Respiratory Research

RESEARCH

Open Access

Automatic lung cancer subtyping using rapid on-site evaluation slides and serum biological markers

Junxiang Chen[1,2,3†], Chunxi Zhang[1,2,3†], Jun Xie[4], Xuebin Zheng[4], Pengchen Gu[4], Shuaiyang Liu[1,2,3], Yongzheng Zhou[1,2,3], Jie Wu[5], Ying Chen[6], Yanli Wang[6], Chuan He[4] and Jiayuan Sun[1,2,3*]

Abstract

Background Rapid on-site evaluation (ROSE) plays an important role during transbronchial sampling, providing an intraoperative cytopathologic evaluation. However, the shortage of cytopathologists limits its wide application. This study aims to develop a deep learning model to automatically analyze ROSE cytological images.

Methods The hierarchical multi-label lung cancer subtyping (HMLCS) model that combines whole slide images of ROSE slides and serum biological markers was proposed to discriminate between benign and malignant lesions and recognize different subtypes of lung cancer. A dataset of 811 ROSE slides and paired serum biological markers was retrospectively collected between July 2019 and November 2020, and randomly divided to train, validate, and test the HMLCS model. The area under the curve (AUC) and accuracy were calculated to assess the performance of the model, and Cohen's kappa (κ) was calculated to measure the agreement between the model and the annotation. The HMLCS model was also compared with professional staff.

Results The HMLCS model achieved AUC values of 0.9540 (95% confidence interval [CI]: 0.9257–0.9823) in malignant/benign classification, 0.9126 (95% CI: 0.8756–0.9365) in malignancy subtyping (non-small cell lung cancer [NSCLC], small cell lung cancer [SCLC], or other malignancies), and 0.9297 (95% CI: 0.9026–0.9603) in NSCLC subtyping (lung adenocarcinoma [LUAD], lung squamous cell carcinoma [LUSC], or NSCLC not otherwise specified [NSCLC-NOS]), respectively. In total, the model achieved an AUC of 0.8721 (95% CI: 0.7714–0.9258) and an accuracy of 0.7184 in the six-class classification task (benign, LUAD, LUSC, NSCLC-NOS, SCLC, or other malignancies). In addition, the model demonstrated a κ value of 0.6183 with the annotation, which was comparable to cytopathologists and superior to trained bronchoscopists and technicians.

Conclusion The HMLCS model showed promising performance in the multiclassification of lung lesions or intrathoracic lymphadenopathy, with potential application to provide real-time feedback regarding preliminary diagnoses of specimens during transbronchial sampling procedures.

Clinical trial number Not applicable.

†Junxiang Chen and Chunxi Zhang contributed equally to this work.

*Correspondence:
Jiayuan Sun
xkyyjysun@163.com

Full list of author information is available at the end of the article

Keywords Lung cancer, Subtyping, Rapid on-site evaluation, Serum biological markers, Deep learning

Introduction

Lung cancer is the leading cause of cancer death worldwide [1]. Accurate pathological subtyping is essential for guiding personalized therapy strategies, especially for small biopsies and cytology specimens from inoperable lung cancer patients [2]. Transbronchial sampling is the most common modality to obtain specimens of lung lesions and intrathoracic lymphadenopathy for the accurate diagnosis, and has proved to be safe and effective [3–6]. However, whether the specimens obtained were satisfactory and sufficient was indeterminate during the procedure. If the specimens obtained fail to meet the needs for an accurate diagnosis, patients may need to undergo additional invasive examinations. To overcome this problem, rapid on-site evaluation (ROSE) has emerged as a crucial auxiliary technique for transbronchial sampling [7–10]. With ROSE, the specimen undergoes an immediate cytopathologic evaluation, and the bronchoscopist receives feedback regarding specimen quality and preliminary diagnosis.

