SUPPLEMENTARY MATERIAL

SUPPLEMENTARY FIGURES

Algorithm 1 Auto Film Layout

```
1:  Images ← dicomImageSet
2:  Images ← lungFieldSegmentation
3:  Nodules ← predictedResults
4:  Nodule ← Nodules.sort(importance)[0]
5:  Image ← Images[Nodule.sliceId]
6:  procedure AUTOLAYOUT(Images, Image, Nodule)
7:      layout = []
8:      Image.renderNoduleAsRect(Nodule)
9:      Image.resizeToNodule(Nodule)
10:     layout.push(Image)
11:     layout.push(Image.drawMeasurement())
12:     layout.push(Image.mediastinalwindow())
13:     layout.push(Image.renderMPR())
14:     for each image i in Images do
15:         if index%gap ≠ 0 then
16:             continue
17:         end if
18:         if layout.length ≥ 35 then
19:             break
20:         end if
21:         layout.push(i)
22:     end for
23:     for each image i in Images do
24:         if index%gap ≠ 0 then
25:             continue
26:         end if
27:         if layout.length ≥ upperLimit then
28:             break
29:         end if
30:         layout.push(i.mediastinalWindow())
31:     end for
32:     Return generatePrintable(layout)
33: end procedure
```

Figure S1. The outline logic of the programming language of the auto film layout program. Line 1-5 prepare data and resources needed for applying our program. Line 1 loads all DICOM images to memory; Line 2 filters out images that do not belong to lung field; Line 3 applies the AI model to detect all possible nodules, and Line 4 sorts detected nodules by importance defined by 1 operator; Line 5 extracts the most important image from the image sets by the most critical nodule. After preparation, the critical image, critical nodule and image set are fed to the program as parameters. Lines 6-33 are the actual logic of the program. Lines 7-13 define the overall layout and generate the first row of the layout, which corresponds to the most important nodule with its associated image for both the patient and doctor to review. The row will include a lung window image, a lung window with measurements indicated, a mediastinal window image and two images generated by applying the MPR algorithm. After the first row, the rest of the program from line 14 to 33 are two iterations that fill the layout with 30 lung window images followed by mediastinal windows for the rest. In each iteration, the program will automatically distribute images equally to fulfill the total images needed in the layout, and the inner logic can be found in lines 15-16 and 24-25.

Algorithm 1 Generate CT scan structured report

```
1:  Images ← dicomImageSet
2:  Nodules ← predictedResults
3:  procedure GENERATEREPORT(Images, Nodules)
4:      Nodules.sort(importance)
5:      report = []
6:      for each nodule i in Nodules do
7:          image = Images[n.sliceId]
8:          image.renderNoduleAsRect()
9:          image.resizeToNodule()
10:         report.push([image, n])
11:     end for
12:     report ← ImageFindings
13:     report ← DiagnosticImpression
14:     report ← PatientInformation
15:     report ← HospitalInformation
16:     report.generate()
17:     Return report
18: end procedure
```

Figure S2. The outline logic of the programming language of the auto structured report generating program. Lines 1-2 prepare data for the report generation process. Line 1 loads all DICOM images to memory, and line 2 performs the deep learning algorithm on the image set to detect all possible nodules. Then, the results are taken as parameters in the main function in lines 3-18. Line 4 in the main procedure will sort all nodules by their corresponding importance. Lines 5-11 will insert each nodule with its associated image in the report. The iteration starting from line 6 to line 11 iterates through a limited number of nodules (predefined by operator), enlarges the image with its nodule position, and then inserts the image into the report. Lines 12-16 fill out the findings, impression, patient and hospital information to the report's data structure. Lines 16-17 generate the report and output in PDF format for the patients and doctors to review.

