



INTERNATIONAL TELECOMMUNICATION UNION

**TELECOMMUNICATION
STANDARDIZATION SECTOR**

STUDY PERIOD 2022-2024

FG-AI4H-P-009-A01

ITU-T Focus Group on AI for Health

Original: English

WG(s): Plenary Helsinki, 20-22 September 2022

DOCUMENT

Source: TG-DiagnosticCT Topic Driver

Title: Att.1 – TDD update (TG-DiagnosticCT)

Purpose: Discussion

Contact: Kuan Chen
InferVision, China

Email: ckuan@infervision.com

Contact: Rongguo Zhang
InferVision, China

Email: zrongguo@infervision.com

Abstract: This document contains the updated topic description document (TDD) on AI for volumetric chest computed tomography (CT) for this meeting.

Table of Contents

Table of Contents	2
1 Introduction.....	3
1.1 Document Structure.....	3
1.2 Topic Description	3
1.2.1 Topic group: AI for Volumetric Chest CT.....	3
1.3 Ethical Considerations.....	3
1.4 Existing AI Solutions	3
1.5 Existing work on benchmarking.....	4
2 FG-AI4H Topic Group	4
2.1 Topic group structure	4
2.1.1 Topic and Relevance:	4
3 Method.....	5
3.1 AI Input Data Structure and Output Data Structure	5
3.2 Test Data Labels	5
3.3 Score and Metrics	5
3.3.1 Standalone Testing	5
3.3.2 Primary Benchmarking Metrics	5
3.3.3 Secondary Benchmarking Metrics	7
3.4 Available Public Data and Undisclosed Test Data Set Collection	8
3.5 Reporting Methodology.....	9
4 Results.....	10
5 Discussion.....	13
6 Declaration of Conflict of Interest	14

1 Introduction

With the advent of artificial intelligence (AI) and its strength in faster and more accurate disease detection and diagnosis, it is imperative to establish a standardized assessment framework for the evaluation of AI-based methods for medical applications. This benchmarking process would benefit towards more timely and widespread adoption of AI technologies in health improvement for human beings.

1.1 Document Structure

This document is intended to propose a benchmark for AI in volumetric Chest CT which includes data format, desired data for AI training and testing as well as AI performance evaluation methodologies. The document would include a general review of a specific area of Chest CT lung cancer detection and relevant existing AI solutions followed by data format such as input data format and output data format. AI performance evaluation metrics were also included.

1.2 Topic Description

This topic is intended to build agreement on benchmarking for AI in volumetric chest CT.

1.2.1 Topic group: AI for Volumetric Chest CT

1. A topic group is a community of stakeholders from the medical and AI communities with a shared interest in a topic. The objectives of the topic groups are manifold: to provide a forum for open communication among various stakeholders,
2. to agree upon the benchmarking tasks of this topic and scoring metrics,
3. to facilitate the collection of high-quality labelled test data from different sources,
4. to clarify the input and output format of the test data,
5. to define and set up the technical benchmarking infrastructure, and
6. to coordinate the benchmarking process in collaboration with the Focus Group management and working groups.

The primary output of a topic group is one document that describes all aspects of how to perform the benchmarking for this topic. (The document will be developed cooperatively by suggesting changes as input documents for the next FG-AI4H meeting that will then be discussed and integrated into an official output document of this meeting. The process will continue over several meetings until the topic description document is ready for performing the first benchmarking.)

This topic group is dedicated to AI for volumetric chest CT.

1.3 Ethical Considerations

It is necessary to collect massive data for AI solution development, however, ethical considerations such as patient safety concerns should be taken into careful consideration and relevant regulations should be followed.

1.4 Existing AI Solutions

There have been several AI solutions provided by companies such as Infervision Medical Technology, VoxelCloud Technology, Deepwise, etc.

1.5 Existing work on benchmarking

There have been several benchmarking available in different regions such as the US, Europe, Japan, China, etc. However, almost all of them are not disclosed to the public.

2 FG-AI4H Topic Group

Each topic group agrees upon representative benchmarking tasks in a pragmatic, best-practice approach, which can later be scaled and expanded to similar tasks. Every benchmarking task should address a health problem of relevance (e.g., impacting a large and diverse part of the global population or challenging to treat) and for which AI technology would provide a tangible improvement relative to the current practice (e.g., better care, results, and/or cost/time effectiveness).

