EBioMedicine 44 (2019) 162–181

Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.ebiomedicine.com

EBioMedicine
Published by THE LANCET

# IILS: Intelligent imaging layout system for automatic imaging report standardization and intra-interdisciplinary clinical workflow optimization

Yang Wang [a,1], Fangrong Yan [b,1], Xiaofan Lu [b,1], Guanming Zheng [c], Xin Zhang [a], Chen Wang [a], Kefeng Zhou [d], Yingwei Zhang [e], Hui Li [e], Qi Zhao [e], Hu Zhu [f], Fei Chen [g], Cailiang Gao [h], Zhao Qing [a], Jing Ye [i], Aijing Li [j], Xiaoyan Xin [a], Danyan Li [a], Han Wang [a], Hongming Yu [a], Lu Cao [k], Chaowei Zhao [k], Rui Deng [k], Libo Tan [k], Yong Chen [l], Lihua Yuan [a], Zhuping Zhou [a], Wen Yang [a], Mingran Shao [a], Xin Dou [a], Nan Zhou [a], Fei Zhou [a], Yue Zhu [b], Guangming Lu [m], Bing Zhang [a,*]

[a] *Department of Radiology, the Affiliated Nanjing Drum Tower Hospital of Nanjing University Medical School, Nanjing 210008, China*
[b] *Research Center of Biostatistics and Computational Pharmacy, China Pharmaceutical University, Nanjing, China*
[c] *Department of Statistics, University of Michigan, Ann arbor 48105, USA*
[d] *Department of Radiology, NanJing GaoChun People's Hospital, No.9 Chunzhong Road, GaoChun, NanJing, China*
[e] *Department of Respiratory, the Affiliated Nanjing Drum Tower Hospital of Nanjing University Medical School, Nanjing 210008, China*
[f] *College of Telecommunication and Information Engineering, Nanjing University of Posts and Telecommunications, No.66 Xin Mofan Road, Nanjing, China*
[g] *Department of Radiology, Affiliated Yancheng Hospital, School of Medicine, Southeast University, Yancheng, Jiangsu, China*
[h] *Department of Radiology, Chongqing Three Gorges Central Hospital, Chongqing 404000, China*
[i] *Department of Radiology, Northern Jiangsu People's Hospital, No.98 Nantong West Road, Yangzhou, Jiangsu 225001, China*
[j] *Department of Radiology, Ningbo No. 2 Hospital, No. 41, Xibei street, Haishu District 315010, Zhejiang, China*
[k] *FL 8, Ocean International Center E, Chaoyang Rd Side Rd, ShiLiPu, Chaoyang Qu, 100000 Beijing Shi, China*
[l] *Department of Medical Administration, the Affiliated Nanjing Drum Tower Hospital of Nanjing University Medical School, Nanjing 210008, China*
[m] *Department of Medical Imaging, Jinling Hospital, School of Medicine, Nanjing University, Nanjing, 210002, Jiangsu, China*

## ARTICLE INFO

*Article history:*
Received 28 March 2019
Received in revised form 15 May 2019
Accepted 15 May 2019
Available online 23 May 2019

*Keywords:*
Lung nodule
Artificial intelligence
Deep learning algorithms
Intelligent image layout system
Standardized e-film and visualized structured report
Clinical workflow

## ABSTRACT

*Background:* To achieve imaging report standardization and improve the quality and efficiency of the intra-interdisciplinary clinical workflow, we proposed an intelligent imaging layout system (IILS) for a clinical decision support system-based ubiquitous healthcare service, which is a lung nodule management system using medical images.

*Methods:* We created a lung IILS based on deep learning for imaging report standardization and workflow optimization for the identification of nodules. Our IILS utilized a deep learning plus adaptive auto layout tool, which trained and tested a neural network with imaging data from all the main CT manufacturers from 11,205 patients. Model performance was evaluated by the receiver operating characteristic curve (ROC) and calculating the corresponding area under the curve (AUC). The clinical application value for our IILS was assessed by a comprehensive comparison of multiple aspects.

*Findings:* Our IILS is clinically applicable due to the consistency with nodules detected by IILS, with its highest consistency of 0·94 and an AUC of 90·6% for malignant pulmonary nodules versus benign nodules with a sensitivity of 76·5% and specificity of 89·1%. Applying this IILS to a dataset of chest CT images, we demonstrate performance comparable to that of human experts in providing a better layout and aiding in diagnosis in 100% valid images and nodule display. The IILS was superior to the traditional manual system in performance, such as reducing the number of clicks from 14·45 ± 0·38 to 2, time consumed from 16·87 ± 0·38 s to 6·92 ± 0·10 s, number of invalid images from 7·06 ± 0·24 to 0, and missing lung nodules from 46·8% to 0%.

*Interpretation:* This IILS might achieve imaging report standardization, and improve the clinical workflow therefore opening a new window for clinical application of artificial intelligence.

*Fund:* The National Natural Science Foundation of China.

[*] Corresponding author.
*E-mail address:* zhangbing_nanjing@vip.163.com (B. Zhang).

[1] Yang Wang, Fangrong Yan, Xiaofan Lu all gave the same contribution to the paper and were recommended as co-first authors.

https://doi.org/10.1016/j.ebiom.2019.05.040

**Research in context**

*Evidence before this study*

The implementation of traditional clinical decision support results comprising images and reports for radiology and respiratory departments faces challenges with reliability and interpretability.

*Added value of this study*

Through a new entry point, a new work process will be established, and some relevant operators will be unnecessary due to the invention of the IILS, which includes two deep learning models first applied to clinical medicine, Faster RCNN and ResNet. Superior to the traditional manual system, the IILS will be promoted and applied to other imaging methods, such as magnetic resonance imaging and imaging of other parts of the body. The IILS will be integrated into radiology workflows by series connection instead of parallel connection to considerably simplify and optimize the clinical workflow and to benefit more doctors and even patients during regular follow-up.

*Implications of all the available evidence*

With the advent of the information era, over the next decades, more goals could be fulfilled by the IILS when supporting regions with insufficient healthcare resources. Therefore, the IILS opens a new window for clinical application of artificial intelligence.

## 1. Introduction

Lung cancer is one of the most common cancers and one of the leading causes of cancer death [1]. In 2018, approximately 234,030 new cases and 154,050 deaths of lung cancer occurred [2]. Pulmonary nodules are the most common manifestation of lung cancer [3]. When a tumor is detected on imaging, it was likely present as microscopic disease for a longer duration [4]. Therefore, low-dose computed tomography (CT) is recommended because it can greatly improve the likelihood of detecting small nodules; thus, lung cancer will be detected at an earlier stage or a potentially more curable stage [5]. Over the last two to three decades, the demand for imaging services has increased at an unprecedented rate, and the amount of imaging has increased dramatically [6]. However, at present, the layout of all medical images is still conducted manually or there is no layout before all images are uploaded to the Picture Archiving and Communication Systems (PACS). The currently applied methods have caused difficulties for many clinicians. After entering the 21st century, the common expectation of almost all radiologists worldwide was that filmless radiology departments in completely digital regional hospitals would be established [7]. However, medical dry laser images are still widely used. For example, based on the data of AGFA, the world's third largest medical imaging film company, 2·5 billion graphic sheets were sold in 2017 (Supplementary Fig. S1), which may be related to the global imbalance of medical resources (Video 1). A typical example of a clinical task is to sort and generate the layout of many chest CT images that are closely associated with the diagnosis of lung nodules. In the screening detection and follow-up period, five problems remain within the current daily workflow. First, a lack of imaging report standardization for radiologists and clinicians: since there was no standardized, scientifically validated approach to the evaluation of nodules, trial radiologists developed guidelines for diagnostic follow-up, but no specific evaluation approach was mandated [8] (Fig. 1). Second, missing nodules: if dry laser film is used as an imaging information carrier, the failure to display nodules on the images corresponding to the description on the report would be the most common complaint (Fig. 2). Third, a lack of key images: after the image acquisitions from CT scanners, a huge number of images are all entered in the PACS without any selection. Moreover, many clinicians are relatively unfamiliar with imaging knowledge. Even a senior doctor facing complicated image information that lacks key images would be required to spend much time and effort in flipping through the images, not to mention using a smartphone or tablet to check these images. In addition, many invalid images often appear in a series of images (Fig. 2). Therefore, the treatment process is extremely inefficient [9]. Fourth, difficulties in accessing images from other hospitals: if a patient requests his or her own images, the images are usually burned on a compact disc (CD) or are transferred by a portable hard disk drive. However, many modern computers are not equipped with a CD drive or the universal serial bus (USB) interface of the computer is forbidden at hospitals. Therefore, imaging has to be completed for patients in different hospitals. Fifth, a lack of consideration for the needs of clinicians and patients (Fig. 2): as a radiologist, supporting other doctors and patient-centered metrics are described in paper [10]. In fact, the opportunities to help others more easily read and understand imaging results have not yet been fully exploited. Thus, obtaining a standardized e-film with key images and a visualized structured report is urgent to solve these problems (Video 1, see Materials and Methods for more details). With the advent of the third wave of artificial intelligence (AI) technology [11], there has been substantial progress in the use of AI in the medical field [12–14]. Currently, the majority of AI use in radiology focuses on the diagnosis, prediction and evaluation of treatment outcomes [15,16] (Video 1). However, the current applications of AI seem to ignore two facts. First, standardized images with high quality are the basis for AI development, and second, there are simple and repetitive actions that AI could take over [17].

In this study, we sought to develop an intelligent film layout system (IILS) based on a fusion of AI technology and an adaptive layout tool to establish a new work process for daily work and obtain both standardized images and reports for radiologists and clinicians. The primary application of our machine learning algorithm was in the detection and classification of pulmonary nodules, sorting and comparison with pathological results, and ultimately providing an impact assessment. At the same time, we assess the comprehensive strength of IILS from three perspectives, including i) the comparison of the diagnostic efficiency for nodules between IILS and clinical experts; ii) the degree to which IILS could optimize clinical workflow; iii) the cross-manufacture applicability of IILS. Taken together, we argue that the AI technology could be integrated into the radiology workflow by series connection instead of following the traditional workflow based on a simple parallel relationship. The efficiency of the workflow could be improved, resulting in reduced medical costs (Fig. 3).

## 2. Materials and methods

The key resources are shown in Table 4.

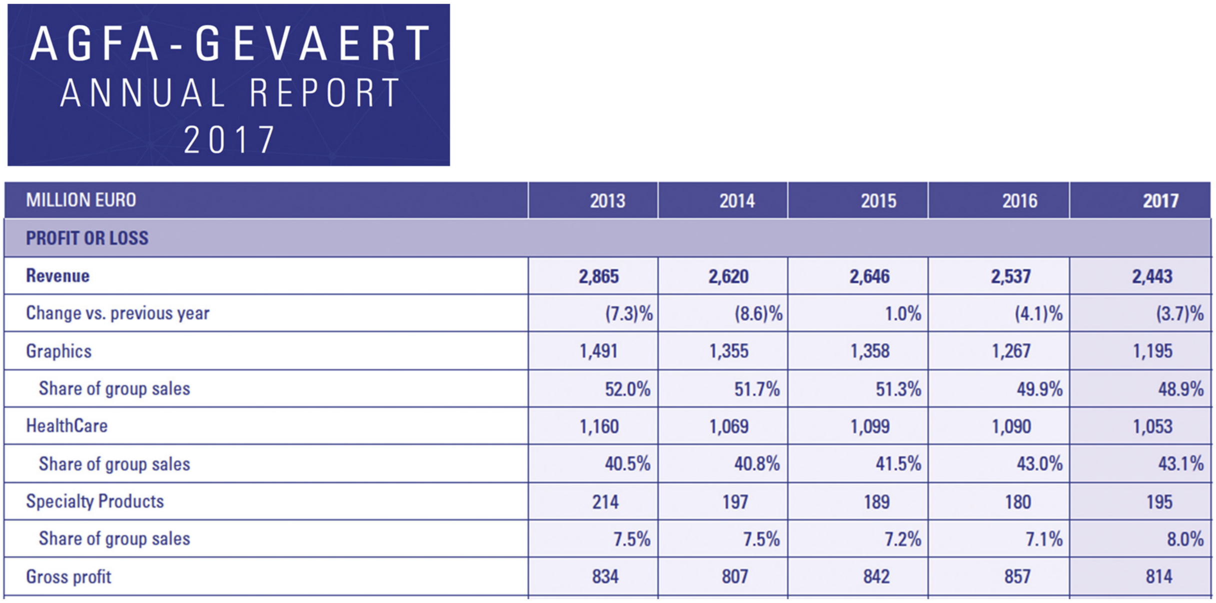
### 2.1. Experimental software and hardware

The models in this article are all trained on the DGX1 platform (NVIDIA DGX1 system, 8× Tesla V100 GPUs, 128 GB total system GPU Memory, dual 20-core Intel Xeon E5–2698 CPU v4 2.2 GHz, Santa Clara, California, USA).

### 2.2. Experimental model and subject details

#### 2.2.1. Images from human subjects and acquisition

The study was approved by the institutional review board of the University Medical Center. Institutional Review Board (IRB)/Ethics



AGFA-GEVAERT
ANNUAL REPORT
2017

| MILLION EURO | 2013 | 2014 | 2015 | 2016 | 2017 |
|---|---|---|---|---|---|
| PROFIT OR LOSS | | | | | |
| **Revenue** | **2,865** | **2,620** | **2,646** | **2,537** | **2,443** |
| Change vs. previous year | (7.3)% | (8.6)% | 1.0% | (4.1)% | (3.7)% |
| Graphics | 1,491 | 1,355 | 1,358 | 1,267 | 1,195 |
| Share of group sales | 52.0% | 51.7% | 51.3% | 49.9% | 48.9% |
| HealthCare | 1,160 | 1,069 | 1,099 | 1,090 | 1,053 |
| Share of group sales | 40.5% | 40.8% | 41.5% | 43.0% | 43.1% |
| Specialty Products | 214 | 197 | 189 | 180 | 195 |
| Share of group sales | 7.5% | 7.5% | 7.2% | 7.1% | 8.0% |
| Gross profit | 834 | 807 | 842 | 857 | 814 |

**Fig. 1.** Agfa-Gevaert annual report 2017. Based on the financial statements of AGFA, a listed company in the past 5 years, financial income remained basically stable, and Euro 1053 million sales were obtained in the healthcare segment, including 2·5 billion global medical graphic sheets sold in 2017 (see more details on websites http://www.agfa.com/corporate/investor-relations/key-figures/ or http://www.agfa.com/movies/annual_report_2017/).