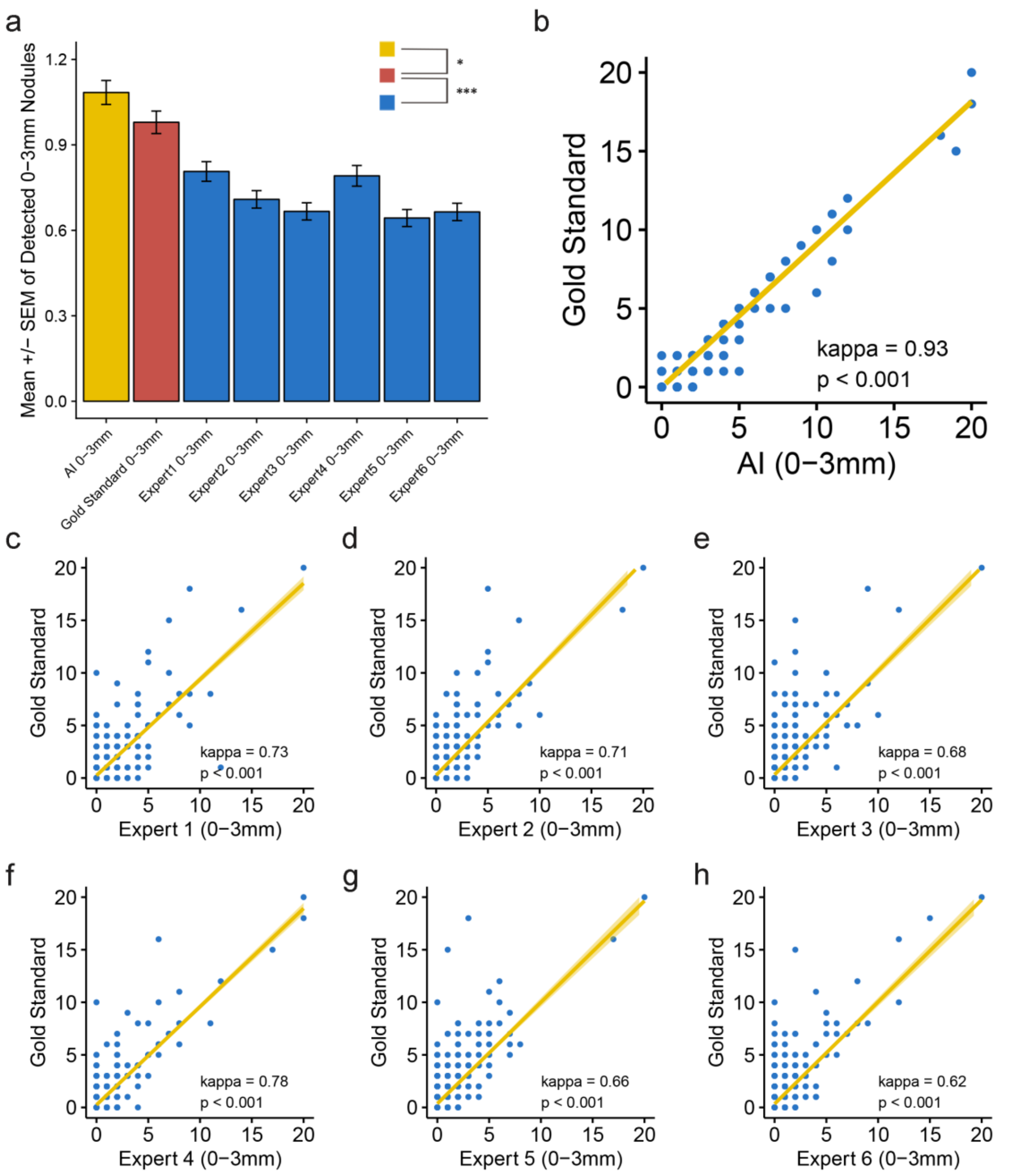

Figure S3. Consistency analysis among AI, human experts and the gold standard in detecting lung nodules with a size of 0-3 mm. Using the gold standard as a reference, (a) concluded that differences existed in all pairwise Wilcoxon tests (p<0 05 for AI and p<0 001 for human experts). (b-h) demonstrated that both AI and human experts were highly significantly consistent with the gold standard (kappa coefficient range from 0 62-0 78, p<0 001) when detecting 0-3 mm lung nodules. The horizontal and vertical coordinates for (b-h) indicate the detected nodule number. Statistical significance is labeled as follows: for <0 1, * for <0 05, ** for <0 01, *** for <0 005 and NS for no significance.