For a rigorous and sound evaluation, undisclosed test data sets must be available (or have to be collected) for each task. All data must be of high quality and compliant with ethical and legal standards. In addition, the data must originate from a variety of sources so that it can be determined whether an AI algorithm can generalize across different conditions, locations, or settings (e.g., across different people, hospitals, and/or measurement devices). The format/properties of the data serving as input to the AI and of the output expected from the AI, as well as the benchmarking metrics are agreed upon and specified by the topic group.

Finally, the AI-to-be-evaluated will be benchmarked with the undisclosed test data on FG-AI4H computing infrastructure. Here, the AI will process single samples of the undisclosed test data set and predict output variables, which will be compared with the "ground truth." The results of the benchmarking will be provided to the AI developers and will appear on a (potentially anonymized) leader board.

2.1 Topic group structure

2.1.1 Topic and Relevance:

According to a World Health Organization (WHO) cancer report, lung cancer remains the leading cause of morbidity and mortality with an estimated 2.09 million new cases and 1.76 million deaths worldwide in 2018. The overall five-year survival rate is merely 10-15%, early detection is crucial because stage I lung cancer is reported to have a 10-year survival rate of ~90% if surgical resections were conducted. Early lung cancer characterization (including differentiating benign from malignant nodules) is also critical because this would dramatically affect treatment selection and patient outcome. Annual screening by low-dose volumetric chest computed tomography has thus been recommended by National Lung Screening Trial (NLST) for early lung cancer detection to reduce lung cancer mortality. However, large numbers of thin slices (300-400) were generated in one lung CT scan and the interpretation is tedious and time-consuming and small nodules are often overlooked while misdiagnoses are not uncommon. In China, lung cancer is the leading cause of mortality with a recorded number of 610,200 lung cancer deaths in 2015 and it is projected that lung cancer mortality in China may increase by approximately 40% between 2015 and 2030.

The two biggest issues in the Chinese medical system are shortage and centralization of medical resources. There is a serious shortage of qualified radiologists with an annual growth rate of ~4% while the medical imaging demand increase rate was nearly 30%. Patients tend to seek medical care in large tertiary hospitals in top-tier cities while community hospitals, village health centers and clinics are largely underused. As a result, the workloads of radiologists were immensely intensified especially for big hospitals since they must examine hundreds of images every day while rural city radiologists remain less skilled, which makes early lung cancer detection and characterization more difficult.

3 Method

The method for AI benchmarking includes data format requirement of input data and output data, testing data label quality control as well as testing metrologies and scoring matrixes.

3.1 AI Input Data Structure and Output Data Structure

Chest CT volumetric CT images obtained in first scans or follow-ups were eligible for AI benchmarking. Original DICOM data together with labelling information were required. Image resolution should be no less than 512*512 pixels. The AI output should include information about lesions position, long-axis diameter, classification, density, and malignancy probability.

3.2 Test Data Labels

1. For clinical evaluation of AI algorithm for lung nodule detection, gold standard labelling is obtained from a panel of experts. The panel should involve at least three experts with 10 years of practice in thoracic radiology or above. One of the experts serves as an arbitrator to make final decisions on disputed annotations from the other experts. All labelling experts would receive prior specialized training regarding how to annotate nodules, delineation of the nodule boundaries, long-axis diameter at maximum cross-section measurement, etc.
2. For clinical evaluation of AI for malignant nodule diagnosis, the gold standard comes from pathological or CT follow-ups. As for pathological results, both pathological type, stage and pathology details (whether obtained by liquid biopsy or fine-needle aspiration biopsy, whether confirmed by histopathology or cytology) should be recorded.
3. The gold standard annotation of testing data should be strictly monitored and managed to preserve the confidentiality of gold standard testing data results.

3.3 Score and Metrics

3.3.1 Standalone Testing

In the standalone test which evaluates AI's performance independently, influencing factors such as nodule size, nodule type, scanning and image reconstruction protocols, imaging devices, population characteristics were tested by dividing testing data into corresponding category groups. The participants submit the algorithms through an online portal, after which the algorithms will be run against the undisclosed dataset. Inputs and outputs of the test will be defined as follow.

3.3.2 Primary Benchmarking Metrics

Primary testing includes nodule detection with the overall dataset (including public data and undisclosed data). Detected nodules by standalone AI system, were compared with pre-labelled nodules to determine the true positive nodules and false-positive nodules, benchmarking metrics including nodule-based sensitivity, false positive rate, Alternative Free-Response ROC (Receiver Operating Characteristic), F-score, time spent reading the cases were calculated. For the malignancy classification task, sensitivity, specificity, ROC, time spent reading the cases were calculated.


Nodule detection includes positioning. Nodule classification can be divided into binary classification and multi-class classification.