Committee approvals were obtained. The work was conducted in a manner compliant with the People's Republic of China Health Insurance Portability and Accountability Act (HIPAA) and was adherent to the tenets of the Declaration of Helsinki. All chest CT scanning imaging was performed as part of the patients' routine clinical care. There were no exclusion criteria based on age or gender. All patient scans were downloaded in the DICOM image format according to the manufacturer's software and instructions. In the training data set, the images used in our study were obtained between October 2016 and May 2018. In the independent testing data set, all the chest CT scanning images were selected from retrospective cohorts of adult patients from Nanjing Drum Tower Hospital, Northern Jiangsu People's Hospital, Ningbo No.2 Hospital, and NanJing GaoChun People's Hospital between October 2016 and November 2018. Low-dose chest CT examinations

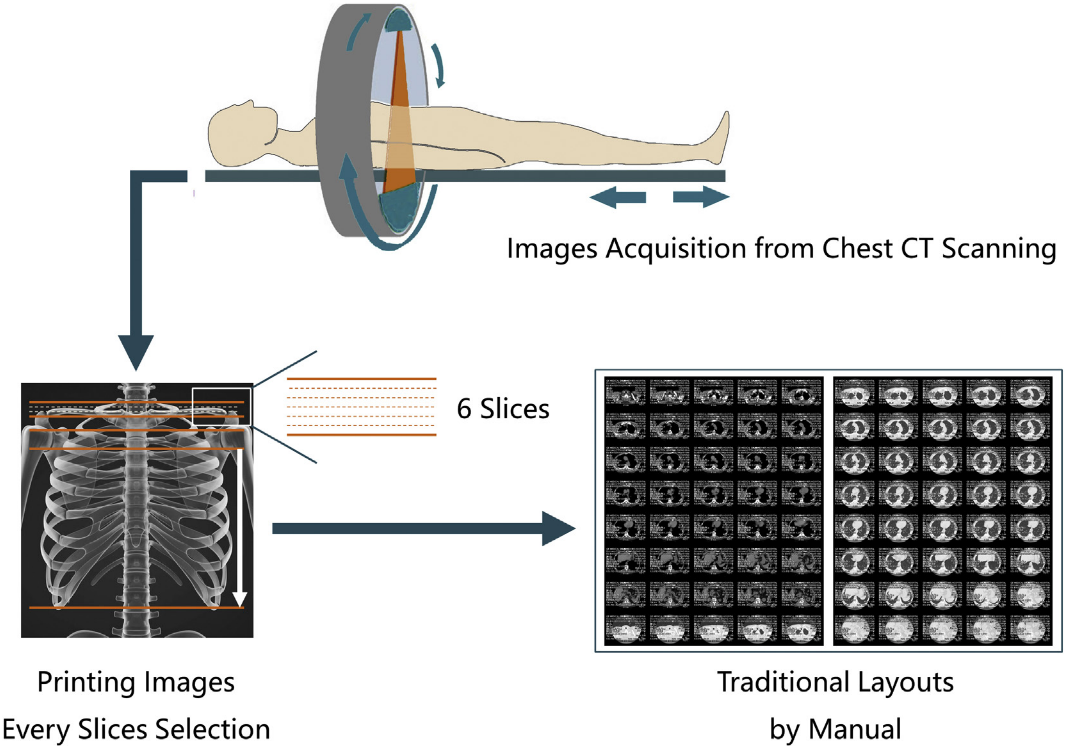


**Fig. 2.** The current film-selecting process by manual and related problems in daily work. For example, a typical chest CT scan of an adult could acquire approximately three hundred images. However, the final layouts in a sheet of film are approximately forty images only. Thus, approximately 87·7% of the images have been ignored. The impression on the corresponding report could not be 100% matched with the layout results, especially when the diameter of the lung nodules is <1 cm. Using GE's CT scanning method as an example, lung tissue in the range of 6 (layers) × 1·25 mm (thickness) = 7·5 mm is usually ignored when using the manual image layout in daily work. Therefore, when doctors obtain the final image layouts, the following problems could be found: a lack of imaging report standardization, missing nodules, lack of key images, and lack of consideration for the needs of clinicians and patients.

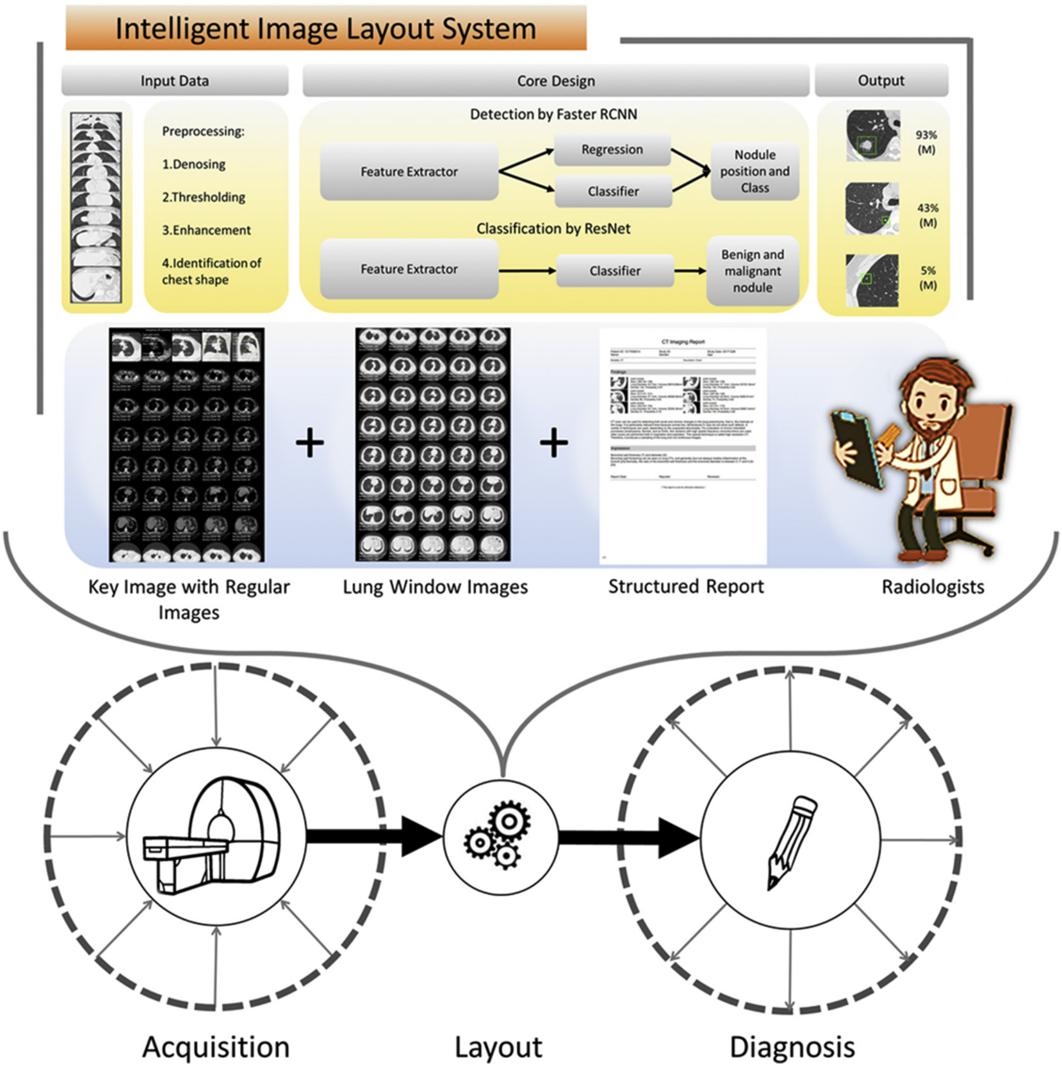


**Fig. 3.** The composition of the IILS and how to integrate it into the current imaging process. The new IILS includes the following parts: one is AI lung nodule detection and classification, and the other is an adaptive layout tool including auto films and visualized structured report generation that was invented by our team. To doubly ensure that the quality of the images and the results could be controlled, we had a radiologist who is usually responsible for writing the report double check the automatically generated structured report and image layout results. The entire process of daily work in the imaging department includes the following key steps: i) acquisition: collecting image information from patients from different clinical departments; ii) layout: inclusion of manual layout and image management for daily work; iii) diagnosis: image diagnosis, prediction and evaluation by radiologists. The application of the new intelligent system is integrated into radiology workflow by series connection instead of parallel connection. A new radiology work process has been developed.

were performed with or without contrast material for clinical purposes. All CT images in cohorts were acquired by using 16, 64, 80, 128 or 256-detector row CT scanners (GE LightSpeed VCT 64, GE Healthcare, Boston, USA; GE Discovery CT 750 HD 64, GE Healthcare, Boston, USA; Philips Brilliance ICT 256-slice, Philips Medical Systems, Best, the Netherlands; Somatom Definition, 64-slice, Siemens Medical Solutions, Forchheim, Germany; Toshiba Aquilion, 16-slice, Toshiba Medical Systems, Tokyo, Japan; United-Imaging uCT760 80-slice, United-Imaging Healthcare Company, Shanghai, China) and a low-dose radiation protocol. Data were acquired by using 16, 64, 80, 128 or 256 × 0·5, 0·625 or 0·75 mm collimation; a rotation time between 0·27 and 1 s; a tube current time product of 30 mAs; and a tube voltage between 80 and 140 kVp, dependent on the weight of the patient. The reconstructed section thickness was between 1 and 1·5 mm, with a reconstruction increment between 1 and 1·5 mm. A moderate lung reconstruction kernel was used. The smallest field of view included the outer rib margins at the widest dimension of the thorax and a 512 × 512 matrix. In this study, we initially obtained 11,205 patients with 3,527,048 chest CT scanning images from October 2016 to November 2018 from five main different CT manufacturers. During the training process, 9240 patients were divided into two cohorts. Training cohort 1 contained 8120 patients for nodule detection. Training cohort 2 contained 1120 patients for benign and malignant classification. In the testing process, another 1965 patients were assigned to the independent testing cohort (Fig. 4). We labeled nodules in sets to train our models. The models were tested with nodules, and 152 nodules were malignant (1880 nodules with diameter ≤ 3 mm, 6461 nodules with diameter 3–6 mm, 2195 nodules with diameter 6–10 mm and 923 nodules with diameter 10 mm ~ 3 cm) in an independent testing dataset. Among patients with nodules, the rate of malignancy in the independent testing dataset was 7·68%.

### *2.2.2. Expert comparisons and reproducibility*

Eight experts participated as observers. Three were members of the American Thoracic Society with significant clinical physician experience in respiratory (Y, Z, H. L, and Q. Z), three radiologists (F. C, K. Z, and N. Z) were involved in the IILS trial readings, and another three (X. D, W. Y, and Z. Z) were general radiologists with a specific interest in thoracic radiology. Experience with interpretation of chest CT images ranged from 3 years to >25 years. To evaluate our convolutional neural network

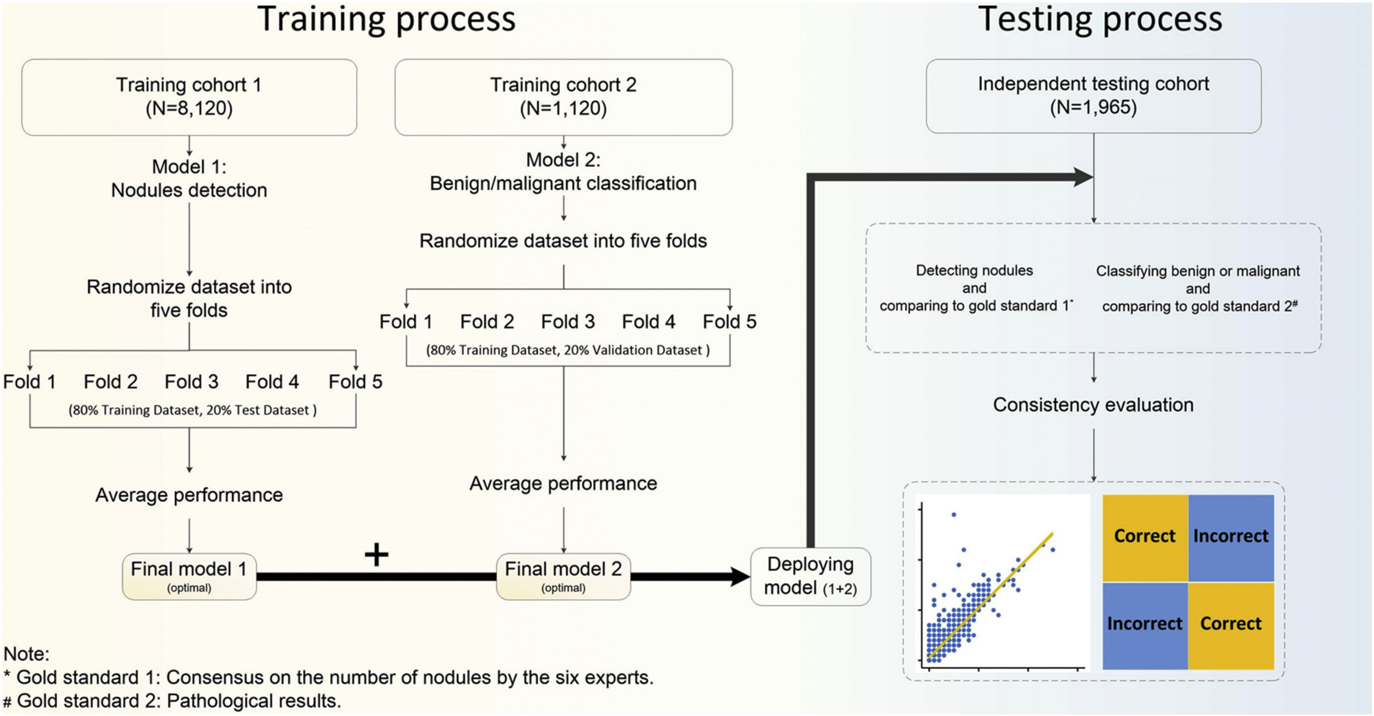

Training process
Testing process
Training cohort 1 (N=8,120)
Training cohort 2 (N=1,120)
Independent testing cohort (N=1,965)
Model 1: Nodules detection
Model 2: Benign/malignant classification
Randomize dataset into five folds
Randomize dataset into five folds
Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
(80% Training Dataset, 20% Validation Dataset )
Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
(80% Training Dataset, 20% Test Dataset )
Average performance
Average performance
Final model 1 (optimal)
+
Final model 2 (optimal)
Deploying model (1+2)
Detecting nodules and comparing to gold standard 1*
Classifying benign or malignant and comparing to gold standard 2#
Consistency evaluation
Correct
Incorrect
Incorrect
Correct
Note:
* Gold standard 1: Consensus on the number of nodules by the six experts.
# Gold standard 2: Pathological results.

**Fig. 4.** Data flow diagram showing our approach to detect nodules and classify benign or malignant cases. A total of 11,205 patients were used in this study. The training process was divided into two parts with two separate training cohorts. Two models derived from a convolutional neural network (CNN) were accessed for performance evaluation by 5-fold cross-validation and subsequently merged to form the first layer of the IILS, that is, the screening part used to detect nodules and classify cases. We deployed the final two models to an independent cohort that contained 1965 cases, including almost all major CT platforms on the market to manifest the credibility of our IILS by consistency analysis under specific consensus on the number of nodules by the six clinical experts. The pathological gold standards were the determination of nodules as benign or malignant on biopsy or surgical resection.

in the context of clinical experts, we used the independent testing set of 1965 patients to compare our network decisions with the decisions made by human experts. The 1965 patients as subjects were randomly selected for the inter-reader and inter-reader reproducibility study. Determination of the number of all lung nodules and judgment of benign and malignant nodules were checked twice by six experts with a 1-month time interval to minimize memory bias. All decisions were made by six experts for inter-reader reproducibility analysis. Weighted error scoring was used to reflect that a false negative result (failing to make decisions) is more detrimental than a false positive result. Using these weighted penalty points, error rates were computed for the model and for each of the human experts.