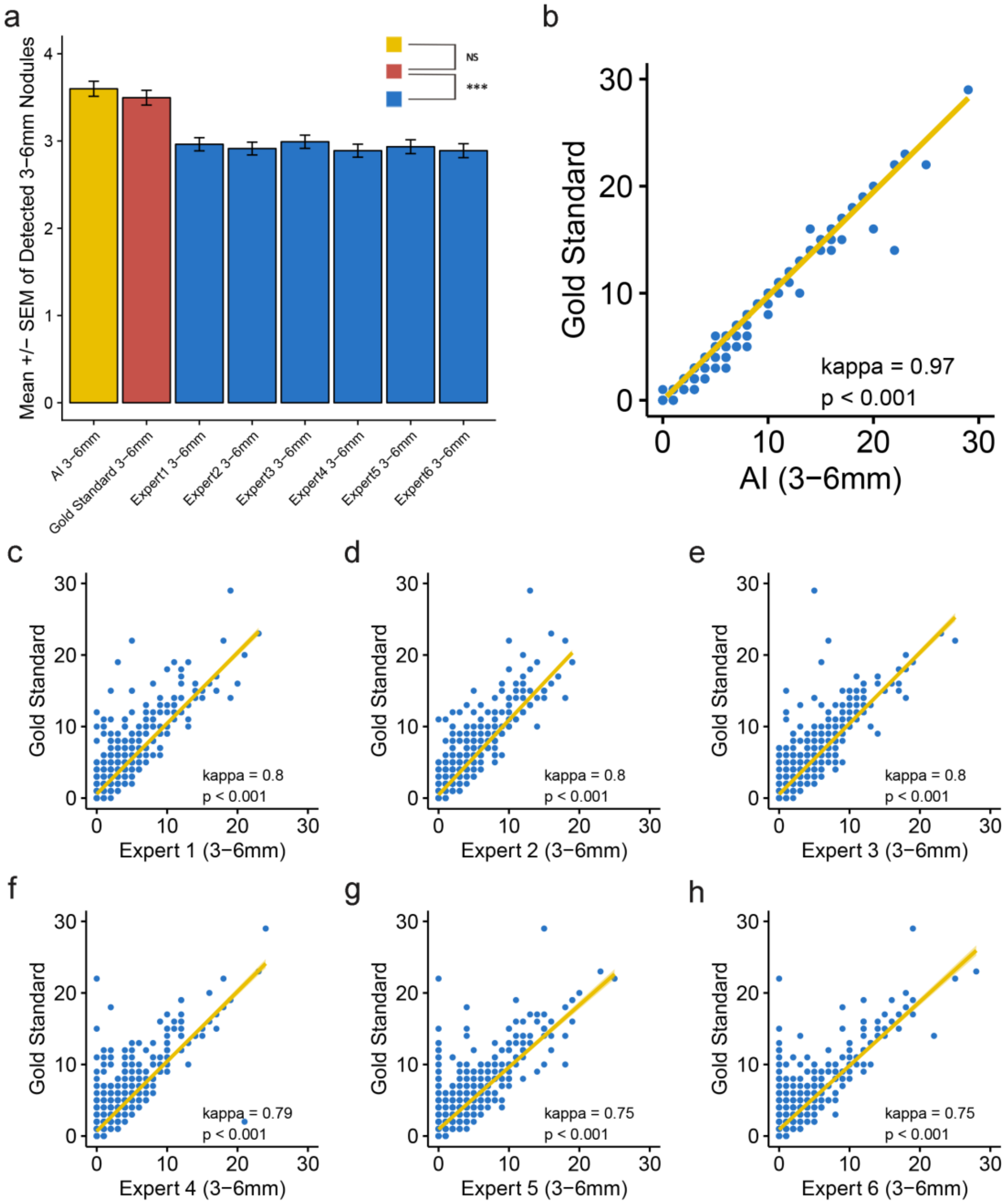

Figure S4. Consistency analysis among AI, human experts and the gold standard in detecting lung nodules with a size of 3-6 mm. Using the gold standard as reference, (a) concluded that differences existed in all pairwise Wilcoxon tests except for AI(p=0·28 for AI and p<0·001 for human experts). (b-h) demonstrated that both AI and human experts were highly significantly consistent with the gold standard (kappa coefficient range from 0·75-0·8, p<0·001) when detecting 3-6 mm lung nodules. The horizontal and vertical coordinates for (b-h) indicate the detected nodule number. Statistical significance is labeled as follows: for <0·1, * for <0·05, ** for <0·01, *** for <0·005 and NS for no significance.

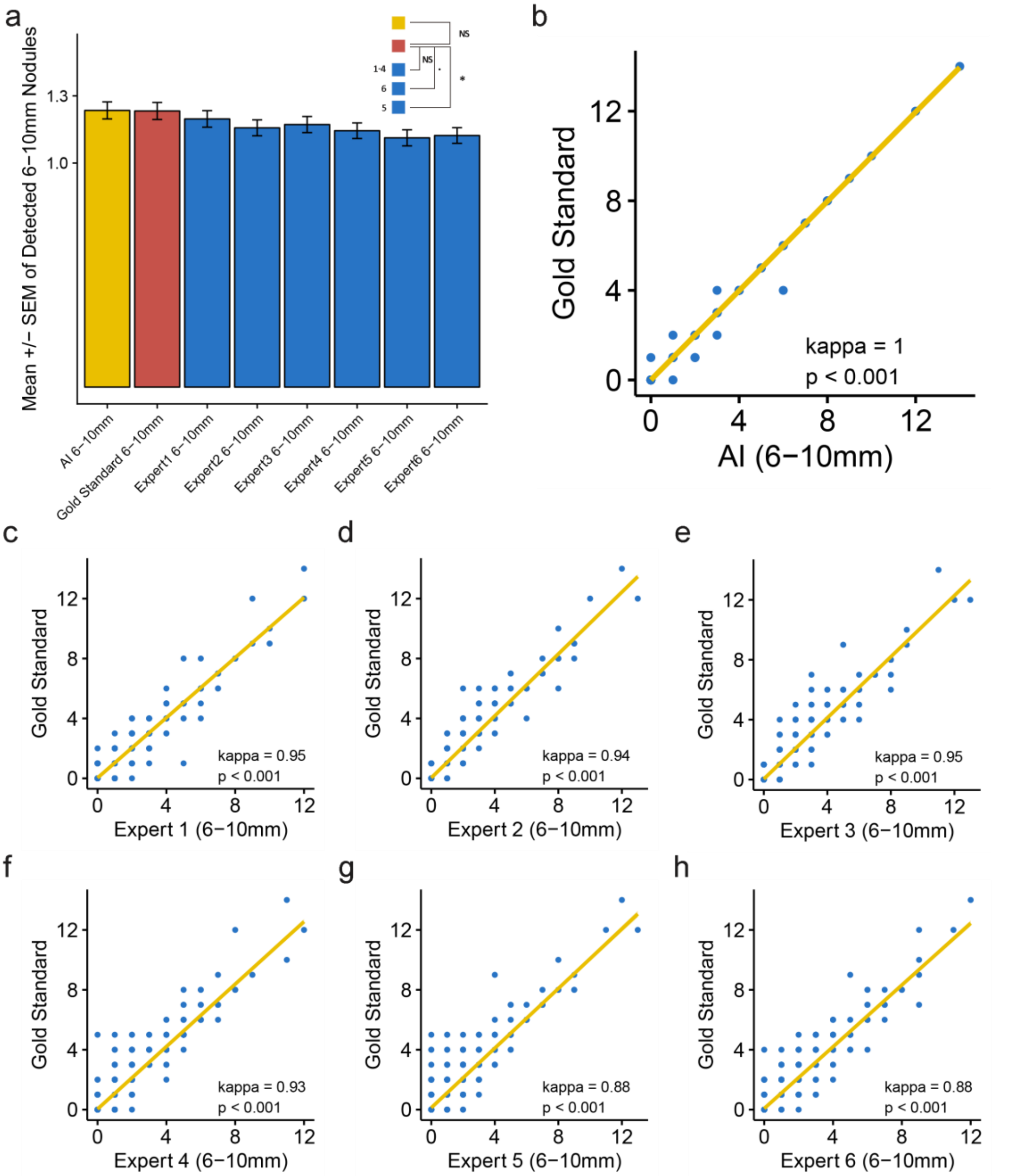

Figure S5. Consistency analysis among AI, human experts and the gold standard in detecting lung nodules with a size of 6-10 mm. Using the gold standard as reference, (a) no significant difference was observed when comparing AI and Expert 1-4 to the gold standard (p=0 97 for AI, p=0 63 for Expert 1, p=0 34 for Expert 2, p=0 46 for Expert 3 and p=0 23 for Expert 4). Detection results derived from Expert 5 may not come from the same distribution of the gold standard (p=0 01). A trend toward a difference existed between Expert 6 and the gold standard (p=0 07). (b-h) demonstrated that both AI and human experts were highly significantly consistent with the gold standard (kappa coefficient range from 0 88-0 95, p<0.001) when detecting 6-10 mm lung nodules. The horizontal and vertical coordinates for (b-h) indicate the detected nodule number. Statistical significance is labeled as follows: for <0 1, * for <0 05, ** for <0 01, *** for <0 005 and NS for no significance.

