Original Article

Implementation of artificial intelligence in the histological assessment of pulmonary subsolid nodules

Jiajun Deng[1#], Mengmeng Zhao[1#], Qiuyuan Li[1#], Yikai Zhang[2], Minjie Ma[3], Chuanyi Li[4], Jun Wang[5], Yunlang She[1], Yan Jiang[1], Yunzeng Zhang[6], Tingting Wang[7], Chunyan Wu[8], Likun Hou[8], Sheng Zhong[9], Shengxi Jin[10], Dahong Qian[5], Dong Xie[1], Yuming Zhu[1], Yasmeen K. Tandon[11], Annemiek Snoeckx[12,13], Feng Jin[14], Bentong Yu[15], Guofang Zhao[16,17], Chang Chen[1,3,18]; on behalf of the MultiomIcs claSSIfier for pulmOnary Nodules (MISSION) Collaborative Group

[1]Department of Thoracic Surgery, Shanghai Pulmonary Hospital, Tongji University School of Medicine, Shanghai, China; [2]School of Information Science and Technology, ShanghaiTech University, Shanghai, China; [3]Department of Thoracic Surgery, The First Hospital of Lanzhou University, Lanzhou, China; [4]Department of Thoracic Surgery, Nantong No. 6 People's Hospital, Nantong, China; [5]School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China; [6]Department of Thoracic Surgery, Shandong Public Health Clinical Center, Jinan, China; [7]Department of Radiology, Shanghai Pulmonary Hospital, Tongji University School of Medicine, Shanghai, China; [8]Department of Pathology, Shanghai Pulmonary Hospital, Tongji University School of Medicine, Shanghai, China; [9]Tailai Biosciences Inc., Shenzhen, China; [10]Dianei Technology, Shanghai, China; [11]Department of Radiology, Mayo Clinic, Rochester, MN, USA; [12]Department of Radiology, Antwerp University Hospital and University of Antwerp, Edegem, Belgium; [13]Faculty of Medicine and Health Sciences, University of Antwerp, Wilrijk, Belgium; [14]Provincial Key Laboratory for Respiratory Infectious Diseases in Shandong, Shandong Provincial Chest Hospital, Cheeloo College of Medicine, Shandong University, Jinan, China; [15]Department of Thoracic Surgery, The First Affiliated Hospital of Nanchang University, Nanchang, China; [16]Department of Cardiothoracic Surgery, Hwa Mei Hospital, University of Chinese Academy of Sciences, Ningbo, China; [17]Ningbo Institute of Life and Health Industry, University of Chinese Academy of Sciences, Ningbo, China; [18]The International Science and Technology Cooperation Base for Development and Application of Key Technologies in Thoracic Surgery, Lanzhou, China

Contributions: (I) Conception and design: J Deng, M Zhao, Y She, C Chen; (II) Administrative support: F Jin, B Yu, G Zhao, C Chen; (III) Provision of study materials or patients: D Xie, Y Zhu, G Zhao, C Chen; (IV) Collection and assembly of data: Q Li, M Ma, C Li, Y She; (V) Data analysis and interpretation: Y Zhang, J Wang, T Wang, C Wu, L Hou, D Qian; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work.

Correspondence to: Feng Jin. Provincial Key Laboratory for Respiratory Infectious Diseases in Shandong, Shandong Provincial Chest Hospital, Cheeloo College of Medicine, Shandong University, 46 Lishan Rd., Lixia District, Jinan 250013, China. Email: jf13791123070@163.com; Bentong Yu. Department of Thoracic Surgery, The First Affiliated Hospital of Nanchang University, 17 Yongwaizheng St., Nanchang 330006, China. Email: yubentong@126.com; Guofang Zhao. Department of Cardiothoracic Surgery, Hwa Mei Hospital, University of Chinese Academy of Sciences, 41 Xibei St., Ningbo 315000, China. Email: guofzhao@hotmail.com; Chang Chen. Department of Thoracic Surgery, Shanghai Pulmonary Hospital, Tongji University School of Medicine, 507 Zhengmin Rd., Shanghai 200433, China, Email: changchenc@tongji.edu.cn.

Background: Clinical management of subsolid nodules (SSNs) is defined by the suspicion of tumor invasiveness. We sought to develop an artificial intelligent (AI) algorithm for invasiveness assessment of lung adenocarcinoma manifesting as radiological SSNs. We investigated the performance of this algorithm in classification of SSNs related to invasiveness.

Methods: A retrospective chest computed tomography (CT) dataset of 1,589 SSNs was constructed to develop (85%) and internally test (15%) the proposed AI diagnostic tool, SSNet. Diagnostic performance was evaluated in the hold-out test set and was further tested in an external cohort of 102 SSNs. Three thoracic surgeons and three radiologists were required to evaluate the invasiveness of SSNs on both test datasets to investigate the clinical utility of the proposed SSNet.

Results: In the differentiation of invasive adenocarcinoma (IA), SSNet achieved a similar area under the curve [AUC; 0.914, 95% confidence interval (CI): 0.813–0.987] with that of the 6 doctors (0.900, 95% CI: 0.867–0.922). When interpreting with the assistance of SSNet, the sensitivity of junior doctors, specificity

of senior doctor, and their accuracy were significantly improved. In the external test, SSNet (AUC: 0.949, 95% CI: 0.884–1.000) achieved a better AUC than doctors (AUC: 0.883, 95% CI: 0.826–0.939) whose AUC increased (AUC: 0.908, 95% CI: 0.847–0.982) with SSNet assistance. In the histological subtype classifications, SSNet achieved better performance than practicing doctors. The AUCs of doctors were significantly improved with the assistance of SSNet in both 4-category and 3-category classifications to 0.836 (95% CI: 0.811–0.862) and 0.852 (95% CI: 0.825–0.882), respectively.

Conclusions: The AI diagnostic system achieved non-inferior performance to doctors, and will potentially improve diagnostic performance and efficiency in SSN evaluation.

Keywords: Artificial intelligence (AI); pulmonary subsolid nodules (SSNs); lung adenocarcinoma; computed tomography (CT)

Submitted Oct 27, 2021. Accepted for publication Dec 23, 2021.
doi: 10.21037/tlcr-21-971
View this article at: https://dx.doi.org/10.21037/tlcr-21-971

Introduction

Previously, it has been reported that a reduction of mortality with low-dose computed tomography (CT) in a number of lung cancer screening trials (1-3). Consequently, lung cancer screening is more and more being implemented in the past two decades. With the increased use of CT and pulmonary subsolid nodules (SSNs), SSNs are increasingly being detected. Imaging assessment of invasiveness of SSNs is essential in the clinical management of patients. However, the histological prediction of SSNs, which has been reported to have a 9% detection rate in screening trials, poses several challenges (4,5). The degree of invasiveness is used as the basis for clinical management decisions. Lung adenocarcinoma appearing as SSN can present with a variety of morphological and imaging features, which can be related to different degrees of invasiveness and prognosis. Reported evidence of high intra-observer and interobserver variability in the invasiveness classification of SSNs has highlighted concerns about undertreatment and overtreatment. Therefore, an accurate diagnostic system or assistant tool can have a beneficial clinical impact (6).

To overcome these diagnostic challenges, a number of solutions for malignancy evaluation have been previously proposed (7-11), including radiological density, morphological features, and clinical features. Risk-assessing tools based on clinical and radiological features have been used to determine cancer risk and standardize clinical management recommendations (8,12). Additionally, quantitative analyses have been carried out to evaluate malignancy depending on accurate delineation of nodule borders and feature engineering (13-16). However, the application of previous methods relies on subjective interpretation or manual segmentation, indicating the implementation of automated approaches remains unsolved.

Recently advanced AI models have demonstrated specialist-level classification performance in medical image diagnosis (17-24). AI models which automatically correspond representative features from medical image data to specific task, have recently been introduced as a novel technique (25,26). The development of an accurate AI system could reduce the inconsistency among doctors with different expertise and provide management decision support. There is limited research on developing AI algorithm classifying invasiveness of pulmonary nodules (7,27,28). Attempts at assessing SSN invasiveness have been limited to binary classification or simple comparison with doctors (27,29). Previous researches constructed algorithms based on 2D image rather than 3D volume, which has limited the performance of AI techniques. Rare evidence has been reported in external validation in developing invasiveness classification AI system. Nevertheless, the clinical utility of AI-assisted diagnostic models needs to be investigated (6).

In the present study, we aimed to elucidate the applicability and reliability of a 3D AI algorithm to assess the invasiveness of SSNs by comparing both against the diagnostic performance of chest radiologists and thoracic surgeons and our previously developed feature-based radiomic signature (10). We investigated the practicality of our proposed AI algorithm by evaluating the improvement of prediction performance when the proposed method served as a second opinion. To further investigate the clinical utility, the proposed AI system was validated in

an external cohort with chest radiologists and thoracic surgeons. To our knowledge, this is the first investigation of how AI assists doctors in SSN malignancy evaluation. We present the following article in accordance with the STARD reporting checklist (available at https://dx.doi.org/10.21037/tlcr-21-971).