### 2.2.3. Nodule selection and characteristics

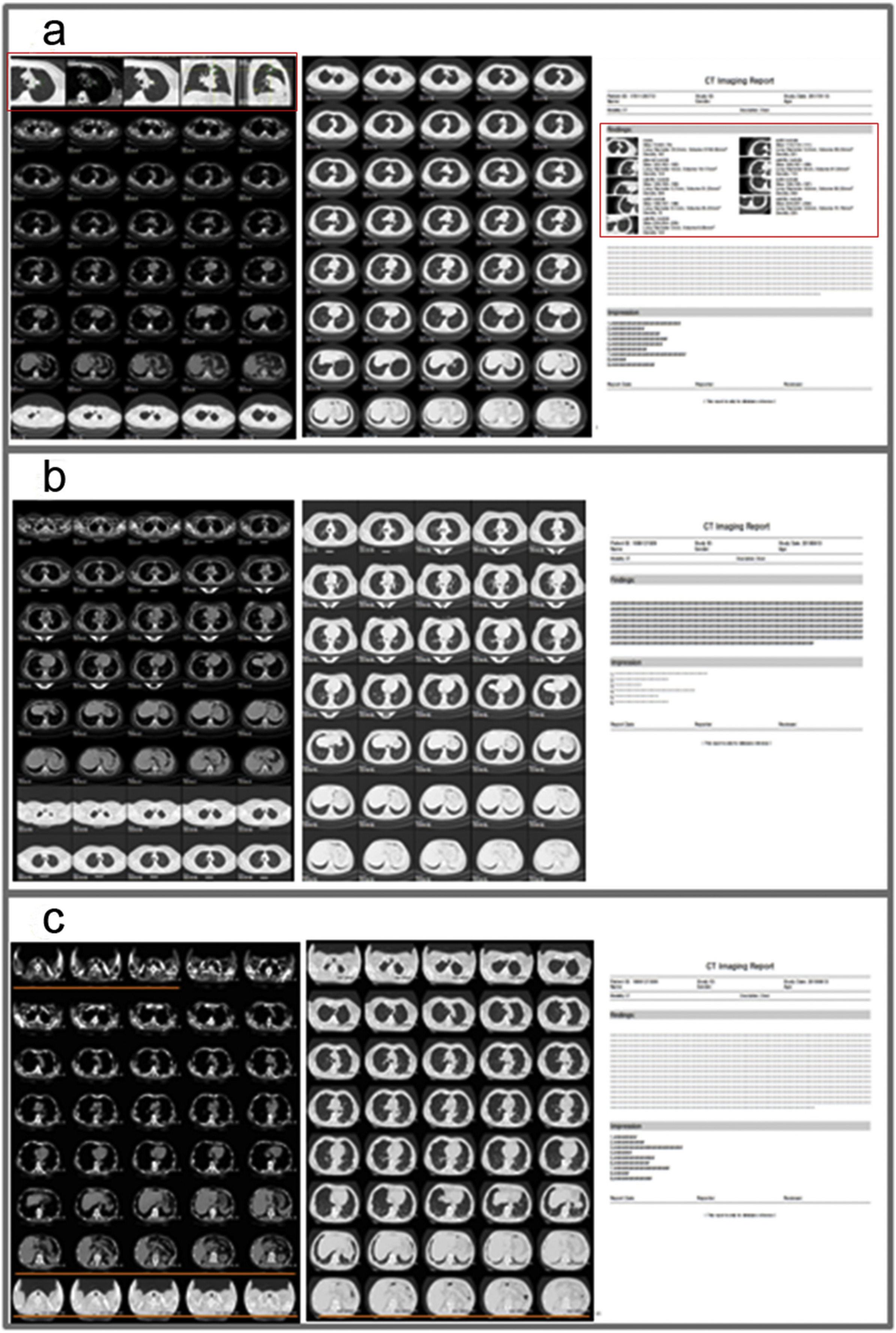
Even though the patients also had larger lesions, we included only nodules smaller than 30 mm, a size that corresponds to a mean diameter of approximately 30 mm, because the definition of lung nodules is a lesion smaller than 3 cm in diameter. We included nodules suspected of being metastases as well as nodules that could potentially have benign histologic features. However, miliary tuberculosis, interstitial lesions, sarcoidosis and severe pneumonia were excluded. The following parameters were used to assess the effect of nodule characteristics and image quality on observer agreement: total nodule size (largest diameter in millimeters), nodule type, benign or malignant, and the density of nodules in the lung parenchyma. The parameters nodule size, benign or malignant and type were extracted from the database. The density of nodules was measured by experts (H. Y and H. W) who did not participate in the reading process. Two approximately 1 cm$^2$ regions of interest were placed in two homogeneous regions within the nodule, and the standard deviation of Hounsfield units averaged over the two measurements was the measure for density.

## 2.3. Patient classification

Randomization was performed by using pseudorandom numbers generated from the random function in the Python Standard Library (Python 3.6.13, Python Software Foundation, Wilmington, Del). Patients in the training process were randomly split into an 80%: 20% ratio in both the training set and validation set (Fig. 4). The training set was used to train the algorithm, the validation set was used for model selection, and the test set was used for assessment of the final chosen model. In deciding the percent split, the goal is to retain enough data for the algorithms to train from but have enough validation and test cases to maintain a reasonable confidence interval of the accuracy of the model [18]. The dataset represents the most common patients with medical solid, calcified or ground-glass nodule(s) presenting and receiving treatment at all participating clinics.

## 2.4. Image labeling

Before training, each image went through a tiered grading system consisting of multiple layers of trained graders of increasing expertise for verification and correction of image labels. Each image imported into the database started with a label matching the most recent diagnosis of the patient. The first tier of graders comprised residents who had basic knowledge of the respiratory system and imaging. This first tier of graders conducted initial quality control and excluded chest CT images containing severe artifacts or significant image resolution reductions. The second tier of graders comprised two experts who independently graded each image that had passed the first tier. The presence or absence of solid, calcified or ground-glass nodule(s) and other pathologies visible on chest CT images were recorded. Finally, a third tier of two senior independent respiratory and imaging experts, each with over 15 years of clinical respiratory and imaging experience, verified the


a
CT Imaging Report
Findings
Impression
b
CT Imaging Report
Findings
Impression
c
CT Imaging Report
Findings
Impression

true labels for each image. A validation subset of all images was graded separately by two expert graders, with disagreement in clinical labels arbitrated by a senior expert, to account for human error in grading.

### 2.5. Auto film layout and structured report

The development of the software system was carried out under the Linux Ubuntu 18.04 environment (Ubuntu 18.04.1 LTS, Bionic Beaver, Boston, Massachusetts, USA). Pycharm (Pycharm 2018.1, JetBrains, Czech Republic) and VS Code (VS Code 1.28, Microsoft, USA) were used as IDEs for development. Chrome debugger was used for testing and debugging UI/UX. The implementation details of the software are confidential, and the following section mainly describes the design and logic of the implementation.

### 2.6. Auto film layout

The auto film layout program was designed to make a productive, quality assured, unattended film printing process (Supplementary Fig. S1). The program first captures the most important nodule in AI (can be overwritten by an operator) findings in different forms and fills the rest of the film with images in both mediastinal and lung windows. The program achieves productivity and quality assurance by enabling traceability. Each image on the film can be traced by its slice ID and redirected to its original location in our image set. This process is performed by separating the filming output process into the following sub tasks: 1, verification; and 2, export. In the verification task, our program first processes the most important nodule, generates five enlarged output images focusing on the nodule along with a highlighted rectangular shape, indicating the position of the nodule in the forms of the lung window, long diameter measuring, mediastinal window and two MPR perspectives. The five output images are placed in the first row of the film, followed by 30 mediastinal window images and the rest in the lung window. Specifically, the first five grids are the automatic layout of a single nodule with the highest risk of malignant probability, which is predicted by the AI. The five pictures can also be verified and overwritten by radiologists. The outputs from the adaptive layout tool include two e-films and a structured report composed of four sets of images, Fig. 5: Set 1, Layout of the key nodule using five images. The first five small cells would be occupied by the nodule with double confirmation by both the AI prediction (the risk of malignancy ≥50%) and a radiologist. Set 2, Layout of the mediastinal window sequence, which is a regular sequence of chest CT images. Mediastinal window images account for 30 grids. Set 3, Layout of the lung window sequence, which is also regular sequence of chest CT images. The remaining 45 grids are all allocated to lung window images. Set 4, Structured report. The structured report is added to the image information and relates the description information of each nodule. As a comparison to traditional reports, the IILS provides the following information: i) basic information display: patient information, check information, radiologist information, etc. ii) findings (double confirmation both from the AI prediction and a radiologist): standardized description of lung nodule images including the nodular location, morphology and density, the number of layers of image information, long diameter nodules, nodule volume, mean CT value for nodules, and malignant probability of nodules. In addition, we have reserved enough space for radiologists to write regular reports for other lesions. iii) Diagnostic impression: diagnostic advice written by the radiologists, Fig. 5. The suitability of the range from the start to the end point of the lung field is the main concern, which means that whether the five images attached to the first set, that is, one set to show the five forms of the nodule with the highest AI predicted score to be malignant, were in line with our predesign will be highly valued. The second and third sets were designed to display images belonging to mediastinum or lung windows to simulate traditional film layout. The fourth set is to show the image with the largest layer of pulmonary nodules in the structured report. Each image can be verified by the operator by tracing its source in the original image set. After verification, the film can then be exported to a printable format to grant visualization to both radiologists and patients alongside automatically generated structure reports. We also reasoned that a good film layout system mainly includes the following three main contents: 1) any hinting key images with any credible, objective measurement data; 2) a series of images to display tumor characteristics, including shape, number, density, size, enhancement, multiangle observation, and follow-up comparison. 3) There is continuous display of chest longitudinal window and lung window images (Fig. 5a–b). Moreover, we show a picture of the current layout form by hand that is very common in daily work as a comparison (Fig. 5c).

### 2.7. Structured report

The Structured Report Generating program was designed to fulfill a complete work flow in a common CT scan scenario (Supplementary Fig. S2). As a comparison to the traditional report, our program provides visualizations of images and findings to both radiologists and patients. The program proceeds mainly as the following three steps: 1, gathering resources; 2, rendering images; 3, exporting. We will now describe each step in detail. For gathering resources, we need to load multiple resources into our program, including the DICOM image set, AI predicted nodules, and patient/hospital information, as well as capturing radiologist findings and diagnostic impressions. After gathering necessary assets, we proceed to the rendering section. The program will first sort nodules by its importance (defined by AI but can be overwritten by operator), then render each nodule using rectangular shape on corresponding images. The program also enlarges the image and sets its center, focusing on the nodule itself. After rendering and transforming, a special event listener is triggered to notify the program to capture the rendered data. Finally, the program generates a predefined printable output.

### 2.8. Quantification and statistical analysis

ROC curves plot the true positive rate (TPR, sensitivity) versus the false positive rate (1-specificity). ROC curves were generated using classification probabilities of malignant nodules and the true labels of each test image and the ROC function of the Python Standard Library (Python 3.6.13, Python Software Foundation, Wilmington, Del). The area under the ROC curve is a measure of performance, and the

**Fig. 5.** An example of the layout plus the visualized structured report and a comparison with the traditional layout plus the report. (a) With the new image layout after IILS selection, the new layout films are divided into three parts (the areas of the two long red boxes that represent the areas where the key images are located). Obviously, there are no invalid images in the films compared with invalid images marked with orange underline in (c). The first part that includes the first five small cells of the beginning on the layout shows in turn: 1) the image of the largest cross-sectional slice of the nodule under pulmonary window conditions (WW: 1500; WL: −500), 2) the image with long and short diameter measurement data, 3) the image of the nodule under mediastinum window conditions (WW: 350, WL: 50), 4) the coronal image reconstruction of the nodule, 5) the sagittal image reconstruction of the nodule. The second part is a group of images per layer interval under mediastinum window conditions. The last part is a group of thin layers of lung tissue images approximately six layers apart to fill all remaining cells. Another convenience is that each image in any cells on the films can be traced by its slice id and redirected to its original location in image set by double clicking the mouse. The visualized structured report related to the films is also automatically generated. See Video 2 for more details. (b): If the patient has no pulmonary nodules, the layout and report given by the IILS will be similar those given by traditional systems. (c): With the traditional manual layout form, the form is divided into two parts. The front part includes the mediastinum tissue images, and the latter part is the lung tissue images. The main problems of traditional layout format are a lack of key images, various invalid images (some images with the orange underline), and a lack of linked function. The related report is filled with text and has no structured report generation.

TPR (sensitivity) at some chosen true negative rate (TNR or specificity) on the ROC curve is the probability that the classifier will rank a randomly chosen “the highest risk of malignancy” higher than a randomly chosen normal. Accuracy was measured by dividing the number of correctly labeled images by the total number of test images. Sensitivity and specificity were determined by dividing the total number of correctly labeled malignant nodules and the total number of correctly labeled benign nodules, respectively, by the total number of test images.

Continuous variables are described as the mean ± standard error of the mean (SEM), and the categorical variables are presented as characteristics such as B/M for benign/malignancy. The clinical characteristics between the traditional image layout group and the intelligent system group and normal controls were compared with the Mann-Whitney *U* test, Chi-square test, or Fisher's exact test when appropriate. The differentiations were compared between the traditional layout group and the intelligent layout group and normal control group using a two-sample Mann–Whitney U test. Kappa statistics were used to measure the degree of consistency between two appraisers, that is, AI and human expert. A kappa value of at least 0·75 indicates good agreement based on the literature [19]. However, we reasoned that larger kappa values, such as 0·90, are preferred. A two-tailed *P* value <0·05 was considered statistically significant. All statistical analyses were executed by R/3.5.0 (https://www.r-project.org/).

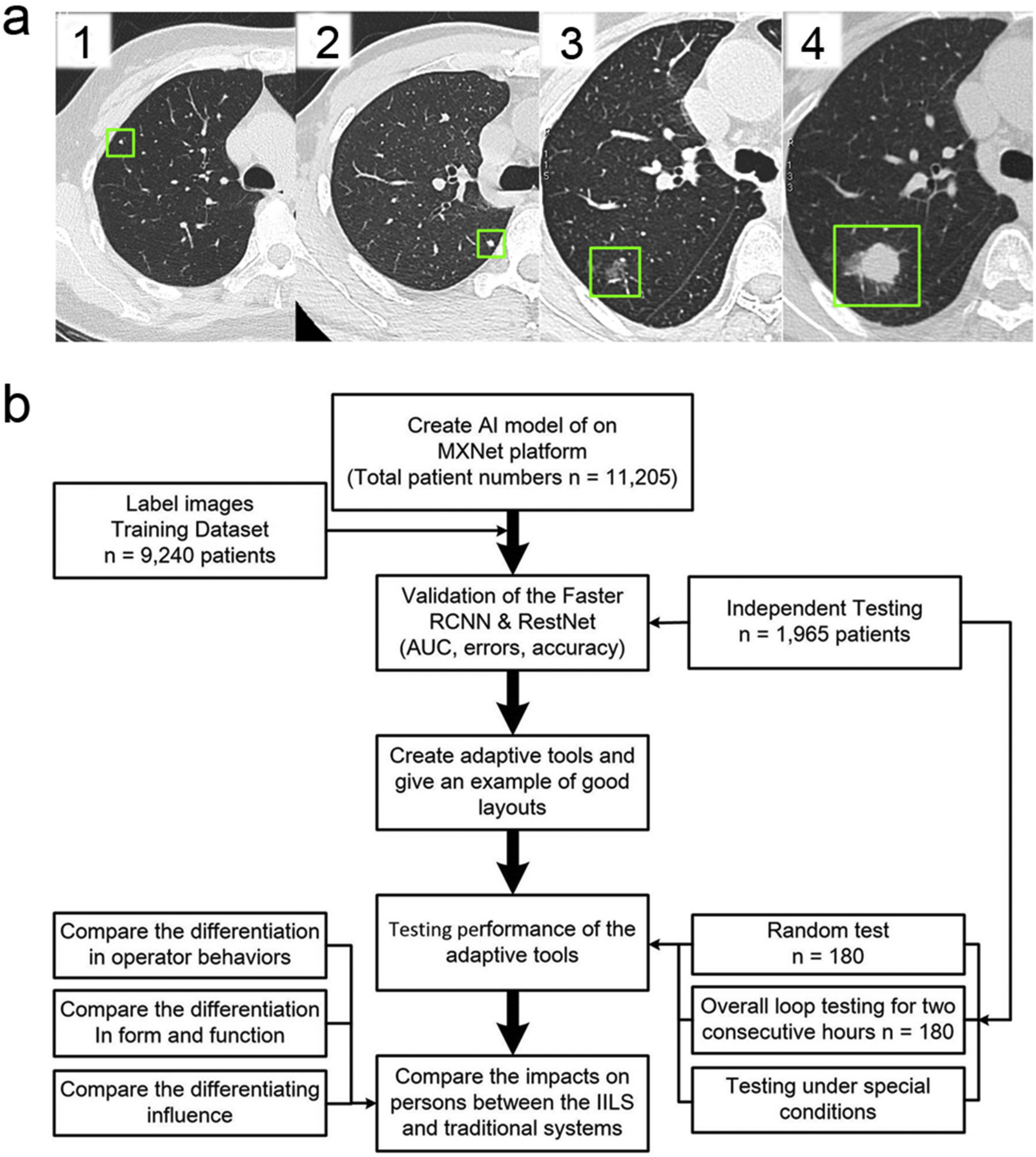

a
1
2
3
4
b
Create AI model of on MXNet platform (Total patient numbers n = 11,205)
Label images Training Dataset n = 9,240 patients
Validation of the Faster RCNN & RestNet (AUC, errors, accuracy)
Independent Testing n = 1,965 patients
Create adaptive tools and give an example of good layouts
Testing performance of the adaptive tools
Random test n = 180
Overall loop testing for two consecutive hours n = 180
Testing under special conditions
Compare the differentiation in operator behaviors
Compare the differentiation In form and function
Compare the differentiating influence
Compare the impacts on persons between the IILS and traditional systems

