to lung cancer detection. To reduce computational expense, we next segment the lung region using a threshold-based segmentation technique[2] which isolates the lung and allows us to ignore the extraneous regions of the original scan. Figure 2b depicts the lungs following segmentation. A bounding box is created circumscribing the lung, and this region of interest is input into the CG DNN.

To date, only one study has demonstrated the use of DNNs for the segmentation of pulmonary nodules [38], and DNNs have not been previously used in CG. The U-Net architecture [58] is a DNN architecture used for generating high quality segmentations of medical imagery from small training datasets using heavy augmentation. The U-Net architecture has also been extended to 3D imaging applications successfully [46]. In this study, we adapt and apply this 3D U-Net architecture specifically for the segmentation of relatively small volumes in CT scans like pulmonary nodules. The architecture, depicted in Figure 3a, comprises 135.6 million trainable parameters.

The U-Net inspired DNN architecture that we use for CG acts as a volume-to-volume prediction (*i.e.* segmentation) network which takes an input volume of 64x64x64 and outputs a volume of 32x32x32. CT scans and the regions of interest generated from the lung segmentation are much larger than 64x64x64. Since the output is 32x32x32, a 16-unit buffer was created on all sides of the region of interest in order to generate an output volume corresponding with every voxel in the region of interest. A sliding window technique was used with a step size of 32 for each axis to iteratively progress through the region of interest, extracting 64x64x64 volumes for input to the DNN. This required the application of the DNN nearly a thousand times in generating candidate nodules for a single scan. The output volume was an array of probabilities. A threshold was applied setting all values less than 0.1[3] to zero. The remaining

[2] Detailed explanation of the lung segmentation technique can be found in Appendix B.

[3] We observed a large number of voxels with predicted values of magnitude lower than the maximum probability of 1.0, and to reduce noise we ignored all values that were more than an order of magnitude below 1.0.

probabilities were converted to a binary array, and labels were generated for the adjacent voxels, yielding an array of numbered volumes. Centroids were computed from these labeled volumes for evaluation and for generating output candidate nodules. During this process, a variety of features were also extracted from nodules including the volume in voxels of each nodule candidate.

.

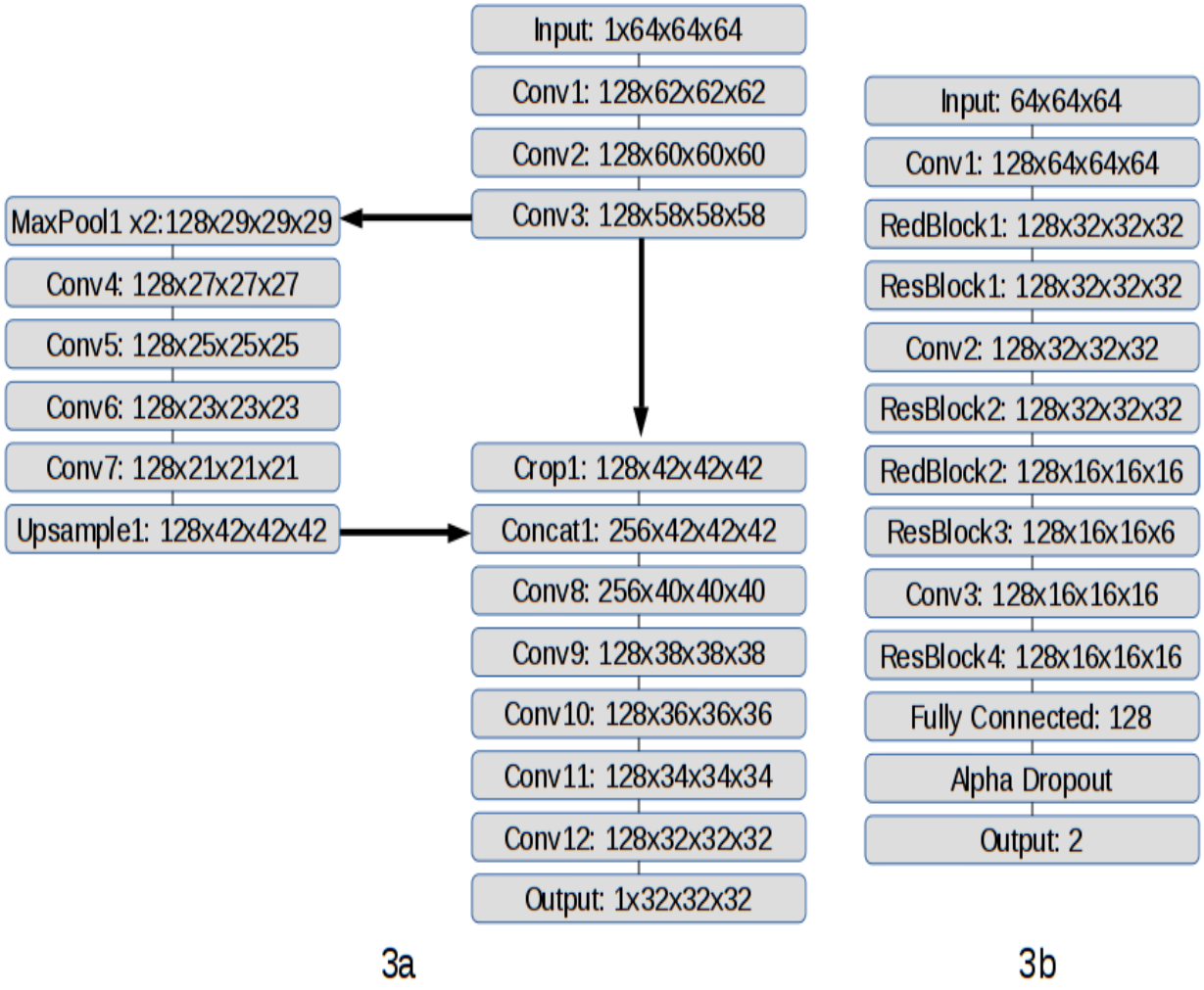

Figure 3[4]: a) The modified 3D U-Net architecture used for the CG segmentation. b) The 3D residual network architecture used for FPR.

[4] The notation for each layer in the architectures presented in Figure 3 indicates the type of layer and its shape. Each layer is represented by a 4D tensor for the 4 dimensions: the width of the network layer, and the x, y and z dimensions of the input volume.

The features extracted during the centroid computation for each of the candidate nodules were used as threshold as an initial FPR technique. The only feature selected for use in this step was nodule volume due to its strength in discriminating among positive and false positive nodules. The optimal threshold value was determined empirically to be $8mm^3$. Nodule candidates smaller in volume than $8mm^3$ were deemed false positives. Examples are shown in Figure 4.