Methods

Patient selection and study materials

Consecutive patients who underwent pulmonary resection for lung adenocarcinoma between January 2013 and December 2015 in Shanghai Pulmonary Hospital were retrospectively collected. Using the descriptive events including "subsolid nodule", "part-solid nodule", "non-solid nodule", "mixed nodule", "ground-glass nodule" or "ground glass opacity", we retrieved the preoperative CT examinations of patients and 4,679 scans were confirmed. The CT scans were reviewed and SSNs were included under the following criteria: (I) the maximum diameter of lesion ≤3 cm on thin-section CT images (<1.5 mm) within 2 weeks prior to the surgery; (II) pulmonary nodules were histopathologically confirmed as atypical adenomatous hyperplasia (AAH), adenocarcinoma *in situ* (AIS), minimally invasive adenocarcinoma (MIA), or invasive adenocarcinoma (IA) according to the lung tumor classification; (III) patients without a history of malignancy or surgery. For patients with multiple lesions, cases without corresponding confirmed pathological diagnosis were excluded. A total of 1,471 patients with 1,589 SSNs from Shanghai Pulmonary Hospital (Shanghai, China) were included in the present study. A total of 1,349 (85%) nodules from 1,262 patients comprised the development set, including a training subset (n=1,191, 75%) and an internal subset (n=158, 10%); 240 nodules (15%) from 209 patients comprised a hold-out test dataset, of which data were unseen during the training course. To independently test the diagnostic value of the proposed framework, an external test dataset of 100 patients with 102 SSNs was included according to the same criteria from Hwa Mei Hospital (Ningbo, China). The workflow of patient inclusion is shown in Figure S1. In the internal test dataset, 14 (5.8%) patients were diagnosed with AAH, 67 (27.9%) with AIS, 55 (22.9%) with MIA, and 104 (43.4%) of with IA. In the external test dataset, 5 (4.9%) patients were diagnosed with AAH, 25 (24.5%) with AIS, 24 (23.5%) with MIA, and 48 (47.1%) with IA. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). This retrospective study was approved by the Shanghai Pulmonary Hospital Institutional Review Board (No. L20-344). The need for informed consent was waived.

Data extraction and annotations

Chest CT images were acquired on two different scanners: Somatom Definition AS+ (Siemens Medical Systems, Germany, n=1,263) and iCT256 (Philips Medical Systems, Netherlands, n=308). All image data were reconstructed with slice thickness of <1.5 mm (30) and matrix of 512 mm × 512 mm. All CT scans were download form our picture archiving and communications systems (PACS) as digital imaging and communications in medicine (DICOM) images. The personal information of patients in CT images including name, medical number and hospital name were eliminated and images were transformed into NIfTI (NII) format by using an in-house software. The lung CT NII format images were imported into 3D slicer (version 4.8.0, Brigham and Women's Hospital) for labelling. The region of interest (ROI) of SSNs was annotated with a bounding box including the SSN by 2 junior thoracic surgery doctors (Y.S. and J.D. with 4 and 2 years of experience, respectively), then the consensus of ROI was obtained by discussion with an expert chest radiologist (J.S. with 28 years of experience). Bronchi and pulmonary vessels were excluded as far as possible from the ROI. Then the image data of ROI was extracted in the "rcsv" format for further analysis. Each segmented ROI was annotated by a specific histopathologic label according to the specific histologic subtype of AAH, AIS, MIA, and IA. The histologic slides and results were reviewed by two experienced pathologists (L.H. and C.W.) separately with hematoxylin and eosin slides of the study cases in the absence of any clinical or radiologic information. And these histologic labels were reported in accordance with the updated classification of lung cancer (31).

Study design

As illustrated in *Figure 1*, an AI algorithm for invasiveness assessment, SSNet, was constructed for the histological classification of lung adenocarcinoma appearing as pulmonary SSNs on CT scans (Details of algorithm development are shown in Appendix 1). SSNet was developed and tested by using a retrospective dataset with patients with available CT images and corresponding

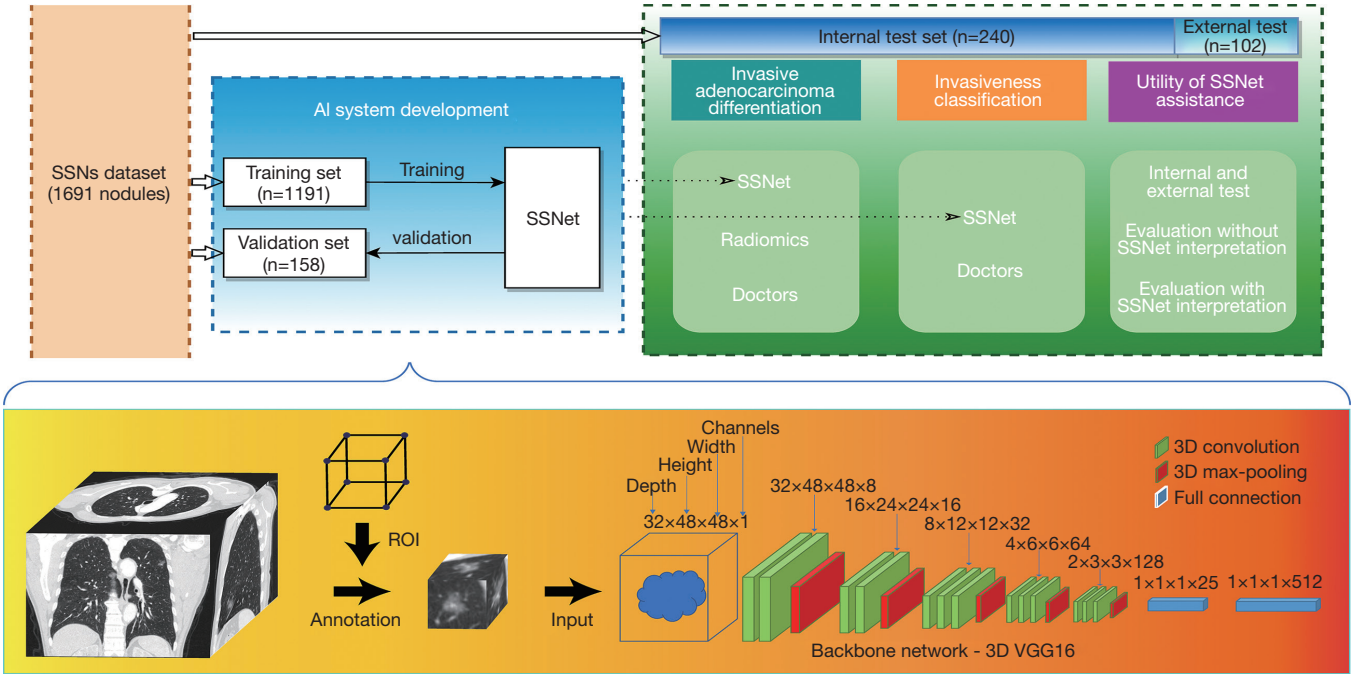

Figure 1 Flowchart of the study design. Artificial intelligence diagnostic tool, SSNet, was first developed and validated using retrospective datasets, then evaluated in an external dataset for its clinical utility. SSNs, subsolid nodules. ROI, region of interest.

histological diagnoses (31). The diagnostic performance of SSNet was evaluated by comparing with those of our previously reported radiomic feature-based signature (10) (Appendix 1) and three chest radiologists and three thoracic surgeons with experience ranging from junior to senior (clinical experience of three to more than 20 years). The 6 doctors were asked to evaluate the SSNs again with the prediction of SSNet to investigate the clinical utility based on performance improvements. Both the diagnostic accuracy and clinical utility for SSNs were further examined in an external test cohort.

Clinical interpretation of CT images

All included cases were reviewed independently in a blinded fashion by six independent doctors from junior to senior degree in thoracic surgery and imaging (Y.S. and T.W. of junior degree with less than 5 years of experience, J.M. and J.S. of intermediate degree with 5–15 years of experience, D.X. and X.S. of senior degree with more than 25 years of experience in thoracic imaging, respectively). All readers subjectively were asked to provide categories of four histological subtypes, then the predicted labels were regrouped according to different tasks for evaluation. To evaluate the increments of diagnostic performance assisted by the SSNet, the six doctors were asked to re-evaluate the SSNs at least 4 weeks after the first evaluation with the predictions of SSNet as a second opinion. Inter-observer variability and diagnostic performance were compared against those from the first evaluation. Clinical interpretation was done without time constrain in RadiAnt Viewer (version 4.6.9, Medixant, https://www.radiantviewer.com). Raters were free to adjust the display window setting and use electronic caliper provided in the software.

Performance testing and statistical analysis

In the discrimination of IA from non-IA, including AAH, AIS, and MIA, the diagnostic performance of SSNet was compared to practicing doctors and radiomic signature using area under receiver-operating characteristic (ROC) curve (AUC) metric (Appendix 1). Comparisons of the diagnostic performance between SSNet and the practicing doctors were also done in 3- and 4-category classifications in a similar manner. Statistically significant differences in AUCs were assessed with Bonferroni-corrected confidence intervals (CIs; 1–0.05/n). Interobserver variability in participant level was evaluated by kappa concordance index.