**Fig. 6.** The demonstrations of the process of detecting different sizes of lung nodules by AI and the workflow diagram of the overall experimental design. (a): The four cases of different sizes of lung nodules with different characteristics and evolution show the importance of follow-up. (1): The upper lobe nodule (diameter: 2 mm, the green square frame) with tags to the pleural surface demonstrates some features of benign and stability at baseline; (2): A solid lung nodule (diameter: 4 mm, the green square frame) also demonstrates some features of benign and stable characteristics; (3, 4): The evolution of a small subsolid nodule in the right lung during the follow-up. (3): Inconspicuous small irregular nodule (diameter: 9 mm, the green square frame) adjacent to the right major fissure demonstrates acute margins to the fissure and does not satisfy criteria for an intrapulmonary lymph node. (4): Significant growth (diameter: 20 mm, the green square frame) is noted in the lesion approximately three months later due to progressive adenocarcinoma. (b): Workflow diagram showing the overall experimental design describing the flow of lung CT images through the labeling and grading process followed by creation of the IILS, which then underwent training and subsequent testing. The training dataset included images that passed sufficient quality standards from the clinical dataset. Subsequently, the output of the IILS was tested and compared with that of the traditional system. Finally, the impact of the process caused by the IILS was also assessed.

**Table 1**
Patient characteristics, number of CT images (patients) from different manufacturers and number of nodules.

| Parameters | Patient (Images) metric |
|---|---|
| Patient characteristics | |
| Number of patients (Training 1 and 2/Independent Testing Datasets) | 9240/1965 |
| Female to male (Training/Independent Testing Datasets) | 1.07:1 (4774:4466)/1.02:1 (993:972) |
| Age (y) (Training/Independent Testing Datasets) | 56 ± 24/55 ± 16 |
| | |
| Different manufacturers | Number of patients in training/testing datasets |
| GE | 1767/399 |
| Philips | 2045/386 |
| Siemens | 1730/181 |
| Toshiba | 1525/290 |
| United Imaging | 1119/378 |
| Control Group (No nodules reported) | 1054/331 (mixed manufactures) |
| Total Number of Chest Images | 2,982,742/544,306 |
| | |
| Different sizes of nodules | Number of nodules |
| ≤ 3 mm | 8080/1880 |
| 3–6 mm | 6127/6461 |
| 6–10 mm | 3544/2195 |
| 10 mm ~ 3 cm | 1460/923 |

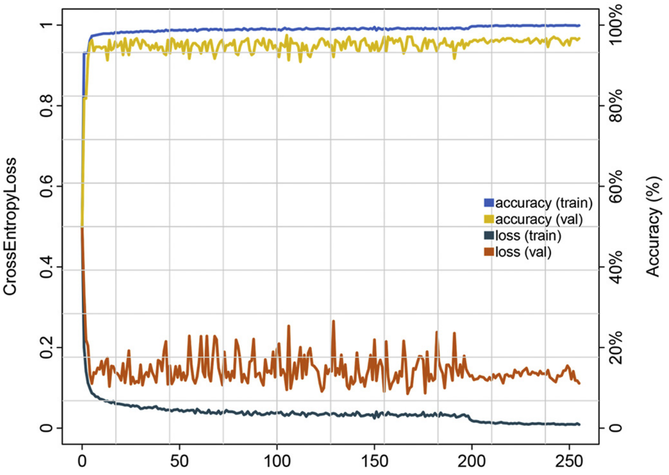


**Fig. 8.** Plot showing the performance of classification in benign and malignant nodules on chest CT images in the training and validation datasets using ResNet. Accuracy is plotted against the training step, and cross-entropy loss is plotted against the training step during the length of the training of the multiclass classifier over the course of >250 epochs. The validation set accuracy and loss show good performance. For model accuracy, the validation set curve converges to 97% (100% for the training process); for the loss function, the validation set curve approaches 0·11 (0 for the training process).

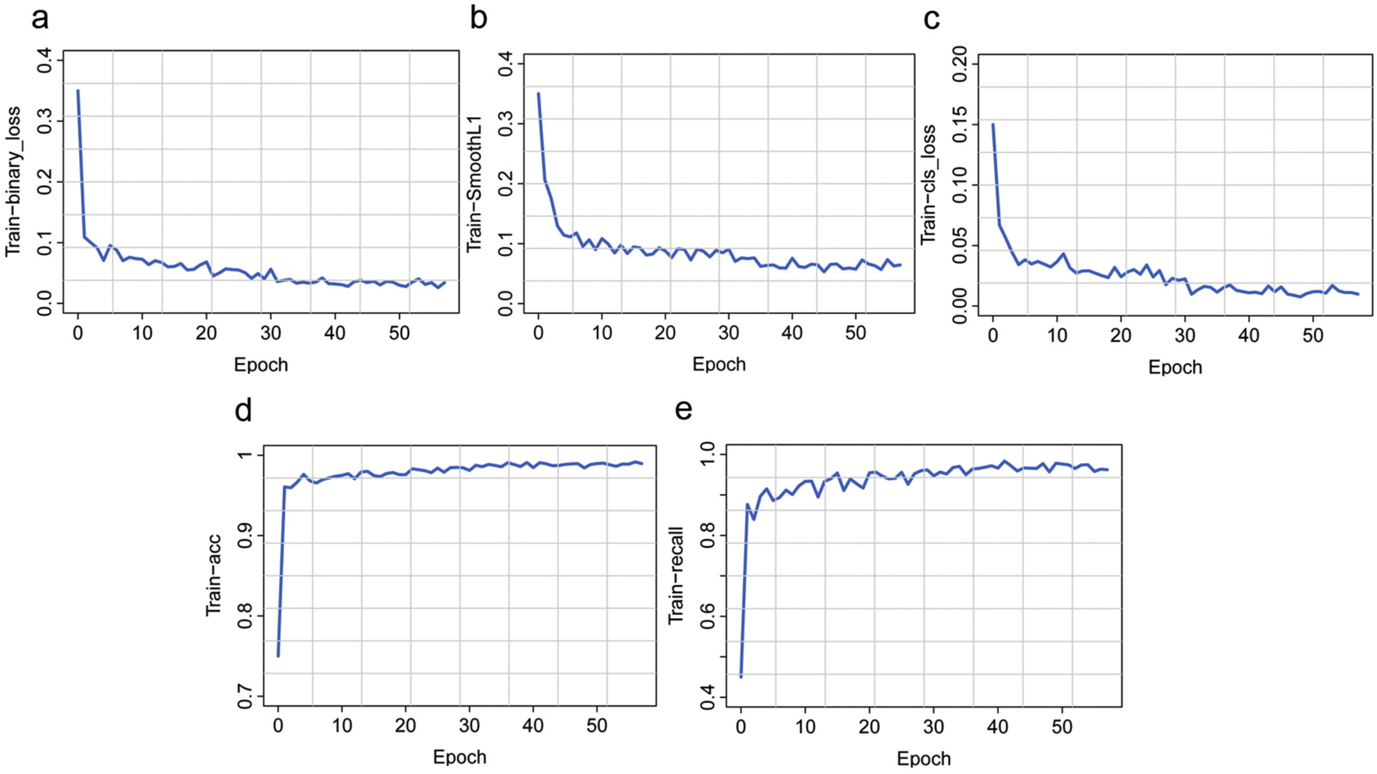


**Fig. 7.** Performance for the training process of detecting nodules. (a): The dichotomy loss embodies the ability to judge whether nodules existed in the mass of anchor frames generated by the detection model. If the overlapping area of the anchor frame and the real nodule box was larger than a certain threshold, a nodule in the anchor frame shall be considered. (b): Position regression loss is harnessed to assess the accuracy of the detection frame position, aiming to make the model-based detection frame close to the real nodule frame, with the same premise as nodule classification loss. (c): The classification loss reflects the ability to determine the nodule category, such as 0–3 mm nodules and 3–6 mm nodules. The premise is that at least one nodule existed in the anchor box and failed otherwise. (d): The accuracy of nodule detection refers to the dichotomic ability of accurately distinguishing nodules and backgrounds. (e): Nodule detection recall rate indicates the number of nodules found in the model compared to the total number, i.e., the recall rate of the model.

### 2.9. Contact for reagent and resource sharing

Further information and requests for resources and data should be directed to and will be fulfilled by the Lead Contact, Bing Zhang (zhangbing_nanjing@vip.163.com). There are no restrictions on the use of the independent testing cohort materials disclosed.

## 3. Results

### 3.1. Patient and image characteristics

Cases with four different nodule sizes along with their characteristics and evolution show the importance of follow-up (Fig. 6a). The characteristics of the patient cohort and nodules used for training, validation

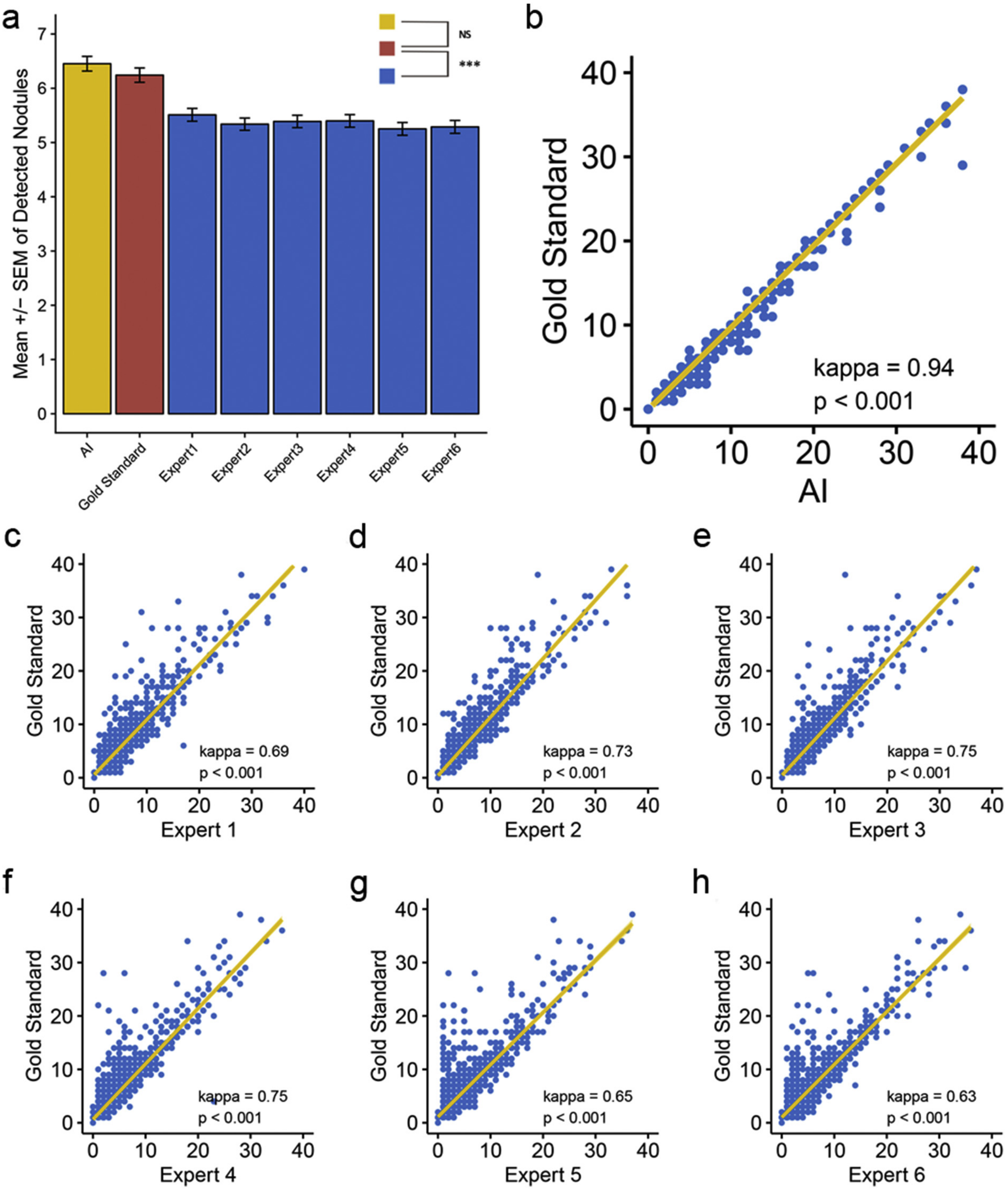


**Fig. 9.** Consistency analysis among AI, human experts and the gold standard in detecting lung nodules. Using the gold standard as a reference, (a) concluded that differences existed in all pairwise Mann–Whitney *U* tests except for AI. Although all kappa consistency analyses were statistically significant (*p* < 0·001), (b-h) demonstrated that AI outperformed human experts in the degree of agreement with the gold standard by a kappa coefficient of 0·94. 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 <0·0.01, *** for <0·005 and NS for no significance.

and independent testing datasets are summarized in Table 1. Nodule type was classified to be solid, calcified or ground glass according to the literature [20]. The overall experimental design of the workflow diagram is shown in Fig. 6b.

### 3.2. Model design and performance evaluation

The core design of the IILS system was the deep learning model, which was divided into two parts, Faster RCNN and ResNet. Faster RCNN is primarily responsible for the detection and localization of pulmonary nodules. Faster RCNN also aids in classifying pulmonary nodules into the following classes: 0–3 mm, 3–6 mm, 6–10 mm, 10–30 mm pulmonary nodules, solid nodules, ground-glass nodules (GGNs) and calcified nodules. The second part is ResNet, and its main responsibility is to classify benign and malignant pulmonary nodules. In part one, regarding feature map extraction, we used layer conv4_x in ResNet-50 as the CNN output in the base of Faster RCNN. In our case, layer conv4_x in ResNet-50 exhibited the best performance in detection. In the region proposal network (RPN), we used binary cross entropy as the classification loss function and selected the smooth L1 loss function as the regression loss function. The training process of the model was perfect, and all the curves reached convergence. The training process of RPN is reflected in Fig. 7a-b, and all curves converged to zero. This finding also indicates that our model can distinguish the foreground and background well and provide a precise bounding box of the foreground. The curve also converged to zero (Fig. 7c). The convergence of this curve means that the model can distinguish seven classes of pulmonary nodules very well. The curves converge to 1 in Figure7-d-e, reflecting that our detection model could distinguish nodules and background with high precision and accurately identify nodules. In part two, we classified benign and malignant nodules by ResNet. The network layer became deeper but not because the model became more accurate. In contrast, our model was inaccurate, and a series of problems occurred. Thus, to obtain a more accurate model and avoid problems such as gradient dispersion, we chose ResNet. In the IILS system, ResNet acted as a good classifier to finish the job. The convergence of the curves represents the success of our classification job; for model accuracy, both training and validation curves approached 100% (100% for the training process and 97% for the validating process). In the loss function part, the curves also show that the model performed very well in classifying benign and malignant nodules. The training process converged to 0·11 for the validating process (Fig. 8).