3.3.2.1 Nodule detection

Nodule detection refers to positioning nodules based on radiological results. The indicators include:

Indicators	Explanation and formula
True positive rate	$\frac{\text{Detected nodule}(s)}{\text{All nodules}} \times 100\%$
Average false positive number	$\frac{\text{False positioning number}}{\text{Radiological cases}}$
AFROC Curve	Vertical axis is true positive rate bases on nodules. Lateral axis is false positive rate based on cases.

3.3.2.2 Clinical Performance indicators by binary classification

Performance indicators	Explanation
Accuracy rate (ACC)	$\frac{TP + TN}{P + N} \times 100\%$
Sensitivity rate (True positive rate, TPR)	$\frac{TP}{TP + FN} \times 100\%$
Specification rate (True negative rate, TNR)	$\frac{TN}{TN + FP} \times 100\% = 1 - FPR$
Precision rate (Positive predictive value, PPV)	$\frac{TP}{TP + FP} \times 100\%$
Fall-out (False positive rate, FPR)	$\frac{FP}{FP + TN} \times 100\% = 1 - TNR$
F-Score	$(1 + \beta^2) \times \frac{PPV \times TPR}{(\beta^2 \times PPV) + TPR}$ β for balancing rate, usually the number is 1.
ROC Curve	 <p>The ROC curve graph plots the True Positive Rate (TPR) on the y-axis against the False Positive Rate (FPR) on the x-axis, both ranging from 0.0 to 1.0. A blue curve represents a 'Perfect classifier' starting at (0,0) and ending at (1,1). A green curve represents a 'Better' classifier. An orange curve represents a 'Random classifier'. A red dashed diagonal line represents the 'Worse' classifier. Arrows indicate the direction of performance: 'Better' towards the top-left and 'Worse' towards the bottom-right.</p>
AUC Score	$\frac{1}{2} \sum_{i=1}^{m-1} (x_{i+1} - x_i) \times (y_i + y_{i+1})$

3.3.2.3 Clinical Performance indicators by multi-class classification

Multi-class classification including grading classification and comprehensive classification. Grading classification can transfer classification to sets of binary classification. Comprehensive classification refers to indicators like Cohen's kappa coefficient (κ).

3.3.3 Secondary Benchmarking Metrics

In addition to the primary benchmarking metrics, other performance metrics relevant to the task at hand should be included as secondary benchmark metrics.

- 1) Nodule characterization: size measurement, density, classification, and malignancy

In nodule size measurement, error rate which is the absolute value of error between AI measured nodule long axis diameter and radiologist's labelling nodule long axis diameter divided by radiologist's labelling diameter should be calculated. In nodule density characterization (which classify nodules into solid, part-solid, ground-glass opacity and calcified nodules), accuracy (defined as true positive nodules divided by gold standard nodules) was used to depict nodule characterization capability. For malignancy characterization, nodules were classified as benign and malignant, sensitivity, specificity, accuracy, and AUC were used to evaluate testing performance.

Plan: indicators for size measurement, segmentation, outline, nodule types and malignancy

- 2) Subtle case performances: e.g., easily missed but important lesions like GGN, nodules near vessels, nodules near cavity etc.

Subtle cases are subsets of the test data which held great clinical significance and shall be tested independently. Subtle cases should be selected from the original pool of testing data, preferably including more than 100 cases or more for each subset. Performance metrics can be adopted from primary benchmarking.

- 3) Other indicators:

3.1) Robustness:

Manufacturers should evaluate various factors that may interfere with algorithm performance during the clinical use phase based on product risk analysis and clinical deployment environment characteristics. Manufacturers need to obtain or simulate relevant data to form a dedicated test set to conduct adversarial testing of algorithm performance. The changes of each index need to be analyzed to form robust research data.

3.1.1) Adversarial testing for hardware changes

Testers should consider the diversity of medical imaging hardware equipment and parameter settings, collect or simulate more image data, as an extension to the test set, to verify the robustness of the algorithm to image acquisition hardware equipment. The diversity of parameter settings includes: physical resolution, pixel resolution, brightness, focus, ray quality, etc. The image data generated by the simulation should not affect the annotation conclusions.

3.1.2) Adversarial testing for software pre-processing

Testers should consider the diversity of software pre-processing, collect, or simulate more image data, as an expansion of the test set, to verify the robustness of the algorithm in the face of software pre-processing. The diversity of software pre-processing includes: background cropping, image compression, background filling, smoothing pre-processing, reconstruction operators, etc. The image data generated by the simulation should not affect the annotation conclusions.

