Towards radiologist-level malignancy detection on chest CT scans: a comparative study of the performance of convolutional neural networks and four thoracic radiologists

Aims and objectives

The demonstration of a 20% reduction in lung cancer mortality in the USA National Lung Screening Trial (NLST) [1] and the subsequent decision by the U.S. Centers for Medicare and Medicaid Services to provide Medicare coverage for lung cancer screening has paved the way for nationwide lung cancer screening in the USA. Additionally, results of the NELSON trial also confirmed the value of low-dose CT screening with decreased mortality by 26% in high-risk men and 61% in high-risk women over a 10-year period. [2]

Lung cancer screening programs have subsequently initiated big data analysis projects on chest CTs. The National Lung Screening Trial (NLST) dataset, in particular, has longitudinal data of high-risk patients to closely monitor potentially malignant lung nodules and provide the opportunity for the development of Computer-aided Detection (CADe) in detecting lung nodules and Computer-aided Diagnosis (CADx) systems in characterizing lung nodules to assist radiologists in reporting high volumes of chest CT scans.

The purpose of this work is to evaluate the performance of a deep learning system based on convolutional neural networks in predicting the presence of malignant lung nodules on chest CT scans. We also attempt to benchmark its performance against four radiologists.

Methods and materials

Data preparation:

In this retrospective study, low-dose chest CTs were taken from the NLST dataset. 1245 CT scans were taken for training and 350 CT scans were taken for validation. Pathologically proven malignancy status of lung cancers was taken as ground-truth. Lung nodule annotations from 4 radiologists were taken from 888 CT scans from the publicly available LIDC-IDRI [3] dataset. CT scans with slice-thickness > 2.5mm were excluded to avoid partial-volume effect as recommended by Ginneken *et al* [4] and Setio *et al* [5].

Nodule detection is a volumetric detection task and hence all CT scans were resampled to isotropic voxel spacing of [1.0, 1.0, 1.0] mm in each direction to leverage the computational capacity of 3D convolutions and HU windowing was done from -1200 HU to 600 HU to visualize the lung fields effectively.

Training:

A deep learning system based on convolutional neural networks was trained to predict the malignancy status from CT scans of the chest. The deep learning system comprises of a nodule detector and a malignancy estimator. The nodule detection system was trained and validated to pick up pulmonary nodules >= 3mm on 554 CT scans from NLST and 888 CT scans from LIDC-IDRI. The malignancy estimator was trained on the 1245 CT scans and validated on 350 CT scans from NLST.

CADe (Nodule Detector):

The CADe system is an ensemble of 3 single-shot 3D Feature Pyramid Networks (FPN) [5][6] which is trained to detect lung nodules from the CT scans. The 3D FPN is built with a U-Net encoder-decoder architecture, composed of 3D convolutions, to maximize the effective receptive field and fuse multi-scale information.

Multi-scale information is essential for differentiating pulmonary nodules from vasculature present in the organs. The network takes in a 3D patch of size 128x128x128 as input and gives out 32x32x32x3x5 as output, with 3 anchor boxes of varying size limits for each network in the ensemble. During inference, the ensemble is rolled over the CT scan with 128x128x128 overlapping patches and the predicted bounding boxes are fused with non-maximum suppression to provide the final candidates for lung nodules.

The CADe system is trained on 711 CTs from LIDC-IDRI and 554 CTs from NLST with Adam optimizer and a learning rate of 0.0001, a weight decay of 0.0005 and a dropout of 0.5. Validation was done on 177 CTs from LIDC-IDRI. The data was augmented with nodules of different sizes to ensure training was not biased towards detecting small nodules.

CADx (Malignancy Estimator):

The CADx system is a leaky Noisy-OR gate [6] based on deep convolutional neural networks. The noisy-OR model operates on 96x96x96 patches from each detected nodule, fuses the information from each nodule and gives the probability of the patient being affected by lung cancer.

During training, top 5 nodule candidates, based on their nodule probabilities, are taken from the CADe system and fed to the noisy-OR model. A leakage probability is assigned to the CADx model during training to account for missed primary nodules/masses by the CADe system. The CADx model shares the same backbone as the CADe model with the convolutional layers sharing their weights to avoid over-fitting.

The CADx system was trained on 1245 CT scans and validated on 350 CT scans from NLST. During inference, all the detected nodules are considered to compute the overall malignancy risk at a scan-level.

Evaluation:

100 unseen low-dose CT scans from the validation set were chosen at random and predictions were generated from the deep learning system. Studies were randomized and presented to 4 thoracic radiologists with 2, 5, 8- and 15-years' experience to characterize the chest CT scans. The radiologists were asked to assess the probability of malignancy in the scans on a Likert scale of 1 (highly unlikely) to 5 (highly suspicious). The ROC curves were analysed for the AI and the radiologists. Post-analysis, 4 CT scans without lung nodules but marked malignant in the NLST EMR were removed from the study.
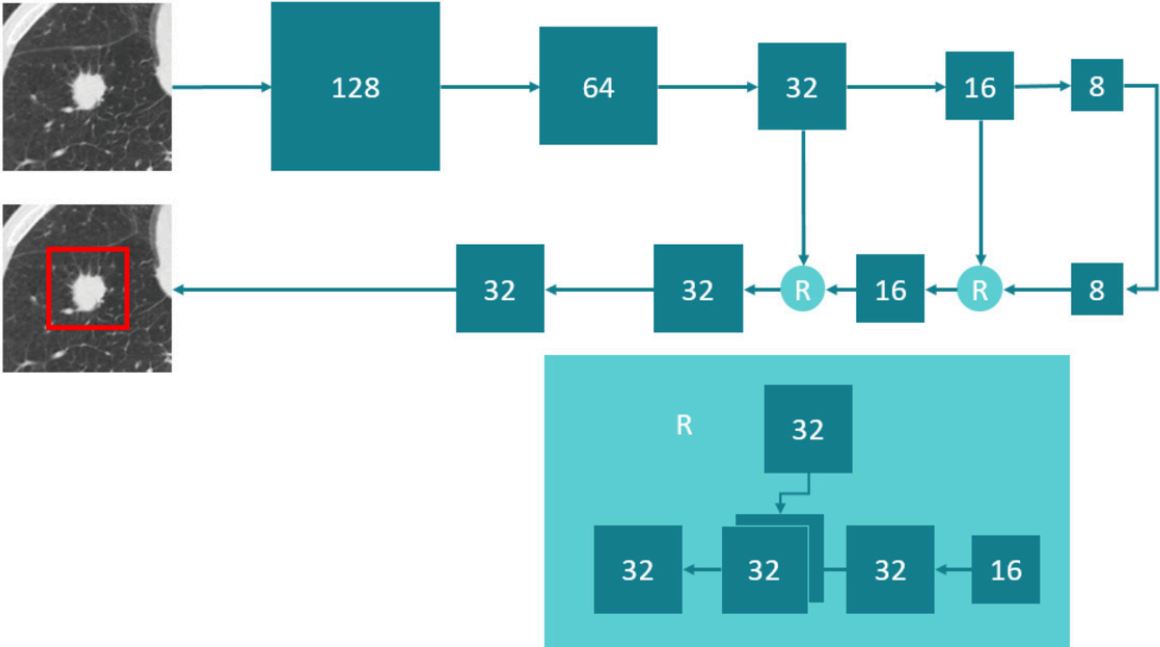
Images for this section:

Fig. 1: A Feature Pyramid Network used as nodule detector in the CADe system.
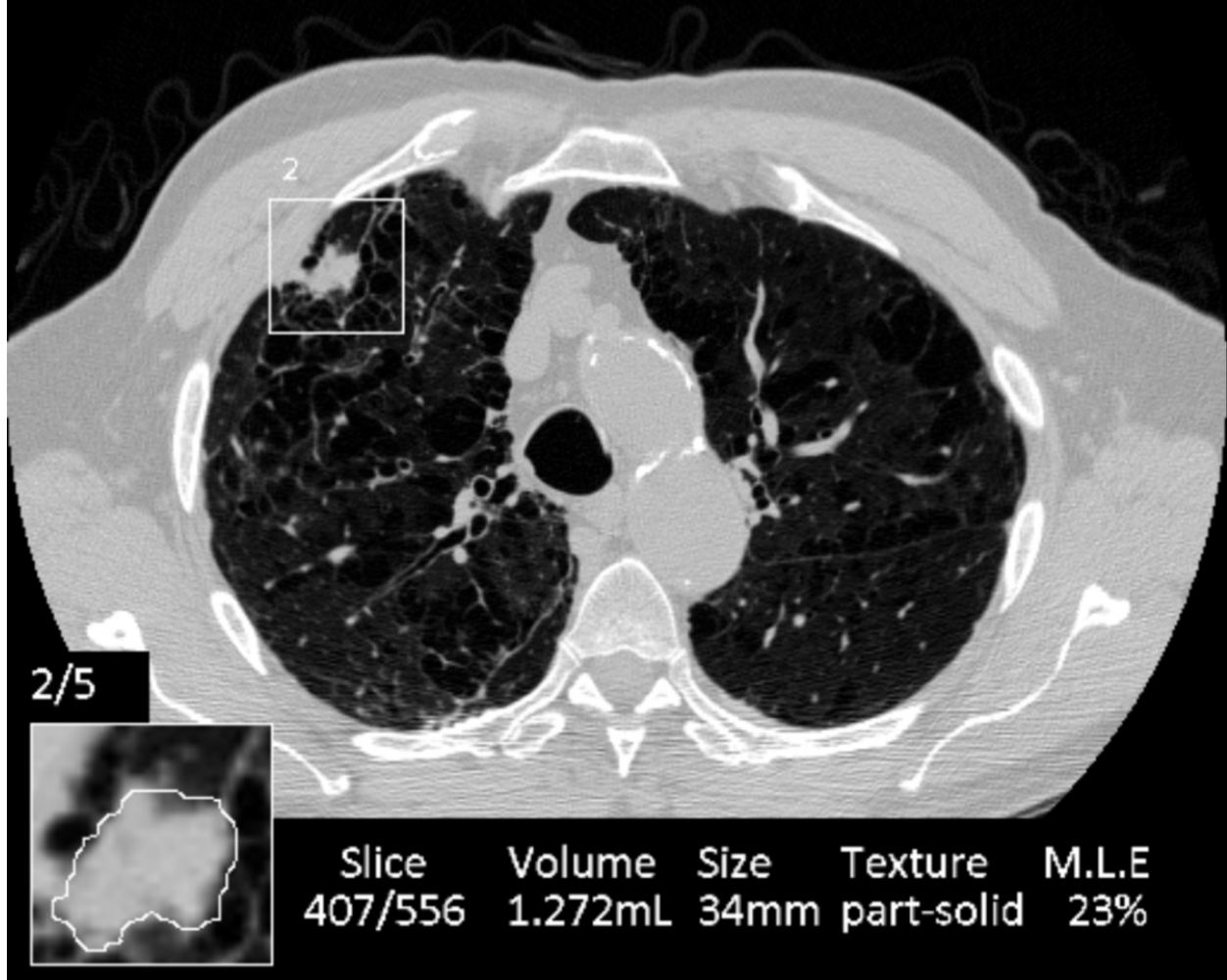
© Predible health