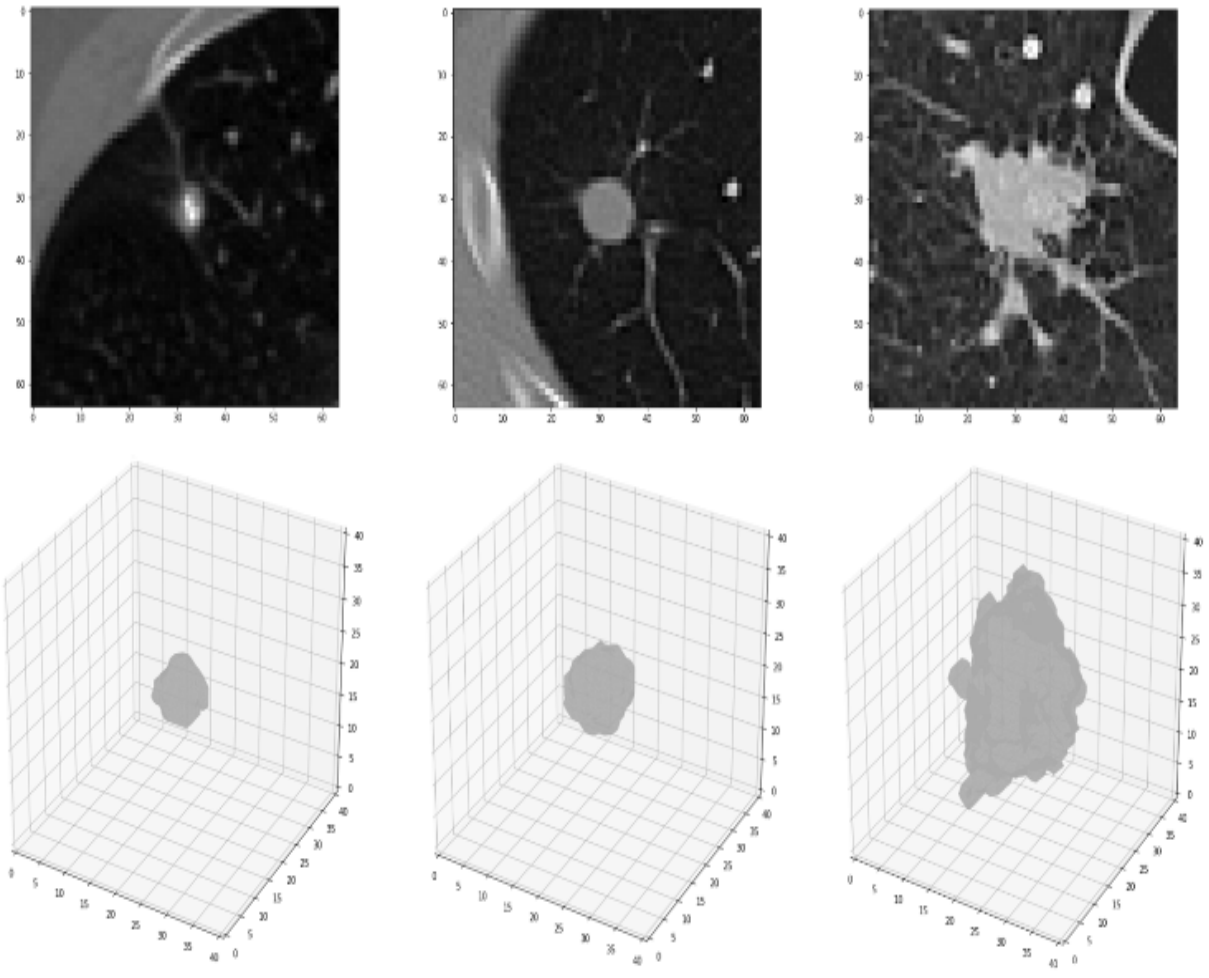

Figure 4: Top: actual nodules. Bottom: corresponding segmented volumes of generated candidates for diameter sizes 6mm, 12mm and 20mm.

Following the features-based FPR step, remaining candidate nodules were processed with the DNN FPR module. The DNN model used here involved a complex DNN architecture for the simple task of binary classification. The architecture was based on the work by Szegedy *et al*. [59] in combining the inception network architecture [41] with residual learning elements [42]. The architecture used in our proposed model is adapted from the Inception-ResNet architectures proposed by Szegedy *et al*., albeit adapted to 3D images. Rather than using leaky rectified linear units as activation functions, scaled exponential linear units, recently demonstrated as better than standard rectified linear units for very deep networks [47], were used as initial testing indicated performance improvements. Consequently, Gaussian initialization and alpha dropout were used to improve convergence and performance.

The reduction and residual blocks (Figure 5a and 5b), provide further details on the architecture of the network (Figure 3b). The reduction block in Figure 5a was adapted from the wider "Reduction-B" block from Inception-ResNet-v1, one of three Inception-ResNet architectures proposed by Szegedy *et al*. [48]. The residual block was adapted from the "Inception-ResNet-A" block of Inception-ResNet-v1 with an extra convolution layer similar to the residual inception blocks of Inception-ResNet-v4. Unlike Inception-ResNet-v1, the entire architecture of our proposed model was built from only two of these reduction blocks and four of these residual blocks. Each of the convolution layers depicted in the overall architecture apply 1x1x1 convolutions. These are included between concurrent residual blocks to maintain a constant number of features. A smaller input size of 32x32x32 was used to speed up training as it did not affect the results. The entire architecture was comprised of 1.28 million trainable parameters.

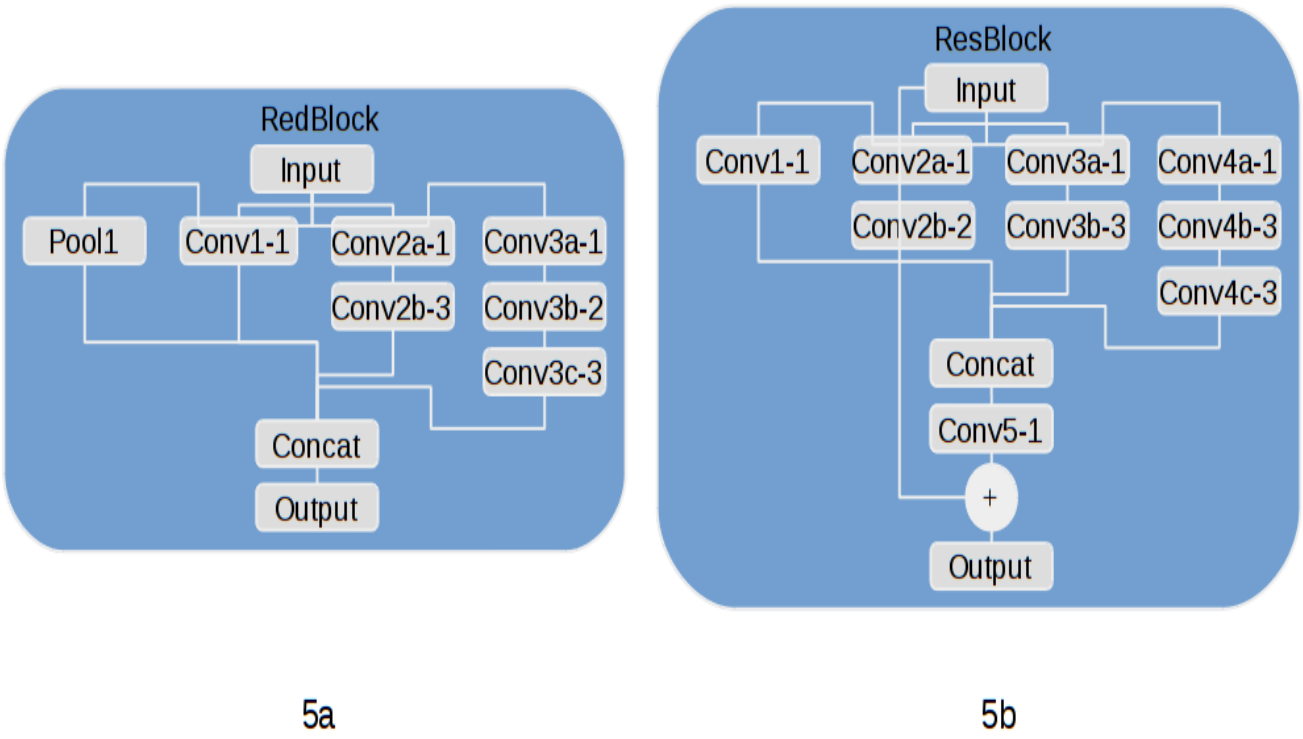

Figure 5: Architectures for a) 3D reduction blocks and b) 3D residual blocks. Convolutions were uniform in 3 dimensions of size denoted by the hyphenated number in the diagrams.

Training[5][6][7]

We used 10-fold cross-validation consistent with the cross-validation experimental design used by Setio *et al*. [37]. For each of the ten folds of the cross-validation, 9-fold cross-validation was conducted and the 10th bin was held out as test data. The best performing model from the 9-fold cross-validation was evaluated on the test data. This enabled the use of all 888 scans as test data.

Training for the CG network was conducted with all available positive nodules[8] for each of the 9-fold cross-validations conducted. The nodule candidates generated for each of the validation bins during each

[5] The source code used for both training and evaluation is available at https://github.com/rossgritz/cad17.
[6] DNN training was conducted using Keras with a Tensorflow backend.
[7] Further details regarding training are presented in Appendix B.
[8] No negative examples were used because the Dice coefficient could only be computed from positive examples.

cross-validation were used to train the FPR network. Over 1000 positive examples were used for each of the 10 9-fold cross-validations.

A 3D array of training labels was required as the CG network was a volume-to-volume prediction network. Approximate binary labels for each nodule were generated as spheres based on the diameter of each nodule. Heavy data augmentation was employed during training, involving random translation up to 14-units along each axis, random rotation about each axis up to 90°, and random flipping about each axis. Implementation of this augmentation required the application of the same random transformations to both the input image and its associated labeled array.