For ROSE, comprehensive evaluation of specimens is typically performed by experienced cytopathologists directly in the bronchoscopy suite. However, the widespread adoption of ROSE faces a significant challenge due to the shortage of cytopathologists, especially in areas with less medical resources. Meanwhile, the manual evaluation process is time-consuming and depends on the experience of cytopathologists [11]. With the development of artificial intelligence (AI), various methods have been proposed for automatic pathological image assessments [12–15]. Nevertheless, only a few studies have explored AI-aided ROSE for transbronchial sampling [16–19]. Besides, existing studies only focused on the differential diagnosis of benign and malignant, and none have investigated the classification of lung cancer subtypes. Moreover, it is essential to acknowledge that the diagnosis of lung cancer should not solely rely on image analysis. Recently, some researches have proved that AI models that combine medical imaging with serum biological markers can promote the accuracy of lung cancer subtyping [20, 21].

In this study, we proposed a hierarchical multi-label lung cancer subtyping (HMLCS) model that combines whole slide images (WSIs) of ROSE slides and serum biological markers, with the purpose of achieving performance comparable to cytopathologists in discriminating between benign and malignant lesions and recognizing different subtypes of lung cancer.

Materials and methods

Study design

This is a retrospective observational study. The overview of this study is illustrated in Fig. 1. Lung cancer is mainly divided into non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC). About 80–85% of lung cancer cases are NSCLC, and the subtypes of NSCLC include lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), and some other types. To achieve better performance in classifying the lung cancer subtypes as mentioned above, we proposed the HMLCS model which could be divided into three modules: benign or malignant classification (B/M) module, malignancy subtyping (M-sub) module, and NSCLC subtyping (NSCLC-sub) module. Specifically, WSIs of ROSE slides and paired serum biological markers were input to the HMLCS model. We first classified them as benign or malignant with the B/M module. The classification stopped here if the output of this step was benign. Otherwise, the input was forwarded to the M-sub module and was classified into three sub-categories: NSCLC, SCLC or other malignancies. Finally, if the output of the second step was NSCLC, we took one more step with the NSCLC-sub module to classify the input into three deeper sub-categories: LUAD, LUSC or NSCLC not otherwise specified (NSCLC-NOS). WSIs of ROSE slides and paired serum biological markers were retrospectively collected and randomly divided into training, validation and test sets to develop and test our model. In addition, we conducted a comprehensive comparison between our model and professional staff (including cytopathologists, trained bronchoscopists and technicians) to further evaluate the performance of our model. This study was approved by the Ethics Committee of Shanghai Chest Hospital (No. KS2023), and the requirement for informed consent was waived. The study was performed in accordance with the Declaration of Helsinki.

Data collection

The ROSE slides of patients with lung lesions or intrathoracic lymphadenopathy who underwent transbronchial sampling in Shanghai Chest Hospital between July 2019 and November 2020 were retrospectively collected. One of the following transbronchial sampling techniques was performed: transbronchial needle aspiration (TBNA), transbronchial lung biopsy (TBLB), and transbronchial biopsy (TBB). TBB was performed in endobronchial lung lesions, TBLB was performed in lung lesions invisible during bronchoscopy, and TBNA was performed in intrathoracic lymphadenopathy and central lung lesions adjacent to airways. During the procedure, the specimen was