3.1.3) Adversarial testing for spoofing attacks

Spoofing attack is an attack method that adds perturbations that are difficult for people to detect to fool the model. Testers can use a white-box attack (Projected Gradient Descent, PGD) to generate perturbations with a limited maximum norm (such as less than 8/256), and Perturbations are inserted into the original images, and the model is then tested on these perturbed images to verify that the model is resistant to malicious spoofing attacks.

Testers should determine the applicability of spoofing attacks and the configuration of test parameters based on the product's network security capabilities and risk analysis documents. The perturbed data should be confirmed by the annotator and used for testing.

3.1.4) Stress test

Stress testing is to test the performance, reliability, stability, etc. of an algorithm model in the simulation of long-term extreme inputs or environments (different loads, limit values, boundary values, bulk data, erroneous data, rare data, etc.) that may be encountered in practical applications.

Definition of Stress Test Sample

Stress test samples refer to samples with extremely large or extremely small feature capacity within the calibration range of an algorithm model. Stress test samples should not interfere with the normal judgment of the physician.

Selection of stress test samples

The selection of stress test samples can follow the following principles:

- a) images with older subjects;
- b) imaging of specific diseases;
- c) images with artifacts but meeting data quality requirements;
- d) the layer thickness of the image is extremely large or small;
- e) the image series contains a very large number of images;
- f) with implants (interfering items);
- g) with complications;
- h) multiple, diffuse lesions.

3.2) Reproducibility: Testers should repeat the test on the same test set, and the number of tests should not be less than three times.

3.3) Generalization: Generalization ability refers to the ability of the algorithm to adapt to unfamiliar samples. Manufacturers should analyze the differences between the training set used for product development and real-world unfamiliar samples according to the intended use and deployment environment of the product, and form a document as a basis for configuring the test set. In the actual test, the manufacturer should verify the generalization ability of the algorithm through the diversity and variability of the test set.

3.4 Available Public Data and Undisclosed Test Data Set Collection

1. To assess algorithm robustness, sufficient and diversified data from multiple heterogeneous sources (e.g., imaging device models, imaging protocols, patient demographics, clinical conditions, etc.) should be used for testing to verify the generalization capacity of AI. Public and real-world undisclosed data (desensitized) should be enrolled.
2. Public database may include Lung Image Database Consortium Image Collection from Cancer Imaging Archive (LIDC/LDRI) and NLST.
3. To avoid prior finetuning and overfitting on the public data by algorithms, mostly real-world undisclosed data shall be used for the test.

4. Undisclosed test data should include geographic differences (terrain such as plain and plateau), gender differences, multiple countries (from >3 continents) with different socioeconomics and the resulting lifestyles (≥ 5 countries), multiple hospitals (≥ 20) including big hospitals (> 800 beds) and small hospitals and different races.
5. Undisclosed test data should come from various CT vendors (GE, Siemens, Philips, Canon, United Imaging, etc.) reconstructed with various thin slices (0.625mm, 1mm, 1.25mm, 2mm, 3mm, etc.) and thick slices (5mm, 7mm, etc.). Various reconstruction kernels such as lung kernel, bone kernel, the standard kernel should also be included.
6. Undisclosed data should include low dose CT scans (with effective doses of 1.5~2.5 mSv) and normal dose CT scans (2.5 mSv~5 mSv).
7. All testing data would preferably have pathological finding confirmation of nodule existence and property.
8. It would be good for undisclosed test data pool size to be as large as possible (preferably $> 5,000$ cases)
9. Case-control subsets of testing data (subtle cases) should also be included for comprehensive evaluation of AI performance in various perspectives. Several subsets were recommended: e.g., easily missed but important lesions like GGN, nodules in proximity with vessels, nodules near cavity etc.
10. An arbitration commission comprising of chest CT experts will examine labelled undisclosed test patient data to confirm data sufficiency, heterogeneity, labelling quality and conformity to ethical and legal standards.

3.5 Reporting Methodology

Reporting metrics should include nodule-based sensitivity, false positive rate, Alternative Free-Response ROC (Receiver Operating Characteristic), F-score, time spent reading the cases were calculated. For the malignancy classification task, sensitivity, specificity, ROC, and time spent reading the cases were calculated.

4 Results

4.1 Dataset we built

we collect the largest and most diverse dataset named PN9 for pulmonary nodule detection by far. Specifically, it contains 8,798 CT scans and 40,439 annotated nodules from 9 common classes. We further propose a slice-aware network (SANet) for pulmonary nodule detection.