The Sorensen-Dice coefficient [60 61], or simply the Dice coefficient (*DC*), was used to measure the spatial overlap between the predicted volume and the ground truth for the purposes of optimization. It is defined as:

$$DC(G,P) = \frac{2|G \cap P|}{|G| + |P|}$$

where P is the set of prediction results and G is the set of ground truth. It ranges in value from 0 to 1, with 1 indicating perfect prediction accuracy. In order to optimize the network architecture, the objective function, $J(\theta)$, was defined as:

$$J(\theta) = 1 - DC(G,P)$$

The Adam optimizer [62], with an initial learning rate of $1e^{-5}$, was used for minimizing $J(\theta)$. In order to minimize overfitting, the dropout technique [63] and batch normalization [64] were used for regularization. As noted, the CG architecture included over 100 million trainable parameters. This effectively restricted the batch size to two, significantly slowing training[9]. Due to this, each epoch required training 400 minibatches. Models were trained for 210 epochs. Upon completion of the training,

[9] More information regarding training can be found in Appendix B.

the models exhibiting lowest $J(\theta)$ values were evaluated systematically in order to identify the best models at a minimal computational expense. Results were recorded, and the process was repeated until performance gains plateaued.

The FPR network was trained as a binary classifier with the candidates generated from the best performing models of each validation bin during cross-validation. These candidate nodules were labeled as either nodule or non-nodule[10]. Data augmentation was used involving random translation along each axis up to 2-units, random rotation of up to 15° about each axis and random flipping about each axis. Data was centered by subtracting the image mean and normalized by the image standard deviation.

Categorical cross-entropy was used as an objective function and was minimized using the AdaDelta optimizer [65]. Again, to minimize overfitting, dropout [63] and batch normalization [64] were used. With the architecture including over a million trainable parameters a batch size of 64 were trained per epoch. During training, only positive examples were included in the validation dataset in order to prioritize sensitivity. Consequently, a large number of well performing models were saved, and specificity was assessed among these.

RESULTS

By conducting 10 9-fold cross-validations, each scan and nodule in the dataset was able to be evaluated as test data. To evaluate test data, scans were first evaluated with the CG module. The resulting candidate nodules were determined to be either a direct hit, a near hit, or a miss. A candidate nodule was classified a 'direct hit' when the center of the true nodule was within the segmented volume of the candidate nodule. A nodule was treated as 'near hit' when the center of the candidate nodule was within 0.75D of the true

[10] There were substantially more negative examples generated during CG. To handle this, the system was designed such that approximately equal numbers of positive and negative examples were used for each training epoch. The examples used for each epoch were sampled randomly from a single directory containing all negative examples and *n* copies of all positive examples prior. *n* was selected such that the total number of positive and negative examples in the directory were roughly equal.

nodule [15]. Direct hits and near hits were taken as positive candidates. All candidate nodules generated were then evaluated with the FPR module.

The results from the cross-validation and the test data are reported in Table I. This table only reports mean values of sensitivity, false positives per scan (FPPS) and receiver operating characteristic area under the curve (ROC AUC). The full test results for each of the 10 cross-validations are reported in Appendix A in Table III. The cross-validation results are depicted for comparison only and, unless otherwise noted, all references to results in the paper refer to the test results. The full cross-validation results for the eight values reported in Table I are presented in Tables IV through XI in Appendix A, which also reports the cross-validation results for each of the 90 cross-validation systems.

Table I: Results

Task	Dataset	Sensitivity	FPPS	ROC AUC
CG	Cross-Validation	0.9516	30.68	-
	Test	0.9477	30.39	-
FPR	Cross-Validation	0.9594	1.815	0.9849
	Test	0.9421	1.789	0.9835
Complete System	Cross-Validation	0.9129	1.815	0.9372
	Test	0.8929	1.789	0.9324

The free-response ROC curve (FROC) is commonly used in lieu of the traditional ROC curve for cases such as pulmonary nodule detection in which one or more abnormalities may be present for each evaluated case [66 67]. Figure 6 depicts an FROC curve for the system plotted on a logarithmic scale. Recent work on pulmonary nodule CAD systems [6 33 37] has featured a metric referred to as the 'competition performance metric' (CPM) that is based on the FROC curve for evaluating CAD systems [68]. This metric is computed by taking the mean sensitivity value for each of the seven labeled positions on the x axis in Figure 6. The CPM for the system was 0.8135.

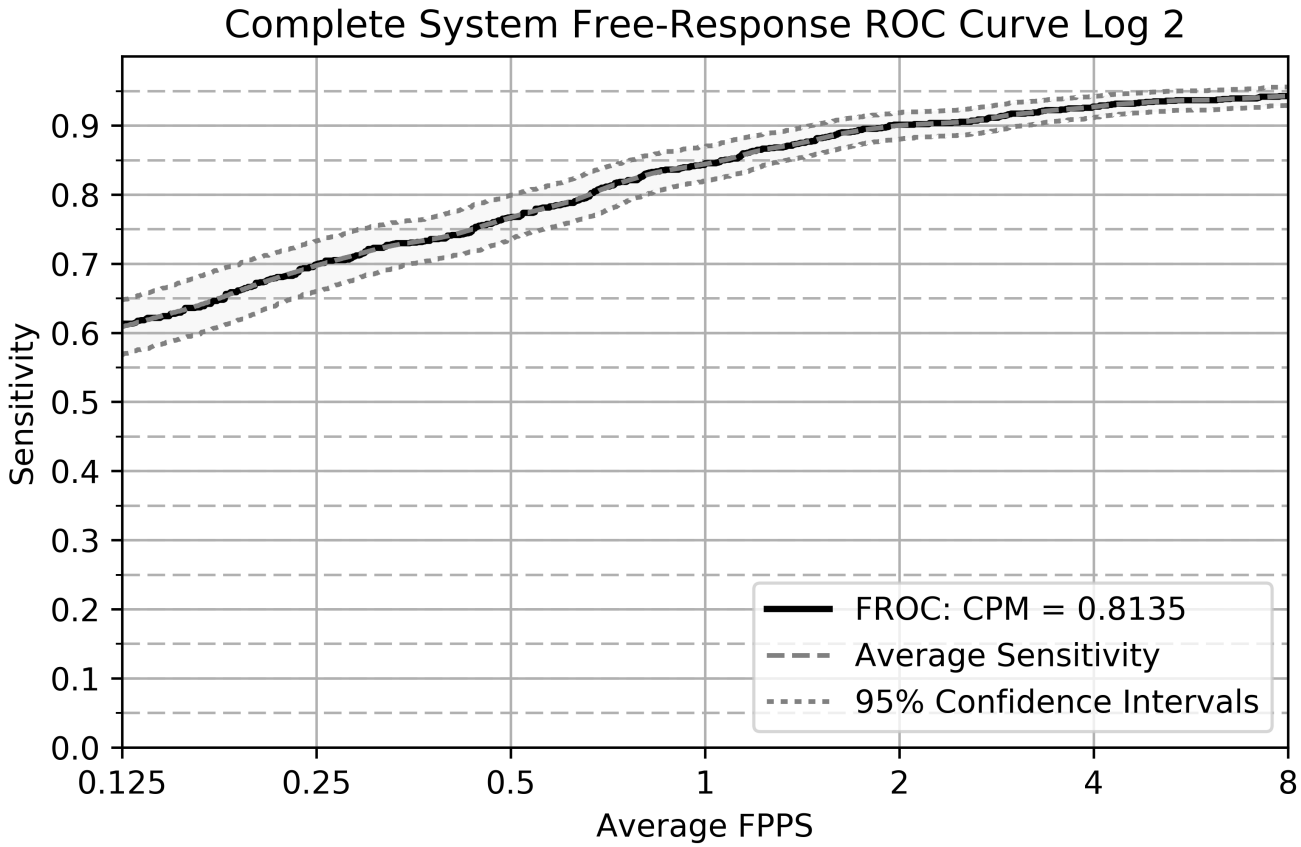

***Figure 6**:* An FROC curve for the complete system and depicted on a logarithmic scale[11] of base 2 from 0.125 to 8.0.