### 3.3. The comparison of the diagnostic efficiency for nodules between IILS and human experts

We evaluated our model in detecting and classifying the most common pulmonary nodules. This model detected and classified images with nodules of different grades of benign and malignant tumors as a “primary layout nodule”. These conditions would demand relatively urgent referral to related respiratory physicians or thoracic surgeons for definitive treatment. The system categorized images with benign lung nodules or false positive nodules, which have a low probability of becoming a malignant tumor, as “only shown in the visualized structured report”. Microscopic nodules that are very common in clinical work are not indicated for malignant tumors; therefore, referral to a related expert for treatment is less urgent.

Here, we sought to decipher the advantages of AI in detecting lung nodules compared to human experts. In this study, we resorted to a simple and intuitive way, that is, evaluating the degree of agreement between the detected nodules and those screened by the gold standard. We conducted consistency analysis, using the pathological gold

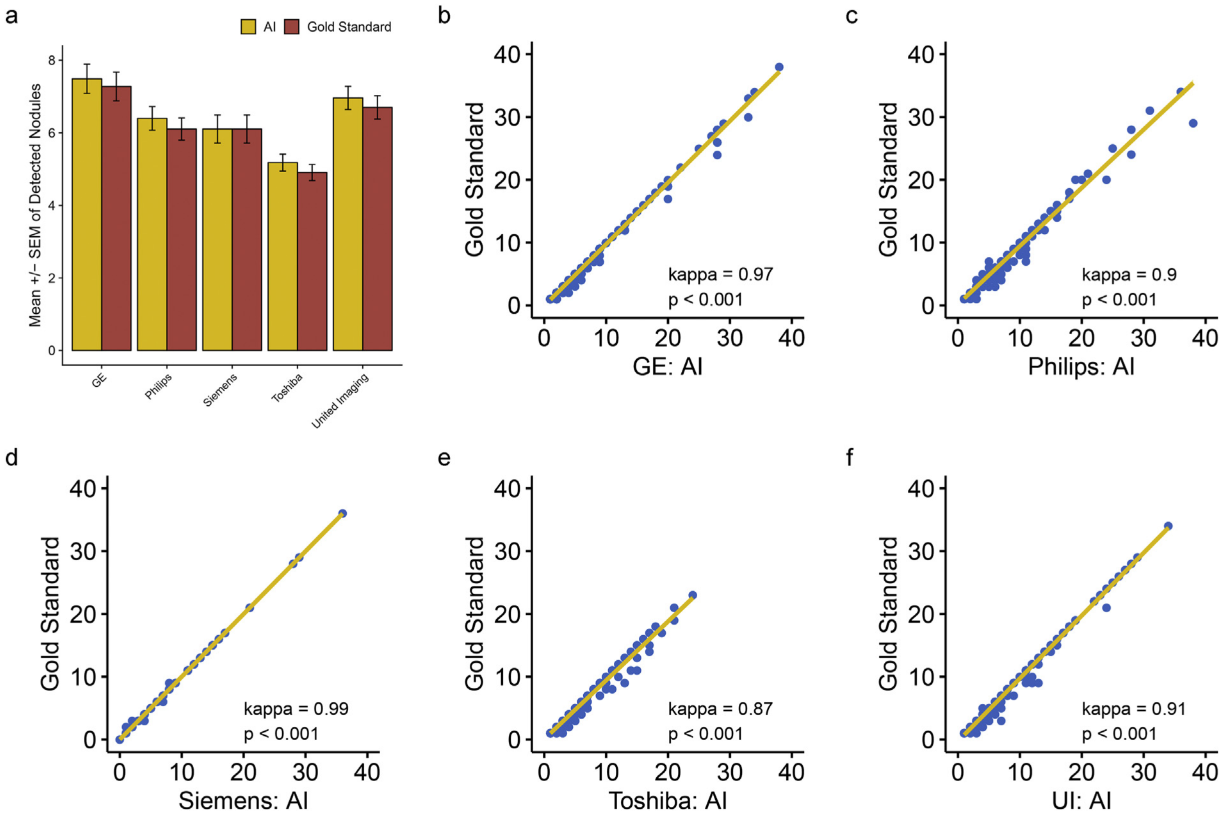


**Fig. 10.** Performance of AI in the consistency of lung nodule diagnosis when applied to imaging equipment from five different manufacturers. Using the gold standard as a reference, (a) no significant difference was observed regardless of the type of manufacturer ($p > 0·05$). (b-f) demonstrated that in all kinds of manufacturers, AI represented highly significant consistency with the gold standard (kappa coefficient range from 0·87–0·99, $p < 0·001$). The horizontal and vertical coordinates for (b-f) indicate the detected nodule number.

standard as a reference, by the kappa consistency coefficient and two-sample Mann–Whitney *U* tests separately. Strikingly, a difference existed in all pairwise comparisons except for AI ($p = 0{\cdot}138$ for AI, $p < 0{\cdot}001$ for other comparisons, see Fig. 9a and Supplementary Table S1). Compared with human experts, AI can also be significantly consistent with lung nodules detected by the gold standard, whereas AI stood out due to its highest consistency coefficient (kappa = 0·94 for AI, $p < 0{\cdot}001$ for all comparisons, see Fig. 9b-h and Supplementary Table S1). We further compared the consistency of the detected nodules in different size ranges in more detail and demonstrated that regardless of the nodule size range, AI showed a much more favorable consistency with the gold standard that exceeded human experts. Specifically, kappa = 0·93 for 0–3 mm nodule detection, kappa = 0·97 for 3–6 mm nodule detection, kappa = 1 for 6–10 mm and 10–30 mm (all $p < 0{\cdot}001$). AI was only significantly different ($p = 0{\cdot}013$) from the gold standard in the diagnosis of nodules of 0–3 mm, whereas the human experts showed significant differences in all sizes of nodules, detecting these nodules to varying degrees (Supplementary Fig. S3–6).

### *3.4. The cross-manufacture applicability of IILS*

Now that AI has been confirmed to be superior to human experts in detecting nodules regardless of size, it is necessary to judge the applicability of AI from another angle. Essentially, diagnosis by AI depends on the images produced by the existing manufacturers; thus, evaluating the influence of image output from different manufacturers on nodules detected by AI is reasonable. For the sake of exploring the adaptability of AI to different imaging manufacturers under the condition in which the gold standard was referenced, we further assessed the consistency of AI with the gold standard in diagnosing nodules in various size ranges on different manufacturers by pairwise Mann–Whitney U tests and kappa consistency analysis. Overall, AI was well configured on five manufacturers with no difference compared to the gold standard ($p = 0{\cdot}576$ for GE, $p = 0{\cdot}472$ for Philips, $p = 0{\cdot}988$ for Siemens, $p = 0{\cdot}376$ for Toshiba and $p = 0{\cdot}343$ for United Imaging (UI)). In addition, high consistency was achieved when referring to the gold standard with a kappa coefficient ranging from 0·87 to 0·99 (all $p < 0{\cdot}001$, Fig. 10). In regard

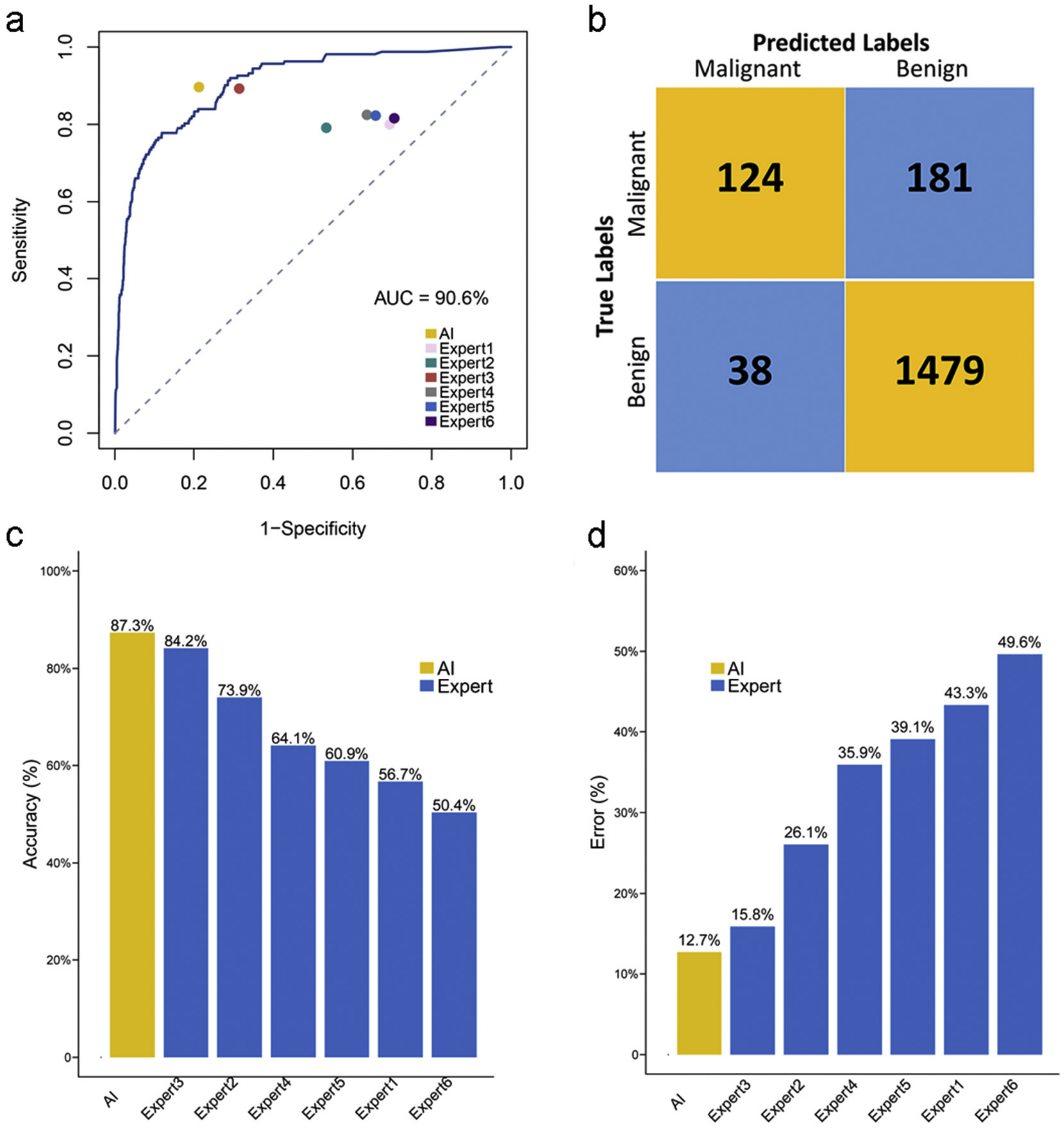


**Fig. 11.** Evaluation of performance for AI in recognizing benign or malignant lesions. (a) The corresponding area under the ROC curve for the graphs is 90·6% for malignant pulmonary nodules versus benign nodules. The comparison between the predictive performance of model and human expert on another independent cohort of 284 patients was shown around the curve with corresponding false positive rate (TPR, sensitivity) and true positive rate (FPR, 1-specificity). (b) Contingency table for predicted labels and true labels of malignant and benign based on a cutoff score of our original model. (c-d) Ordered accuracy and error bar to assess the performance of AI when diagnosing the lesion status of lung nodules.

**Table 2**
Five-point scales for evaluation of consistency and accuracy of layout.

| Scales | Scores[a] | Description |
|---|---|---|
| Range of lung field[1] | 1/0 | Consistency and accuracy of the starting and ending position of the lung field |
| The first set[2] | 1/0 | Consistency and accuracy of showing the most suspicious malignant nodule in five morphological forms (no nodule no count) |
| The second set[3] | 1/0 | Consistency and accuracy of showing the images of mediastinum |
| The third set[4] | 1/0 | Consistency and accuracy of showing the images of lung |
| The fourth set[5] | 1/0 | Consistency and accuracy of showing lung nodules in the structured report |

[a] Record 1 point if it meets the requirements for layout, otherwise record 0 points with the same weight for all five scales.
[1] Range of lung field: range from the beginning to the end of lung fields.
[2] The first set: for the one nodule most suspicious detected and predicted by the IILS.
[3] The second set: for displaying mediastinal tissue.
[4] The third set: for displaying lung.
[5] The fourth set: For displaying images in the structured report.

to a specific size range, interestingly, a significant false positive rate was observed when detecting 0–3 mm lung nodules with only UI ($p = 0{\cdot}006$) (Supplementary Fig. S7). Consistency remained high regardless of the kind of nodule detected across different manufacturers. Specifically, the kappa coefficient ranged from 0·86 to 0·99 for 0–3 mm nodules, 0·95 to 1 for 3–6 mm nodules, and 0·99–1 for 6–10 mm and 10–30 mm nodules (all $p < 0{\cdot}001$, Supplementary Fig. S7–10). This finding indicates that overall, AI performs well with each manufacturer (Supplementary Table S2).