Dataset	Year	Scans	Nodules	Class	Available
ANODE09	2010	55	710	4	Yes
LIDC-IDRI	2011	1,108	2,562	3	Yes
LUNA16	2016	888	1,186	2	Yes
DSB 2017	2017	2,101	N/A	2	No
PN9	2021	8,798	40,439	9	Yes

TABLE 1: Comparison with the existing datasets of the pulmonary nodule. 'Scans' indicates the number of CT scans. 'Nodules' denotes the number of labeled nodules. 'Class' means the class number. And 'Available' denotes whether the dataset is available.

In Table 1, we compare the PN9 with several existing pulmonary nodule datasets. Compared to the widely used public dataset LUNA16, PN9 contains over 10 times more CT scans and over 30 times more annotated nodules. As for the class diversity, other datasets only have three categories: nodule ≥ 3 mm, nodule < 3 mm, and non-nodule. Due to these limitations, it is difficult for most of the existing nodule datasets to apply to the practice. However, our PN9 contains many CT scans and 9 classes, which will contribute to the detection and classification tasks of the pulmonary nodules, allowing researchers to design more effective algorithms based on different types of nodules. Besides, there are more pulmonary nodules of small size, like 0-3mm solid nodules and 0-5mm ground-glass nodules. It helps identify small nodules more accurately, then the doctors can diagnose and treat patients earlier. In summary, our dataset not only is larger than the previous datasets, but also has superior diversity and performance.

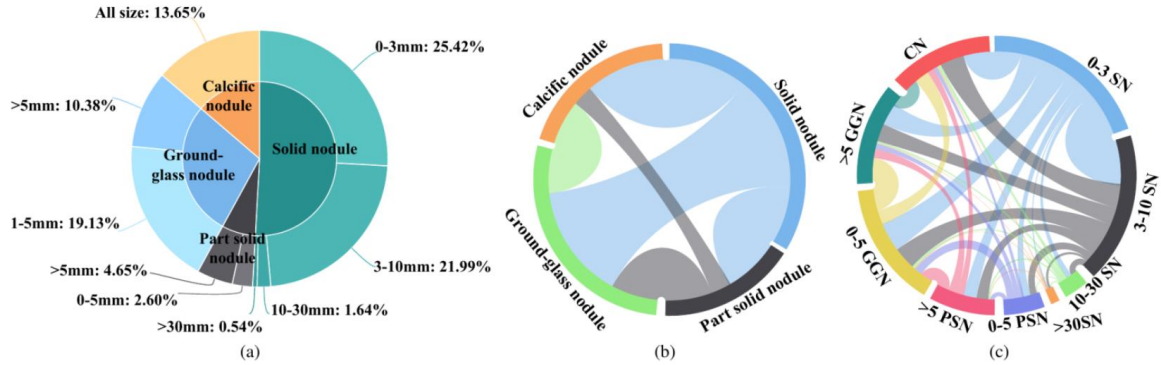


Figure 2. Statistics of class in PN9. (a)Taxonomy of the PN9 dataset. It contains four super-classes and nine sub-classes. The percentage represents the proportion of a certain class of nodules to all nodules. (b) Mutual dependencies among super-classes. (c) Mutual dependencies among sub-classes.

Our PN9 has a hierarchical class structure, and its detailed taxonomy is shown in Fig. 2. According to the property of the pulmonary nodules, all nodules in our dataset are first divided into four upper-level classes (denoted as super-class), including solid nodule (SN), part-solid nodule (PSN), ground-glass nodule (GGN), and calcific nodule (CN). Meanwhile, To satisfy the practical demands of doctors and hospitals, we further subdivide the super-class referring to the medical guidelines. Each nodule is assigned with a subordinate class (denoted as sub-class) belonging to a certain super-class based on the nodule size. For example, sub-class 0-3mm solid nodules (denoted as 0-3SN) are defined as any nodules identified to be super-class solid nodules with the most significant

in-plane dimension in the range of 0-3 mm. And 9 different sub-classes are finally obtained. The statistics of nodules in each class are shown in Fig. 2 (a). In Fig. 2 (b-c), we show the mutual dependencies among super-classes and sub-classes, respectively. The larger width of a link between two classes indicates a higher probability for the two classes' nodules appearing in one patient simultaneously. For example, a patient diagnosed with ground-glass nodules is also likely to have solid nodules.