Fig. 2: Snapshot of a 3.4cm nodule on the right upper lobe detected by the system.

© Predible Health
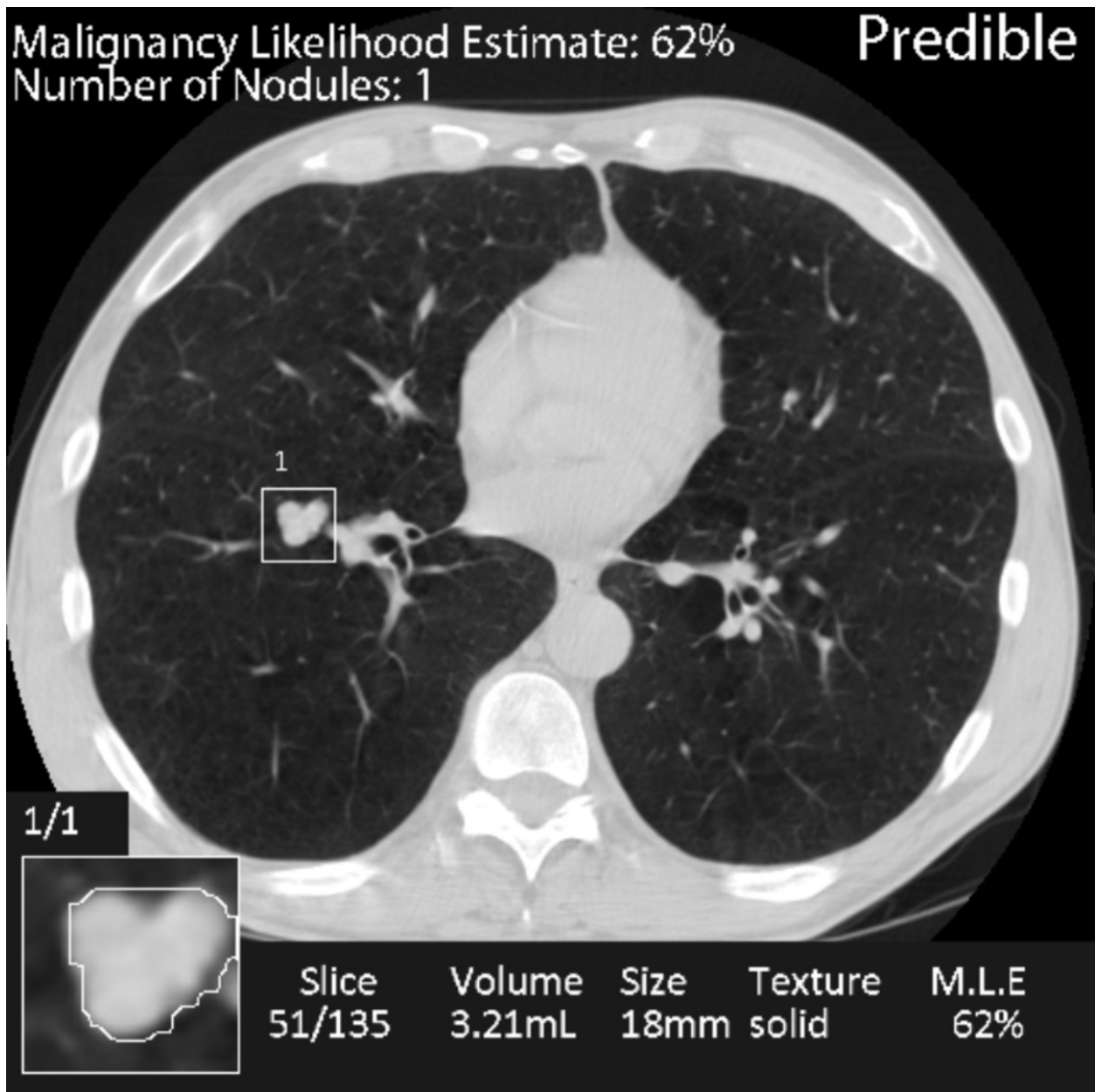
Fig. 3: A suspicious lobulated nodule detected by the system in the right lower lobe
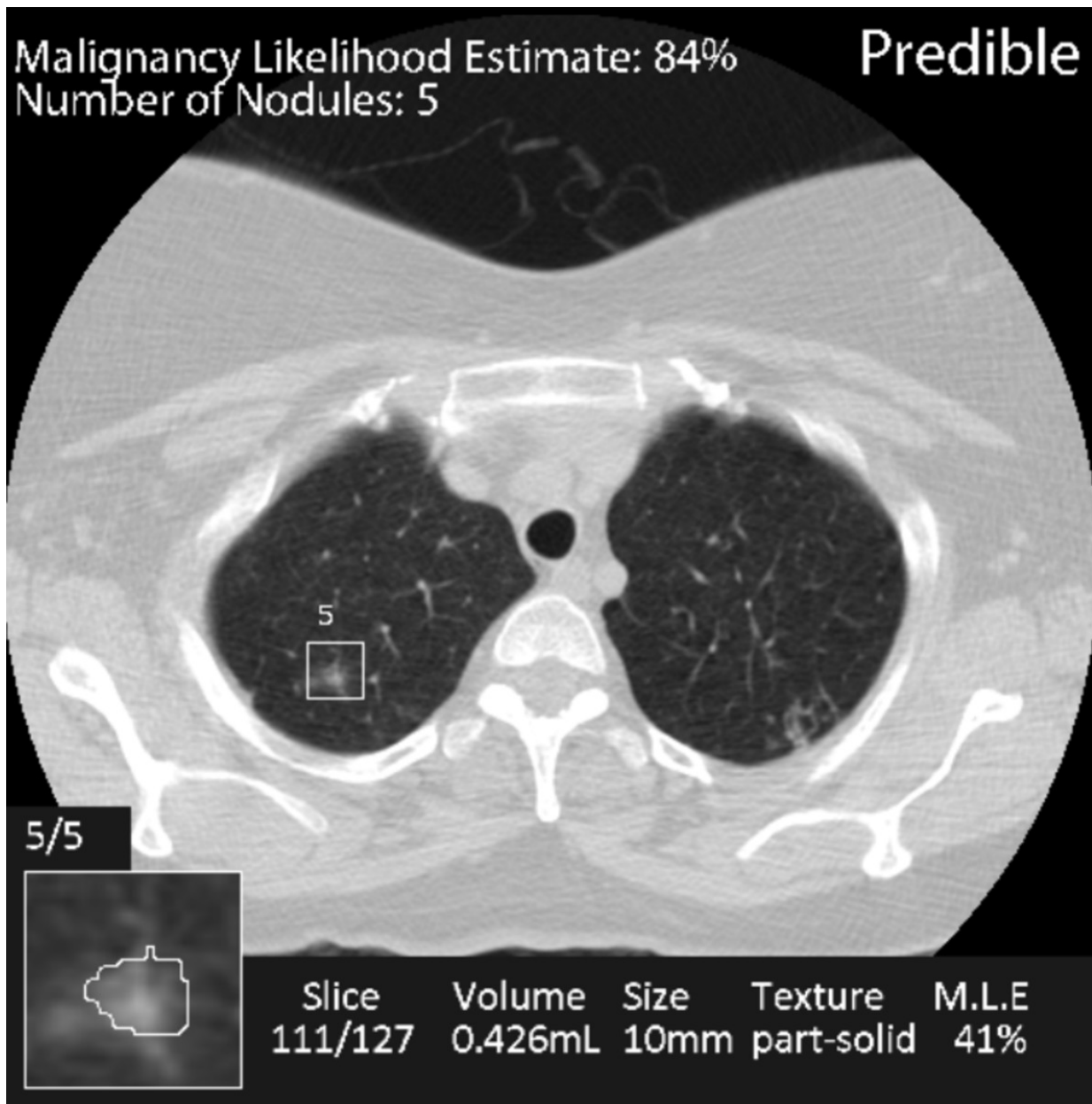
© Predible Health

