CLINICALLY APPLICABLE DEEP LEARNING STRATEGY FOR PULMONARY NODULE RISK PREDICTION: INSIGHTS INTO HONORS

INDEX OF SUPPLEMENTARY MATERIALS

SUPPLEMENTARY MATERIALS AND METHODS. . . . . Page 2

SUPPLEMENTARY TABLES S1, 2, 3, 4, 5 and 6. . . . . Page 4

SUPPLEMENTARY FIGURES S1, 2, 3, 4 and 5. . . . . .Page 8

SUPPLEMENTARY MATERIALS AND METHODS

Inclusion and exclusion criteria

The inclusion criteria of Jinling Hospital dataset and multi-center dataset were as follows: (1) lesions manifested as pulmonary nodules on CT scans; (2) nodules measured between 4 and 30 mm in diameter on CT; (3) underwent surgical resection or biopsy within 15 days after CT scans. The exclusion criteria of Jinling Hospital dataset and multi-center dataset were as follows: (1) poor imaging quality due to motion artifact; (2) the nodule was pathologically proved metastatic tumor or intermediate tumor; (3) patients have accepted therapies before CT scans.

NLST recruited individuals who were aged 55–74 years and at high-risk for lung cancer from 33 medical centers in the United States for lung cancer screening[1]. The mPNs were proved pathologically while bPNs were proved pathologically or if no change was found over the 2-year follow-up period. Images scanned by Siemens and with reconstruction thickness $\leq$ 2.5mm were included for further analysis.

Pulmonary nodule detection network

We developed a novel two-stage nodule detector network that integrated both image and feature pyramid for nodule detection. Firstly, we extended Feature Pyramid Network (FPN)[2] to 3D as our nodule proposal network to avoid the detail information missing in upper layers of the network. Given one CT image as input, FPN generated rich semantics feature maps at different resolution by fusing both high and low features, enabling nodule detection in the proper resolution. Secondly, an image pyramid was designed for further false positive reduction. Due to lack of knowledge of object's size, traditional image pyramid consisted of a set of images with different scales[3]. In conclusion, we used FPN to produce rough size information of the proposal and detect the proposals in a proper resolution.

Feature visualization of FGP-Net

Given the black box property of deep learning, we further conducted a two-way feature interpretation to explore whether FGP-NET learned solid and effective features. Firstly, we applied T-distributed Stochastic Neighborhood Embedding (t-SNE), a popular nonlinear dimensionality reduction method for exploring and visualizing high-dimensional data, to visualize the global feature in 2 dimensions. The perplexity was set to 40, learning rate to 200 and 1000 iterations.

Secondly, we generated probability heat-map to visualize the local features. We extracted the feature maps just before DFL modules with two different scales as probability plot, and normalized them by using gamma transformation. The probability graph was then mapped to a color scheme, and overlaid with the original input images.

Comparison of HONORS with radiologists

The FGPNet was analyzed at the HSpe point which valued the specificity as 99%. The radiologists rated the nodules from 1 to 4 (4 denotes the highest malignancy probability). The corresponding sets were defined as S1, S2, S3 and S4, respectively. When assessing the radiologists, S4 were defined as malignant and S1, S2, S3 were defined as benign. We assumed that the power of radiologists was 75% when they scored the nodule as S4 or S1. To keep assistance with FGPNet, we determine that only if the nodule was scored by over 75% of radiologists, it would be reclassified as S4 or S1, otherwise it would be reclassified as S2 or S3. After recalculating the score rated by the radiologists, we evaluated the performance of radiologists by sensitivity, specificity.

SUPPLEMENTARY TABLES

Table S1: The scanning and reconstruction parameters of CT images from different centers.

Site	Jinling Hospital	Wuxi People's Hospital	Southeast University Zhongda Hospital	Second Affiliated Hospital of Nantong University	NLST
CT scanner	SOMATOM Definition Flash/ SOMATOM Definition/ SOMATOM Emotion/ SOMATOM Perspective	SOMATOM Definition/ SOMATOM Definition Flash	Discovery CT750 HD/ SOMATOM Sensation 64/ Revolution CT	SOMATOM Definition Flash/ Sensation 16/ iCT 256/ Perspective/ SOMATOM Force Brilliance 64	Emotion 16/ Sensation 16/ Sensation 4/ Volume Zoom
Tube voltage, kVp	120/130	120	120	100/110/120/130	120/130/140
Tube current, mA	62-663	42-682	56-569	127-680	37.5-160
Detector collimation	128*0.6/64*0.6/6*1.0	128*0.6/64*0.6	128*0.625/256*0.625	-	-
Gantry speed	0.5/0.6/0.8	0.5	-	-	-
Slice thickness, mm	1/1.25	1	0.75/1/1.25/1.5	1	1/2
Kernel	I70f/ I50s/ B50f/ B50s	B70f	LUNG/ STANDARD /B31f/ B80f	C/ YB/ B70f/ B60f /B70s	B50f/ B50s/ B60/ B60s/ B70f
Matrix	512*512	512*512	512*512	512*512	-