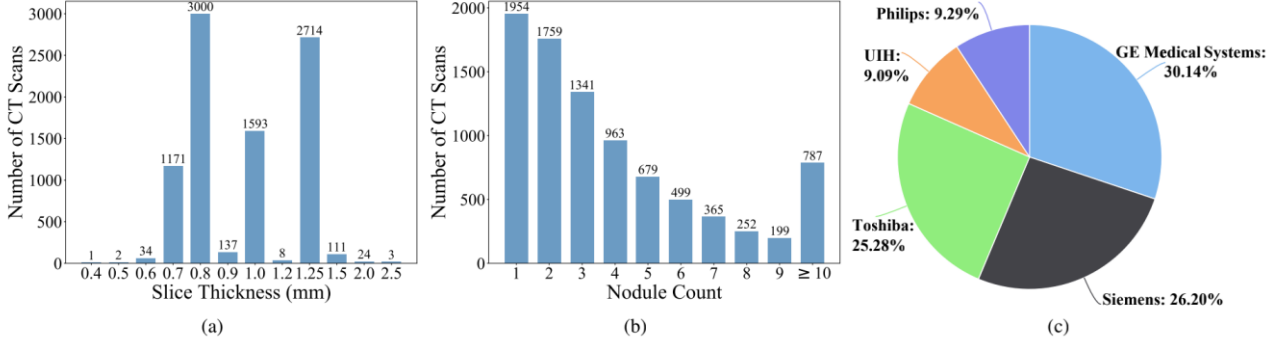


Fig. 4: Statistics of the proposed PN9 dataset. (a) Slice thickness distribution of CT scans. (b) Distribution of nodule count in one patient. (c) Percentage of CT manufacturer.

The CT scans in PN9 are obtained by a series of CT manufacturers and corresponding models, as shown in Fig. 4 (c). PN9 includes 2,652 scans from ten different GE Medical Systems scanner models, 2,305 scans from eleven different Siemens scanner models, 2,224 scans from three different Toshiba scanner models, 800 scans from two different United Imaging Healthcare (UIH) scanner models, and 817 scans from six different Philips scanner models. Since the images of thick slice are not optimal for CAD analysis, we mainly collect the CT scans with thin-slice. As illustrated in Fig. 4 (a), slice thickness ranges from 0.4 mm to 2.5 mm, and most are located at 0.7, 0.8, 1.0, and 1.25 mm. Besides, the pixel spacing ranges from 0.310 mm to 1.091 mm, with a mean of 0.706 mm. In Fig. 4 (b), we illustrate the distribution of nodule count in one patient. We observe that approximately 68 % of patients have nodules less than 5 in our PN9. However, there are about 9 % of patients with more than 10 nodules, which may be difficult to detect.

4.2 Results we got

For the dataset PN9, we split the 8,798 CT scans into 6,707 scans for training and 2,091 scans for testing. During training, we separate 670 CT scans from the training set as the validation set to monitor the convergence of the model. There are three preprocessing steps for the raw CT images. First, all raw data are converted into the Hounsfield Unit (HU) since HU is a standard quantitative value describing radiodensity. Then, the data is clipped into $[-1200, 600]$. Finally, we transform the data range linearly into $[0, 255]$.

Comparison based on FROC. We first evaluate the FROC score defined in the LUNA16 dataset. The experiment results are listed in Table 2, and the FROC curves are illustrated in Fig. 6.

TABLE 2: Comparison of our SANet and other methods in terms of FROC on dataset PN9. The values are pulmonary nodule detection sensitivities (unit: %) with each column representing the average number of false positives per CT scan.

Method	0.125	0.25	0.5	1.0	2.0	4.0	8.0	Average
2D CNN-Based Methods:								
Faster R-CNN [34]	10.79	15.78	23.22	32.88	46.57	61.94	75.52	38.10
RetinaNet [38]	8.42	13.01	20.13	29.06	40.41	52.52	65.42	32.71
SSD512 [12]	12.26	18.78	28.00	40.32	56.89	73.18	86.48	45.13
3D CNN-Based Methods:								
Leaky Noisy-OR [7]	28.08	36.42	46.99	56.72	66.08	73.77	81.71	55.68
3D Faster R-CNN [23]	27.57	36.59	46.76	58.00	70.00	80.02	88.32	58.18
DeepLung [23]	28.59	39.08	50.17	62.28	72.60	82.00	88.64	60.48
NodeNet (N ₂) [71]	27.33	38.25	49.40	61.09	73.11	83.28	89.83	60.33
I3DR-Net [43]	23.99	34.37	46.80	60.04	72.88	83.60	89.57	58.75
DeepSEED [39]	29.21	40.64	51.15	62.20	73.82	83.24	89.70	61.42
SANet	38.08	45.05	54.46	64.50	75.33	83.86	89.96	64.46