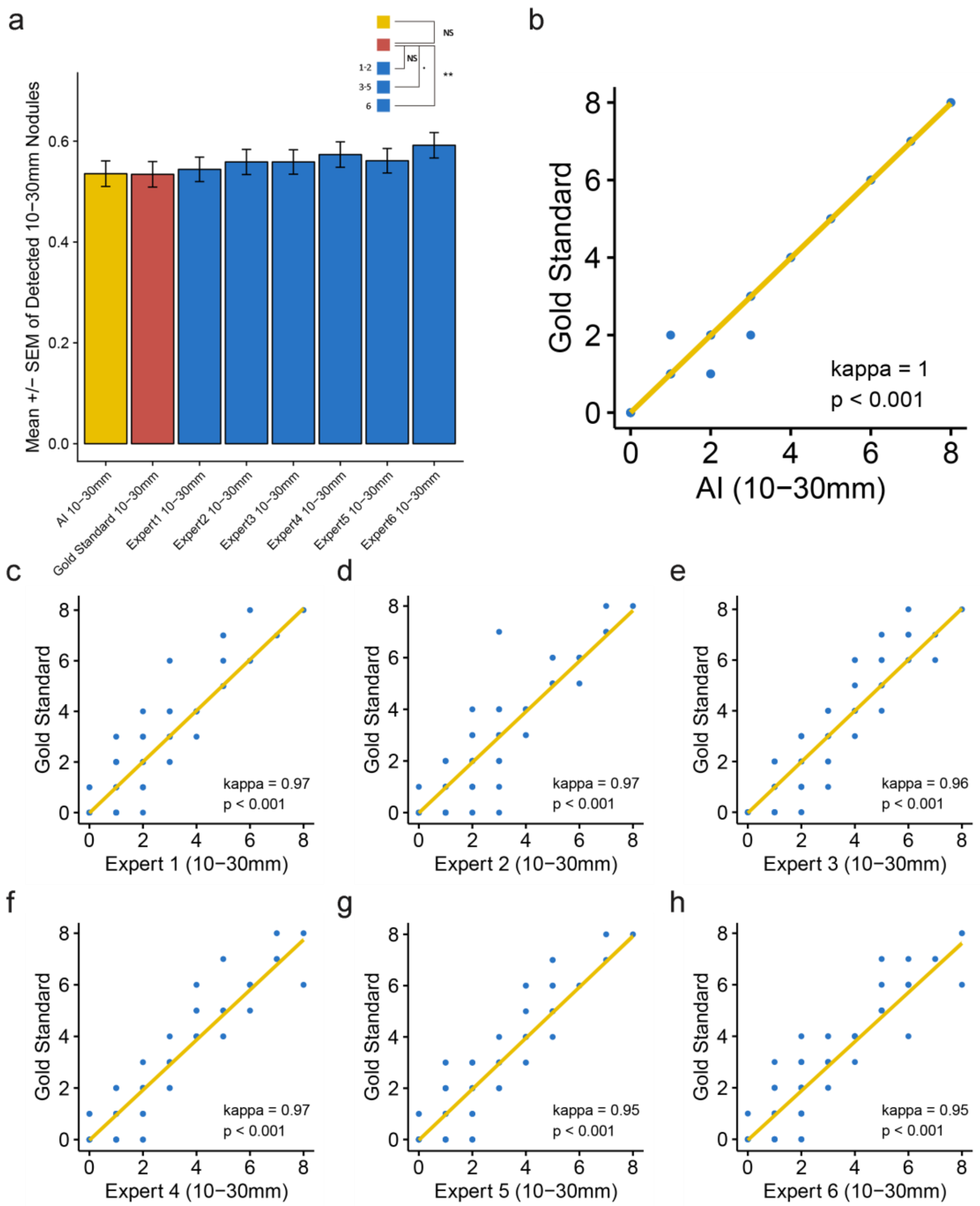

Figure S6. Consistency analysis among AI, human experts and the gold standard in detecting lung nodules with a size of 10-30 mm. Using the gold standard as reference, (a) no significant difference was observed when comparing AI and Expert1-2 to the gold standard (p=0.99 for AI, p=0.30 for Expert1, p=0.21 for Expert2, p=0.46 for Expert3 and p=0.23 for Expert4). Detection results derived from Expert6 may not come from the same distribution of the gold standard (p=0.005). A trend toward a difference existed between Expert3-5 and the gold standard (p=0.08 for Expert3, p=0.05 for Expert4 and p=0.06 for Expert5). (b-h) demonstrated that both AI and human experts were highly significantly consistent with the gold standard (kappa coefficient range from 0.95-0.97, p<0.001) when detecting 10-30 mm lung nodules. The horizontal and vertical coordinates for (b-h) indicate the detected nodule number. Statistical significance was labeled as follows: for <0.1, * for <0.05, ** for <0.01, *** for <0.005 and NS for no significance.