Fig. 4: A suspicious part-solid nodule detected by the system in the right upper lobe

© Predible Health

Results

The results of radiologists on the Likert scales 1 and 2 were considered as negative for malignancy and 3,4 and 5 were considered to be positive for the presence of malignancy. For the AI, a predicted probability > 0.25 was considered to be positive for the presence of malignancy.

On the 96 chest CT scans reviewed by the radiologists, they had AUCs of 0.82, 0.81, 0.83 and 0.83 for predicting the risk of malignancy whereas the AI had an AUC of 0.91. Individually, radiologists' accuracy varied from 76 to 77% and AI's accuracy was 83%. The difference in the radiologist's interpretation was not found to be statistically significant as the one - way ANOVA revealed p-value is 0.77311.

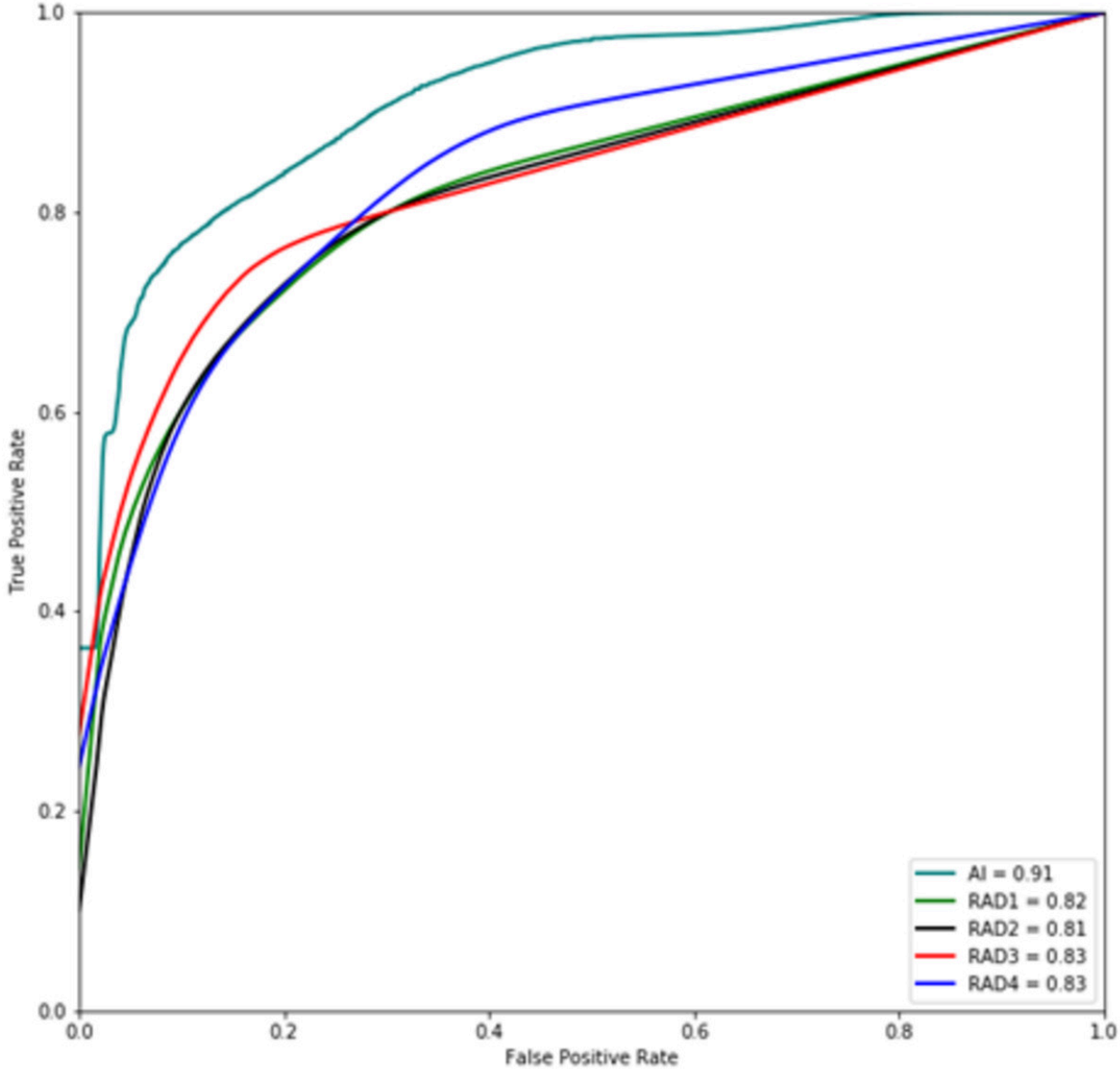| User	Accuracy	Sensitivity	Specificity
AI	83%	73%	87%
RAD1	76%	65%	86%
RAD2	76%	65%	86%
RAD3	77%	65%	88%
RAD4	76%	65%	86%

Images for this section:

Fig. 5: ROC plots for AI and the 4 radiologists in predicting the presence of malignancy in 96 low-dose Chest CT scans from the NLST dataset.
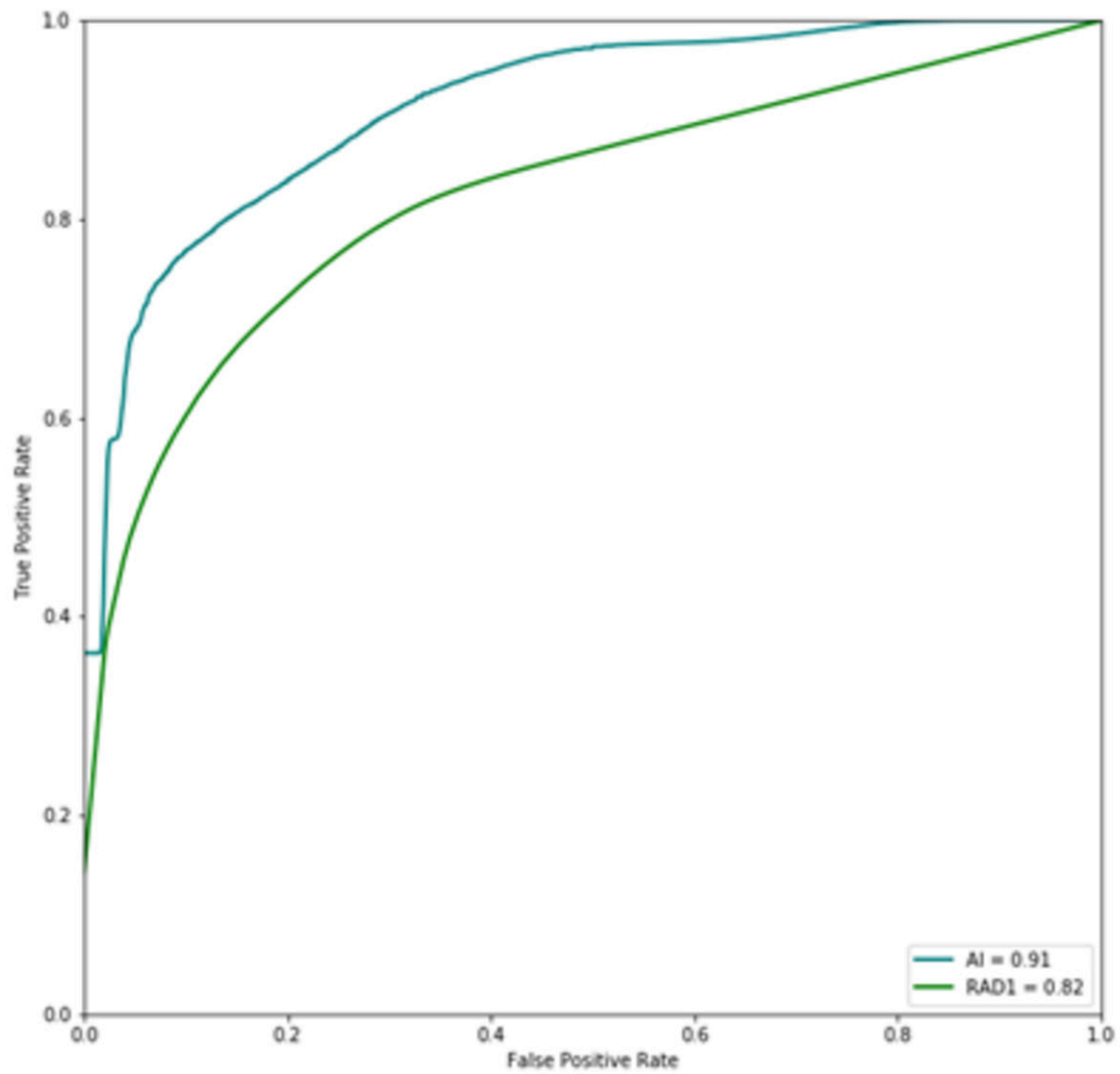
© Predible health