Abbreviations: CT, computed tomography; NLST, National Lung Cancer Screening Trial.

Table S2: The pathological types of nodules in training, validation and test sets.

Pathological type	Training set (n = 1606)	Validation set (n = 200)	JLH test set (n = 100)	NLST test set (n =200)	Multi-center set (n =242)
Malignant					
Total, No (%)	845 (100)	86 (100)	75 (100)	22 (100)	187 (100)
Adenocarcinoma	769 (91.0)	77 (89.5)	71 (94.7)	19 (86.4)	172 (92.0)
Squamous cell carcinoma	47 (5.6)	6 (7.0)	2 (2.7)	1 (4.5)	8 (4.3)
Other malignant types[a]	29 (3.4)	3 (3.5)	2 (2.7)	2 (9.1)	7 (3.7)
Benign					
Total, No (%)	761 (100)	114 (100)	25 (100)	178 (100)	55 (100)
Chronic inflammation	83 (10.9)	3 (2.6)	3 (12.0)	0 (0.0)	14 (25.5)
Granuloma	72 (9.5)	7 (6.1)	10 (40.0)	0 (0.0)	20 (36.4)
Hamartoma	46 (6.0)	6 (5.3)	6 (24.0)	0 (0.0)	4 (7.3)
Sclerosing pneumocytoma	14 (1.8)	2 (1.8)	0 (0.0)	0 (0.0)	7 (12.7)
Inflammatory pseudotumor	22 (2.9)	2 (1.8)	3 (12.0)	0 (0.0)	0 (0.0)
Fungus infection	13 (1.7)	2 (1.8)	0 (0.0)	0 (0.0)	4 (7.3)
Atypical adenomatoid hyperplasia	15 (2.0)	0 (0.0)	1 (4.0)	0 (0.0)	0 (0.0)
Other benign types[b]	36 (4.7)	3 (2.6)	2 (8.0)	0 (0.0)	6 (10.9)
Unknown[c]	460 (60.4)	89 (78.1)	0 (0.0)	178 (100.0)	0 (0.0)

[a]Other malignant types were including small cell carcinoma, large cell neuroendocrine carcinoma, carcinoid tumor, adenosquamous carcinoma and acinar cell carcinoma.

[b]Other benign types were including PEComa, adenoleiomyoma, tumorlet, lymphonodus and so on.

[c]NLST did not provide the pathological results of the benign nodules, thus the benign nodules of the NLST dataset in the table were classified as unknown.

Abbreviations: JLH, Jinling hospital; NLST, National Lung Cancer Screening Trial.

Table S3: The characteristics of nodules in training, validation and test sets.