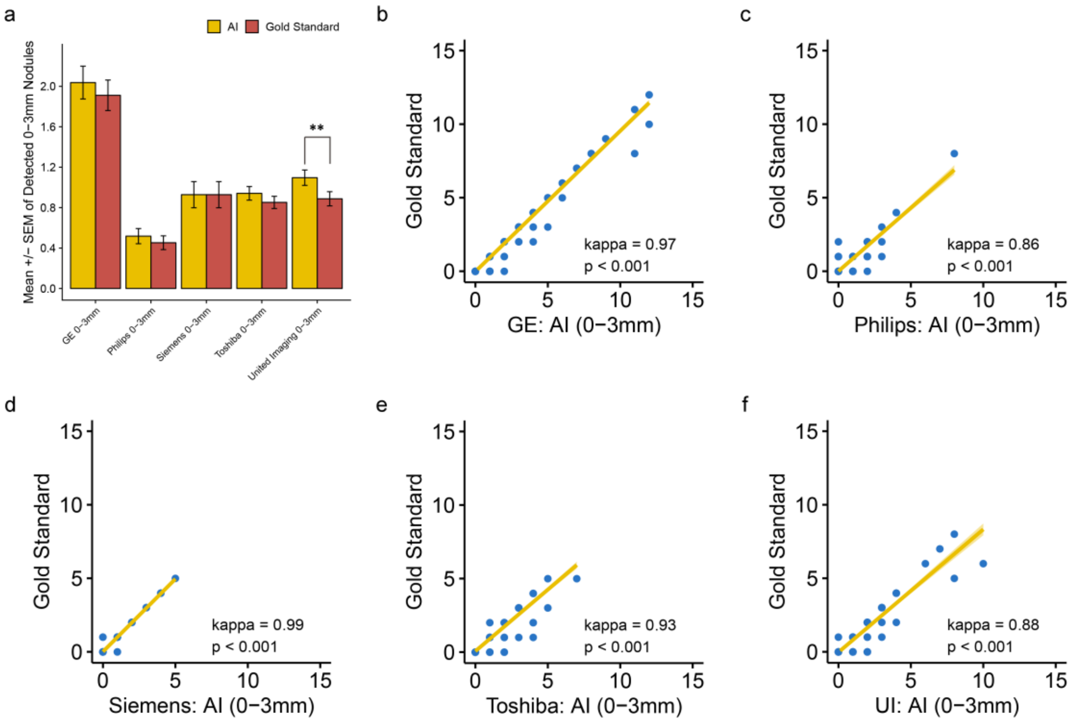

Figure S7. Performance of AI for consistency of 0-3 mm lung nodule diagnosis when deploying imaging equipment from five different manufacturers. Using the gold standard as a reference, (a) no significant difference was observed regardless of the type of manufacturer ($p>0.05$) except for United Imaging ($p=0.006$). (b-f) demonstrated that in all kinds of manufacturers, AI represented highly significant consistency with the gold standard (kappa coefficient range from 0.86-0.99, $p<0.001$). The horizontal and vertical coordinates for (b-f) indicate the detected nodule number. Statistical significance is labeled as follows: ** for <0.01.

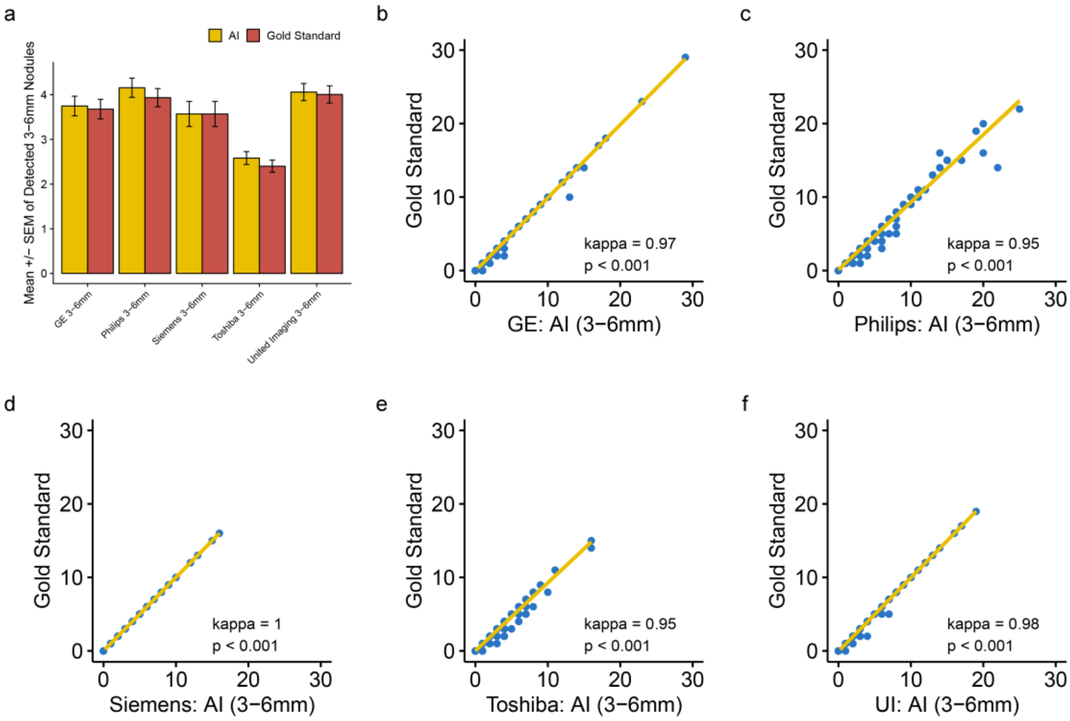

Figure S8. Performance of AI for consistency of 3-6 mm lung nodule diagnosis when applied to imaging equipment from five different manufacturers. Using the gold standard as a reference, (a) no significant difference was observed regardless of the type of manufacturer($p>0·05$). (b-f) demonstrated that in all kinds of manufacturers, AI represented highly significant consistency with the gold standard (kappa coefficient range from 0·95-1, $p<0·001$). The horizontal and vertical coordinates for (b-f) indicate the detected nodule number.