DISCUSSION

A comparison of the features and performance of pulmonary nodule CAD systems is reported in Table II. In this table, a range of values for the proposed system's sensitivity is reported in order to enable comparisons with the various systems. The middle value is the raw result of the FPR module. The top and bottom values are sensitivities taken from the FROC curve for 1.0 and 4.0 FPPS, respectively. This table includes no studies that exclusively use the LUNA16 Challenge evaluation framework due to its

[11] The same data is depicted for log base 10 in Figure 7 of Appendix A.

limitations. Detailed explanation for this is provided in Appendix B. Results reported by Setio et al. [37] and vanGinneken et al. [34] for CAD systems employing deep learning for FPR presented sensitivities based on false positive rates from FROC curves. Table II demonstrates the improvements in the performance of the proposed systems over these two systems. Setio et al. [37] reported a CPM of 0.722 and was the only one of the studies listed in Table II to report this metric. The CPM reported here substantially exceeds this value. Shaukat et al. [15] demonstrated excellent performance for their complete system, which was heavily influenced by the strong performance of their FPR module, and is discussed further in Appendix C. These results suggest that future work could explore the combination of deep learning systems for feature extraction and support vector machines (SVMs) for classification. As reported in Table II, among the studies included for comparison, only Setio et al. [37] and Jacobs et al. [52] used the complete number of scans for test data. While Setio et al. [37] used existing CG systems, Jacobs et al.[52] simply evaluated existing systems. Shaukat et al. [15] did, however, use 30% of their total dataset for test, which is substantial. Of the other studies involving a large number of scans, Hamidian et al. [39] used 25 scans for test while Firmino et al. [12] and vanGinneken et al. [34] strictly reported cross-validation results. We are not aware of another novel system in the literature to have been evaluated as rigorously. A complete comparison and discussion of results for each of the tasks can be found in Appendix C.

Table II: Comparison of Pulmonary Nodule CAD Systems

	# Scans	Sensitivity	FPPS	CG	FPR	Thickness	Small/non-nodules
Proposed System	888	0.8446	1.00	3D DNN	3D DNN	<=3.0mm	Yes
		0.8929	1.79				
		0.9273	4.00				
Shaukat *et al.* 2017	850	0.9269	1.91	filter-based segmentation	specified features w/ SVM classifier	1.0-3.0 mm	Yes
Hamidian *et al.* 2017	509	0.8000	15.28	3D CNN	3D CNN	1.5-3.0 mm	Yes
Firminio *et al.* 2016	420	0.9440	7.04	segmentation & rule-based	HOG features & SVM classifier	<=3.0mm	No
Jacobs *et al.* 2016	888	0.8200	3.10	comparison of 3 systems (commercial & academic)		<=3.0mm	Yes
Setio *et al.* 2016	888	0.7820	1.00	ensemble of existing systems	multi-view CNN	<=3.0mm	Yes
		0.8790	4.00				
vanGinneken *et al.* 2015	865	0.7300	1.00	existing system	comparison of 3 existing DNNs	<=2.5mm	No
		0.7600	4.00				
Choi & Choi 2014	84	0.9545	6.76	dot-enhancement filter	AHSN features & SVM classifier	1.0-3.0 mm	Yes
Choi & Choi 2013	58	0.9276	2.27	block segmentation	specified features w/ SVM classifier	1.0-3.0 mm	Yes

Strengths and Limitations

Strengths of the study include the robust experimental design, the large test dataset and the methodology. Particularly, retaining 30,513 false positive items to evaluate the system strongly distinguishes this study from a large number of current deep learning studies for nodule detection. A further strength of the study is the open availability of the source code and the trained weights for the 180 different models for which results are reported in Appendix A.

Despite the use of a substantial amount of test data, the lack of evaluation on an external dataset remains a significant limitation. Further limitations of the dataset are described by Armato et al. [49] and includes the lack of clinical information and the fact that scans weren't all scored by the same radiologists. Other limitations concern limited analysis of segmentation quality due to the focus on the CAD system and limited computational capacity due to the high computational costs.

CONCLUSION

We present a novel, fully automated system for the detection of pulmonary nodules that was extensively validated using each scan in the dataset for both training and test. As described earlier, the CG and FPR modules each independently demonstrated results that outperform other reported or known systems for their respective tasks. Overall, the results indicate that the system is comparable with the best existing systems.

The proposed CAD system is designed for detecting pulmonary nodules. Extending it to quantify malignancy and offer diagnostic predictions is an important research direction for future work. Such a system could be evaluated rigorously with a large external dataset – a significant limitation of this study – using data from the National Lung Screening Trial [2 69].

FUNDING

Support for this work was provided by Auburn University. This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

CONTRIBUTIONS

RG developed the model and did coding for the model under supervision of AG. RG and AG developed the manuscript and performed revisions. DP edited the manuscript and provided feedback.

COMPETING INTERESTS

Authors have no competing interests to declare.

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. The following institutions were responsible for the LIDC/IDRI: Weill Cornell Medical College, the University of California, Los Angeles, the University of Chicago, the University of Iowa and the University of Michigan.

REFERENCES

1. Ferlay J, Soerjomataram I, Dikshit R, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. International Journal of Cancer 2015;136(5).

2. Team NLSTR. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 2011;(365):395-409.

3. Singh H, Sethi S, Raber M, Petersen LA. Errors in cancer diagnosis: current understanding and future directions. Journal of Clinical Oncology 2007;25(31):5009-18.

4. Bach PB, Mirkin JN, Oliver TK, et al. Benefits and harms of CT screening for lung cancer: a systematic review. JAMA 2012;307(22):2418-29.

5. Peldschus K, Herzog P, Wood SA, et al. Computer-aided diagnosis as a second reader: spectrum of findings in CT studies of the chest interpreted as normal. Chest Journal 2005;128(3):1517-23.

6. Setio AAA, Traverso A, de Bel T, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. Medical Image Analysis 2017;42:1-13.

7. Messay T, Hardie RC, Rogers SK. A new computationally efficient CAD system for pulmonary nodule detection in CT imagery. Medical Image Analysis 2010;14(3):390-406.

8. Choi W-J, Choi T-S. Genetic programming-based feature transform and classification for the automatic detection of pulmonary nodules on computed tomography images. Information Sciences 2012;212:57-78.

9. Armato SG, Giger ML, Moran CJ, et al. Computerized detection of pulmonary nodules on CT scans. Radiographics 1999;19(5):1303-11.

10. Dehmeshki J, Ye X, Lin X, et al. Automated detection of lung nodules in CT images using shape-based genetic algorithm. Computerized Medical Imaging and Graphics 2007;31(6):408-17.

11. Ye X, Lin X, Dehmeshki J, et al. Shape-based computer-aided detection of lung nodules in thoracic CT images. IEEE Transactions on Biomedical Engineering 2009;56(7):1810-20.

12. Firmino M, Angelo G, Morais H, Dantas MR, Valentim R. Computer-aided detection (CADe) and diagnosis (CADx) system for lung cancer with likelihood of malignancy. Biomedical Engineering Online 2016;15(1):2.

13. Suárez-Cuenca JJ, Tahoces PG, Souto M, et al. Application of the iris filter for automatic detection of pulmonary nodules on computed tomography images. Computers in Biology and Medicine 2009;39(10):921-33.

14. Li Q, Sone S. Selective enhancement filters for nodules, vessels, and airway walls in two-and three-dimensional CT scans. Med Phys. 2003;30(8):2040-51.

15. Shaukat F, Raja G, Gooya A, et al. Fully automatic and accurate detection of lung nodules in CT images using a hybrid feature set. Med Phys. 2017;44(7):3615-3629.

16. Tan M, Deklerck R, Jansen B, Bister M, Cornelis J. A novel computer-aided lung nodule detection system for CT images. Med Phys. 2011;38(10):5630-45.

17. Ge Z, Sahiner B, Chan HP, et al. Computer-aided detection of lung nodules: False positive reduction using a 3D gradient field method and 3D ellipsoid fitting. Med Phys. 2005;32(8):2443-54.

18. Boroczky L, Zhao L, Lee KP. Feature subset selection for improving the performance of false positive reduction in lung nodule CAD. IEEE Trans Inf Technol Biomed. 2006;10(3):504-11.

19. Gurcan MN, Sahiner B, Petrick N, et al. Lung nodule detection on thoracic computed tomography images: Preliminary evaluation of a computer-aided diagnosis system. Med Phys. 2002;29(11):2552-58.

20. Wu YC, Doi K, Giger ML, et al. Reduction of false positives in computerized detection of lung nodules in chest radiographs using artificial neural networks, discriminant analysis, and a rule-based scheme. J Digit Imaging. 1994;7(4):196-207.

21. Lo S-C, Lou S-L, Lin J-S, et al. Artificial convolution neural network techniques and applications for lung nodule detection. IEEE Trans Med Imaging. 1995;14(4):711-18.