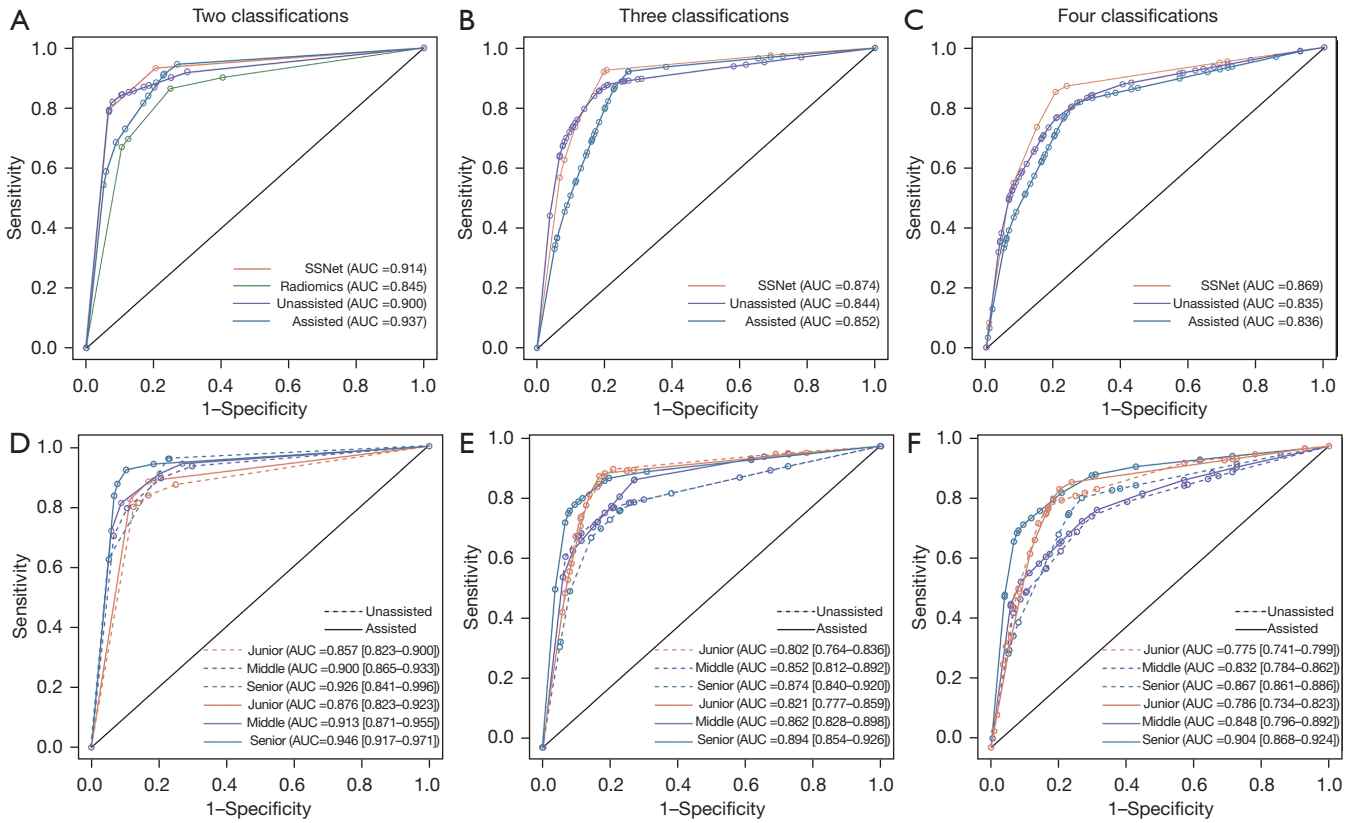

Figure 2 ROC curves showing the diagnostic performance in binary (A,D), 3-category (B,E), and 4-category (C,F) classifications. (A-C) ROC curves measure performance on the methodology-level, including practicing doctors with and without SSNet served as a second viewer. (D-F) ROC curves measure performance on the participant-level of practicing doctors, indicating the performance improvement with the assistance of SSNet. ROC, receiver-operating characteristic; AUC, area under ROC curve.

Performance metrics of sensitivity, specificity, accuracy, positive predictive value, and negative predictive value of each method were measured. Area under the precision-recall curve (AUPRC) and the F1 score were used to evaluate the multiple-category classification performance and reported as the macro average and micro average. Performance-evaluation metrics of practicing doctors were reported in group level and participant level, respectively. McNemar's test was used to compare the statistical difference of sensitivity, specificity, and accuracy between the performance of SSNet and that of practicing doctors, as well as between the performance of practicing doctors with and without the interpretation assistance of SSNet in the binary classification task. Statistical analyses were performed in MedCalc (version 15.2.0; Mariakerke, Belgium), SPSS (version 23.0; IBM, Armonk, NY, USA), and R software (version 3.6.2; https://www.r-project.org/). P<0.05 was considered statistically significant.

Results

Baseline information

The internal dataset consisted of 1,589 SSNs from 471 patients (median age: 57 years, range, 23–82 years) and the external test dataset cohort included 102 SSNs from 100 patients (median age: 56 years, range, 28–75 years). The distribution of histological subtypes was similar between the 2 test datasets. There was no significant difference in age and sex of the 2 cohorts (Table S1).

Diagnostic performance in invasive classification

In the differentiation of IA from minimally invasive/pre-invasive lesions, the ROC curves for the 3 approaches are illustrated in *Figure 2A*, and comparisons of AUCs are reported in *Table 1*. The SSNet algorithm (AUC: 0.914, 95% CI: 0.813–0.987) performed as well as practicing

Table 1 Diagnostic performance and clinical utility in the internal and external test

Tasks	AUC	95% CI	Difference (Bonferroni corrected CI)	Advantage
Internal test				
Two classifications				
SSNet	0.914	0.813–0.987	–	–
Human (unassisted)	0.900	0.867–0.922	–0.014 (–0.090 to 0.060)*	No difference
Human (assisted)	0.937	0.911–0.970	0.037 (–0.078 to –0.014)†	Human (assisted)
Radiomics	0.845	0.806–0.883	0.067 (–0.034 to 0.145)*	No difference
			0.071 (0.032–0.110)‡	Human (unassisted)
Three classifications				
SSNet	0.874	0.832–0.909	–	–
Human (unassisted)	0.844	0.816–0.864	–0.030 (0.000–0.087)*	SSNet
Human (assisted)	0.852	0.825–0.882	0.008 (-0.015-0.042)†	No difference
Four classifications				
SSNet	0.869	0.824–0.892	–	–
Human (unassisted)	0.835	0.817–0.862	–0.034 (0.012–0.098)*	SSNet
Human (assisted)	0.836	0.811–0.862	0.001 (–0.030 to 0.036)†	Human (assisted)
External test				
Two classifications				
SSNet	0.949	0.884–1.000	–	
Human (unassisted)	0.883	0.826–0.939	–0.066 (0.037–0.212)*	SSNet
Human (assisted)	0.908	0.847–0.982	0.025 (–0.092 to 0.029)†	Human (assisted)

*, AUC difference was calculated as the AUC of the algorithm minus the AUC of the doctors (unassisted) or the AUC of radiomics. †, AUC difference was calculated as the AUC of the doctors (assisted) minus the AUC of the doctors (unassisted). ‡, AUC difference was calculated as the AUC of the doctors (unassisted) minus the AUC of the radiomics. To account for multiple hypothesis testing, the Bonferroni corrected CI (1–0.05/n, 97.5% for 2 classifications; 98.3% for 3 classifications; 98.8% for 4 classifications) around the difference was computed. AUC, area under the receiver-operating characteristic curve; CI, confidence interval.

doctors (AUC: 0.900, 95% CI: 0.867–0.922). The radiomic signature was inferior to the practicing doctors, with an AUC of 0.845 (95% CI: 0.806–0.883) and a statistically significant difference.

In the differentiation of pre-invasive, minimally invasive, and invasive lesions, the ROC curves for SSNet and practicing doctors are illustrated in *Figure 2B*, and comparisons of AUCs are reported in *Table 1*. The SSNet algorithm (AUC: 0.874, 95% CI: 0.832–0.909) performed better than that of practicing doctors (AUC: 0.844, 95% CI: 0.816–0.864).

In the differentiation of all 4 histological subtypes of lung adenocarcinoma, the ROC curves for SSNet and practicing doctors are illustrated in *Figure 2C*, and comparisons of AUCs are reported in *Table 1*. The SSNet algorithm (AUC: 0.869, 95% CI: 0.824–0.892) performed better than practicing doctors (AUC: 0.835, 95% CI: 0.817–0.862).

Performance of SSNet in assisting readers

In the performance test using the SSNet algorithm, AUCs of clinicians were 0.937, 0.852, and 0.836 for differentiating IA from non-IA, MIA, AIS, and for differentiating IA from AAH and MIA, AIS, and AAH, respectively (*Figure 2D-2F*; *Table 1*). Compared with the diagnostic performance of subjective evaluation only, increments of AUCs were 0.037, 0.008, and 0.001, respectively, and those in multiple-category classification were statistically significant (*Table 1*).