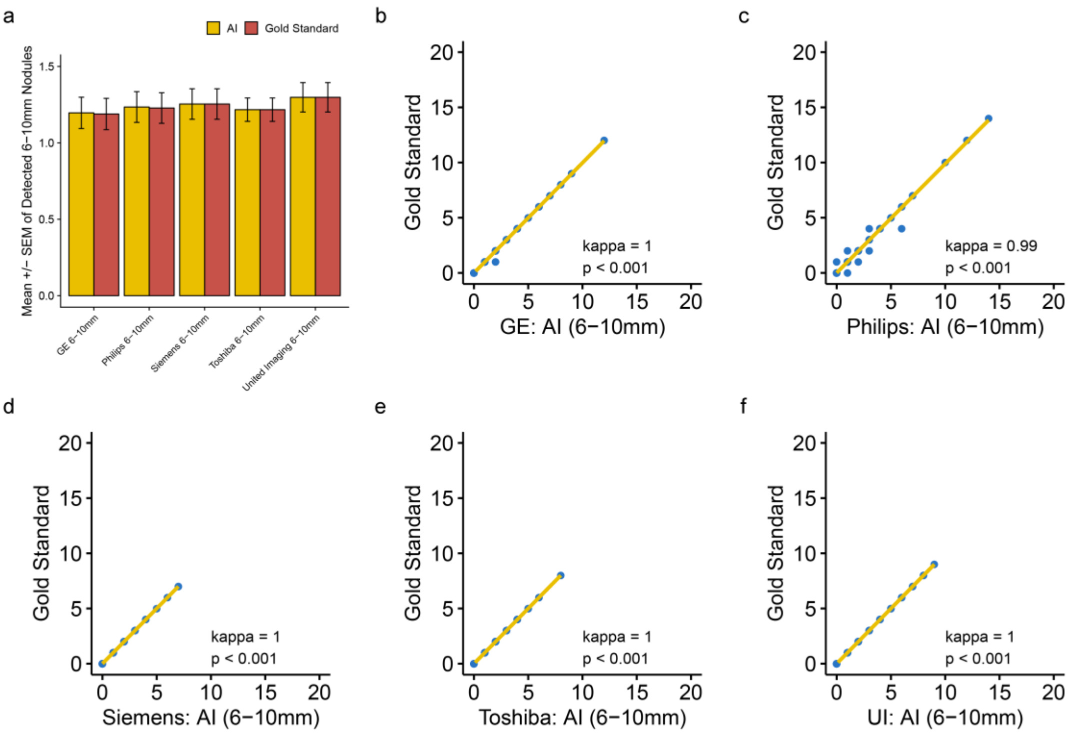

Figure S9. Performance of AI for consistency of 6-10 mm lung nodule diagnosis when applied to imaging equipment from five different manufacturers. Using the gold standard as a reference, (a) no significant difference was observed regardless of the type of manufacturer ($p>0·05$). (b-f) demonstrated that in all kinds of manufacturers, AI represented highly significant consistency with the gold standard (kappa coefficient range from 0·99-1, $p<0·001$). The horizontal and vertical coordinates for (b-f) indicate the detected nodule number.

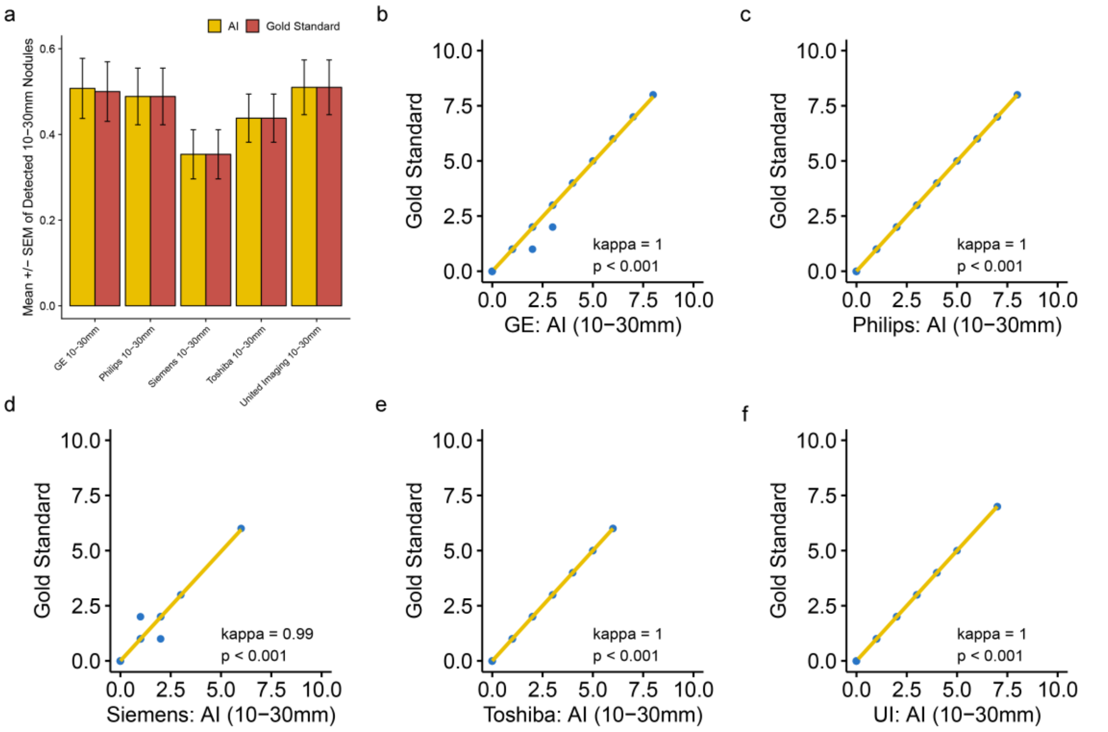

Figure S10. Performance of AI for consistency of 10-30 mm lung nodule diagnosis when applied to imaging equipment from five different manufacturers. Using the gold standard as a reference, (a) no significant difference was observed regardless of the type of manufacturer (p>0 05). (b-f) demonstrated that in all kinds of manufacturers, AI represented highly significant consistency with the gold standard (kappa coefficient range from 0 99-1, p<0 001). The horizontal and vertical coordinates for (b-f) indicate the detected nodule number. Statistical significance is labeled as follows: for < 0 1, * for <0 05, ** for <0 01 and *** for <0 005.