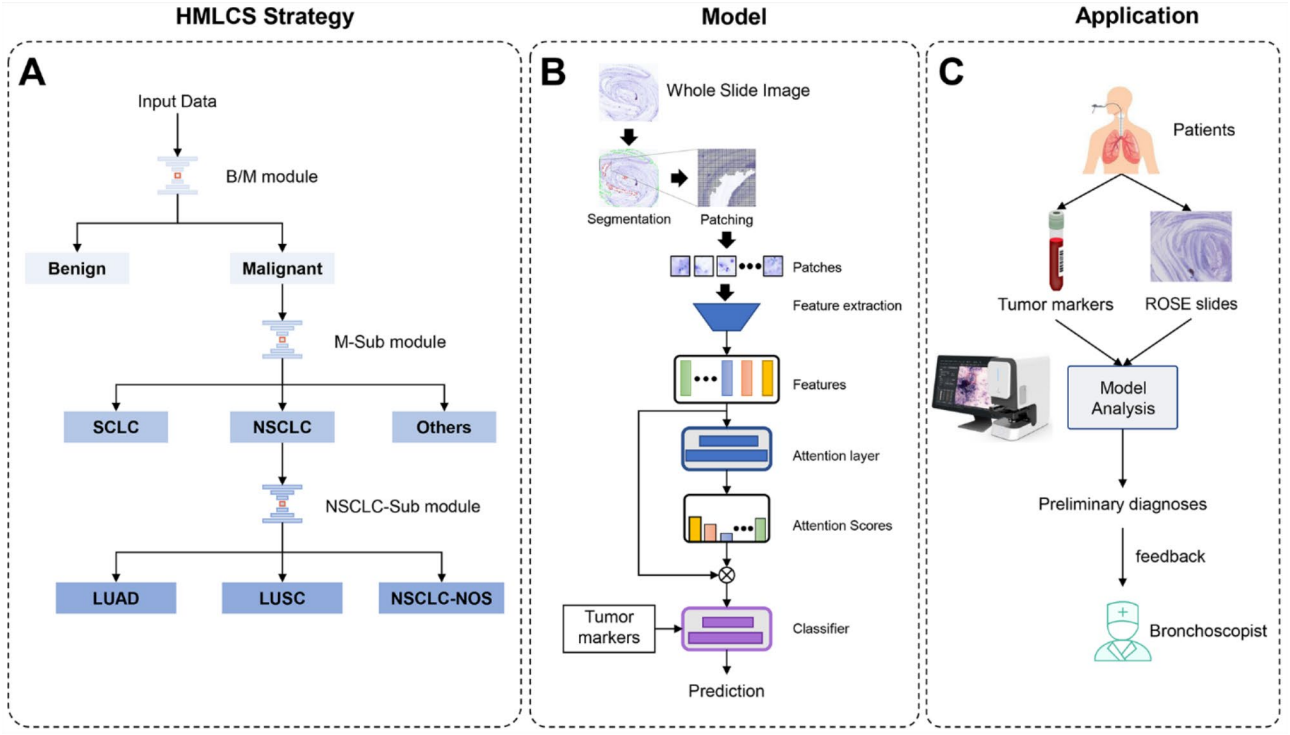

Fig. 1 Overview of this study. (**A**) The strategy of the HMLCS model. The HMLCS model can be divided into three modules: benign or malignant classification (B/M) module, malignancy subtyping (M-sub) module, and NSCLC subtyping (NSCLC-sub) module. (**B**) The details of the HMLCS model. The WSIs were segmented to patches, and then the patches were encoded into features by a pre-trained network. For each module, the attention layer was utilized to calculate attention scores for each patch feature, and the features of patches within a WSI were aggregated according to attention scores. Finally, the aggregated feature was concatenated with five serum biological markers for the classifier to perform the classification. (**C**) The potential application of the HMLCS model. The HMLCS model could provide real-time feedback regarding primary diagnoses of specimens during transbronchial sampling. LUAD: lung adenocarcinoma, LUSC: lung squamous cell carcinoma, NOS: not otherwise specified, NSCLC: non-small cell lung cancer, SCLC: small cell lung cancer, ROSE: rapid on-site evaluation, WSIs: whole slide images

stamped on glass slides for immediate ROSE. For specimens from forceps, the specimens were picked up using tweezers and spread in concentric circles about 1 cm in diameter on the glass slides. For specimens from brushes, the brush tip was pushed out, and the specimens were smeared on the glass slides, forming a rectangle of about 1 cm × 2 cm. For specimens from needles, the tissue was picked up using tweezers and spread in concentric circles, and one drop of the liquid specimen was placed on the glass slide and spread by pressing using another slide. The slides were air-dried and stained using a Diff-Quik stain kit (Baso Diagnostics Inc., Zhuhai, China). The slides were immersed in solution A for 20–30 s and rinsed with phosphate-buffered saline (PBS), followed by immersion in solution B for 20–30 s and rinsed with PBS. Owing to the convenience of Diff-Quik staining, it finds extensive application in ROSE and allows for rapid differentiation of benign and diverse malignancy subtypes (Fig. 2). The remaining specimens after ROSE slide preparation were sent for pathological examinations.