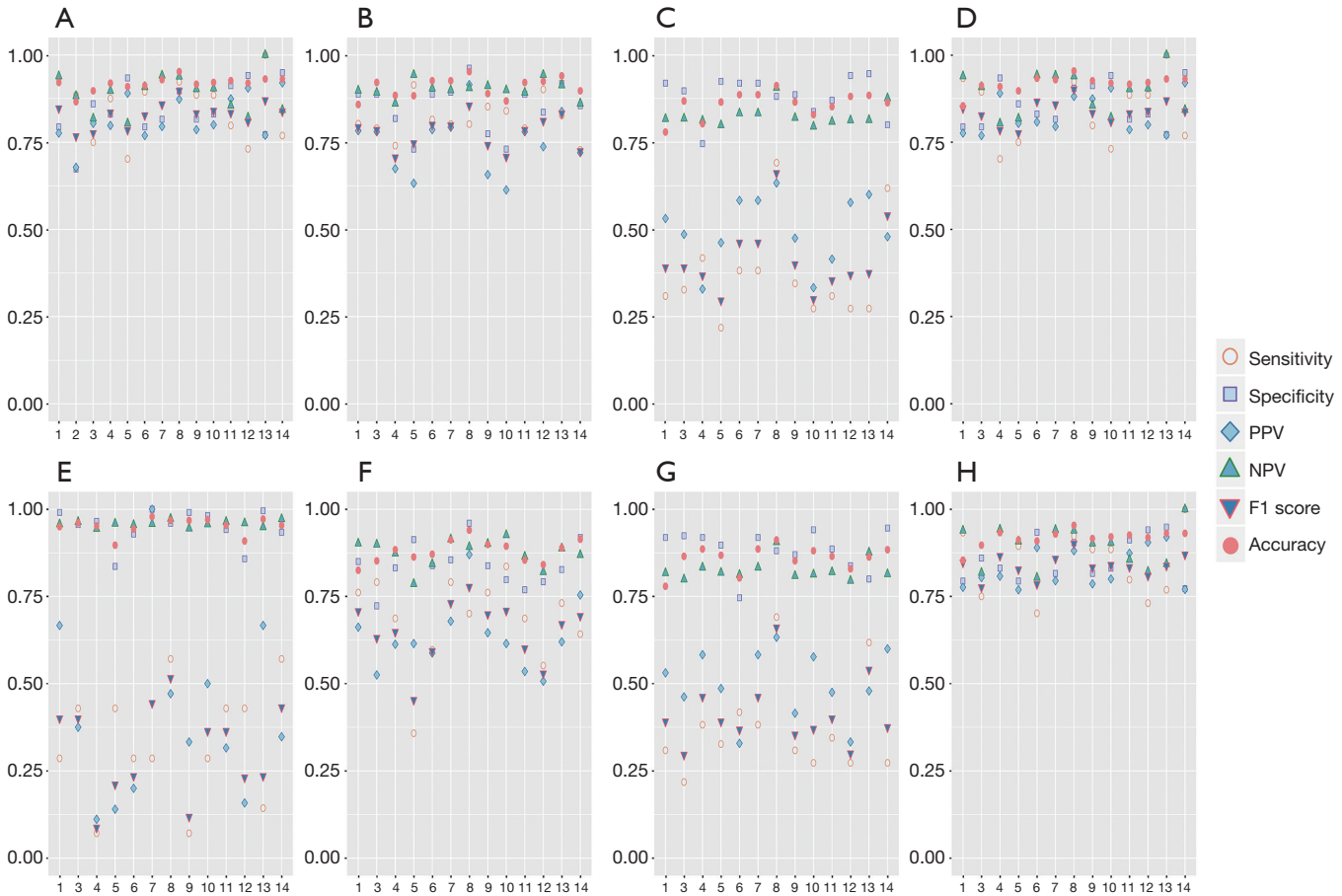

Figure 3 Box graph demonstrating the evaluation metrics in binary (A), 3-category (B-D), and 4-category (E-H) classifications. 1, performance of SSNet; 2, performance of previously constructed radiomic signature; 3–8, performance of practicing doctors without artificial intelligence interpretation; and 9–14, performance of practicing doctors with artificial intelligence interpretation. NPV, negative predictive value; PPV, positive predictive value.

Specifically, in the differentiation of IA, the sensitivity of 1 of the junior doctors improved from 0.750 to 0.885 (P=0.004), and the specificity of 1 of the senior doctors increased from 0.897 to 0.949 (P=0.039). In the multiple-category classification, improvements in multiple statistics were observed in 3 practicing doctors (Tables S2,S3). Improvements were seen in the evaluation metrics in all classes for 3-category classification and were more often in classes of lower-grade invasiveness in the 4-category classification. Kappa statistics improved from 0.480 to 0.496 for 4-category classification but decreased from 0.601 to 0.596 for 3 classifications, respectively.

Performance evaluation in details

In discriminating invasive lesions, performance-evaluating results, including sensitivity, specificity, and accuracy were shown in *Figure 3A* and *Table 2*. In terms of the approach level, SSNet achieved a sensitivity of 0.933, which was higher than the micro-average sensitivity of practicing doctors (0.846), and a radiomic signature of 0.885. SSNet accuracy was 0.921, which was higher than the micro-average accuracy of practicing doctors (0.919) and that of the radiomic signature (0.866). For the 3-category classification, SSNet had better performance for most of the evaluation metrics when compared with practicing doctors as a group and individually (*Figure 3B-3D*; Tables S2,S4). For the 4-category classification, SSNet demonstrated better performance for all evaluation metrics when compared with group or individual performance of practicing doctors (*Figure 3E-3H*; Tables S3,S4). SSNet maintained better performance in terms of micro-average

Table 2 Comparison of SSNet, radiomic signature, and practicing doctors to differentiate invasive adenocarcinoma in the internal and external test

Performance metrics	SSNet	Radiomics	Unassisted							Assisted						
			Junior		Middle		Senior		Micro average	Junior		Middle		Senior		Micro average
			1	2	1	2	1	2		1	2	1	2	1	2	
Retrospective																
Sensitivity	0.933	0.885	0.750	0.875	0.702	0.894	0.933	0.923	0.846	0.885	0.885	0.798	0.731	1.000	0.769	0.845
McNemar's test			<0.001*	0.146*	<0.001*	0.424*	1.000*	1.000*		0.004†	1.000†	0.052†	0.001†	0.016†	0.002†	
Specificity	0.794	0.673	0.860	0.831	0.934	0.794	0.816	0.897	0.855	0.816	0.831	0.912	0.941	0.772	0.949	0.870
McNemar's test			0.049*	0.302*	<0.001*	1.000*	0.250*	0.003*		0.146†	1.000†	0.508†	<0.001†	0.180†	0.039†	
Accuracy	0.921	0.866	0.897	0.919	0.909	0.912	0.929	0.952	0.919	0.916	0.921	0.926	0.919	0.931	0.931	0.880
McNemar's test			<0.001*	0.054	<0.001*	0.596*	0.250*	0.012*		0.001†	1.000†	0.031†	<0.001†	0.007†	<0.001†	
Kappa‡						0.718							0.701			
Prospective																
Sensitivity	0.958		0.708	0.854	0.625	0.875	0.958	0.896	0.819	0.0.875	0.813	0.729	0.667	1.000	0.729	0.802
McNemar's test			<0.001*	0.063*	<0.001*	0.289*	1.000*	0.375*		0.039†	0.688†	0.227†	0.006†	0.500†	0.039†	
Specificity	0.796		0.815	0.852	0.907	0.759	0.815	0.833	0.830	0.759	0.778	0.870	0.944	0.778	0.944	0.846
McNemar's test			1.000*	0.453*	0.031*	0.727*	1.000*	0.688*		0.453†	0.289†	0.727†	0.006†	0.727†	0.031†	
Accuracy	0.932		0.867	0.921	0.873	0.897	0.938	0.926	0.904	0.897	0.885	0.891	0.897	0.938	0.915	0.904
McNemar's test			0.004*	0.039*	0.001*	0.804*	1.000 *	0.227*		0.019†	0.791†	0.167†	<0.001†	0.344†	0.001†	
Kappa‡						0.632							0.649			

1, 2 represents doctors 1 and 2. *, McNemar's test P value for comparison of evaluation metrics between SSNet and practicing doctors alone; †, McNemar's test P-value for comparison of evaluation metrics between practicing doctors with and without the assistance of SSNet; ‡, Kappa value was calculated as Fleiss' kappa for the 6 readers.

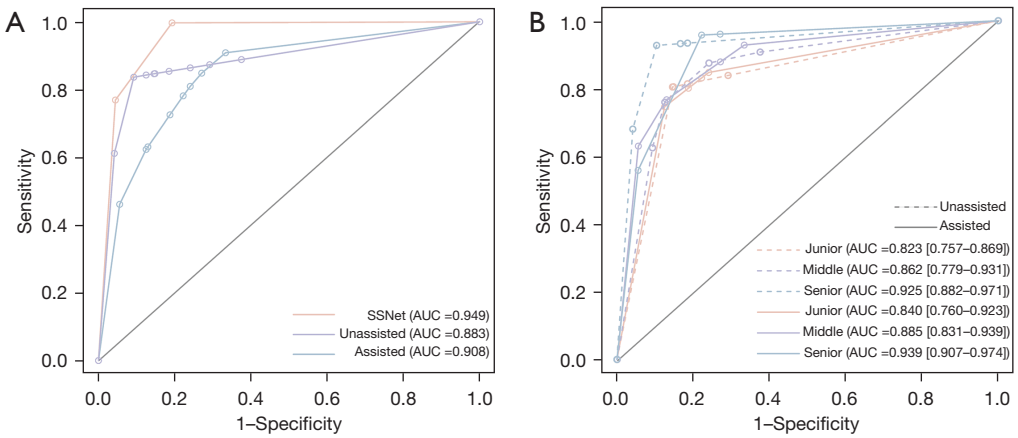

Figure 4 ROC curves showing the diagnostic performance (A) for invasive adenocarcinoma discrimination in prospective validation by SSNet and practicing doctors (B). ROC, receiver-operating characteristic; AUC, area under ROC curve.

AUPRC than junior doctors and 1 of the mid-career doctors. A mid-career and a senior doctor achieved a higher micro-average AUPRC than SSNet.

External test for diagnostic performance

The primary outcome was evaluated for 102 SSNs. SSNet demonstrated excellent diagnostic performance, with an AUC of 0.949 (95% CI: 0.884–1.000), and was better than that of the practicing doctors (AUC: 0.883, 95% CI: 0.826–0.982) (*Table 1*; *Figure 4*). The sensitivity and accuracy of the differentiation for IA by SSNet was 0.958 and 0.932, respectively, which was significantly higher than the micro-average sensitivity (0.819) and accuracy (0.904) of practicing doctors as a group, respectively. In the evaluation of clinical utility, the AUC of practicing doctors was significantly improved from 0.883 to 0.908 with the assistance of SSNet (*Table 1*; *Figure 4*). Sensitivity and specificity micro averages of practicing doctors also improved. For the participant level of performance improvement, the accuracy of a junior doctor significantly improved from 0.867 to 0.897 (*Table 2*). The kappa statistic also improved from 0.632 to 0.649.