22. Suzuki K, Li F, Sone S, Doi K. Computer-aided diagnostic scheme for distinction between benign and malignant nodules in thoracic low-dose CT by use of massive training artificial neural network. IEEE Trans Med Imaging. 2005;24(9):1138-50.

23. Suzuki K, Armato SG, Li F, Sone S. Massive training artificial neural network (MTANN) for reduction of false positives in computerized detection of lung nodules in low-dose computed tomography. Med Phys. 2003;30(7):1602-17.

24. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Bartlett P, Pereira F., Burges C.J.C. et al. eds. Advances in Neural Information Processing Systems: 26th Annual Conference on Neural Information Processing Systems 2012, 2012:1097-105.

25. Cireşan DC, Giusti A, Gambardella LM, et al. Mitosis detection in breast cancer histology images with deep neural networks. Med Image Comput Comput Assist Interv. 2013;16(Pt 2):411-8.

26. Roth HR, Lu L, Seff A, et al. A new 2.5 D representation for lymph node detection using random sets of deep convolutional neural network observations. Med Image Comput Comput Assist Interv. 2014;17(Pt 1):520-7.

27. Chen H, Dou Q, Yu L, et al. VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images. Neuroimage. 2017. pii: S1053-8119(17)30334-8.

28. Prasoon A, Petersen K, Igel C. et al. Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. Med Image Comput Comput Assist Interv. 2013;16(Pt 2):246-53.

29. Ciresan D, Giusti A, Gambardella LM, et al. Deep neural networks segment neuronal membranes in electron microscopy images. In: Bartlett P, Pereira F., Burges C.J.C. et al. eds. Advances in Neural Information Processing Systems: 26th Annual Conference on Neural Information Processing Systems 2012. 2012:2843-51.

30. Yu L, Yang X, Chen H, et al. Volumetric ConvNets with Mixed Residual Connections for Automated Prostate Segmentation from 3D MR Images. AAAI, 2017:66-72.

31. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016;316(22):2402-10.

32. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542(7639):115-18.

33. Dou Q, Chen H, Yu L, et al. Multi-level contextual 3D CNNs for false positive reduction in pulmonary nodule detection. IEEE Trans Biomed Eng. 2017 Jul;64(7):1558-1567.

34. van Ginneken B, Setio AA, Jacobs C, Ciompi F. Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans. Biomedical Imaging (ISBI), IEEE 12th International Symposium on Biomedical Imaging. 2015:286-89.

35. Gruetzemacher R, Gupta A. Using deep learning for pulmonary nodule detection & diagnosis. Americas Conference on Information Systems, 2016.

36. Shen W, Zhou M, Yang F, et al. Multi-scale convolutional neural networks for lung nodule classification. Inf Process Med Imaging. 2015;24:588-99.

37. Setio AAA, Ciompi F, Litjens G, et al. Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks. IEEE Trans Med Imaging. 2016;35(5):1160-69.

38. Wang S, Zhou M, Liu Z, et al. Central focused convolutional neural networks: Developing a data-driven model for lung nodule segmentation. Med Image Anal. 2017;40:172-183

39. Hamidian S, Sahinerb B, Petrickb N, et al. Spring M. 3D Convolutional Neural Network for Automatic Detection of Lung Nodules in Chest CT. Proc SPIE Int Soc Opt Eng. 2017:1013409-09-6.

40. He K, Zhang X, Ren S, et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. Proc IEEE international Conf on Computer Vision, 2015:1026-34.

41. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. Proc. IEEE conf on Computer Vision and Pattern Recognition, 2015:1-9.

42. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. Proc. IEEE conf on computer vision and pattern recognition, 2016:770-78.

43. Choi W-J, Choi T-S. Automated pulmonary nodule detection system in computed tomography images: A hierarchical block classification approach. Entropy 2013;15(2):507-23.

44. Choi W-J, Choi T-S. Automated pulmonary nodule detection based on three-dimensional shape-based feature descriptor. Comput Methods Programs Biomed. 2014;113(1):37-54.

45. LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. Proceedings of the IEEE. 1998;86(11):2278-324.

46. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. International Conference on Medical Image Computing and Computer-Assisted Intervention: Springer, 2016:424-32.

47. Klambauer G, Unterthiner T, Mayr A, et al. Self-Normalizing Neural Networks. arXiv preprint arXiv:1706.02515 2017.

48. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. AAAI, 2017:4278-84.

49. Armato SG, McLennan G, Bidaut L, et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. Med Phys. 2011;38(2):915-31.

50. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging. 2013;26(6):1045-57.

51. Bergtholdt M, Wiemker R, Klinder T. Pulmonary nodule detection using a cascaded SVM classifier. SPIE Medical imaging. 2016;9785:978513.

52. Jacobs C, Rikxoort EM, Murphy K, et al. Computer-aided detection of pulmonary nodules: a comparative study using the public LIDC/IDRI database. Eur Radiol. 2016;26(7):2139-47.

53. Tan M, Deklerck R, Cornelis J, et al. Phased searching with NEAT in a time-scaled framework: experiments on a computer-aided detection system for lung nodules. Artif Intell Med. 2013;59(3):157-67.

54. Samuel G, Armato III GM, Bidaut L, et al. Data from LIDC-IDRI: The cancer imaging archive.

55. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Cognitive Modeling 1988;5(3):1.

56. Rumelhart DE, Hinton GE, Williams RJ. Learning internal representations by error propagation: DTIC Document, 1985.

57. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436-44.

58. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention: Springer, 2015:234-41.

59. Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning. arXiv preprint arXiv:1602.07261 2016.

60. Sørensen T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. Biol. Skr. 1948;5:1-34.

61. Dice LR. Measures of the amount of ecologic association between species. Ecology 1945;26(3):297-302.

62. Kingma D, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 2014.

63. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, et al. Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research. 2014;15(1):1929-58.

64. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 2015.

65. Zeiler MD. ADADELTA: an adaptive learning rate method. arXiv preprint arXiv:1212.5701 2012.

66. DeLuca PM, Wambersie A, Whitmore GF. Extensions to conventional ROC methodology: LROC, FROC, and AFROC. J ICRU. 2008;8:31-5.

67. Bandos AI, Rockette HE, Song T, et al. Area under the Free-Response ROC Curve (FROC) and a Related Summary Index. Biometrics. 2009;65(1):247-56.

68. Niemeijer M, Loog M, Abramoff MD, et al. On combining computer-aided detection systems. IEEE Transactions on Medical Imaging. 2011;30(2):215-23.

69. Team NLSTR. The national lung screening trial: overview and study design. Radiology 2011;258(1):243-53.

APPENDIX A

Additional Results

The complete results for cross-validation are reported in Tables III-XI. Table III depicts the results for each of the test bins evaluated during the cross-validation procedure. The mean test results for the complete system, as reported in the main paper, are listed in the far-right column of Table III. Tables IV-XI show the results for each of the 10 9-fold cross-validation cases. The metrics included in these tables are: candidate generation sensitivities, candidate generation false positives per scan, false positive reduction sensitivities, false positive reduction false positives per scan, false positive reduction ROC AUCs, complete system sensitivities, complete system ROC AUCs and complete system false positives per scan, respectively. The best performing models from each of the cross-validations were selected based on the overall system sensitivity. In the event of a tie, the best performing model was selected based on the complete system ROC AUC. The best performing system from each cross-validation is highlighted in each of the Tables depicting cross-validation results. The model weights for each of the cross-validation bins, as well as the source code, have been made available online through the GitHub link provided in the main paper.