Characteristic	Training set (n = 1606)			Validation set (n =200)			JLH test set (n =100)			NLST test set (n =200)			Multi_center set (n =242)		
	Benign	Malignant	p	Benign	Malignant	p	Benign	Malignant	p	Benign	Malignant	p	Benign	Malignant	p
Longest axial diameter, Mean (SD), cm	0.87 (0.57)	1.62 (0.65)	< 0.001[a]	0.75 (0.50)	1.61 (0.57)	<0.001[a]	1.51 (0.43)	1.70 (0.52)	0.11[a]	0.59 (0.32)	0.68 (0.35)	0.241[a]	1.85 (0.75)	2.06 (0.71)	0.07[a]
Attenuation pattern															
Total, No. (%)	761(100)	845(100)		114(100)	86(100)		25(100)	75(100)		178(100)	22(100)		55(100)	187(100)	
Non solid	156(20.5)	152(18.0)	< 0.001	16(14.0)	10(11.6)	< 0.001	1 (4.0.)	14 (18.7)	< 0.001	41 (23.0)	0 (0.0)	< 0.001	4(7.3)	39 (20.9)	< 0.001
Part solid	51(6.7)	392(46.4)	[b]	6 (5.3)	41(47.7)	[b]	3 (12.0)	37 (49.3)	[b]	5 (2.8)	8 (36.4)	[b]	1(1.8)	46 (24.6)	[b]
Solid	554(72.8)	301(35.6)		92(80.7)	35 (40.7)		21 (84.0)	24 (32.0)		132(74.2)	14(63.6)		50(90.9)	102 (54.5)	
Location															
Total, No. (%)	761(100)	845(100)		114(100)	86(100)		25(100)	75(100)		178(100)	22(100)		55	187	
RUL	196 (25.8)	309 (36.6)	< 0.001	27 (23.7)	28 (32.6)	0.29 [b]	7 (28.0)	22 (29.3)	0.33 [b]	40 (22.5)	7 (31.8)	0.04 [b]	16(29.1)	76 (40.6)	0.09 [b]
RML	85(11.2)	59(7.0)	[b]	11 (9.6)	4 (4.7)		1 (4.0)	6 (8.0)		20 (11.2)	0 (0.0)		4(7.3)	11 (5.9)	
RLL	169(22.2)	143(16.9)		23(20.2)	22 (25.6)		4 (16.0)	11(14.7)		42(23.6)	5 (22.7)		14 (25.5)	25 (13.4)	
LUL	172 (22.6)	213 (25.2)		28(24.6)	19 (22.1)		4 (16.0)	23 (30.7)		29(16.3)	8 (36.4)		10(18.2)	51(27.3)	
LLL	139(18.3)	121(14.3)		25(21.9)	13(15.1)		9 (36.0)	13 (17.3)		47 (26.4)	2 (9.1)		11 (20.0)	24(12.8)	

[a] This P value was calculated using independent-sample t test or Mann-Whitney U test.

[b] This P value was calculated using Chi-Squared test or Fisher's exact test.

Abbreviations: JLH, Jinling hospital; LLL, left lower lung; LUL, left upper lung; NLST, National Lung Cancer Screening Trial; RLL, right lower lung; RML, right middle lung; RUL, right upper lung; SD, standard deviation.

Table S4: The average performance of 126 radiologists and three groups with different experience.

Group	Sensitivity	Specificity	Accuracy	PPV	NPV
Average radiologist	72.2%	71.7%	72.1%	88.8%	50.2%
Resident	71.2%	68.5%	70.5%	87.5%	48.3%
Fellow	71.4%	73.1%	71.9%	89.1%	49.6%
Consultant	74.0%	73.4%	73.8%	89.7%	52.7%

Abbreviations: NPV, negative predictive value; PPV, positive predictive value.

Table S5: The performance of the HONORS in the three-step way in the screened nodules.

Scenario		Dataset	Operating point	AUC	Sensitivity	Specificity	NPV
Screen	Step-1	NLST (n=200)	HSen	0.963	95.5%	72.5%	99.2%
	Step-2	NLST_ subset (n=70)	Hspe	0.946	33.3%	95.9%	77.0%
	Step-3	NLST_ subset (n=61)	Youden	0.965	100.0%	76.6%	100.0%

Abbreviations: AUC, area under the curve; Hsen, high sensitivity; Hspe, high specificity; NLST, National Lung Cancer Screening Trial; NPV, negative predictive value.

Table S6: The performance of the HONORS in the three-step way in the incidentally detected nodules.

Scenario		Dataset	Operating point	AUC	Sensitivity	Specificity	PPV
Diagnosis	Step-1	Multi (n=242)	Hsen	0.855	100.0%	5.5%	78.2%
	Step-2	Multi _ subset (n=239)	Hspe	0.846	41.7%	94.2%	96.3%
	Step-3	Multi_ subset (n=158)	Youden	0.797	93.6%	46.9%	79.7%

Abbreviations: AUC, area under the curve; Hsen, high sensitivity; Hspe, high specificity; PPV, positive predictive value.