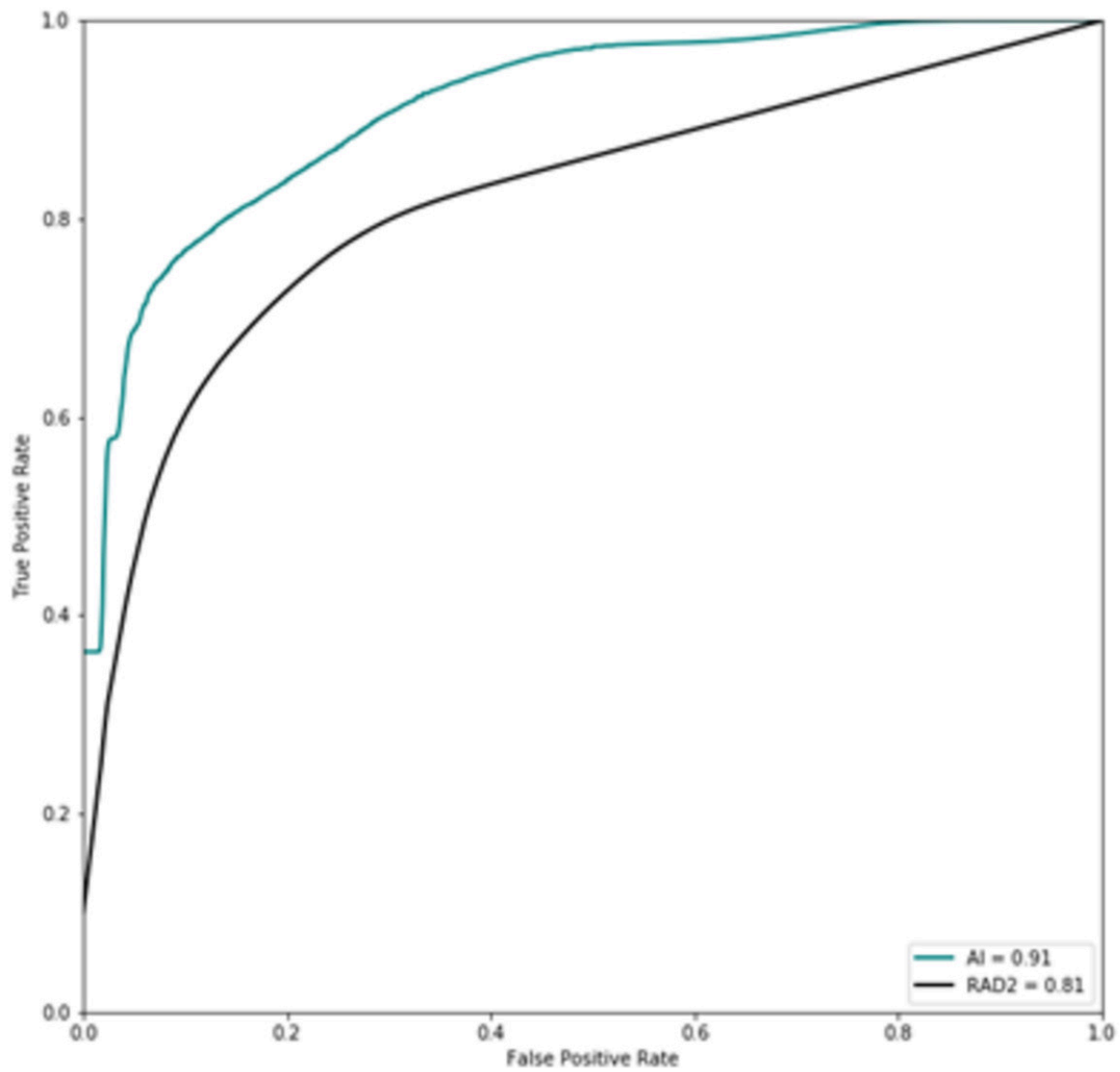**Fig. 6:** ROC plots for AI and RAD 1

© Predible health

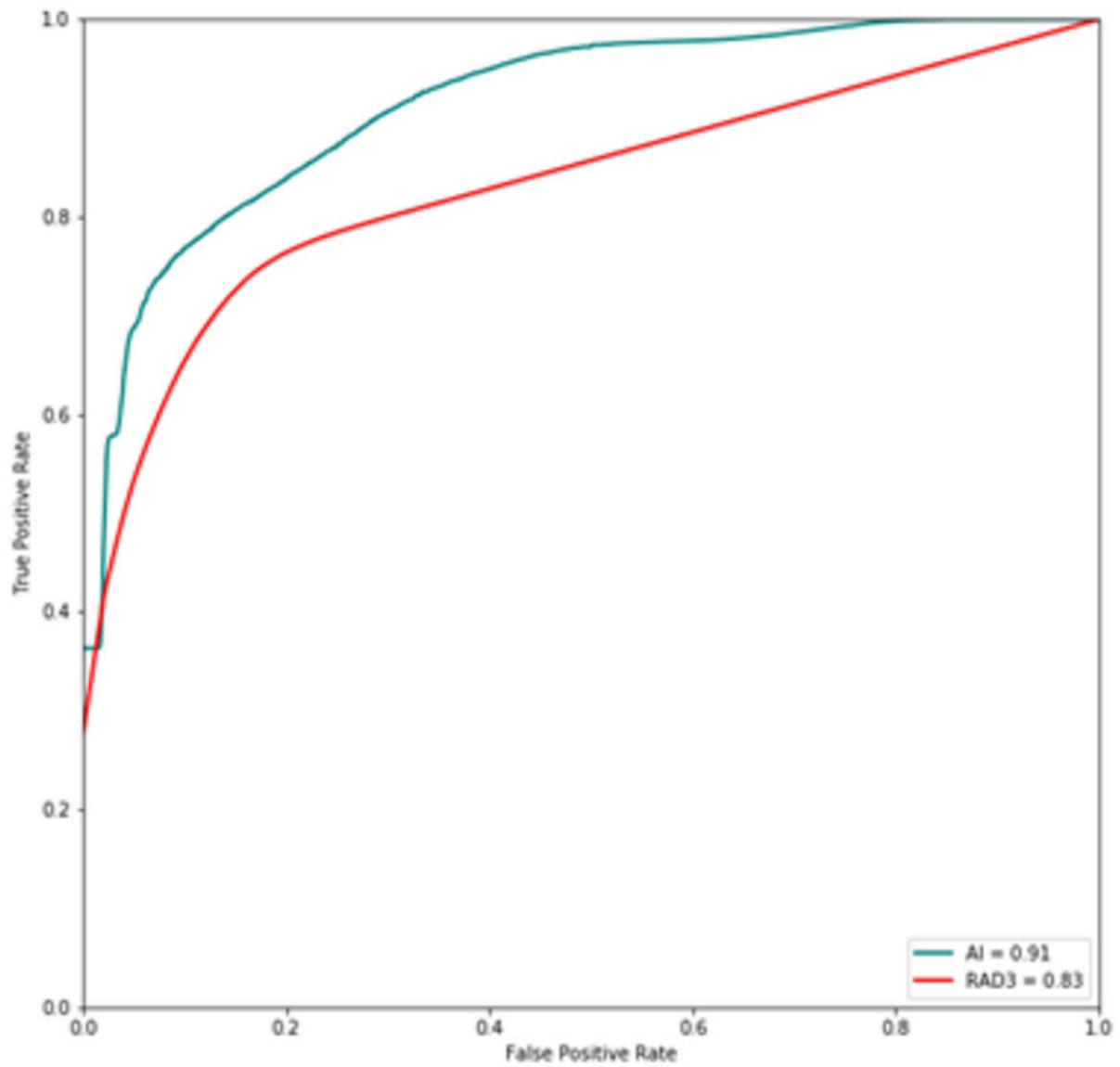**Fig. 7:** ROC plots of AI and RAD 2

© Predible health

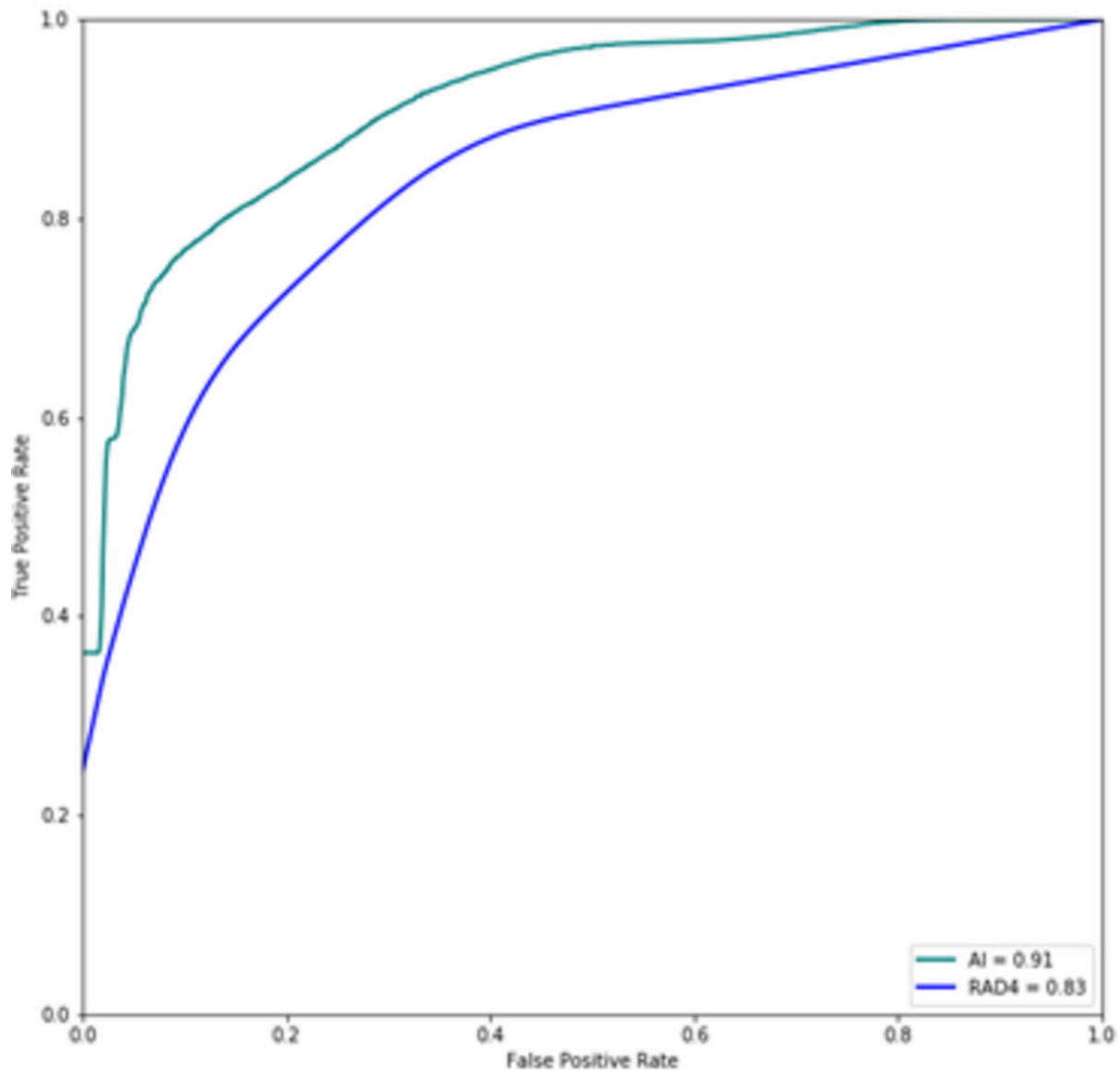**Fig. 8:** ROC plots for AI and RAD 3

© Predible health

Fig. 9: ROC plots for AI and RAD 4

© Predible health

Conclusion

The deep learning system shows better performance than experienced radiologists, individually and in aggregate, in predicting the presence of malignant nodules on the 96 CT scans obtained from the NLST dataset. The difference in the interpretation of radiologists were not found to be statistically significant.