Discussion

Clinical management for SSNs between invasive and pre-invasive lesions is different. Therefore, the use of a risk-prediction tool that distinguishes invasive lesions from pre-invasive ones is significant. In the present study, we demonstrated that a simple 3D AI diagnostic tool, SSNet, based on CT images enabled the differentiation of IA from pre-invasive/minimally invasive lesions and histological subtype classification in lesions that appeared as SSNs (<3 cm) on chest CT. Performance was equal between SSNet, radiomic signature and doctors in binary discrimination on an internal test but better for SSNet than doctors on external test in the classification of more than two categories. In addition, the use of SSNet enhanced doctors' SSN interpretations.

Evaluation by radiomic signature, which was previously developed for discriminating IAs, did not reach its optimal performance in our internal test. A possible reason for performance discrepancy would be the spectrum effect (25,32). The population used for the construction of a radiomic signature had a higher proportion of invasive lesions (>50%), while the rate of IA was relatively low in the present study (<50%) (10). A similar performance drop was also seen in another validation experiment using a similar population (33). Previous models developed to differentiate invasive lesions by CT features had an AUC of 0.64–0.91 (7,10,34-38). However, these models have not been validated in an external cohort. The performance of the SSN evaluation model based on predefined features was limited by the subjectivity and proficiency in CT interpretation, whereas the AI-based evaluation model is able to learn representative features from raw medical image without specifying radiological features. Recently, Wang *et al.* proposed an AI system using a 3D convolutional neural network for differentiating pre-invasive lesions from IA appearing as SSNs no larger than 3 cm (29). In their study, the proposed architecture with an AUC of 0.892 outperformed the performances of 4 radiologists,

who yielded an AUC between 0.805 and 0.867. However, the model was not designed for further discrimination of specific lung adenocarcinoma histological subtypes and the model was not fully investigated or externally validated for its clinical utility. In our study, the 3D SSNet utilized the volumetric data of thin-section CT from 1,471 patients and the proposed AI system achieved a competitive AUC of 0.914 in terms of differentiating IA from pre-invasive and minimally invasive lesions compared with doctors. In addition, the external diagnostic evaluation found that SSNet outperformed practicing doctors in IA discrimination.

Concerns remain regarding the actual help of an AI-based evaluation system in clinical practice. To date, limited studies have investigated the benefits of an AI system in assessing invasiveness (7,20,39,40). In the present study, we investigated the improvement with the assistance of AI interpretation as a reference. Based on our results, the performance of practicing doctors improved with the assistance of SSNet in invasive discrimination and multiple-category invasiveness assessments. Although the AUPRC, a metric evaluating a classifier's performance in imbalanced data, of SSNet was lower than some of the practicing doctors in terms of multiple-category classification, the accuracy of doctors improved with SSNet. These findings support the role of the AI system as a second viewer that would increase the AUPRC in diagnosing cases that doctors might misinterpret or miss.

In the present study, SSNet achieved better performance than practicing doctors in identifying lesions with lower grade of invasiveness, which was shown in the 3- and 4-category classifications. The F1 scores for AAH class and MIA class were relatively low due to the number of samples, which could limit the performance of the AI system. It is recommended that patients with AAH are routinely followed up or undergo resection after comprehensive evaluation. MIA is defined as small nodules with ≤5 mm predominantly lepidic invasion, Lim *et al.* inferred that invasion ≤5 mm might not greatly contribute in the emergence of increased attenuation on CT scan (41). Therefore, MIA appearing as SSNs can easily be misclassified as AIS or IA by doctors and the AI system (Figure S2).

In the present study, there was no incorrect prediction of SSNet as AAH being misclassified as IA or IA as AAH (Figure S2). SSNet achieved a higher AUPRC in differentiating histological subtypes than junior doctor in multiple-category classification. Although the AUPRC of SSNet was not higher than those of intermediate and senior doctors, the reduced time in interpretation of SSNs and the improved diagnostic performance would streamline the workflow and reduce subjectivity bias. Additionally, several underestimated predictions can be corrected by the AI system, as radiological features indicating malignancy could be absent or not fully identified (Figure S2). This was also proved by the highest sensitivity by SSNet and the confusion matrices. The SSN evaluation process is simple. Only a cuboid box that fully embraces SSN at its mass center is required for the invasiveness evaluation. Therefore, the incorporation of SSNet into the workflow would improve efficiency and accuracy in identifying SSNs and the workload of radiologists.

The present study has several limitations. First, data of this study only reflects this particular study population and cannot extrapolated to a screening setting. Secondly, though the proposed AI algorithm showed discriminative ability in classifying invasiveness of SSNs, the relevant image features that dominate model for decision-making were not reflected due to the black box nature of deep learning algorithm. Further explainable deep learning modes with good performance are necessary to improve the transparency and re reliability for humans. Additionally, the generalizability of our proposed method should be confirmed with more external and prospective validation. In the present study, the test dataset was split prior to model development and treated as a hold-out dataset to evaluate diagnostic performance. An external test cohort was used for validation. It should be acknowledged that although geographically external validation as applied, the SSN interpretation was done in a non-clinical environment and would not change clinical decisions that happened to the included patients. There is a potential decision threshold shift that could bias the actual clinical utility of the AI diagnostic tool. Therefore, further prospective studies for clinical decision making with long-time follow-up are needed to warrant the clinical utility of this diagnostic AI system and the value of improving patient outcome for such an AI diagnostic system designed for clinical management decision support. Multi-institutional randomized controlled trials are critical to test the benefit of incorporating AI into workflow.

In conclusion, the proposed SSNet is helpful in evaluating SSNs and can be used to assess invasiveness. The implementation of SSNet in practice has the potential to improve patient care workflow and optimize clinical decision support. The safety and feasibility of AI-assisted tools in supporting clinical decisions for SSNs are warranted

in long-term and multi-institutional trials.

Acknowledgments

The authors appreciate the academic support from the AME Thoracic Surgery Collaborative Group. The authors would like to thank biostatistician, Professor Aihong Zhang (Department of Medical Statistics, Tongji University School of Medicine, Shanghai, China), for guidance on the statistical analysis in this research.
Funding: This study was supported by National Natural Science Foundation of China (No. 8210071009); Shanghai Science and Technology Commission (No. 21YF1438200); Shanghai Municipal Health Commission (No. 2019SY072); the Science-Technology Foundation for Young Scientists of Gansu Province (No. 18JR3RA305, 21JR1RA107); and the Natural Science Foundation of Gansu Province (No. 21JR1RA118, 21JR1RA092).

Footnote

Reporting Checklist: The authors have completed the STARD reporting checklist. Available at https://dx.doi.org/10.21037/tlcr-21-971

Data Sharing Statement: Available at https://dx.doi.org/10.21037/tlcr-21-971

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://dx.doi.org/10.21037/tlcr-21-971). SZ reports employment at Tailai Biosciences Inc., Shenzhen, Chinia. SJ reports employment at Dianei Technology, Shanghai, China. The other authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). This retrospective study was approved by the Shanghai Pulmonary Hospital Institutional Review Board (No. L20-344). The need for informed consent was waived.

References

1. Field JK, Duffy SW, Baldwin DR, et al. UK Lung Cancer RCT Pilot Screening Trial: baseline findings from the screening arm provide evidence for the potential implementation of lung cancer screening. Thorax 2016;71:161-70.
2. de Koning HJ, van der Aalst CM, de Jong PA, et al. Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial. N Engl J Med 2020;382:503-13.
3. Rzyman W, Szurowska E, Adamek M. Implementation of lung cancer screening at the national level: Polish example. Transl Lung Cancer Res 2019;8:S95-105.
4. Yankelevitz DF, Yip R, Smith JP, et al. CT Screening for Lung Cancer: Nonsolid Nodules in Baseline and Annual Repeat Rounds. Radiology 2015;277:555-64.
5. Henschke CI, Yip R, Smith JP, et al. CT Screening for Lung Cancer: Part-Solid Nodules in Baseline and Annual Repeat Rounds. AJR Am J Roentgenol 2016;207:1176-84.
6. Kauczor HU, von Stackelberg O. Subsolid Lung Nodules: Potential for Overdiagnosis. Radiology 2019;293:449-50.
7. Varghese C, Rajagopalan S, Karwoski RA, et al. Computed Tomography-Based Score Indicative of Lung Cancer Aggression (SILA) Predicts the Degree of Histologic Tissue Invasion and Patient Survival in Lung Adenocarcinoma Spectrum. J Thorac Oncol 2019;14:1419-29.
8. McWilliams A, Tammemagi MC, Mayo JR, et al. Probability of cancer in pulmonary nodules detected on first screening CT. N Engl J Med 2013;369:910-9.
9. Li Y, Chen KZ, Wang J. Development and validation of a clinical prediction model to estimate the probability of malignancy in solitary pulmonary nodules in Chinese people. Clin Lung Cancer 2011;12:313-9.
10. She Y, Zhang L, Zhu H, et al. The predictive value of CT-based radiomics in differentiating indolent from invasive lung adenocarcinoma in patients with pulmonary nodules. Eur Radiol 2018;28:5121-8.
11. Maldonado F, Boland JM, Raghunath S, et al. Noninvasive characterization of the histopathologic features of