SUPPLEMENTARY FIGURES

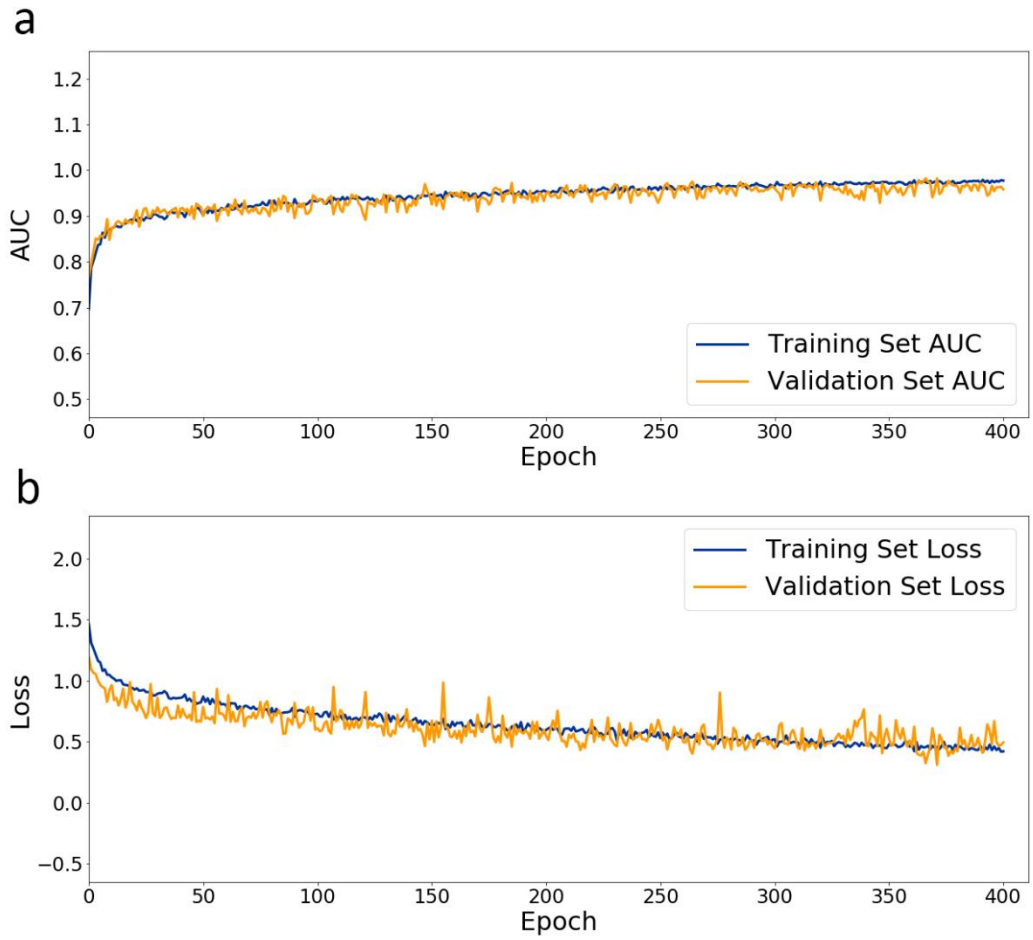


Figure S1: Plot showing performance in the training set and validation set. Plot showing the performance of pulmonary nodule risk stratification on CT images in the training and validation set using FGP-NET. a) AUC is plotted against the training step, and b) cross-entropy loss is plotted against the training step. The validation set AUC and loss show good performance. For AUC, the validation set curve converges to 95.8% (97.7% for the training process); for the loss function, the validation set curve approaches 0·49 (0·42 for the training process).