COMPLETE CROSS VALIDATION RESULTS												
Task	Metric	TEST BIN										Overall Mean
		0	1	2	3	4	5	6	7	8	9	
Candidate Generation	Sensitivity	0.9590	0.9375	0.9766	0.9664	0.9297	0.9141	0.9643	0.9640	0.9322	0.9333	0.9477
	FPPS	22.92	30.40	39.14	31.78	39.41	25.76	30.62	27.21	37.86	18.77	30.39
False Positive Reduction	Sensitivity	0.9532	0.9333	0.9360	0.9739	0.9328	0.9381	0.9520	0.9252	0.9273	0.9490	0.9421
	FPPS	2.213	1.045	1.764	1.775	1.899	1.798	1.809	1.445	1.932	2.239	1.789
	ROC AUC	0.9777	0.9861	0.9884	0.9878	0.9893	0.9831	0.9877	0.9793	0.9870	0.9682	0.9835
Complete System	Sensitivity	0.9141	0.8750	0.9141	0.9412	0.8672	0.8575	0.9180	0.8919	0.8644	0.8857	0.8929
	FPPS	2.213	1.045	1.764	1.775	1.899	1.798	1.809	1.416	1.932	2.239	1.789

Table IV: Segmentation Cross-Validation Sensitivities

		TEST										
	BIN	0	1	2	3	4	5	6	7	8	9	Average
VAL	0		0.9554	0.9821	0.9464	0.9732	0.9464	0.9643	0.9554	0.9643	0.9554	0.9603
	1	0.9766		0.9609	0.9688	0.9609	0.9688	0.9609	0.9609	0.9609	0.9766	0.9661
	2	0.9609	0.9688		0.9688	0.9688	0.9766	0.9609	0.9609	0.9766	0.9609	0.9670
	3	0.9748	0.9663	0.9832		0.9580	0.9832	0.9748	0.9663	0.9832	0.9832	0.9748
	4	0.9453	0.9375	0.9297	0.9375		0.9375	0.9531	0.9375	0.9375	0.9531	0.9410
	5	0.9074	0.9167	0.9259	0.9352	0.9074		0.9259	0.9167	0.9259	0.9259	0.9208
	6	0.9612	0.9690	0.9690	0.9612	0.9767	0.9612		0.9612	0.9612	0.9535	0.9638
	7	0.9729	0.9550	0.9550	0.9550	0.9820	0.9640	0.9550		0.9550	0.9640	0.9620
	8	0.9237	0.9407	0.9237	0.9407	0.9322	0.9407	0.9407	0.9492		0.9237	0.9350
	9	0.9428	0.9238	0.9143	0.9333	0.8952	0.9333	0.9238	0.9333	0.9238		0.9248
Average		0.9517	0.9481	0.9493	0.9497	0.9505	0.9569	0.9510	0.9490	0.9543	0.9551	0.9516

Table V: Segmentation Cross-Validation False Positives per Scan

		TEST										
	BIN	0	1	2	3	4	5	6	7	8	9	Average
VAL	0		33.35	24.85	27.28	44.80	34.38	33.21	29.26	23.35	38.95	32.16
	1	21.44		17.85	30.79	17.54	19.52	21.08	16.23	23.75	17.70	20.66
	2	28.22	38.79		39.75	40.68	29.05	29.75	30.34	34.55	34.71	33.98
	3	31.35	32.40	17.88		40.73	24.40	23.97	30.18	37.23	23.20	29.04
	4	35.48	34.52	25.86	19.92		26.43	42.49	39.60	32.05	36.98	32.59
	5	31.24	26.22	28.30	33.57	18.02		34.15	32.63	35.06	28.63	29.76
	6	21.92	51.95	37.38	17.84	37.33	18.59		39.15	35.21	39.71	33.23
	7	22.94	38.17	25.37	29.22	50.98	37.67	28.72		37.72	30.52	33.48
	8	44.98	31.30	24.78	26.22	23.21	40.12	29.28	32.78		29.31	31.33
	9	30.88	32.22	16.39	35.05	27.10	35.64	26.15	31.22	40.41		30.56
Average		29.83	35.44	24.30	28.85	33.38	29.53	29.87	31.27	33.26	31.08	30.68

Table VI: False Positive Reduction Cross-Validation Sensitivities

		TEST										
	BIN	0	1	2	3	4	5	6	7	8	9	Average
VAL	0		0.9626	0.9182	0.9717	0.9633	0.9717	0.9903	0.9813	0.9722	0.9626	0.9660
	1	0.9680		0.9593	0.9677	0.9593	0.9435	0.9672	0.9350	0.9593	0.9680	0.9586
	2	0.9756	0.9274		0.9597	0.9597	0.9760	0.9593	0.9756	0.9554	0.9756	0.9627
	3	0.9428	0.9478	0.9487		0.9649	0.9573	0.9741	0.9569	0.9658	0.9573	0.9573
	4	0.9586	0.9091	0.9580	0.9500		0.9667	0.9590	0.9583	0.9417	0.9508	0.9502
	5	0.9694	0.9083	0.9700	0.9406	0.9694		0.9400	0.9596	0.9500	0.9400	0.9497
	6	0.9758	0.9120	0.9840	0.9516	0.9762	0.9917		0.9758	0.9758	0.9756	0.9687
	7	0.9534	0.9340	0.9245	0.9151	0.9541	0.9533	0.9811		0.9434	0.9626	0.9468
	8	0.9285	0.9423	0.9450	0.9369	0.9545	0.9820	0.9730	0.9554		0.9908	0.9565
	9	0.9798	0.9691	0.9583	0.9796	0.9787	0.9898	0.9796	0.9796	0.9794		0.9771
Average		0.9613	0.9347	0.9518	0.9525	0.9645	0.9702	0.9693	0.9642	0.9603	0.9648	0.9594

Table VII: False-Positive Reduction Cross-Validation False Positives Per Scan

		TEST										
	BIN	0	1	2	3	4	5	6	7	8	9	Average
VAL	0		2.000	1.045	1.348	1.742	1.909	1.864	1.485	1.152	1.606	1.572
	1	1.787		1.984	1.803	1.410	1.459	1.918	1.361	1.885	1.967	1.730
	2	1.714	1.982		1.607	1.607	1.875	1.929	1.269	1.821	1.875	1.742
	3	1.277	1.923	1.938		1.877	1.415	2.031	1.754	1.877	1.554	1.738
	4	2.100	2.685	1.937	1.937		1.857	1.952	1.873	1.937	1.841	2.013
	5	1.907	2.159	1.815	1.778	1.556		1.370	1.907	2.019	2.000	1.835
	6	1.667	1.746	1.587	1.076	1.825	1.777		1.556	1.825	1.778	1.649
	7	1.796	2.426	2.037	2.093	2.000	1.963	1.611		1.981	1.907	1.979
	8	2.300	2.667	1.983	1.733	3.050	1.917	1.950	1.867		2.900	2.263
	9	2.068	1.847	0.915	1.475	1.610	1.881	1.661	1.559	1.678		1.633
Total		1.846	2.159	1.693	1.650	1.853	1.784	1.810	1.626	1.797	1.936	1.815

Table VIII: False Positive Reduction Cross-Validation ROC AUCs

		TEST										
	BIN	0	1	2	3	4	5	6	7	8	9	Average
VAL	0		0.9917	0.9731	0.9915	0.9925	0.9907	0.9942	0.9939	0.9913	0.9919	0.9901
	1	0.9822		0.9767	0.9854	0.9824	0.9824	0.9815	0.9749	0.9801	0.9799	0.9806
	2	0.9894	0.9819		0.9907	0.9896	0.9861	0.9786	0.9859	0.9840	0.9902	0.9863
	3	0.9864	0.9866	0.9741		0.9909	0.9871	0.9818	0.9822	0.9912	0.9823	0.9847
	4	0.9813	0.9778	0.9808	0.9774		0.9819	0.9868	0.9687	0.9850	0.9887	0.9809
	5	0.9813	0.9773	0.9836	0.9780	0.9837		0.9763	0.9809	0.9840	0.9779	0.9803
	6	0.9907	0.9920	0.9940	0.9847	0.9899	0.9898		0.9915	0.9916	0.9877	0.9902
	7	0.9758	0.9831	0.9786	0.9798	0.9911	0.9893	0.9852		0.9803	0.9785	0.9824
	8	0.9812	0.9816	0.9818	0.9732	0.9797	0.9908	0.9851	0.9870		0.9887	0.9832
	9	0.9876	0.9929	0.9844	0.9920	0.9915	0.9911	0.9902	0.9893	0.9942		0.9904
Average		0.9840	0.9850	0.9808	0.9836	0.9879	0.9877	0.9844	0.9838	0.9869	0.9851	0.9849