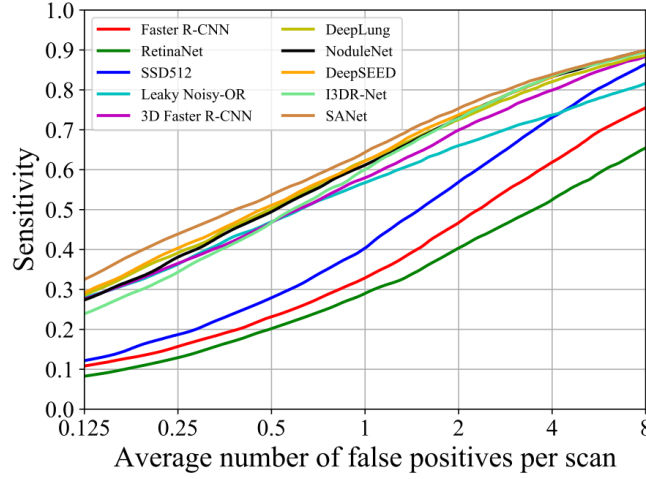


Fig. 6: FROC curves of compared methods and our SANet.

It is noted that our SANet achieves the best results over other methods, which obtains an improvement of 3.04 % on average FROC score over the second-best DeepSEED. And our method especially outperforms the other detection methods by a large margin for the average number of false positives per CT scan smaller than 2.

TABLE 4: Comparison of our SANet and other methods in terms of FROC (%) on the dataset LUNA16 [30].

Method	0.125	0.25	0.5	1.0	2.0	4.0	8.0	Average
Leaky Noisy-OR [7]	59.38	72.66	78.13	84.38	87.50	89.06	89.84	80.13
3D Faster R-CNN [23]	66.20	74.60	81.50	86.40	90.20	91.80	93.20	83.40
DeepLung [23]	69.20	76.90	82.40	86.50	89.30	91.70	93.30	84.20
NoduleNet (N ₂) [71]	65.18	76.79	83.93	87.50	91.07	92.86	93.75	84.43
I3DR-Net [43]	63.56	71.31	79.84	85.27	87.60	89.92	91.47	81.28
DeepSEED [39]	73.90	80.30	85.80	88.80	90.70	91.60	92.00	86.20
SANet	71.17	80.18	86.49	90.09	93.69	94.59	95.50	87.39

Comparison on the LUNA16 dataset.

To further validate the performance of the proposed SANet, we conduct experiments on the widely used dataset LUNA16 with 10-fold crossvalidation. As shown in Table 4, our SANet achieves the best results for pulmonary nodule detection. For example, it obtains an average FROC score of 87.39 %, which improves the secondbest method DeepSEED [39] by 1.19 %. Besides, our SANet outperforms the state-of-the-art nodule detection methods by a large margin for the settings of average number of false positives per CT scan larger than 1.

Comparison with Experienced Doctors.

Furthermore, We compare the detection performance of our SANet and two experienced doctors with at least 10 years' clinical experience. We collect an additional small-scale pulmonary nodule testing dataset that contains 120 CT scans. After annotated accurately by several attending physicians from major hospitals, this testing dataset contains 2,137 annotated nodules with the golden standard. The other two experienced doctors, who never diagnose the small-scale testing dataset, are invited to individually identify lung nodules. As for our SANet, it is trained on dataset PN9 and tested on this small-scale testing dataset, which is evaluated using AP@0.25. The PR curve of our SANet and the detection results of two experienced doctors are shown in Fig. 9. It is noted that the performance of our model is better than two doctors on their individually diagnosed nodules, which validates that SANet surpasses the human-level performance and is suited for pulmonary nodule detection.