pulmonary nodules of the lung adenocarcinoma spectrum using computer-aided nodule assessment and risk yield (CANARY)--a pilot study. J Thorac Oncol 2013;8:452-60.
12. American College of Radiology. Lung-RADS Version 1.1. 2019. Accessed 13 Jan 2019. Available online: https://www.acr.org/-/media/ACR/Files/RADS/Lung-RADS/LungRADSAssessmentCategoriesv1-1.pdf?la=en
13. Lee SM, Park CM, Goo JM, et al. Invasive pulmonary adenocarcinomas versus preinvasive lesions appearing as ground-glass nodules: differentiation by using CT features. Radiology 2013;268:265-73.
14. Heidinger BH, Anderson KR, Nemec U, et al. Lung Adenocarcinoma Manifesting as Pure Ground-Glass Nodules: Correlating CT Size, Volume, Density, and Roundness with Histopathologic Invasion and Size. J Thorac Oncol 2017;12:1288-98.
15. Qi L, Xue K, Li C, et al. Analysis of CT morphologic features and attenuation for differentiating among transient lesions, atypical adenomatous hyperplasia, adenocarcinoma in situ, minimally invasive and invasive adenocarcinoma presenting as pure ground-glass nodules. Sci Rep 2019;9:14586.
16. Kim H, Goo JM, Kim YT, et al. Consolidation-to-tumor ratio and tumor disappearance ratio are not independent prognostic factors for the patients with resected lung adenocarcinomas. Lung Cancer 2019;137:123-8.
17. Esteva A, Kuprel B, Novoa RA, et al. Corrigendum: Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;546:686.
18. Rajpurkar P, Irvin J, Ball RL, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med 2018;15:e1002686.
19. Nam JG, Park S, Hwang EJ, et al. Development and Validation of Deep Learning-based Automatic Detection Algorithm for Malignant Pulmonary Nodules on Chest Radiographs. Radiology 2019;290:218-28.
20. Zhang Y, Wu X, He L, et al. Applications of hyperspectral imaging in the detection and diagnosis of solid tumors. Transl Cancer Res 2020;9:1265-77.
21. Tschandl P, Codella N, Akay BN, et al. Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. Lancet Oncol 2019;20:938-47.
22. Gulshan V, Peng L, Coram M, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA 2016;316:2402-10.
23. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. JAMA 2017;318:2199-210.
24. Li X, Zhang S, Zhang Q, et al. Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. Lancet Oncol 2019;20:193-201.
25. Faes L, Wagner SK, Fu DJ, et al. Automated deep learning design for medical image classification by health-care professionals with no coding experience: a feasibility study. Lancet Digit Health 2019;1:e232-42.
26. Castiglioni I, Rundo L, Codari M, et al. AI applications to medical images: From machine learning to deep learning. Phys Med 2021;83:9-24.
27. Zhao W, Yang J, Sun Y, et al. 3D Deep Learning from CT Scans Predicts Tumor Invasiveness of Subcentimeter Pulmonary Adenocarcinomas. Cancer Res 2018;78:6881-9.
28. Kim H, Lee D, Cho WS, et al. CT-based deep learning model to differentiate invasive pulmonary adenocarcinomas appearing as subsolid nodules among surgical candidates: comparison of the diagnostic performance with a size-based logistic model and radiologists. Eur Radiol 2020;30:3295-305.
29. Wang S, Wang R, Zhang S, et al. 3D convolutional neural network for differentiating pre-invasive lesions from invasive adenocarcinomas appearing as ground-glass nodules with diameters ≤3 cm using HRCT. Quant Imaging Med Surg 2018;8:491-9.
30. Guchlerner L, Wichmann JL, Tischendorf P, et al. Comparison of thick- and thin-slice images in thoracoabdominal trauma CT: a retrospective analysis. Eur J Trauma Emerg Surg 2020;46:187-95.
31. Travis WD, Brambilla E, Nicholson AG, et al. The 2015 World Health Organization Classification of Lung Tumors: Impact of Genetic, Clinical and Radiologic Advances Since the 2004 Classification. J Thorac Oncol 2015;10:1243-60.
32. Riley RD, Ensor J, Snell KI, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ 2016;353:i3140.
33. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med 1978;299:926-30.
34. Liang J, Xu XQ, Xu H, et al. Using the CT features to differentiate invasive pulmonary adenocarcinoma from

pre-invasive lesion appearing as pure or mixed ground-glass nodules. Br J Radiol 2015;88:20140811.
35. Son JY, Lee HY, Kim JH, et al. Quantitative CT analysis of pulmonary ground-glass opacity nodules for distinguishing invasive adenocarcinoma from non-invasive or minimally invasive adenocarcinoma: the added value of using iodine mapping. Eur Radiol 2016;26:43-54.
36. Zhang Y, Shen Y, Qiang JW, et al. HRCT features distinguishing pre-invasive from invasive pulmonary adenocarcinomas appearing as ground-glass nodules. Eur Radiol 2016;26:2921-8.
37. Jin C, Cao J, Cai Y, et al. A nomogram for predicting the risk of invasive pulmonary adenocarcinoma for patients with solitary peripheral subsolid nodules. J Thorac Cardiovasc Surg 2017;153:462-469.e1.
38. Baldwin DR, Gustafson J, Pickup L, et al. External validation of a convolutional neural network artificial intelligence tool to predict malignancy in pulmonary nodules. Thorax 2020;75:306-12.
39. Hwang EJ, Nam JG, Lim WH, et al. Deep Learning for Chest Radiograph Diagnosis in the Emergency Department. Radiology 2019;293:573-80.
40. Hwang EJ, Park S, Jin KN, et al. Development and Validation of a Deep Learning-based Automatic Detection Algorithm for Active Pulmonary Tuberculosis on Chest Radiographs. Clin Infect Dis 2019;69:739-47.
41. Lim HJ, Ahn S, Lee KS, et al. Persistent pure ground-glass opacity lung nodules ≥ 10 mm in diameter at CT scan: histopathologic comparisons and prognostic implications. Chest 2013;144:1291-9.

(English Language Editor: R. Scott)

Cite this article as: Deng J, Zhao M, Li Q, Zhang Y, Ma M, Li C, Wang J, She Y, Jiang Y, Zhang Y, Wang T, Wu C, Hou L, Zhong S, Jin S, Qian D, Xie D, Zhu Y, Tandon YK, Snoeckx A, Jin F, Yu B, Zhao G, Chen C; on behalf of the MultiomIcs claSSIfier for pulmOnary Nodules (MISSION) Collaborative Group. Implementation of artificial intelligence in the histological assessment of pulmonary subsolid nodules. Transl Lung Cancer Res 2021;10(12):4574-4586. doi: 10.21037/tlcr-21-971

Appendix 1

1. Artificial intelligence algorithm

We designed a 3D Deep Learning algorithm, SSNet, with 13 3D convolutional layers, 5 max pooing layers, and 2 fully connected layers (*Figure 1*). The input images were 3D shaped data cropped from the CT scan with a volume of size 32 mm × 48 mm × 48 mm at the mass center of a ROI with a histological label. The output of the proposed algorithm was probabilities for different categories. The artificial intelligence algorithm was trained from scratch for three differentiation tasks: (I) aggressive (IA) or indolent (AAH, AIS, MIA); (II) categories of different invasiveness, pre-invasive (AAH, AIS), minimally invasive (MIA), invasive (IA); (III) categories of four histological subtypes.

2. Algorithm training and interpretation

The training of the algorithm was performed on a computer with an NVIDA GTX 1080 (NVIDIA, Santa Clara, Calif) graphics processing unit (GPU) and used the TensorFlow deep learning framework (Google, Mountain View, CA). Momentum optimizer was used to minimize the Softmax cross-entropy between the outputs and reference labels with a batch size of 64 and initial learning rate of 0.01, decayed every 300 iterations using an exponential rate of 0.99. We augmented the samples by randomly rotating each patch to 0, 90, 180, and 270 degrees along the Z axis, and randomly flipping them in the X, Y, and Z directions. To prevent overfitting, we used L2 regularization during training. Our training loss converged after 3,000 iterations. The model with the lowest validation loss was selected eventually. To increase the understandability and dependability of the proposed SSNet, we adopted class activation mapping method to generate heat maps to indicate invasiveness in input images by using the feature map extracted from the developed network. The heat mapping was done with the "Matplotlib" module and all programming was conducted in Python version 3.6.4.

3. Interpretation by a feature-based machine learning method

To exploit the potential difference from traditional feature-based AI technique in interpretation of nodule aggressiveness, our previously published radiomic signature was utilized (10), and analysis was performed with extracted radiomic features. Tumor segmentation, feature extraction, and inter-/intra-observer variability was reported previously. The malignancy risk was computed according to the input features and classified the nodules into IA and non-IA (binary classification).

4. Receiver operating characteristic curves analysis

Instead of a continuous value describing invasiveness, only a binary label was provided by doctors. Thus, the receiver operating characteristic (ROC) curves were estimated for six practicing doctors as a group, radiomic signature, and AI model using partial least-squares regression with constrained splines as previously described to warrant a fair comparison (18). Then linear interpolation and the composite trapezoidal rule were applied to estimate the area under ROC curve (AUC) for three approaches. At last, the confidential intervals (CI) of AUCs were obtained through 10,000 bootstrap replicates drawn from test set, on which three approaches were measured using the same replicate. The difference between AUCs was calculated on these same replicates by the stringent Bonferroni-corrected CIs of 1–0.05/k (k stands for number of classes). There is evidence of difference when 0 was not included in the interval. Similar way for AUC calculation was introduced by Rajpurkar *et al.* previously (18).