### 3.5. Validation of the outperformance for IILS against human experts in diagnosis by an independent cohort

The data from another independent cohort of 284 patients with pathological results were imported to compare malignant pulmonary nodules with benign nodules using the same datasets to determine the accuracy of the model's performance. We reasoned that our original predictive model is completely clinically applicable since its area under the ROC is up to 90·6% for malignant pulmonary nodules versus benign nodules (Fig. 11a). Under a score cutoff of 0·5, 124 cases were predicted to be true positive and 1479 to be true negative. Approximately 38 cases were labeled as false positive, and 181 were false negative. Thus, a sensitivity of 76·5% and specificity of 89·1% were achieved (Fig. 11b). The comparison between the predictive performance of model and human expert on another 284 patient cohort with pathological gold standard is shown around the curve, which concluded that AI (0·21 for FPR and 0·90 for TPR) outperformed the other six experts in both sensitivity and specificity. In addition, compared with human experts, AI showed the highest accuracy; 248

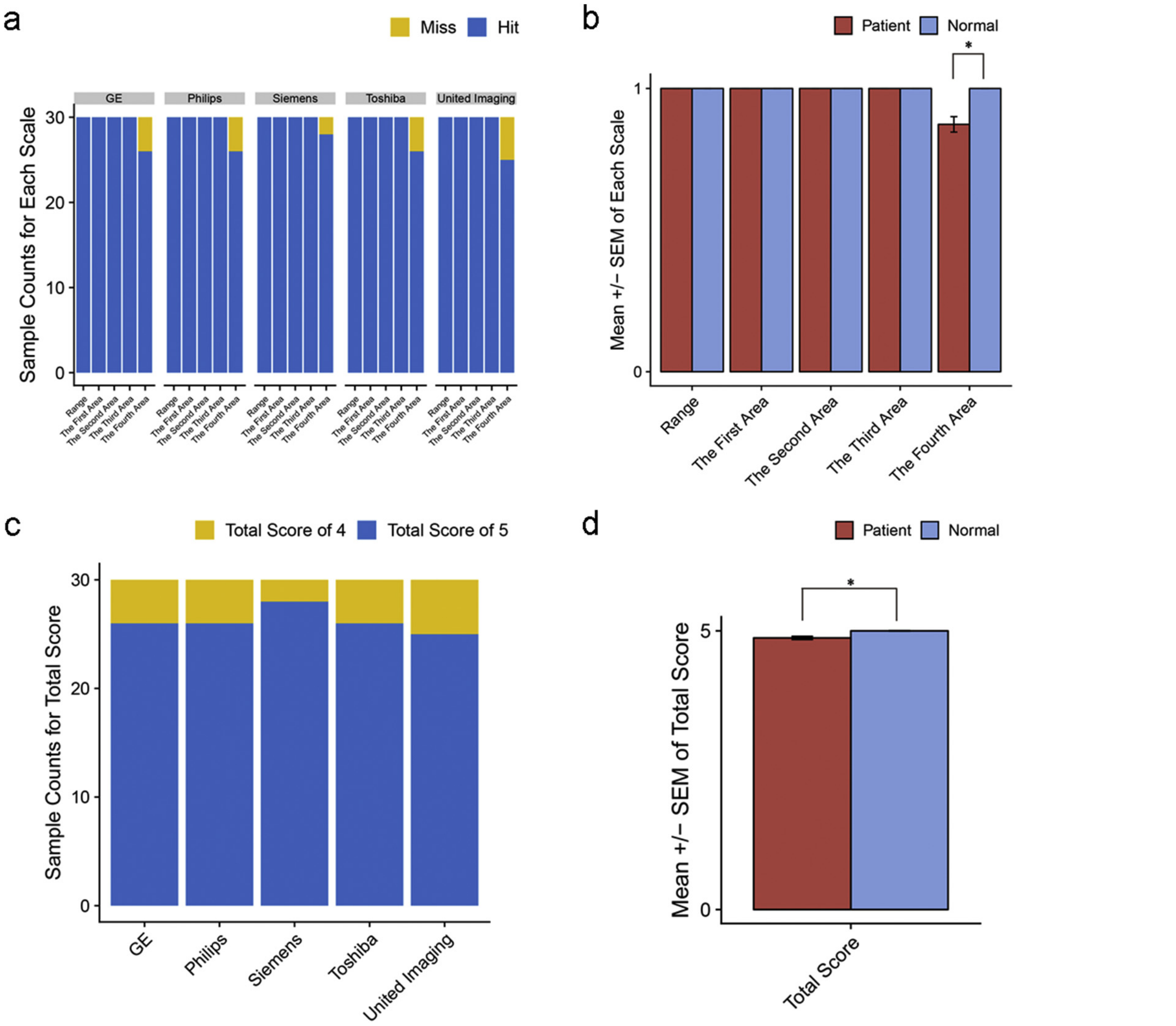


**Fig. 12.** Quantification of IILS deployment in five manufacturers. (a) Histogram showing scores in each scale across 5 manufacturers, and (b) each scale was compared between the patient and normal control groups among all manufacturers. Total score, which was summed by each scale, was compared in the same way as in (c-d). Statistical significance is labeled * for <0·05.

cases were predicted correctly (87·3%, number of patients who were predicted to be true positive or true negative divided by 284), and the corresponding error was the lowest, with 36 mislabeled cases (12·7%, number of patients who were predicted to be false positive or false negative divided by 284) (Fig. 11c-d).

### 3.6. Design and evaluation of automatic adaptive layout tool

After our discussions with six experts (three radiologists and three clinicians), according to the requirements of the 2018 NCCN guideline [21], a final consensus was reached regarding a good image layout form. To simulate the results of daily work after chest CT scanning in the medical imaging department, we designed an automatic adaptive layout tool that produces the “Auto Film Layout and Lung Nodule Structured Report” to connect to the outputs of the CNN network. Automatic adaptive layout tools can export film layouts of key lung nodule images (the nodule with an increased risk of malignancy) and generate a structured report. Both film layouts are used in the fixed format (5 × 8 grids on one film). The chest CT images of 180 patients were mixed and continuously input to the IILS to simulate the condition in which images from different manufacturers enter a workstation in daily work. The characteristics of the adaptive layout tool and whether it could be successfully layout were evaluated. The total number of images for these 180 patients was 60,660, including 12,240 mediastinal window images and 48,420 lung window images. A 5-point scale method was used to eventually judge the layout of e-films, as shown in Table 2.

### 3.7. High quality of IILS by a five-point scales assessment

We harnessed five-point scales to evaluate the quality of our IILS (Table 2). Only a few scores could not hit the fourth scale across all kinds of manufacturers (Fig. 12a). No misses were found with normal layout cases, whereas misses on several cases in the fourth area were observed in patients compared to those in normal controls ($p = 0{\cdot}04$) (Fig. 12b). We also compared the total score derived from adding each scale score, and overall, Siemens worked best when deployed with the IILS (Fig. 12c). The total score indicated that the IILS might be more suitable for nodule-free people than for patients with nodules ($p = 0{\cdot}04$) (Fig. 12d).

### 3.8. Advantages over traditional workstations: less time-consuming, no invalid images and zero omission for IILS

Considering that clicking time is necessary, the average number of clicks of layouts from five main manufacturer devices was 14·45 ± 0·34 (Supplementary Table S3). Specifically, the average number of clicks was 14·37 ± 0·89 for GE, 14·70 ± 0·86 for Philips, 14·57 ± 0·87 for Siemens, 15·77 ± 0·95 for Toshiba and 13·67 ± 0·79 for UI, whereas 2 clicks were observed in the IILS ($p < 0{\cdot}001$) (Fig. 13a, Supplementary Table S4). More clicks were required when using traditional workstations than IILS for both patients ($p < 2{\cdot}2\text{e-}16$) and nodule-free normal people ($p = 1{\cdot}1\text{e-}12$). There was no significant difference in laying out images for patients or normal people between IILS or traditional manufacturers ($p > 0{\cdot}05$) (Fig. 13b-c). With layout images from 250 patients (50 patients for each five manufacturer) by different

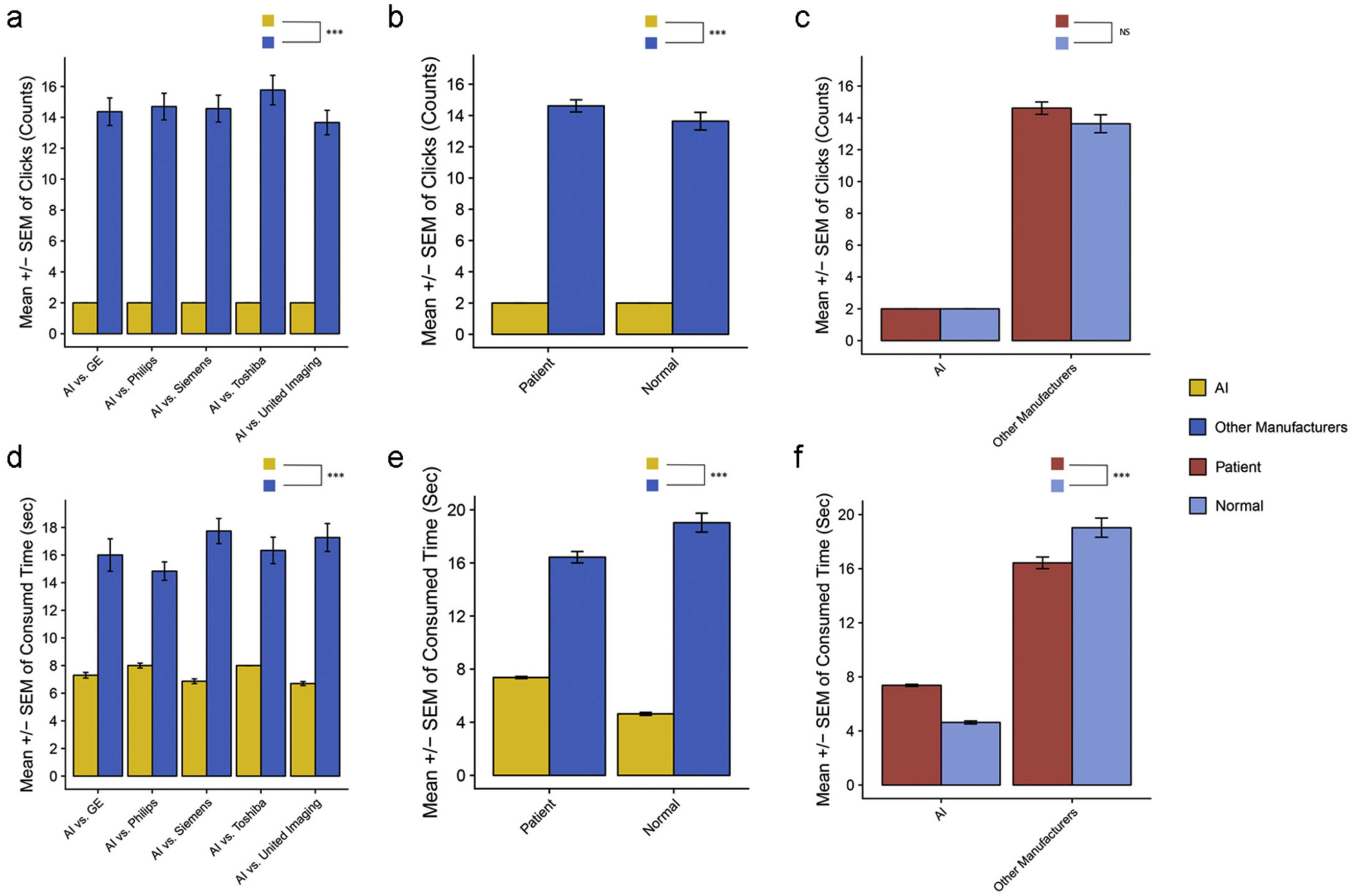


**Fig. 13.** Comparison of the mouse clicks and time consumed between the IILS and traditional manual layout. Error bars for comparing mouse click time between the IILS and each traditional workstation in (a). (b) Click time was compared between IILS and mixed workstations within patient and normal cases separately and was in turn compared between patient and normal cases within the IILS and traditional workstation separately in (c). Same comparison was applied to the time consumed and is shown in (d-f). Statistical significance is labeled *** for <0·005 and NS for no significant difference.

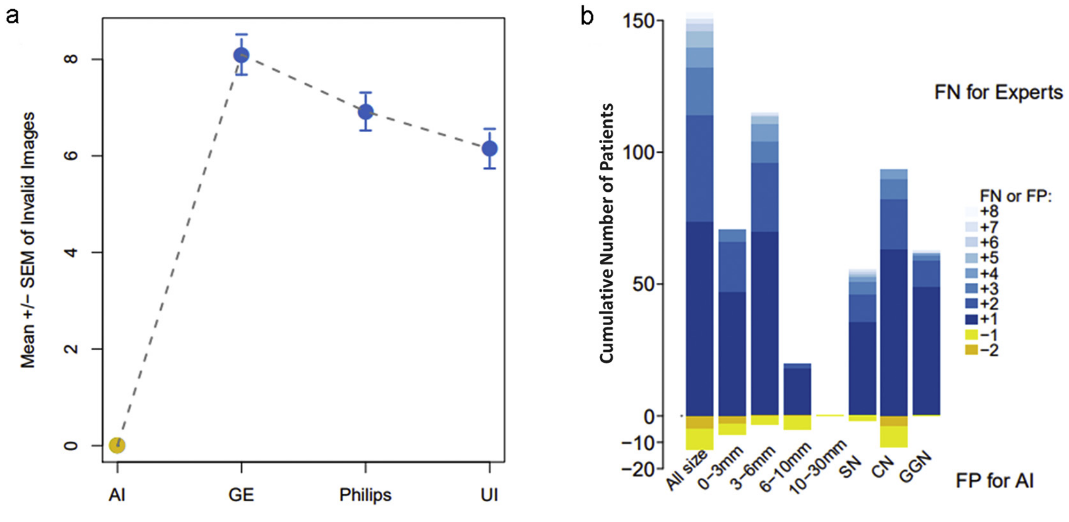


**Fig. 14.** Comparison of the production of invalid images from different layout systems and statistics on the missing lung nodules in clinical imaging reports of missed diagnoses. (a): Comparison of the results derived from the IILS and three traditional layout workstations. No invalid images in AI compared to 8·10 ± 0·42, 6·92 ± 0·39 and 6·15 ± 0·41 invalid images for GE, Philips and United Imaging separately per patient. IILS is marked in yellow, and other workstations are marked in blue with the mean ± SEM. (b): Statistical results after detection and classification by the IILS from randomly selected control group data that were derived from the images of patients who were considered by radiologists to have no pulmonary nodules in their clinical imaging reports. The IILS would not miss any nodules during the entire workflow where the performance was far superior to the traditional method, which approximately missed the lung nodules (false negatives, FNs). However, compared to human experts, the IILS could cause some false positives (FPs).

manufacturer workstations, the amount of time required for 50 patients for each manufacturer was 8 min (GE), 7·42 min (Philips), 8·87 min (Siemens), 8·17 min (Toshiba) and 8·63 min (UI). The average layout time by using the traditional CT workstation is 16·87 s/patient (Supplementary Table S3). In contrast, the IILS requires approximately 6·92 s/patient (Fig. 13d, Supplementary Table S5). Significantly less time was consumed using IILS than using other manufacturer workstations for patients (p < 2·2e-16) and nodule-free (p = 1·6e-11). As we expected, the IILS is more efficient than traditional workstations in both patients and normal cases. Interestingly, the IILS spends less time on nodule-free cases, while traditional workstations take longer (Fig. 13e-f).