Clinically, as low-dose CT scans are non-contrast scans, the classically described contrast enhancement characteristics for diagnosing malignant nodules cannot be used to assess the risk of malignancy in these cases.

The availability of a highly sensitive nodule characterization tool will improve the early cancer detection rates. Radiologists aided by deep learning solutions for malignancy have the potential to identify lung cancer earlier as well as reduce unnecessary biopsies.

References

1. National Lung Screening Trial Research Team, Aberle DR, Adams AM, Berg CD, et al (2011): Reduced lung-cancer screening mortality with low-dose computed tomographic screening. *New England Journal of Medicine, 365*, 395-409.

2. Koning, Harry & Aalst, Carlijn & Haaf, Kevin & Oudkerk, M. (2018). PL02.05 Effects of Volume CT Lung Cancer Screening: Mortality Results of the NELSON Randomised-Controlled Population Based Trial. *Journal of Thoracic Oncology.* 13. S185. 10.1016/j.jtho.2018.08.012.

3. Armato III, Samuel G., McLennan, Geoffrey, Bidaut, Luc, McNitt-Gray, Michael F., Meyer, Charles R., Reeves, Anthony P., … Clarke, Laurence P. (2015). Data from LIDC-IDRI. *The Cancer Imaging Archive*. http://doi.org/10.7937/K9/TCIA.2015.LO9QL9SX

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

5. Arnaud Arindra Adiyoso Setio et al (2016). Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. *CoRR, abs/1612.08012*

6. Tsung-Yi Lin and (2016). Feature Pyramid Networks for Object Detection. *CoRR, abs/1612.03144*

7. Fangzhou Liao and (2017). Evaluate the Malignancy of Pulmonary Nodules Using the 3D Deep Leaky Noisy-or Network. *CoRR, abs/1711.08324*