https://dx.doi.org/10.21037/tlcr-21-971

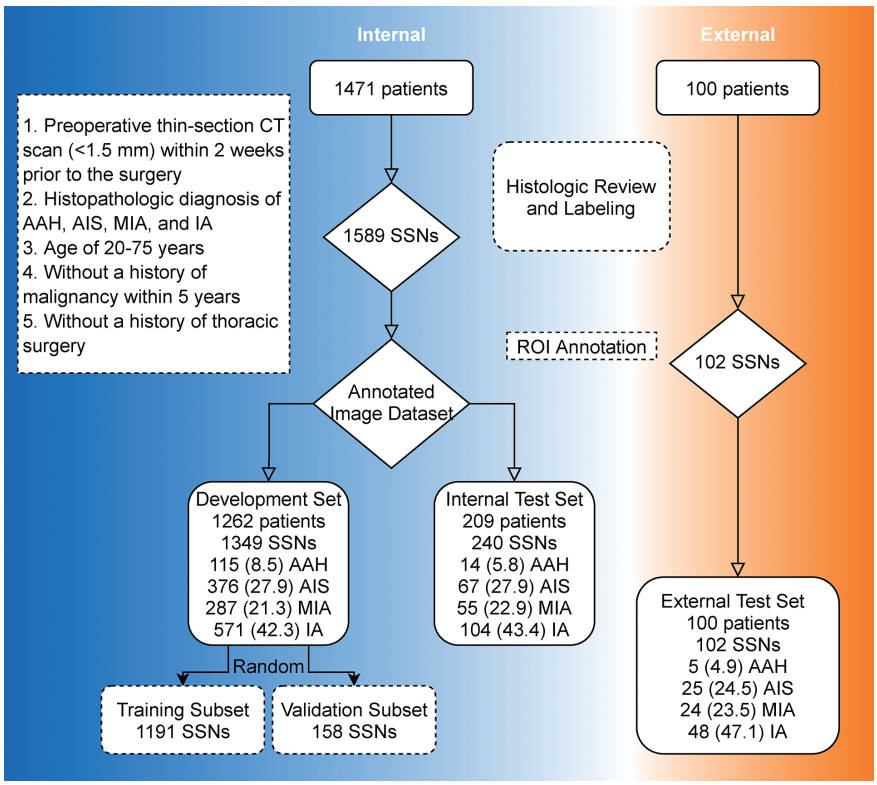

Figure S1 Flowchart of patient allocation in the retrospective dataset and external dataset. Number in parentheses of the left panel represents the percentage of each histological subtype for SSNs. SSN, subsolid nodule; AIS, adenocarcinoma *in situ*; IA, invasive adenocarcinoma; CT, computed tomography MIA, minimally invasive adenocarcinoma; ROI, region of interest.

Table S1 Summary statistics of patients in the Shanghai cohort (training dataset and test dataset) and Ningbo cohort

Characteristics	Development dataset (n=1,262)	Testing dataset (n=209)	External dataset (n=100)	P_1	P_2
Age (years)				0.209	0.692
<65	1,008 (79.9)	159 (76.1)	74 (74.0)		
≥65	254 (20.1)	50 (23.9)	26 (26.0)		
Sex				0.174	0.952
Male	435 (34.5)	62 (29.7)	30 (30.0)		
Female	827 (65.5)	147 (70.3)	70 (70.0)		
Nodule count				0.003	0.956
Solitary	1,168 (92.6)	205 (98.1)	98 (98.0)		
Multiple	94 (7.4)	4 (1.9)	2 (2.0)		

P_1 value, training dataset compared with testing dataset; P_2 value, training dataset compared with external dataset.

https://dx.doi.org/10.21037/tlcr-21-971

Table S2 Comparison of SSNet and practicing doctors to differentiate AAH/AIS, MIA, and IA

Performance metrics	SSNet	Practicing doctors													
		Unassisted							Assisted						
		Junior		Middle		Senior		Micro average	Junior		Middle		Senior		Micro average
		1	2	1	2	1	2		1	2	1	2	1	2	
Sensitivity															
Class 1	0.803	0.790	0.740	0.914	0.815	0.802	0.802	0.811	0.852	0.840	0.790	0.901	0.827	0.728	0.823
Class 2	0.309	0.327	0.418	0.218	0.382	0.382	0.691	0.403	0.345	0.273	0.309	0.273	0.273	0.618	0.348
Class 3	0.933	0.894	0.702	0.750	0.933	0.933	0.923	0.856	0.798	0.731	0.885	0.885	1.000	0.769	0.845
Micro average	0.746	0.782	0.749	0.751	0.808	0.805	0.870	0.734	0.771	0.757	0.777	0.816	0.814	0.797	0.727
Specificity															
Class 1	0.887	0.887	0.818	0.730	0.887	0.893	0.962	0.863	0.774	0.730	0.887	0.836	0.918	0.855	0.833
Class 2	0.919	0.897	0.746	0.924	0.919	0.919	0.881	0.881	0.886	0.838	0.870	0.941	0.946	0.800	0.880
Class 3	0.794	0.794	0.934	0.860	0.831	0.816	0.904	0.857	0.912	0.941	0.816	0.831	0.772	0.949	0.870
Micro average	0.873	0.871	0.818	0.853	0.886	0.884	0.889	0.867	0.865	0.823	0.868	0.858	0.889	0.847	0.863
PPV															
Class 1	0.783	0.674	0.793	0.786	0.780	0.915	0.786	0.750	0.657	0.613	0.780	0.737	0.838	0.720	0.716
Class 2	0.531	0.329	0.583	0.583	0.486	0.633	0.583	0.502	0.475	0.333	0.415	0.577	0.600	0.479	0.464
Class 3	0.776	0.890	0.795	0.808	0.769	0.881	0.808	0.820	0.874	0.905	0.786	0.800	0.770	0.920	0.833
Micro average	0.746	0.705	0.803	0.806	0.780	0.821	0.806	0.747	0.769	0.715	0.775	0.771	0.811	0.752	0.736
NPV															
Class 1	0.898	0.861	0.899	0.904	0.892	0.905	0.904	0.899	0.911	0.899	0.892	0.943	0.913	0.861	0.902
Class 2	0.817	0.812	0.833	0.833	0.818	0.906	0.833	0.832	0.820	0.795	0.809	0.813	0.814	0.876	0.820
Class 3	0.939	0.804	0.941	0.942	0.908	0.939	0.942	0.886	0.855	0.821	0.902	0.904	1.000	0.843	0.880
Micro average	0.873	0.848	0.886	0.888	0.872	0.921	0.888	0.867	0.866	0.853	0.869	0.889	0.891	0.877	0.864
F1 score															
Class 1	0.793	0.706	0.798	0.800	0.785	0.855	0.800	0.779	0.742	0.708	0.785	0.811	0.832	0.724	0.766
Class 2	0.391	0.368	0.462	0.462	0.391	0.661	0.462	0.447	0.400	0.300	0.354	0.370	0.375	0.540	0.398
Class 3	0.847	0.785	0.858	0.866	0.827	0.901	0.866	0.838	0.834	0.809	0.833	0.840	0.870	0.838	0.839
Micro average	0.746	0.726	0.804	0.807	0.781	0.845	0.807	0.779	0.770	0.735	0.776	0.793	0.812	0.774	0.766
Accuracy															
Class 1	0.858	0.884	0.926	0.926	0.921	0.952	0.926	0.916	0.889	0.868	0.921	0.924	0.940	0.897	0.907
Class 2	0.779	0.803	0.886	0.886	0.868	0.912	0.886	0.871	0.865	0.829	0.852	0.881	0.884	0.863	0.863
Class 3	0.854	0.909	0.929	0.933	0.912	0.954	0.933	0.923	0.926	0.919	0.916	0.921	0.931	0.931	0.924
Micro average	0.831	0.884	0.922	0.923	0.945	0.937	0.923	0.924	0.907	0.888	0.910	0.915	0.926	0.906	0.921
AUPRC															
Macro average	0.685	0.668	0.620	0.606	0.709	0.706	0.806		0.659	0.606	0.657	0.674	0.692	0.701	
Micro average	0.750	0.729	0.650	0.683	0.767	0.763	0.829		0.713	0.663	0.721	0.750	0.775	0.721	
Fleiss' kappa					0.601							0.596			

1, 2 represents doctors 1 and 2; class 1 represents AAH/AIS, class 2 represents MIA, and class 3 represents IA. AAH, atypical adenomatous hyperplasia. AIS, adenocarcinoma *in situ*; AUPRC, area under precision-recall curve; IA, invasive adenocarcinoma; MIA, minimally invasive adenocarcinoma; NPV, negative predictive value; PPV, positive predictive value.

https://dx.doi.org/10.21037/tlcr-21-971

Table S3 Comparison of SSNet and practicing doctors to differentiate AAH, AIS, MIA, and IA