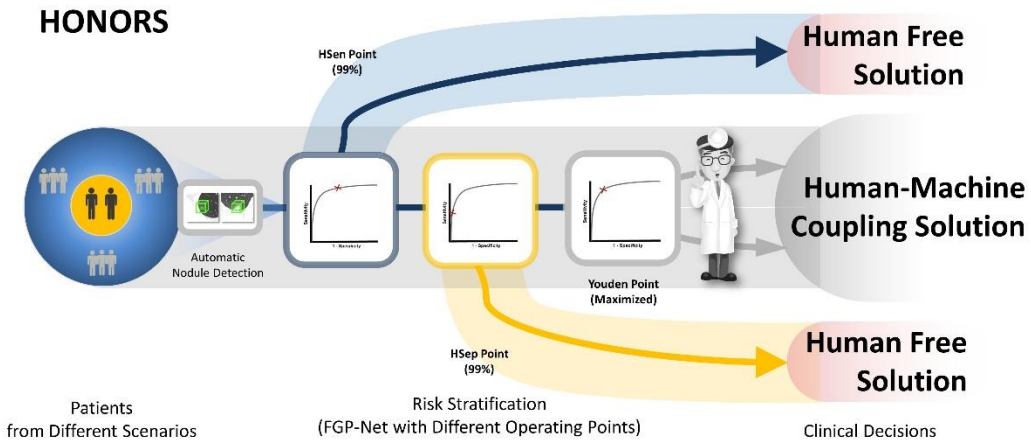


Figure S2: Structure of the three-step of HONORS. An additional three-step way to realize the HONORS. Regardless of the source of nodules, we proposed a three step to stratify them. The first step is the blue translucent paths which focus on sensitivity to stratify the benign nodules with high precision. The second step is the yellow translucent paths which focus on specificity to stratify the lung cancer with high precision. Both of the two steps are “Human Free”. Further, under a Youden point in the third step, the remaining ‘ambiguous nodules’ are differentiated into benign and malignant ones by FGP-NET but require final confirmation by physicians (radiologists), which indicated a “Human-Machine Coupling Solution” mode.

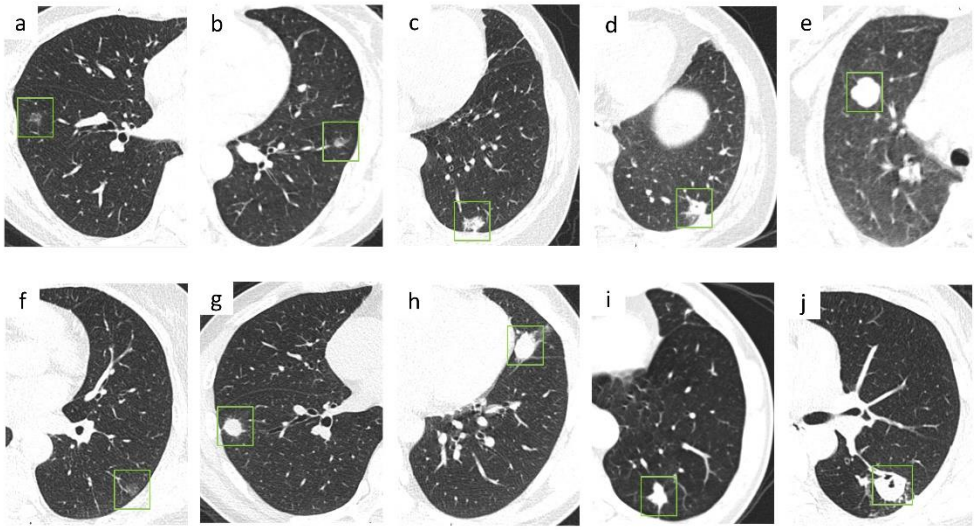
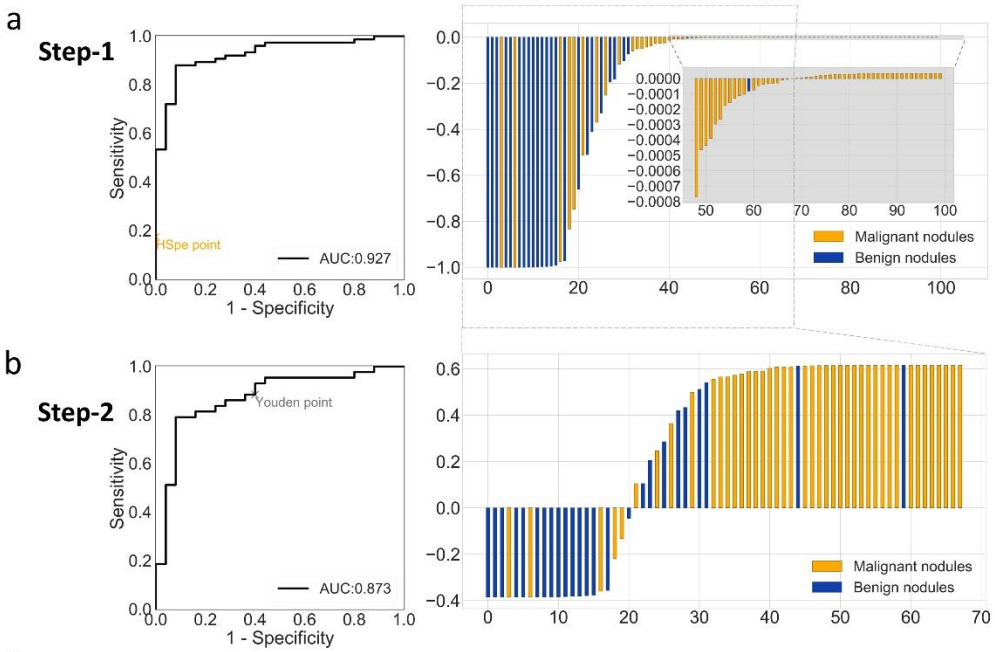