In addition, the differences between two layout forms could be regarded as the second major category and be described from 8 aspects. First, we compared the invalid images derived from two different layout systems. Second, we randomly collected fifty layout results from three manufacturer workstations, with a total of 150 results. With the traditional layout performed manually, the number of invalid images per patients was 8·10 ± 0·42, 6·92 ± 0·39 and 6·15 ± 0·41 for GE, Philips and UI, respectively, compared to 0 per patient for the IILS (p < 2·2e-16 for all) (Fig. 14a). We assessed whether the lung fields in each grid of the film fit each size appropriately. All of the experts subjectively concluded that significant differences existed (Supplementary

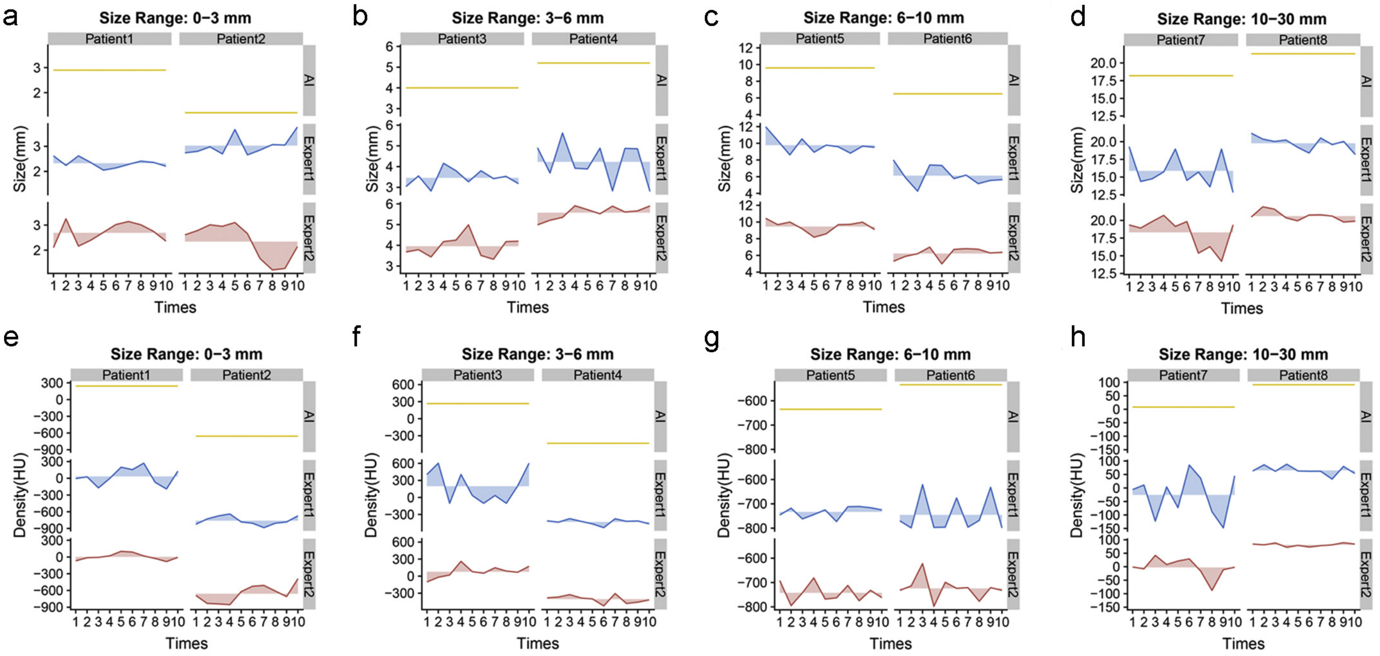


**Fig. 15.** Two types of measurements of lung nodules are shown during repeated testing. The ribbon map was used to describe the stability of repeated measurements or reproducibility. Straight lines, such as the yellow lines for the IILS, indicate that the IILS was always repeatable, and the fluctuations of the blue or red lines, Expert1 or 2, quantify the stability in the opposite direction. That is, greater amplitude resulted in worse reproducibility. (a-d) for size and (e-f) for density.

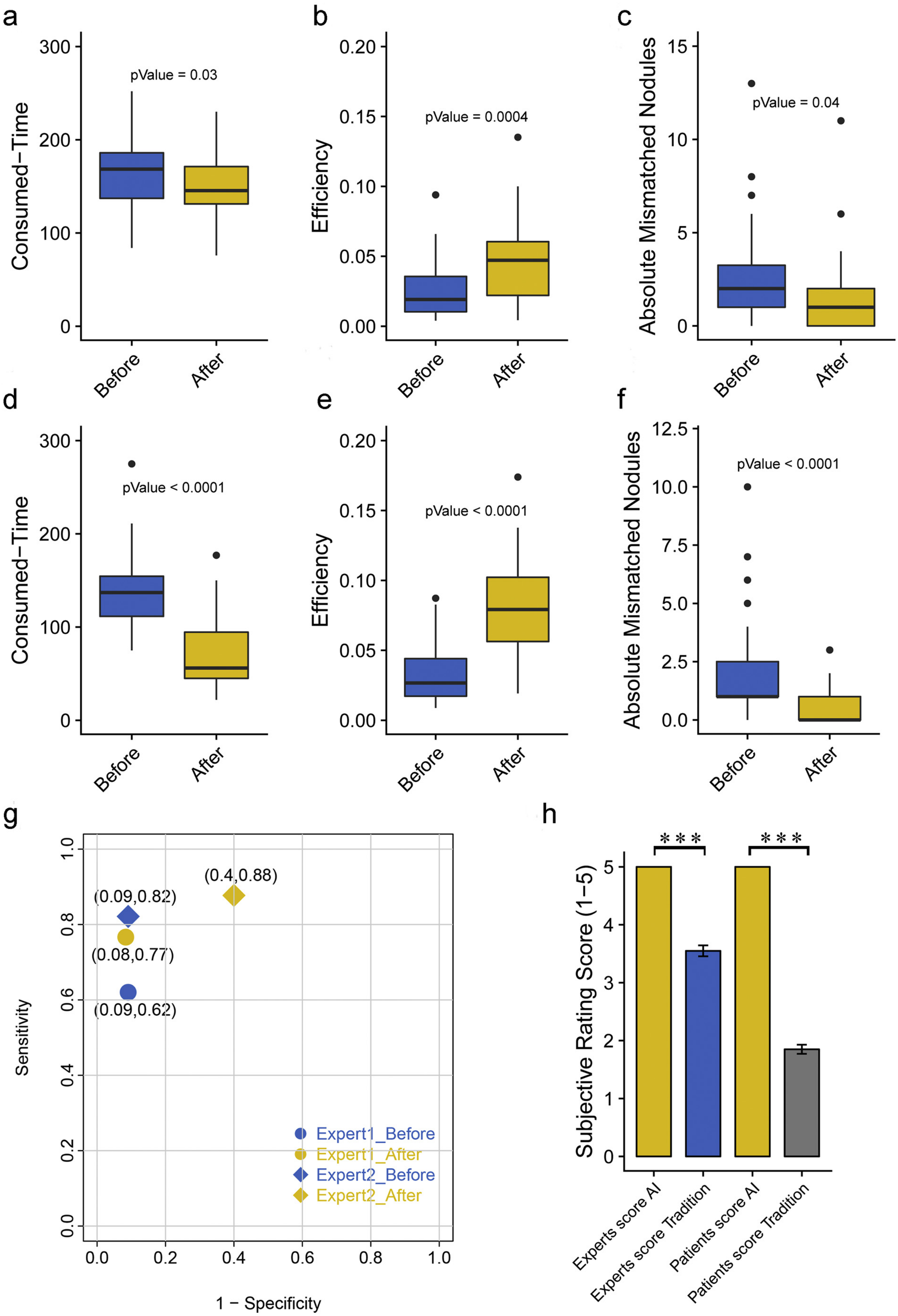

a
Consumed−Time
300
200
100
0
pValue = 0.03
Before
After
b
Efficiency
0.20
0.15
0.10
0.05
0.00
pValue = 0.0004
Before
After
c
Absolute Mismatched Nodules
15
10
5
0
pValue = 0.04
Before
After
d
Consumed−Time
300
200
100
0
pValue < 0.0001
Before
After
e
Efficiency
0.20
0.15
0.10
0.05
0.00
pValue < 0.0001
Before
After
f
Absolute Mismatched Nodules
12.5
10.0
7.5
5.0
2.5
0.0
pValue < 0.0001
Before
After
g
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
(0.4,0.88)
(0.09,0.82)
(0.08,0.77)
(0.09,0.62)
Expert1_Before
Expert1_After
Expert2_Before
Expert2_After
0.0
0.2
0.4
0.6
0.8
1.0
1 − Specificity
h
Subjective Rating Score (1−5)
5
4
3
2
1
0
***
***
Experts score AI
Experts score Tradition
Patients score AI
Patients score Tradition

Table S3). We further investigate whether two layout measures and workflows could cause omissions in lung nodule detection. Two radiologists were required to indicate the locations of nodules derived from two chest CT films and reports on different platforms. The radiologists also recorded the diameter of missing nodules. We found a total of 318 mismatched nodules according to the report description and missed nodules in 46·8% of patients, which is approximately 0·97/patient. Additionally, a diagnosis of GGN was missed in 63 out of 327 patients (19%). As expected, IILS missed zero nodules compared with an expert; however, IILS might generate more false positives than human experts (Fig. 14b, Supplementary Table S3).

### *3.9. A complete reproducibility of IILS against an instability of human measurements*

We compared the results from several aspects in regard to how to display the nodules appropriately (Video 2, Supplementary Table S3). Since the gold standard is lacking, we evaluated which standard is more stable or reproducible. In this regard, a total of eight lung nodules were selected, namely, two nodules in four different sections (size: ≤ 3 mm; 3–6 mm; 6–10 mm; 10 mm ~ 3 cm). Subsequently, two radiologists were required to repeatedly measure the size (Fig. 15a-d) and density (Fig. 15e-f) of these eight nodules ten times in different periods through workstations; the IILS did the same. The IILS measurements were 100% reproducible and represented perfect consistency in multiple measurements, whereas measurements assessed by experts exhibited varying degrees of fluctuations (Supplementary Table S3). It is imperative to embed the positive predictive value of lung nodules and the ability of tracing and redirecting to its original location in an image set by double clicking the mouse on the target image in any cells on the e-films (for more details, see Supplementary Table S3, Method and Video 2).

### *3.10. Human-machine coupled operation requires an adaptation process*

We simulated the normal working scenario of radiologists where two radiologists were asked to make judgments on as many of the 284 patients as possible within two hours, using a traditional diagnosis or a re-diagnosis based on AI judgment after an interval of one month. Significant improvements were found in consumed time, efficiency and absolute mismatched nodules after applying AI judgment as prior information ($p < 0·05$). Specifically, based on the existing judgment of AI, two experts not only reduced the diagnosis time for the same image but also improved the diagnostic efficiency per unit time (detected number of nodules/consumed time). The detection error (absolute number of nodules detected by the gold standard minus human diagnosis) was also significantly decreased (Fig. 16a-f). The detection sensitivity for both experts was improved after using AI, but the specificity for expert 2 was reduced from 99·2% to 60% (Fig. 16g).

### *3.11. The satisfaction of both experts and patients demonstrates friendliness of IILS*

Six experts who blinded to the origin of the results evaluated the results from traditional systems and the IILS. We established a 5-point scoring mechanism for evaluating the layout e-films and reports (or visualized structured reports) produced by two different layout systems (Table 3). Experts gave a significantly better evaluation for IILS with all 5 points than traditional method with approximately 3 or 4 points ($p = 7·674\text{e-}23$). In contrast, patients scored with more extreme points, indicating that the friendliness of the report is very important ($p = 8·164\text{e-}25$) (Fig. 16h).

### *3.12. The performance of IILS evaluated on LUNA16 and LIDC/IDRI benchmark*

The performance of IILS was evaluated on two benchmark databases, that is, LUNA16 and LIDC/IDRI. A total of 888 CT scans of Luna16 (https://luna16.grand-challenge.org/) were tested first and the final score is 0.696 (ranked 18), which is defined as the average sensitivity at 7 predefined false positive rates: 1/8, 1/4, 1/2, 1, 2, 4, and 8 false positives per scan (see more details concerning about the FROC and CAD analysis in Supplementary Fig. S11 and Supplementary Table S7). We further tested our IILS on LIDC/IDRI database (https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI) with 1018 CT scans, which contains more slice thickness and is more similar to real clinical environment. For nodules larger than 3 mm, the recall rate of the model is 88.75%, and the false positive rate is 5.22 false positives per scan.

## 4. Discussion

In this study, by creating and deploying a deep neural network algorithm, our model of the IILS demonstrated competitive performance of chest CT image analysis with a limited need for human behavior. Moreover, the efficacy of the machine learning technique for image analysis likely extends beyond the realm of chest CT images—in principle, the techniques by learning through AI and layout could potentially be employed in a wide range of medical images across multiple disciplines.

A major feature of IILS is almost real-time lung nodule detection. This real-time performance is all due to the Faster RCNN model in the system [22]. The performance of IILS depends highly on the accuracy of detecting and classifying nodules through the trained model. A high agreement with the gold standard was reached. However, there was a significant difference in detecting small nodules by applying AI to UI, which might be caused by small sample size (number of nodules, $n = 1119$) for UI enrolled in the model training. The IILS was confirmed to be superior to six experts in terms of the number of detected nodules and the judgment of benign or malignant status. According to the current constructed model, the area under the curve of the obtained ROC curve was up to 90.6% which was reasoned to be clinically applicable. After rigorous statistical testing, the IILS was confirmed to be superior than six human experts in terms of the number of detected nodules and the judgment of benign or malignant status. There is a Supplementary Table S6 shows some of the relevant work and the results of this comparison. By contrast, the experimental data and the results of the CNN architecture have made some progress, enabling us to be full of hope that the model performance of the IILS is stable, reliable and efficient.

The IILS is designed to be used in the process of daily practical work to accurately detect and classify nodules and to standardize chest CT images and reports. The advantage of this layout is simplification of process where doctors carefully flipped through images to find lung nodules with key images. To optimize the IILS, we evaluated its performance in layout parts. The overwhelming 100% success rate is dependent on the AI output with multi-planar reconstruction program design and is automatically completed. The multi-planar reconstruction is essential for clinicians to observe pulmonary nodules from multiple perspectives, make final diagnosis, evaluate and follow up pulmonary nodules.

**Fig. 16.** Evaluate the impact of AI diagnosis on human expert judgment. Two radiologists (Expert1 was senior and the other was junior) were invited to read images before using AI or after AI diagnosis within a month. Impact was considered in the following aspects: consumed time, efficiency and absolute mismatched nodules. Boxplots show significant differences for all three aspects for expert1 in (a-c) and for expert2 in (d-f). Sensitivity and specificity for each expert shown in (g) were calculated by malignancy status compared to the pathological gold standard. H) Subjective evaluations made by six experts and six patients on the e-films and reports produced by the traditional workstations and the IILS. Scores are shown in error bars with the mean ± SEM.

There are fourteen differences between IILS and the traditional layout system (Supplementary Table S3). Among them, the content was divided into three parts. The first part was focused on benefiting operators. IILS might have opportunities to reduce costs, including increasing the efficiency of utilization of CT, substituting lower cost resources and even replacing some operations. The second part includes the contents from the fourth to the eleventh points. These differences are mainly concentrated on the differences in the final outputs, two e-films plus one corresponding report produced by the two different systems. Although the e-film layouts produced by the IILS were evolved from two parts in the traditional way into three parts in which the first five small grids were used to display only one nodule with the highest risk of malignancy in different presentation forms that would aid in diagnosis, valid images were obtained more often by IILS than traditional way, which improves the effectiveness.

In addition, an interesting phenomenon occurred. To validate the performance of our adaptive tool, we randomly selected 327 cases from the control group who reportedly had no lung nodules in the results of clinical medical reports. However, in the process of retesting, we found that a total of 318 nodules were actually missed in 153 cases (46·8%). Missed nodules were mainly concentrated in the range of 3–6 mm instead of <3 mm, and the main type of missing nodule is calcified nodules rather than GGNs. The likely reason for this issue is that in the traditional native language, calcified nodules might be replaced by "old lesions", whereas there are no synonyms for "ground-glass nodules". In addition, many old lesions at 3–6 mm may cause the above phenomenon. However, GGNs are the type of nodule that requires to follow-up [21].

Multidimensional nodules are displayed in the sixth point; a nodule can be observed and estimated from the difference in nodule diameter between baseline and follow-up CT and the time interval between these two scans in uniform three-dimensional tumor growth [23–25]. However, it is impossible to perform three-dimensional reconstruction of key pulmonary nodules due to the heavy manual labor. Therefore, we added an auto multidimensional observation method to minimize the rate of misdiagnosis. The eighth point comprised nodule size measurements. Assessment of nodule size is commonly performed by manual diameter measurements in images. For guidelines, manual nodule measurements should be based on the average of long- and short-axis diameters, which should be obtained on the same transverse, coronal, or sagittal reconstructed images [21]. Previous studies have indicated good agreement between manual nodule measurements on CT and tomosynthesis [26]. In these studies, the limiting factors include the performance of these measurements on a nonanatomical background, restricting the clinical validity of the results, or on real nodules in clinical images where the true nodule size was unknown, making it difficult to establish any systematic errors in the measurements. Physicians should be aware that size and changes in size over time remain the most important factors determining nodule management. An opacity <3 mm should be referred to as a micro-nodule, which may be well or poorly defined [27].Our results showed that the measurement stability from AI-based pulmonary nodule management was significantly greater than that from manual nodules.