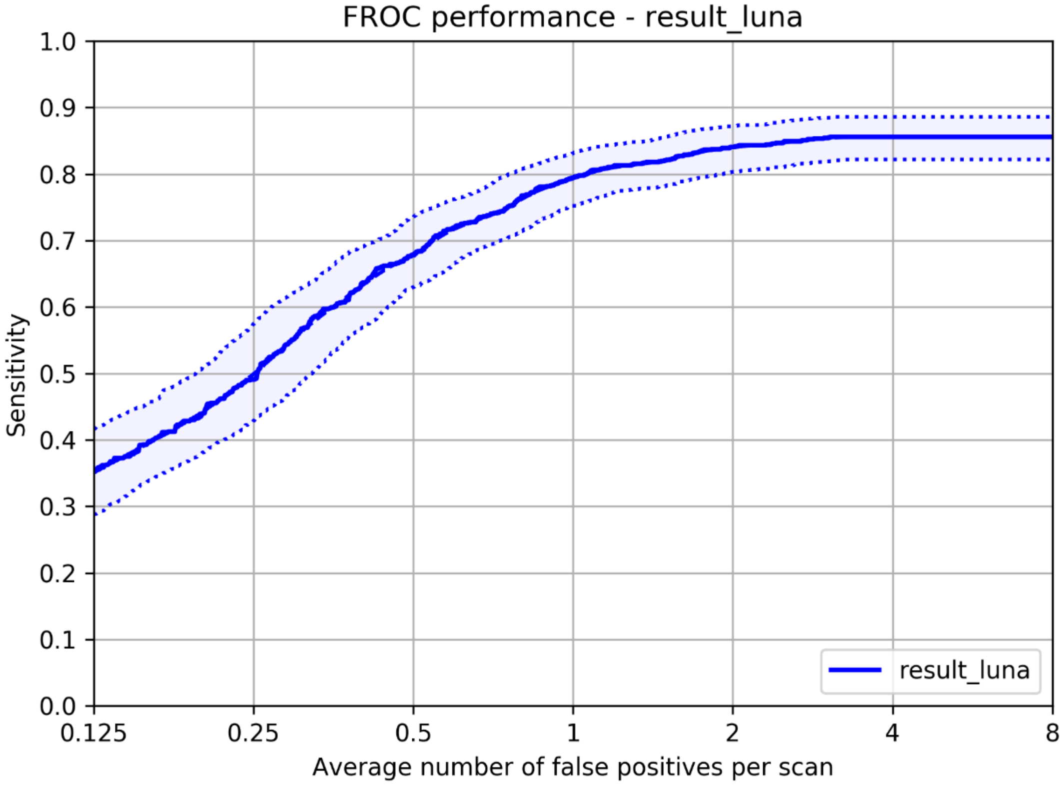

Figure S11. Performance was evaluated by free receiver operating characteristic (FROC) on LUNA16 database. The X-axis is the average false positive nodules per scan of 888 CT scans in Luna16, and the Y-axis is the sensitivity in the case of the average false positive.

Table S1. Consistency analysis among AI, human experts and gold standard in detecting lung nodules (correspond to Figure 9 and Supplementary Figure S3-6)

	Consistency Degree*	P value	
		Kappa	Mann–Whitney U
All			
AI	0.94	<0.001	0.138
Expert1	0.69	<0.001	1.1e-05
Expert2	0.73	<0.001	1.2e-07
Expert3	0.75	<0.001	4.3e-07
Expert4	0.75	<0.001	2.2e-07
Expert5	0.65	<0.001	6.8e-12
Expert6	0.63	<0.001	4.0e-11
Size of 0-3mm			
AI	0.93	<0.001	0.013
Expert1	0.73	<0.001	7.1e-05
Expert2	0.71	<0.001	1.3e-07
Expert3	0.68	<0.001	3.8e-12
Expert4	0.78	<0.001	3.2e-06
Expert5	0.66	<0.001	6.1e-15
Expert6	0.62	<0.001	2.7e-11
Size of 3-6mm			
AI	0.97	<0.001	0.281
Expert1	0.80	<0.001	2.3e-07
Expert2	0.80	<0.001	1.6e-08
Expert3	0.80	<0.001	7.2e-07
Expert4	0.79	<0.001	2.0e-09
Expert5	0.75	<0.001	2.8e-10
Expert6	0.75	<0.001	2.3e-12
Size of 6-10 mm			
AI	1.00	<0.001	0.971
Expert1	0.95	<0.001	0.634
Expert2	0.94	<0.001	0.338
Expert3	0.95	<0.001	0.458
Expert4	0.93	<0.001	0.235
Expert5	0.88	<0.001	0.014
Expert6	0.88	<0.001	0.071
Size of 10-30 mm			
AI	1.00	<0.001	0.989
Expert1	0.97	<0.001	0.304
Expert2	0.97	<0.001	0.212
Expert3	0.96	<0.001	0.080
Expert4	0.97	<0.001	0.052
Expert5	0.95	<0.001	0.057
Expert6	0.95	<0.001	0.005

Note: *Consistency degree is presented by the kappa coefficient.

Table S2. Performance of AI for consistency of lung nodule diagnosis when applied to imaging equipment from five different manufacturers (Corresponds to Supplementary Figure S7-10).

	Consistency Degree*	P value	
		Kappa	Mann–Whitney U
All			
GE	0.97	<0.001	0.576
Philips	0.90	<0.001	0.472
Siemens	0.99	<0.001	0.988
Toshiba	0.87	<0.001	0.376
United Imaging	0.91	<0.001	0.343
Size of 0-3 mm			
GE	0.97	<0.001	0.462
Philips	0.86	<0.001	0.400
Siemens	0.99	<0.001	1.000
Toshiba	0.93	<0.001	0.470
United Imaging	0.88	<0.001	0.006
Size of 3-6 mm			
GE	0.97	<0.001	0.698
Philips	0.95	<0.001	0.439
Siemens	1.00	<0.001	1.000
Toshiba	0.95	<0.001	0.358
United Imaging	0.98	<0.001	0.759
Size of 6-10 mm			
GE	1.00	<0.001	0.948
Philips	0.99	<0.001	0.989
Siemens	1.00	<0.001	1.000
Toshiba	1.00	<0.001	1.000
United Imaging	1.00	<0.001	1.000
Size of 10-30 mm			
GE	1.00	<0.001	0.979
Philips	1.00	<0.001	1.000
Siemens	0.99	<0.001	1.000
Toshiba	1.00	<0.001	1.000
United Imaging	1.00	<0.001	1.000

Note: *Consistency degree is presented by the kappa coefficient.