Performance metrics	SSNet	Practicing doctors													
		Unassisted							Assisted						
		Junior		Middle		Senior		Micro average	Junior		Middle		Senior		Micro average
		1	2	1	2	1	2		1	2	1	2	1	2	
Sensitivity															
Class 1	0.286	0.429	0.286	0.429	0.071	0.286	0.571	0.345	0.429	0.429	0.071	0.286	0.143	0.571	0.321
Class 2	0.761	0.358	0.597	0.791	0.687	0.791	0.701	0.654	0.687	0.552	0.761	0.836	0.731	0.642	0.701
Class 3	0.309	0.327	0.418	0.218	0.382	0.382	0.691	0.403	0.345	0.273	0.309	0.273	0.618	0.273	0.348
Class 4	0.933	0.894	0.702	0.750	0.933	0.933	0.923	0.856	0.798	0.731	0.885	0.885	0.769	1.000	0.845
Micro average	0.704	0.588	0.583	0.621	0.688	0.729	0.788	0.641	0.642	0.559	0.671	0.696	0.688	0.708	0.640
Specificity															
Class 1	0.991	0.836	0.929	0.956	0.965	1.000	0.960	0.941	0.942	0.858	0.991	0.982	0.996	0.934	0.951
Class 2	0.850	0.913	0.838	0.723	0.832	0.855	0.960	0.854	0.769	0.792	0.838	0.798	0.827	0.919	0.824
Class 3	0.919	0.897	0.746	0.924	0.919	0.919	0.881	0.881	0.886	0.838	0.870	0.941	0.800	0.946	0.880
Class 4	0.794	0.794	0.934	0.860	0.831	0.816	0.904	0.857	0.912	0.941	0.816	0.831	0.949	0.772	0.870
Micro average	0.901	0.863	0.861	0.874	0.896	0.91	0.925	0.908	0.881	0.853	0.890	0.899	0.896	0.903	0.913
PPV															
Class 1	0.667	0.140	0.200	0.375	0.111	1.000	0.471	0.266	0.316	0.158	0.333	0.500	0.667	0.348	0.287
Class 2	0.662	0.615	0.588	0.525	0.613	0.679	0.870	0.634	0.535	0.507	0.646	0.615	0.620	0.754	0.606
Class 3	0.531	0.486	0.329	0.462	0.583	0.583	0.633	0.502	0.475	0.333	0.415	0.577	0.479	0.600	0.464
Class 4	0.776	0.769	0.890	0.804	0.808	0.795	0.881	0.820	0.874	0.905	0.786	0.800	0.920	0.770	0.833
Micro average	0.704	0.588	0.583	0.621	0.688	0.729	0.788	0.588	0.642	0.558	0.671	0.696	0.688	0.708	0.587
NPV															
Class 1	0.957	0.959	0.955	0.964	0.944	0.958	0.973	0.959	0.964	0.960	0.945	0.957	0.949	0.972	0.958
Class 2	0.902	0.786	0.843	0.899	0.873	0.914	0.892	0.864	0.864	0.820	0.901	0.926	0.888	0.869	0.877
Class 3	0.817	0.818	0.812	0.799	0.833	0.833	0.906	0.832	0.820	0.795	0.809	0.813	0.876	0.814	0.820
Class 4	0.939	0.908	0.804	0.818	0.942	0.941	0.939	0.886	0.855	0.821	0.902	0.904	0.843	1.000	0.880
Micro average	0.901	0.863	0.861	0.874	0.896	0.910	0.929	0.888	0.881	0.853	0.890	0.899	0.896	0.903	0.888
F1 score															
Class 1	0.400	0.235	0.444	0.087	0.211	0.516	0.301	0.400	0.364	0.231	0.118	0.364	0.235	0.432	0.303
Class 2	0.708	0.593	0.731	0.648	0.453	0.777	0.644	0.631	0.601	0.529	0.699	0.709	0.671	0.694	0.651
Class 3	0.391	0.368	0.462	0.462	0.391	0.661	0.447	0.296	0.400	0.300	0.354	0.370	0.540	0.375	0.398
Class 4	0.847	0.785	0.858	0.866	0.827	0.901	0.838	0.776	0.834	0.809	0.833	0.840	0.838	0.870	0.839
Micro average	0.704	0.583	0.729	0.688	0.588	0.788	0.643	0.621	0.642	0.558	0.671	0.696	0.688	0.708	0.644
Accuracy															
Class 1	0.950	0.943	0.979	0.952	0.897	0.968	0.874	0.961	0.954	0.909	0.968	0.970	0.972	0.954	0.879
Class 2	0.825	0.871	0.912	0.884	0.863	0.940	0.812	0.852	0.854	0.841	0.899	0.894	0.889	0.914	0.807
Class 3	0.779	0.803	0.886	0.886	0.868	0.912	0.796	0.865	0.865	0.829	0.852	0.881	0.863	0.884	0.788
Class 4	0.854	0.909	0.929	0.933	0.912	0.954	0.847	0.897	0.926	0.919	0.916	0.921	0.931	0.931	0.848
Micro average	0.852	0.884	0.927	0.915	0.885	0.944	0.951	0.895	0.902	0.876	0.910	0.918	0.915	0.921	0.955
AUPRC															
Macro average	0.559	0.471	0.495	0.526	0.516	0.624	0.714		0.550	0.467	0.501	0.571	0.571	0.593	
Micro average	0.667	0.588	0.583	0.621	0.688	0.729	0.788		0.642	0.558	0.671	0.696	0.688	0.675	
Fleiss' kappa					0.480							0.496			

1, 2 represents doctors 1 and 2; class 1 represents AAH, class 2 represents AIS, class 3 represents MIA, and class 4 represents IA. AAH, atypical adenomatous hyperplasia. AIS, adenocarcinoma *in situ*; AUPRC, area under precision-recall curve; IA, invasive adenocarcinoma; MIA, minimally invasive adenocarcinoma; NPV, negative predictive value; PPV, positive predictive value.

https://dx.doi.org/10.21037/tlcr-21-971

Table S4 Performance details of different categories in multiclass differentiation on the participant level

AUC	SSNet	Practicing doctors					
		Unassisted			Assisted		
		Junior	Middle	Senior	Junior	Middle	Senior
Three class							
Class 1	0.879	0.841	0.888	0.921	0.829	0.884	0.870
Class 2	0.696	0.652	0.703	0.829	0.641	0.665	0.768
Class 3	0.914	0.900	0.882	0.928	0.913	0.876	0.946
Four class							
Class 1	0.718	0.703	0.752	0.878	0.751	0.718	0.850
Class 2	0.850	0.776	0.796	0.898	0.736	0.828	0.857
Class 3	0.724	0.652	0.703	0.829	0.641	0.665	0.768
Class 4	0.916	0.900	0.882	0.928	0.913	0.876	0.946

In the 3-class differentiation, class 1 represents AAH/AIS, class 2 represents MIA, and class 3 represents IA. In the 4-class differentiation, class 1 represents AAH, class 2 represents AIS, class 3 represents MIA, and class 4 represents IA. AAH, atypical adenomatous hyperplasia; AIS, adenocarcinoma *in situ*; AUC, area under receiver operating characteristic curve; IA, invasive adenocarcinoma; MIA, minimally invasive adenocarcinoma.

https://dx.doi.org/10.21037/tlcr-21-971

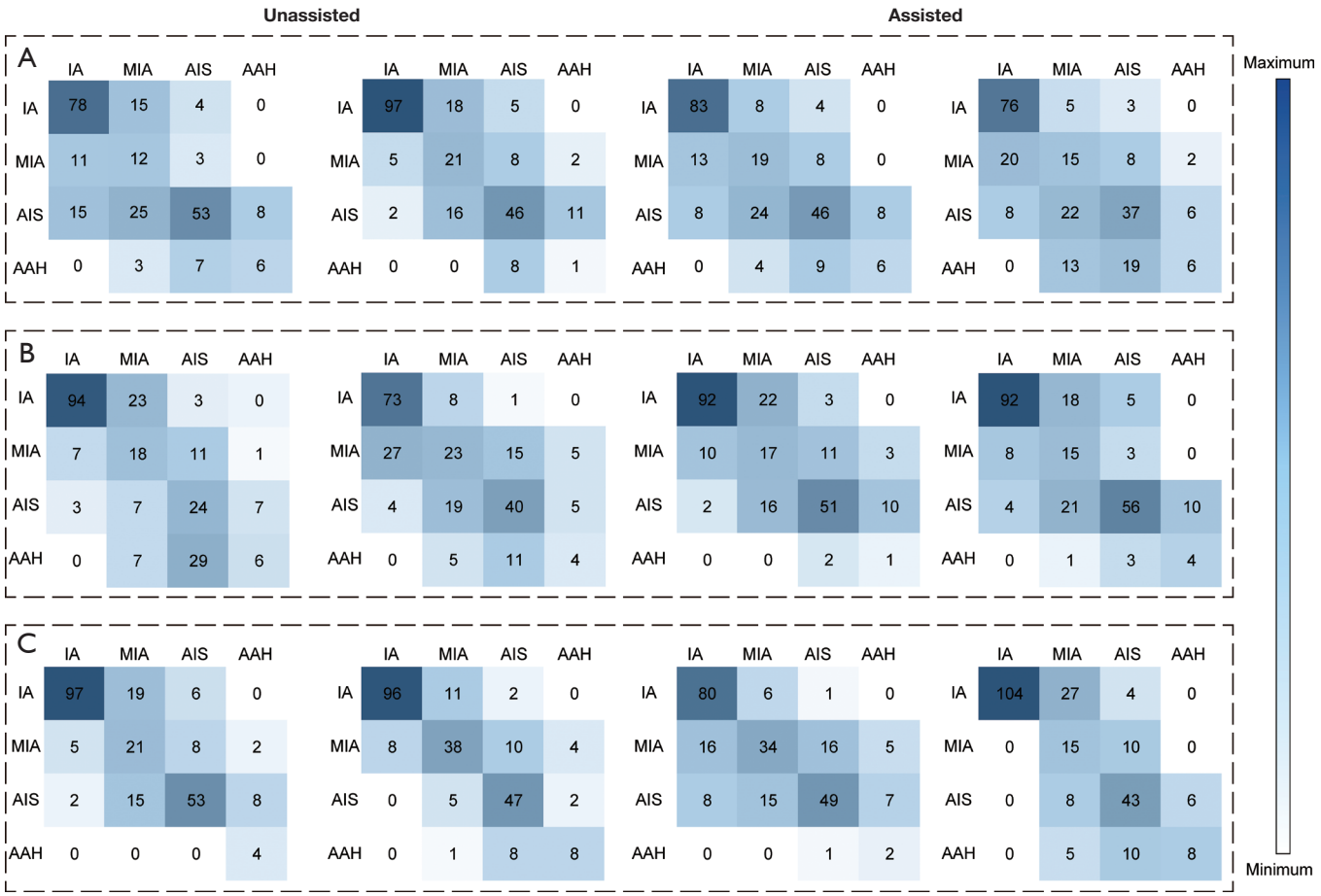

Figure S2 Confusion matrix demonstrating the correlation between prediction (row) and observed (column) labels of subsolid nodules by practicing doctors. (A) Junior rank, (B) middle rank, and (C) senior rank in 4-category classification. AAH, atypical adenomatous hyperplasia; AIS, adenocarcinoma *in situ*; IA, invasive adenocarcinoma; MIA, minimally invasive adenocarcinoma.

https://dx.doi.org/10.21037/tlcr-21-971