The third part included the contents from the twelfth to the fourteenth points, which are a comparison of the impact on all doctors and patients involved between IILS and traditional workstations. Our results show that all doctors and patients are satisfied with the output from IILS. However, in regard to experience-oriented reading habits, the effectiveness of AI in human experts still differs. In our study, a senior radiologist (expert 1) seems to have less confidence in the AI prediction for lung nodules and carefully followed the reading habits to browse images even though the IILS made its decision. A significant difference could be observed before and after applying AI in terms of efficiency, consumed time and absolute mismatched nodules. Interestingly, no obvious improvement in sensitivity or specificity was detected. In contrast, the junior radiologist (expert 2) seems to trust AI to a large extent. We speculated that the human-machine coupled operation might still require an adaptation process.

**Table 3**
Five-point scales for rating of the combination of different types of image layouts.

| Index | Score[a] | Description |
|---|---|---|
| 1 | 1/0 | Poor opacification or missing, non-diagnostic examination |
| 2 | 1/0 | Suboptimal opacification, low confidence in making the diagnosis |
| 3 | 1/0 | Limited opacification but sufficient for diagnosis |
| 4 | 1/0 | Good opacification to the axial and multi-planar image |
| 5 | 1/0 | Excellent opacification to the axial and multi-planar image |

[a] Record 1 point if it meets the requirements for layout, otherwise record 0 points with the same weight for all five indexes.

Although the results are promising, our study has several limitations. In this pilot study, images from patients with deformed thoracic features, such as patients with scoliosis, patients with primary or secondary thoracic deformities, and patients undergoing thoracic surgery, were not included in the training and test set. Therefore, further clinical collection and testing will be needed to assess clinical accuracy for various forms of the thorax. Due to the relatively low incidence of thoracic deformity, the effect would not affect our overall conclusion. The clinical pilot study was performed over the course of two years, and the IILS functioned properly for six months. However, further evaluation of the new system is needed to assess long-term accuracy and stability needs. Additionally, IILS is limited to solving only the problem of CT images of adult lung nodules and not infant cases due to rarely occurring pulmonary nodules in infants and the number of adaptive films. More testing is also needed in a variety of environmental conditions, for example, testing in extremely cold, hot, dry and humid environments. Images with some noise also need to be tested to assess the robustness of the system. In fact, patients with incomplete images were enrolled in the system, ultimately leading to termination and launching the image. Thus, in the processing of special images, such as incomplete images, blank images or incorrect images, the current system still has room for improvement by introducing algorithms such as integrity scanning and grayscale confirmation. Additionally, at present, the IILS can be performed on the chest only. Future work could include applying a device to images of other parts of the body. However, in conclusion, IILS performs better than traditional systems and offers a more affordable and appropriately designed alternative to currently available techniques to optimize the CT layout of lung nodules, saving costs and increasing efficiency. Due to the auto AI-based standardized e-film and visualized structured report generation, one of the new standardizations might be established in the daily workflow and a new radiology work process would be established, and some relevant operators would be unnecessary (Fig. 17).

To provide a benchmark that could be referenced, we evaluated the performance of our IILS on two benchmark databases. As expected, a relative high level of false positives, especially for LUNA16, was

**Table 4**
Key resources for independent testing dataset and models.

| Reagent of resource | Source | Identifier |
|---|---|---|
| Deposited Data | | |
| Images for Training | N/A | Patent Pending |
| Images for Independent Testing | https://pan.baidu.com/s/1OAEMcnlO8uTBK_2cFJ2QqA | N/A |
| Software and Algorithms | | Deep Neural Network |
| MXNet | https://mxnet.apache.org/ | N/A |
| Algorithms | N/A | Patent Pending |

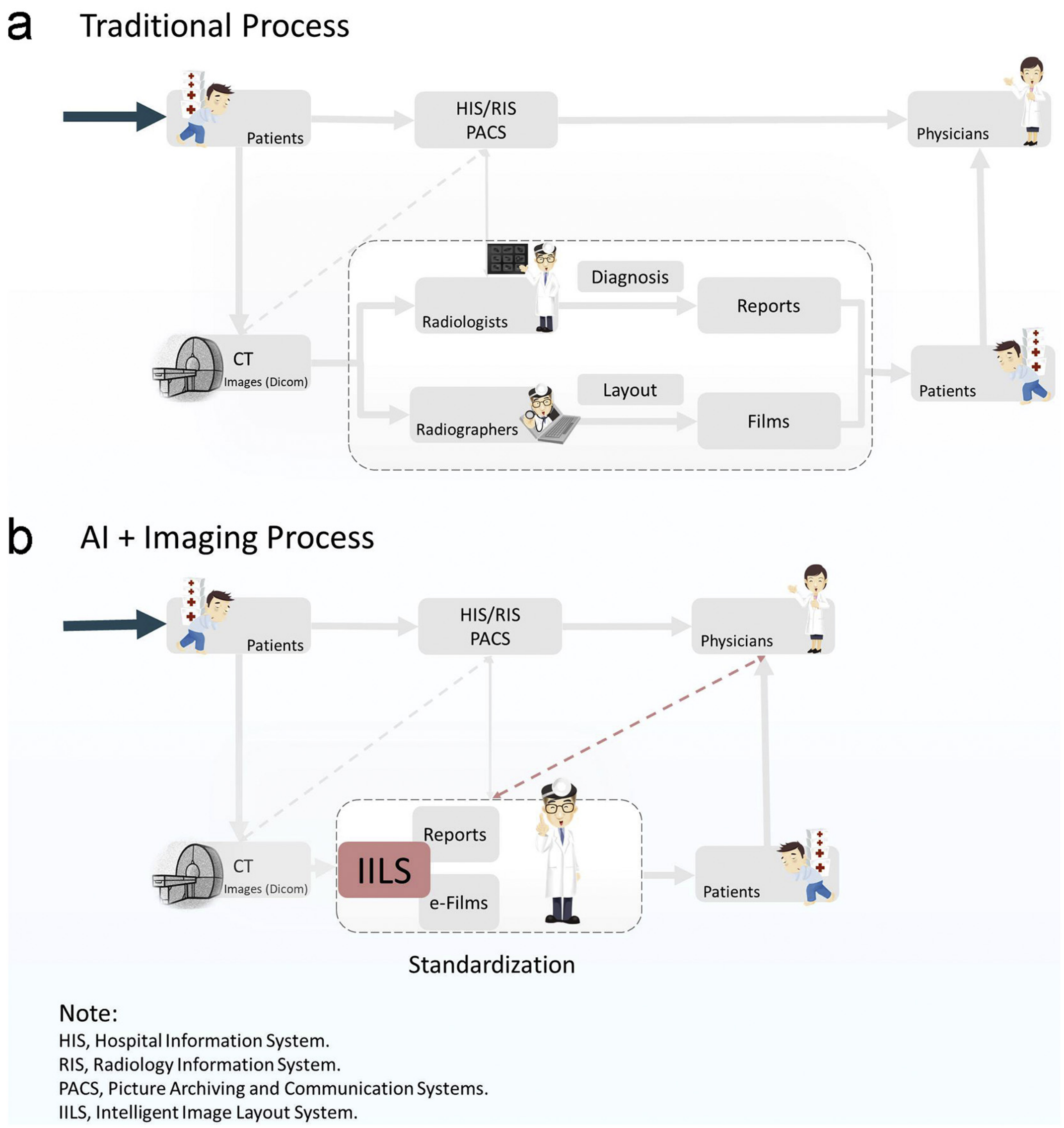


**Fig. 17.** Schematic diagram of the overall process of a patient visiting his doctors at a hospital. The biggest two differences between the two images (a-b) are focused on the internal process of the imaging department (gray dotted square) and the communication between the radiology department and respiratory department (red dotted line). In the patient's operation process, there is no difference between the traditional system and new system, which still contains registration, diagnosis, CT scans, and second diagnosis with a CT report. However, the IILS reduced the CT process by using the human-machine coupled operation instead of the roles of radiographers, efficiently and accurately providing film reports to both radiologists and physicians.

calculated because the nodules with small size were counted as false positives in the prediction. We believe that the design for most diagnostic model is typically cohort-specific, as all training images we adopted are from the Chinese cohort and annotated with detailed records for various size of lung nodules, whereas LUNA16 database collected datasets that were only from the American cohort without any marked record for <3 mm nodules. Although the LIDC/IDRI database contains CT scans with more slice thickness and is more similar to the real clinical environment, the results could not actually reflect the performance of the IILS, because LIDC/IDR1 data training was not used for training, and the real clinical scene that we are in line with the Chinese doctors' reading habits is also very different from the LIDC/IDRI data distribution. We carefully reckoned that such a testing environment may not be consistent with our cohort setting.

Collectively, the IILS offers a simple and accurate method to detect, classify and layout CT images of lung nodules to improve diagnosis on the Chinese population. Therefore, the IILS opens a new window for clinical application of AI and may be an effective way to improve the quality imbalance of medical care worldwide.

## Authors' contribution

Literature research, Y. W, B. Z, X. Z; Experimental and data studies, Y. W, X. Z, C. W, K. Z, Y. Z, H. L, Q. Z, F. C, C. G, Z. Q, J. Y, A. L, X. X, D. L, H. W, H. Y, Y. C, L. Y, Z. Z, W. Y, M. S, X. D, N. Z, F. Z; Statistical analysis, Y. W, F. Y, X. L, Y. Z; Construction of artificial intelligence network, G. Z; H. Z, L. C, C. Z, R. D, L. T, Manuscript editing, Y. W, G. L, B. Z. All authors read and agreed to the final version of the manuscript.

## Declaration of interests

There were no conflicts of interest.

## Acknowledgments

I would like to express my special thanks of gratitude to my teacher and the head of the department (Prof. Bing Zhang) as well as my supervisor (Prof. Guangming Lu) who gave me the golden opportunity to do this wonderful project on the topic, The Application of Artificial

Intelligence in Imaging, which also helped me in doing a lot of research and I came to know about so many new things I am really thankful to them. Secondly I would also like to thank my wife (Simmy) as well as my daughter (Evelyn) who helped me a lot in finalizing this project within the limited time frame.

This work was supported by the National Natural Science Foundation of China (81720108022, 91649116, 81571040); the Social Development Project of Science and Technology Project in Jiangsu Province (BE2016605, BE201707); Key Medical Talents of the Jiangsu Province, the "13th Five-Year" Health Promotion Project of the Jiangsu Province (B.Z.2016-2020); Jiangsu Provincial Key Medical Discipline (Laboratory) (ZDXKA2016020); the Project of the Sixth Peak of Talented People (WSN-138, BZ). National Key R&D Program of China (2016YFC0100100); "Double First-Class" University project (CPU2018GY09). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ebiom.2019.05.040.

## References

[1] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68(6):394–424.

[2] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. CA Cancer J Clin 2018;68(1):7–30.

[3] Reeves AP, Chan AB, Yankelevitz DF, Henschke CI, Kressler B, Kostis WJ. On measuring the change in size of pulmonary nodules. IEEE Trans Med Imaging 2006;25(4):435–50.

[4] Wang J, Mahasittiwat P, Wong KK, Quint LE, Kong FM. Natural growth and disease progression of non-small cell lung cancer evaluated with 18F-fluorodeoxyglucose PET/CT. Lung Cancer 2012;78(1):51–6.

[5] Henschke CI, McCauley DI, Yankelevitz DF, Naidich DP, McGuinness G, Miettinen OS, et al. Early lung cancer action project: overall design and findings from baseline screening. Lancet 1999;354(9173):99–105.

[6] Rego J, Tan K. Advances in imaging-the changing environment for the imaging specialist. Perm J 2006;10(1):26–8.

[7] Nitrosi A, Borasi G, Nicoli F, Modigliani G, Botti A, Bertolini M, et al. A filmless radiology department in a full digital regional hospital: quantitative evaluation of the increased quality and efficiency. J Digit Imaging 2007;20(2):140–8.

[8] Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 2011;365(5):395–409.

[9] Ruparel M, Quaife SL, Navani N, Wardle J, Janes SM, Baldwin DR. Pulmonary nodules and CT screening: the past, present and future. Thorax 2016;71(4):367–75.

[10] Basu PA, Ruiz-Wibbelsmann JA, Spielman SB, Van Dalsem 3rd VF, Rosenberg JK, Glazer GM. Creating a patient-centered imaging service: determining what patients want. AJR Am J Roentgenol 2011;196(3):605–10.

[11] Chilamkurthy S, Ghosh R, Tanamala S, Biviji M, Campeau NG, Venugopal VK, et al. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. Lancet 2018;392(10162):2388–96.

[12] Meyer A, Zverinski D, Pfahringer B, Kempfert J, Kuehne T, Sundermann SH, et al. Machine learning for real-time prediction of complications in critical care: a retrospective study. Lancet Respir Med 2018;6(12):905–14.

[13] Rees CJ, Koo S, Oppong KW. Future directions in diagnostic gastrointestinal endoscopy. Lancet Gastroenterol Hepatol 2018;3(9):595–7.

[14] Ferroni P, Roselli M, Zanzotto FM, Guadagni F. Artificial intelligence for cancer-associated thrombosis risk assessment. Lancet Haematol 2018;5(9):e391.

[15] Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. Radiology 2018;286(3):800–9.

[16] Bonekamp D, Kohl S, Wiesenfarth M, Schelb P, Radtke JP, Gotz M, et al. Radiomic machine learning for characterization of prostate lesions with MRI: comparison to ADC values. Radiology 2018;289(1):128–37.

[17] Choy G, Khalilzadeh O, Michalski M, Do S, Samir AE, Pianykh OS, et al. Current applications and future impact of machine learning in radiology. Radiology 2018;288(2):318–28.

[18] Hastie T, Friedman J, Tibshirani R. Model assessment and selection. The elements of statistical learning. Springer; 2001. p. 193–224.

[19] AIAG A. Measurement systems analysis-reference manual. Troy, MI: The Atomotive Industries Action Group; 2002.

[20] McWilliams A, Tammemagi MC, Mayo JR, Roberts H, Liu G, Soghrati K, et al. Probability of cancer in pulmonary nodules detected on first screening CT. N Engl J Med 2013;369(10):910–9.

[21] Wood DE, Kazerooni EA, Baum SL, Eapen GA, Ettinger DS, Hou L, et al. Lung cancer screening, version 3.2018, NCCN clinical practice guidelines in oncology. J Natl Compr Cancer Network 2018;16(4):412–41.

[22] Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 2017;39(6):1137–49.

[23] Ko JP, Berman EJ, Kaur M, Babb JS, Bomsztyk E, Greenberg AK, et al. Pulmonary nodules: growth rate assessment in patients by using serial CT and three-dimensional volumetry. Radiology 2012;262(2):662–71.

[24] Callister ME, Baldwin DR, Akram AR, Barnard S, Cane P, Draffan J, et al. British Thoracic Society guidelines for the investigation and management of pulmonary nodules. Thorax 2015;70(Suppl. 2):ii1–ii54.

[25] Wang J, Engelmann R, Li Q. Segmentation of pulmonary nodules in three-dimensional CT images by use of a spiral-scanning technique. Med Phys 2007;34(12):4678–89.

[26] Johnsson AA, Fagman E, Vikgren J, Fischella VA, Boijsen M, Flinck A, et al. Pulmonary nodule size evaluation with chest tomosynthesis. Radiology 2012;265(1):273–82.

[27] Hansell DM, Bankier AA, MacMahon H, McLoud TC, Muller NL, Remy J. Fleischner society: glossary of terms for thoracic imaging. Radiology 2008;246(3):697–722.