Table IX: Complete System Cross-Validation Sensitivities

		TEST										
	BIN	0	1	2	3	4	5	6	7	8	9	Average
VAL	0		0.9197	0.9018	0.9196	0.9375	0.9196	0.9549	0.9375	0.9375	0.9197	0.9275
	1	0.9453		0.9218	0.9375	0.9218	0.9141	0.9294	0.8984	0.9218	0.9453	0.9262
	2	0.9375	0.8985		0.9298	0.9298	0.9532	0.9218	0.9375	0.9330	0.9375	0.9309
	3	0.9190	0.9159	0.9328		0.9244	0.9412	0.9496	0.9246	0.9496	0.9412	0.9331
	4	0.9062	0.8523	0.8907	0.8906		0.9063	0.9140	0.8984	0.8828	0.9062	0.8942
	5	0.8796	0.8327	0.8981	0.8796	0.8796		0.8703	0.8797	0.8796	0.8703	0.8744
	6	0.9379	0.8837	0.9535	0.9147	0.9535	0.9532		0.9379	0.9379	0.9302	0.9336
	7	0.9276	0.8919	0.8829	0.8739	0.9369	0.9190	0.9370		0.9009	0.9279	0.9109
	8	0.8577	0.8864	0.8729	0.8813	0.8898	0.9238	0.9153	0.9068		0.9152	0.8944
	9	0.9238	0.8952	0.8762	0.9143	0.8761	0.9238	0.9049	0.9143	0.9048		0.9037
Average		0.9150	0.8863	0.9034	0.9046	0.9166	0.9282	0.9219	0.9150	0.9164	0.9215	0.91289

Table X: Complete System Cross-Validation ROC AUCs

		TEST										
	BIN	0	1	2	3	4	5	6	7	8	9	Average
VAL	0		0.9475	0.9557	0.9384	0.9659	0.9376	0.9587	0.9496	0.9559	0.9477	0.9508
	1	0.9592		0.9385	0.9547	0.9440	0.9517	0.9431	0.9368	0.9418	0.9570	0.9474
	2	0.9508	0.9513		0.9598	0.9587	0.9630	0.9403	0.9474	0.9610	0.9515	0.9537
	3	0.9615	0.9534	0.9577		0.9493	0.9705	0.9571	0.9491	0.9745	0.9658	0.9599
	4	0.9276	0.9167	0.9118	0.9163		0.9205	0.9405	0.9082	0.9234	0.9423	0.9230
	5	0.8904	0.8959	0.9107	0.9146	0.8926		0.9040	0.8992	0.9111	0.9054	0.9027
	6	0.9523	0.9612	0.9632	0.9465	0.9668	0.9514		0.9530	0.9531	0.9418	0.9544
	7	0.9494	0.9389	0.9346	0.9357	0.9733	0.9537	0.9409		0.9362	0.9433	0.9451
	8	0.9063	0.9234	0.9069	0.9155	0.9133	0.9320	0.9267	0.9369		0.9133	0.9194
	9	0.9311	0.9172	0.9000	0.9258	0.8876	0.9250	0.9147	0.9233	0.9185		0.9159
Average		0.9365	0.9339	0.9310	0.9341	0.9391	0.9451	0.9362	0.9337	0.9417	0.9409	0.9372

Table XI: Complete System Cross-Validation False Positives

		TEST										
	BIN	0	1	2	3	4	5	6	7	8	9	Average
VAL	0		2.000	1.045	1.348	1.742	1.909	1.864	1.485	1.152	1.606	1.572
	1	1.787		1.984	1.803	1.410	1.459	1.918	1.361	1.885	1.967	1.730
	2	1.714	1.982		1.607	1.607	1.875	1.929	1.269	1.821	1.875	1.742
	3	1.277	1.923	1.938		1.877	1.415	2.031	1.754	1.877	1.554	1.738
	4	2.100	2.685	1.937	1.937		1.857	1.952	1.873	1.937	1.841	2.013
	5	1.907	2.159	1.815	1.778	1.556		1.370	1.907	2.019	2.000	1.835
	6	1.667	1.746	1.587	1.076	1.825	1.777		1.556	1.825	1.778	1.649
	7	1.796	2.426	2.037	2.093	2.000	1.963	1.611		1.981	1.907	1.979
	8	2.300	2.667	1.983	1.733	3.050	1.917	1.950	1.867		2.900	2.263
	9	2.068	1.847	0.915	1.475	1.610	1.881	1.661	1.559	1.678		1.633
Total		1.846	2.159	1.693	1.650	1.853	1.784	1.810	1.626	1.797	1.936	1.815

Figure 7 depicts a free-response ROC curve for the complete system on a logarithmic scale of base 10 from 0.01 to 10. This has been included to allow for a standard comparative criteria with Setio et al. 2016 [1]. A log 2 plot of the free-response ROC curve was used in Setio et al. 2017 [2]. We, therefore, selected a log 2 plot of the complete system free-response ROC curve for inclusion in the main document since it better depicts the performance over the entire range of false positives per scan of our system. Furthermore, it is more appropriate for display when using the competition performance metric (CPM) [3]. It can be seen in Figure 7 that the curve ends before 10. This is due to the fact that 100% of true nodule candidates can be detected with less than 10 false positives per scan.

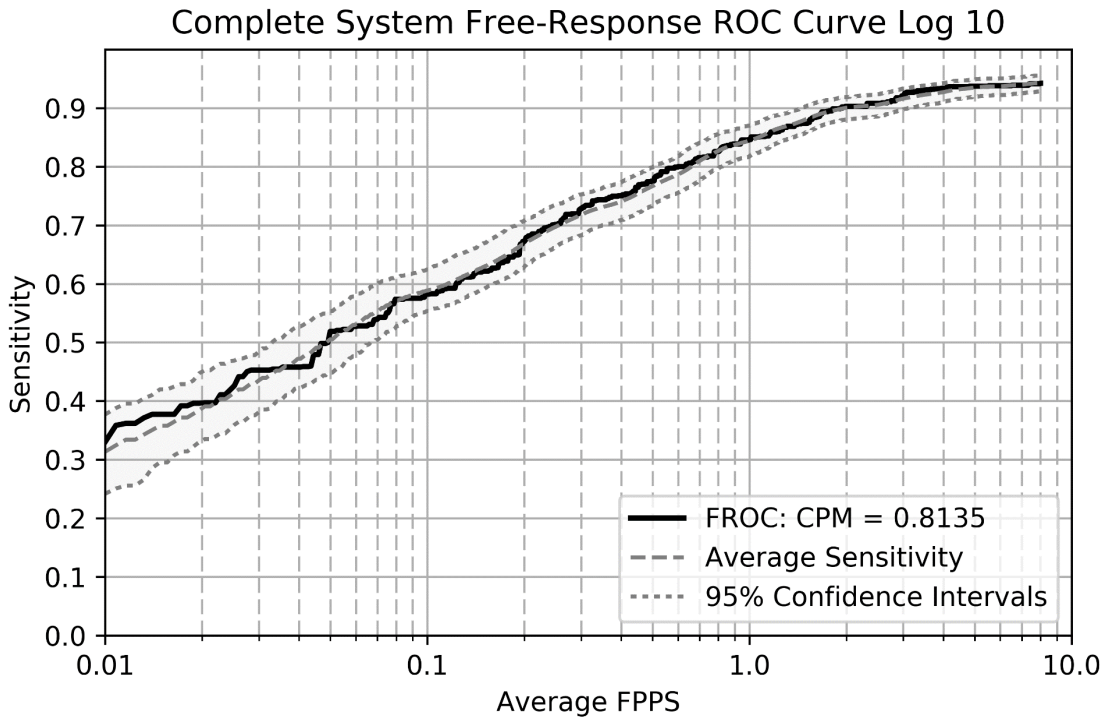

Figure 7: A free-response ROC curve for the complete system depicted on a logarithmic scale of base 10 from 0.01 to 10.

References

1. Setio AAA, Ciompi F, Litjens G, et al. Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks. IEEE transactions on medical imaging 2016;**35**(5):1160-69.
2. Setio AAA, Traverso A, de Bel T, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. Medical Image Analysis 2017;**42**:1-13.
3. Niemeijer M, Loog M, Abramoff MD, Viergever MA, Prokop M, van Ginneken B. On combining computer-aided detection systems. IEEE Transactions on Medical Imaging. 2011 Feb;30(2):215-23.

APPENDIX B

Experimental Design

Each of the 10 9-fold cross-validation cases began by splitting the data into two by setting one bin aside for testing. The test bin was used only for final evaluation and was not used until training was complete. For the purposes of training, validation data was necessary to tune parameters and hyperparameters and selection of the best performing models. These models, selected based on the complete system sensitivity, are highlighted in Tables V-VIII in Appendix A. This experimental design was conducted, creating 10 separate 9-fold cross-validations and 10 separate models, in order to demonstrate the robustness of the results and generalizability of the system by using the entire LUNA16 dataset for testing. This experimental design is consistent with experimental design for artificial intelligence applications [1] and a similar design has been used before for evaluation of a deep learning false positive reduction system by Setio et al. [2]. It was employed in this study because it is superior to the simple cross-validation evaluation framework used in the LUNA16 Challenge [3]. The results presented in this paper are robust and, based on the test results, appear to generalize well to unseen data.