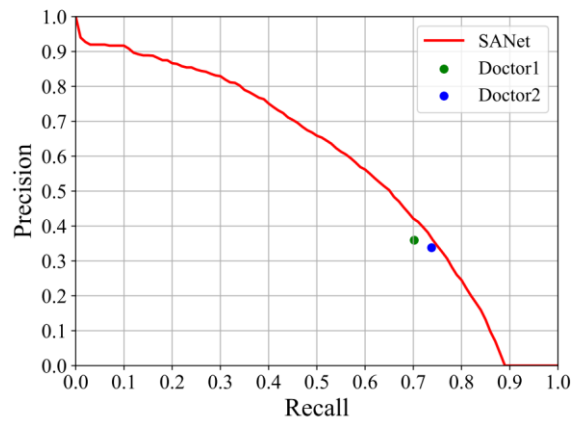
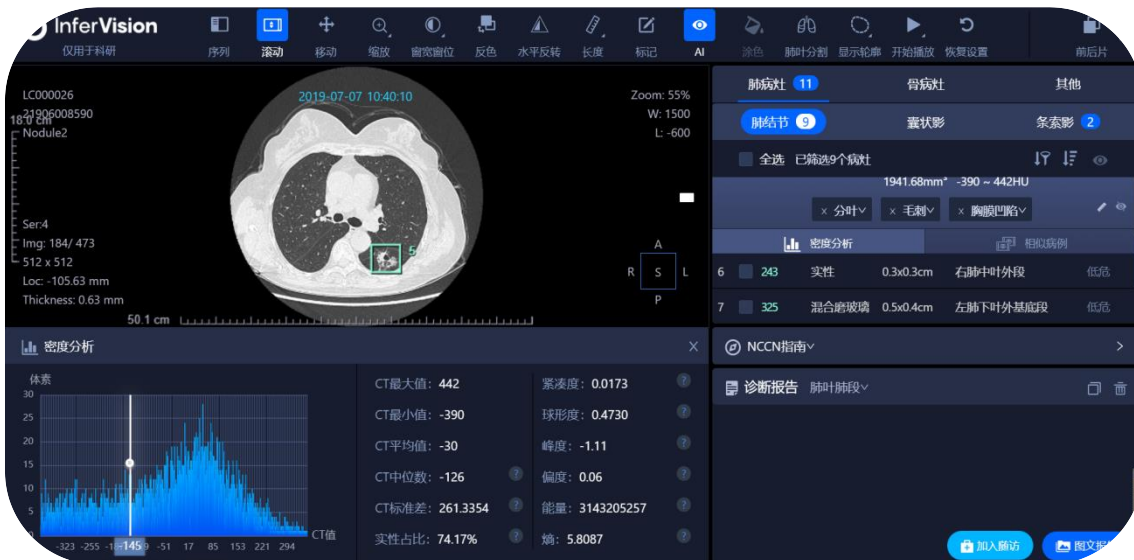


Fig. 9: The precision-recall curve for the nodule detection on the small-scale pulmonary nodule testing dataset. The 'Doctor1' and 'Doctor2' denote the detection results of two experienced doctors.

CASE: Nodule characterization: size measurement, density, classification, and malignancy



InferRead® CT Lung is an FDA-cleared, CE-marked AI solution to assist radiologists with their chest CT image analysis, serving as a second pair of eyes. It is capable of identifying various types of lung nodules, providing quantification for each lesion, and generating radiological reports. The application has been trained with hundreds of thousands of exams to ensure its accuracy, robustness, and generalizability.

Validated through our retrospective, multi-reader multi-case (MRMC) studies, InferRead® CT Lung has shown its ability to reduce nearly 15% image reading time and increase over 40% in the sensitivity of detection. InferRead® CT Lung is currently in use at hundreds of hospitals and imaging centers globally. Millions of patients have already benefited. It is highly compatible with legacy systems and accepts chest CT images from PACS, RIS, or directly from a CT scanner.

"The tremendous potential for lung cancer screening to reduce mortality in the US is very much unrealized due to a combination of reasons. Based on our experience reviewing the algorithm for the past several months and my observations of its extensive use and testing in China, I believe that InferVision's InferRead® CT Lung application can serve as a robust lung nodule "spell-checker" with the potential to improve diagnostic accuracy, reduce reading times, and integrate with the image review workflow. "

--- Eliot Siegel, MD, Professor and Vice-Chair of Research Information Systems in Radiology, University of Maryland School of Medicine

5 Discussion

As shown above, our proposed SANet obtains the best performance on our PN9 and public dataset LUNA16 compared with other state-of-the-art detection methods. Besides, the performance of SANet is better than two experienced doctors. However, there are still some failure cases when identifying nodules with class 0-3 SN. Since the nodules of PSN usually have a fuzzy border, SANet may not identify the entire nodules of PSN and produce smaller bounding boxes than the ground truth. In the future, we will consider the attributes of different categories to detect pulmonary nodules better. In the future, we will compare the results of doctor's independent image reading with those of AI-assisted reading.

6 Declaration of Conflict of Interest

There is no conflict of interest.

References

- [1] Mei J , Cheng M M , Xu G , et al. SANet: A Slice-Aware Network for Pulmonary Nodule Detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, PP(99):1-1.
-