Figure S3: Nodules with inconsistent diagnosis by the radiologist' majority opinion and FGP-NET. A total of unique 18 inconsistent nodules were misdiagnosed by radiologists' majority opinion and FGP-NET, accounting for 9 each. Typical cases of them are presented. a-e) nodules misdiagnosed by radiologists. All of them were pathologically proved to be adenocarcinoma. f-j) nodules misdiagnosed by FGP-NET. f) chronic inflammation. g) inflammatory pseudotumor. h-j) granuloma. Most of them manifested as solid nodules on CT.



c

Performance by the HONORS compared with radiologists in the routine diagnostic scenario

Scenario	Steps	Dataset	Operating point	Sensitivity	Specificity	PPV
Diagnostic FGP-NET	Step-1	JLH (n=100)	Hspe point	42.7%	100.0%	100.0%
	Step-2	JLH_ ambiguous (n=68)	Youden point	88.4%	64.0%	80.9%
Diagnostic Radiologists	Step-1	JLH (n=100)	S1,S2,S3 vs S4	14.7%	96.0%	91.7%
	Step-2	JLH_ ambiguous (n=88)	S1,S2 vs S3	78.1%	91.7%	96.2%

Figure S4: Performance of HONORS in diagnostic scenario using JLH validation set. a) Each line represents the relative malignant score (y-axis) of one nodule in a) and x-axis represents the nodule index, and nodules with scores greater than 0.999965 were directly considered as malignant nodules. b) Each line represents the corresponding malignant score that extracted by Youden point of 0.384642. Nodules with scores greater than zero were predicted as malignant nodules and benign ones for the rest. c) Corresponding statistics used to evaluate the performance of HONORS under diagnostic scenario, including AUC, sensitivity, specificity and PPV.

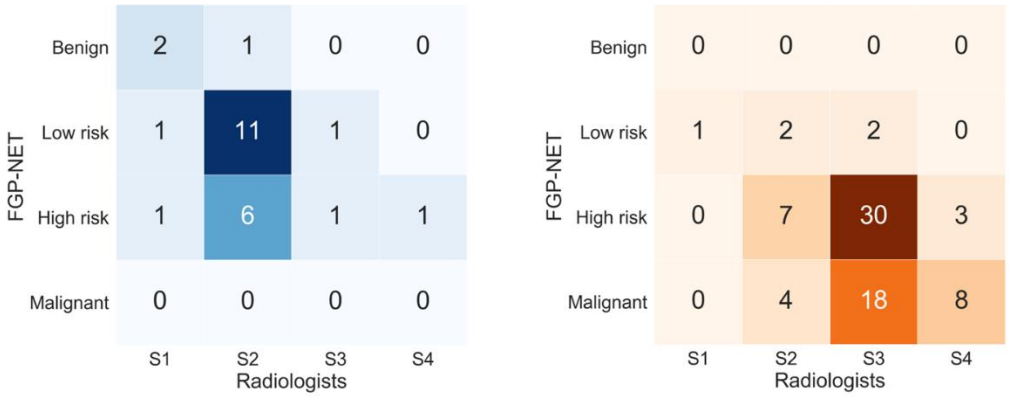


Figure S5: Confusion matrix showing the frequency of FGP-NET prediction with respect to the 126 radiologists on JLH dataset. The ground truth of nodules in left matrix was benign and the ground truth of nodules in right matrix was malignant. FGP-NET diagnosed 30 nodules as lung cancer and made no mistake; while radiologists diagnosed 12 nodules as lung cancer with one out of them misdiagnosed.

Supplementary references

1. Gatsonis, C.A. & Natl Lung Screening Trial Res, T. The National Lung Screening Trial: Overview and Study Design. *Radiology* **258**, 243-253 (2011).

2. Lin, T.-Y.*, et al.* Feature pyramid networks for object detection. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2117-2125 (2017).

3. Adelson, E.H., Anderson, C.H., Bergen, J.R., Burt, P.J. & Ogden, J.M. Pyramid methods in image processing. *RCA engineer* **29**, 33-41 (1984).