Table S3. Traditional layout system vs. Intelligent Layout in chest CT images.

Index	Points of Comparison	Traditional Layout	Intelligent Layout	P values
Differentiation in Operator Behaviors				
1	Need operation or not	Necessary	Unnecessary	
2[a]	Number of clicks (times)	14.45±0.34	2±0	< 2.2e-16
3[b]	Average time consumed (seconds/patient)	16.87±0.38	6.92±0.10	< 2.2e-16
Differentiation on Typographic Layouts				
4[c]	Invalid images	7.06±0.24	0	< 2.2e-16
5[d]	Appropriate size for grid	Unstable	Stable	
6[e]	Missing lung nodules	46.8%	0%	< 2.2e-16
7	Repeatable nodule size	Unstable	Stable	
8	Repeatable CT value	Unstable	Stable	
9	Multidimensional display	No occurrence	100% show	
10[f]	Predictive value	No occurrence	100% show	
11[g]	Traceability	No occurrence	100% show	
Differentiation of Impact on Relevant Persons				
12	Impact on radiologists	Obscure	Clear/Helpful	
13	Impact on physicians	Inconvenient	Convenience/Helpful	
14	Impact on patients	Careless	Convenience/Helpful	

Note:

a. **"Number of clicks":** Range from selecting the patient directory to the end of layout.

b. **"Average time operator consumed"**: Skilled operator with more than five years of work experience spent in operating.

c. **"Invalid images":** No diagnostic value for lung lesions including mediastinum and lung window images, which would appear in any of the cells.

d. **"Appropriate size for grid":** Zoom in or out to the most appropriate state of grid.

e. **"Missing lung nodules":** Take GE's workstation as an example, nodules with a diameter (≤7.5 mm). The thickness of each slice was 1.25 mm. If you choose the interval layout, it takes spacing 6-7 layers of images in order to finish a complete layout for a normal adult. Therefore, for small nodules of 6-8 mm, it is easy to not be selected in image layouts.

f. **"Predictive value":** For benign and malignant lesions of the pulmonary nodules.

g. **"Traceability":** Each image on the film can be traced by its slice id and redirected to its original location in image set by double clicking the mouse.

95% confidence intervals are enclosed in parentheses.

Table S4. Comparing click time (counts) of AI with five manufacturers.

	Click (counts)	
	Mean ±SEM	**P value***
AI (reference)	2.00±0.00	
GE	14.37±0.89	1.2e-12
Philips	14.70±0.86	1.2e-12
Siemens	14.57±0.87	1.2e-12
Toshiba	15.77±0.95	1.0e-12
United Imaging	13.67±0.79	1.2e-12

Note: *Mann–Whitney U test P value was calculated by comparing the corresponding manufacturer with AI.

Table S5. Comparing consumed time (sec) of AI with five manufacturers.

	Consumed time (sec)	
	Mean ±SEM	**P value***
AI (compared with GE)	7.30±0.20	5.3e-09
GE	16.00±1.17	
AI (compared with Philips)	8.00±0.17	8.3e-10
Philips	14.83±0.66	
AI (compared with Siemens)	6.87±0.17	2.1e-11
Siemens	17.73±0.91	
AI (compared with Toshiba)	8.00±0.00	7.2e-10
Toshiba	16.33±0.96	
AI (compared with United Imaging)	6.70±0.15	2.0e-11
United Imaging	17.27±1.01	

Note: * Mann–Whitney U test P value was calculated by comparing pairwise manufacturer with AI.

Table S6. Comparison of different model performance.

Work	**Database (Samples)**	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**
Orozco & Villegas (1)	NBIA-ELCAP	N/A	96.2	52.2
Hua et al. (2)	LIDC	N/A	73.3	78.7
Kumar et al. (3)	LIDC	75.0	83.4	N/A
Da Silva (4)	LIDC-IDRI	82.3	79.4	83.8
CNN (5)	LIDC-IDRI	84.2	84.0	84.3
DNN(5)	LIDC-IDRI	82.4	80.7	83.9
SAE(5)	LIDC-IDRI	82.6	84.0	81.4
IILS (This Paper)	OUR DATASET	87.3	76.5	89.1

Table S7. CAD analysis for LUNA16.

Candidate detection index	Values
True positives	1015
False positives	2752
False negatives	171
True negatives	0
Total number of candidates	5368
Total number of nodules	1186
Ignored candidates on excluded nodules	1591
Ignored candidates which were double detections on a nodule	10
Sensitivity	0.856
Average number of candidates per scan	6.045

REFERENCES

1. Orozco HM, Villegas OOV, Dom ńguez HdJO, Sanchez VGC, editors. Lung nodule classification in CT thorax images using support vector machines. Artificial Intelligence (MICAI), 2013 12th Mexican International Conference on; 2013: IEEE.
2. Hua K-L, Hsu C-H, Hidayati SC, Cheng W-H, Chen Y-J. Computer-aided classification of lung nodules on computed tomography images via deep learning technique. OncoTargets and therapy. 2015;8.
3. Kumar D, Wong A, Clausi DA, editors. Lung nodule classification using deep features in CT images. Computer and Robot Vision (CRV), 2015 12th Conference on; 2015: IEEE.
4. da Silva GL, Silva AC, de Paiva AC, Gattass M. Classification of Malignancy of Lung Nodules in CT Images Using Convolutional Neural Network. OncoTargets and therapy. 2016.
5. Song Q, Zhao L, Luo X, Dou X. Using deep learning for classification of lung nodules on computed tomography images. Journal of healthcare engineering. 2017.