This study also addresses another limitation of the LUNA16 competition evaluation framework: the fact that 30,513 items in the dataset that were marked by radiologists were excluded for evaluation purposes [3]. These included 19,004 lesions labeled non-nodules and 11,509 lesions labeled as nodules less than 3mm in diameter. All nodules annotated as large nodules by three or four radiologists were considered as positive examples, resulting in 1186 nodules. Nodules labeled as large nodules by either one or two radiologists were considered inconclusive and were not considered as false positives in the evaluation.

This decision was justified since these other lesions could be useful in clinical diagnoses other than lung cancer. However, the system presented in this paper was designed with the sole intention of detecting potentially malignant nodules for assisting radiologists in the diagnosis of lung cancer. For reasons outlined in the Introduction, the overwhelming benefit of future pulmonary nodule CAD systems to patients will be through assisting radiologists with lung cancer screening. Consequently, these 30,513 items were all considered relevant, and, when identified by the system, were considered as false positives. This is a major design decision because we treat these objects as negative samples and use them in training the system. Systems that do not count these as false positives may be trained to classify them as positive in order to boost sensitivity of false positive reduction without the cost of a penalty for their prediction. Thus, simply reporting the results for our system without counting these items as false positives is not a sufficient to ensure an impartial comparison. For this reason, and due to the superior experimental design of this study, objective comparisons with studies conducted using the LUNA16 competition framework are not possible. Each of the distinguishing characteristics of this study diminish the quantitative results but more accurately reflect the challenges to deploying these systems in the most widely applicable clinical settings. Consequently, this LUNA16 Challenge trend toward maximizing quantitative results is concerning as it stands to diminish the value of future CAD research to the medical community, clinicians and patients. To mitigate this, we suggest that future studies exclude items of malignant ambiguity and use a substantial test dataset. This presents a greater challenge to researchers and in doing so may benefit the quality of future research.

Lung Segmentation Procedure

Lung segmentation was required prior to nodules segmentation in order to reduce computational costs, and to a lesser degree reduce false positives. To segment the lung, the original CT scan was converted into a binary image using a threshold value of -550 Hounsfield units (HU). Next, labels were generated for the connected components in the binary image. The labels corresponding to the air surrounding the body was identified and set equal to the label for all components greater than -550 HU. The remaining connected components were pockets of air in the body; the lungs were isolated by identifying the largest of these volumes. The dimensions of the resulting volume were used to determine algorithm's success. In all cases where the first attempt was unsuccessful, the seed points from the binary labels were moved from the image origin and the second largest volume was correctly isolated the lungs.

Model Training Procedure

Training and evaluation were both computationally intensive and time consuming processes. Training and evaluation were conducted using the compute clusters with a total of 12 Nvidia GTX1080Ti graphics processing units (GPUs) and 4 Nvidia GTX1070 GPUs. Training was conducted via task parallelism using the 16 GPUs simultaneously for training or evaluating between 16 to 32 separate models concurrently. Data augmentation for the segmentation training was conducted in parallel with 4 threads at the beginning of each minibatch. Data augmentation for the false positive reduction training was conducted in parallel using 2 threads at the beginning of each minibatch.

Deep learning training is typically very computationally expensive and a time-consuming process. Training for each segmentation model took roughly 12 hours to complete (210 epochs). The

learned weights for the model were saved for each epoch and the Dice Coefficients were recorded to a file. After models had been trained the models with the lowest 10 Dice coefficients were identified and were evaluated using the validation data. Complete evaluation of each of these models would take approximately 6 hours, so an automated process was developed to evaluate and eliminate poor performing models based on selected scans from each bin that contained nodules that were more challenging to detect. During the bulk of the candidate generation training one entire node was used solely for this purpose, conducting two evaluations on each of the four GPUs simultaneously, while the three remaining nodes used each of their four GPUs for training segmentation models. False positive reduction training was significantly less time consuming as two or three models could be trained on each GPU. This complete training was conducted for two months, with three quarters of this time devoted entirely to the training models for candidate generation. The complete system results were assessed at the end of this period and the best performing models were selected. The evaluation of the test data was conducted on a single node using four GPUs in roughly 6 hours. The average processing time for a single scan ranged from 2 minutes to 10 minutes based on the resolution of the scanner.

The experimental design required the training of 90 separate segmentation models and 90 separate false positive reduction models. For each segmentation model approximately 10 models were trained and the best performing of these models was selected. For false positive reduction roughly 10 models were trained as well, and the best performing model was selected. Managing the training and evaluation of these models was a large and time-consuming task. Python and bash scripting were used to develop an automated big data workflow for managing these tasks simultaneously with minimal supervision.

References

1. Russell S, Norvig P. *Artificial Intelligence: A modern approach*. 3rd ed, 2011.
2. Setio AAA, Ciompi F, Litjens G, et al. Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks. IEEE transactions on medical imaging 2016;**35**(5):1160-69
3. Setio AAA, Traverso A, de Bel T, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. Medical Image Analysis 2017;**42**:1-13

APPENDIX C

Comparative Performance

Tables XII and XIII are included here for comparison of existing candidate generation and false positive reduction systems, respectively. For the task of candidate generation, only three systems have reasonable numbers of false positives per scan, with the proposed system reporting the least. Of these, only the proposed system and that of Shaukat et al. [1] were evaluated on reasonable numbers of scans. Between these two, the proposed system has a third fewer false positives per scan and small edge in sensitivity.

Table XII: Comparison of Candidate Generation Systems

	Year	Number of Scans	Sensitivity	FPPS
Proposed System	2018	888	0.9477	30.4
Shaukat et al [1]	2017	850	0.9420	45.5
Firmino et al [2]	2016	420	0.9700	N/A
LUNA (Tan et al) [3]	2016	888	0.9290	333.0
Choi & Choi [5]	2014	84	0.9790	270.8
Choi & Choi [4]	2013	58	0.9735	60.2

Shaukat et al. [1] report excellent results for both candidate generation and false positive reduction, resulting in a strong performing system overall. However, for the task of false positive reduction their results stand apart from the rest independent of dataset size. The proposed system demonstrates lower false positives per scan for the raw results, but even at 4 false positives per scan it the system still cannot outperform that of Shaukat et al. with respect to sensitivity. This reflects strongly on the methods used by Shaukat et al.: hand-crafted features with an SVM classifier. This appears to suggest strong potential for combination of SVM classifiers with deep learning models for feature generation.

Table XIII: Comparison of False Positive Reduction Systems

	Year	Number of Scans	Sensitivity	FPPS
Proposed System	2018	888	0.9421	1.79
Shaukat et al [1]	2017	850	0.9815	2.19
Firmino et al [2]	2016	420	0.9440	7.04
LUNA (Dou et al) [6]	2017	888	0.9070	4.00
Choi & Choi [5]	2014	84	0.9750	6.76
Choi & Choi [4]	2015	58	0.9789	2.27

References

1. Shaukat F, Raja G, Gooya A, et al. Fully automatic and accurate detection of lung nodules in CT images using a hybrid feature set. Med Phys. 2017;44(7):3615-3629.

2. Firmino M, Angelo G, Morais H, Dantas MR, Valentim R. Computer-aided detection (CADe) and diagnosis (CADx) system for lung cancer with likelihood of malignancy. Biomedical Engineering Online 2016;15(1):2.

3. Tan M, Deklerck R, Jansen B, Bister M, Cornelis J. A novel computer-aided lung nodule detection system for CT images. Med Phys. 2011;38(10):5630-45.

4. Choi W-J, Choi T-S. Automated pulmonary nodule detection system in computed tomography images: A hierarchical block classification approach. Entropy 2013;15(2):507-23.

5. Choi W-J, Choi T-S. Automated pulmonary nodule detection based on three-dimensional shape-based feature descriptor. Comput Methods Programs Biomed. 2014;113(1):37-54.

6. Dou Q, Chen H, Yu L, et al. Multi-level contextual 3D CNNs for false positive reduction in pulmonary nodule detection. IEEE Trans Biomed Eng. 2017 Jul;64(7):1558-1567.