Only patients with a diagnostic transbronchial sampling procedure were included in this study. A biopsy that resulted in a malignant or specific benign (e.g., tuberculosis, fungal infection, etc.) process was considered diagnostic. Biopsy specimens with non-specific benign findings were considered diagnostic only if: (1) the diagnosis was confirmed by a subsequent surgery, mediastinoscopy, or CT-guided biopsy; or (2) follow-up imaging demonstrated stability or improvement of the lesion. All patients received follow up for at least 6 months. Patients with non-diagnostic transbronchial sampling procedures were excluded. All specimens in this study had undergone immunohistochemistry examinations. Specifically, cell block pathology of TBNA and histopathology of TBLB or TBB were confirmed by immunohistochemistry. Patients with unclear pathological subtypes were excluded in this study.

The ROSE slides were scanned using the Digital Micro Image Analysis System (Shanghai Aitrox Technology Corporation Limited, Shanghai, China) with a ×20 objective lens to produce WSIs. The WSIs were saved by the Microscope Image Information System (Shanghai Aitrox Technology Corporation Limited, Shanghai, China) in SVS format. WSIs with insufficient scan clarity were excluded. Since not all specimens were used for ROSE preparation and the specimens were only stamped on the

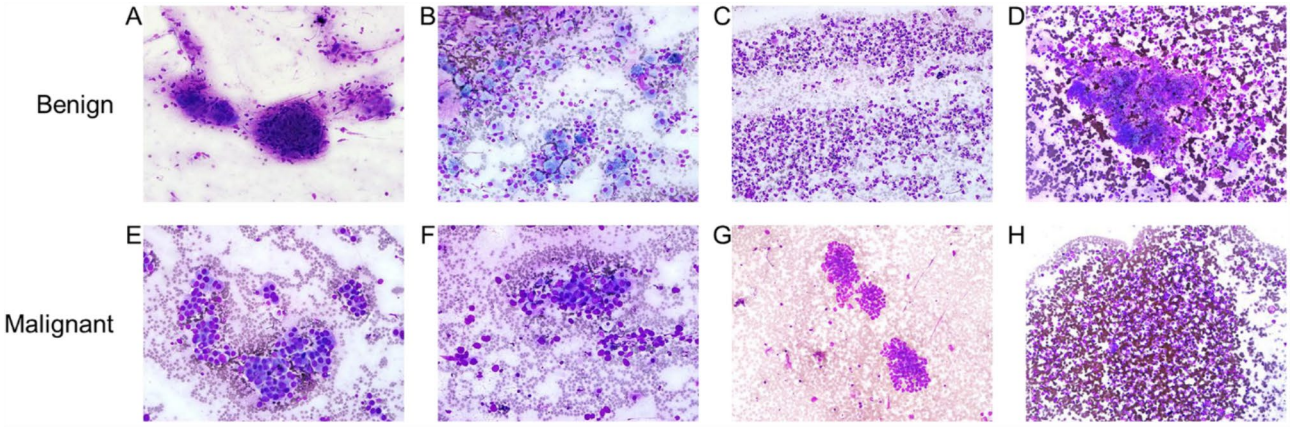

Fig. 2 Typical cytological images of different benign and malignant lesions from rapid on-site evaluation slides staining with Diff-Quik (×200). (**A**) Granulomatous structure with necrosis. (**B**) Pigment-laden macrophages. (**C**) Dense inflammatory background with neutrophil. (**D**) Anthracotic. (**E**) Lung adenocarcinoma. (**F**) Lung squamous cell carcinoma. (**G**) Small cell lung cancer. (**H**) Lymphoma

slides during ROSE preparation, the diagnosis of ROSE slide and the diagnosis of corresponding transbronchial sampling procedure could be inconsistent. For example, a lung lesion was diagnosed as adenocarcinoma by transbronchial sampling, but its ROSE slide could be negative. All WSIs were labeled by one experienced cytopathologist under the guidance of final diagnoses of transbronchial sampling procedures. To guarantee annotation accuracy, the annotation for one WSI was accepted only when it was consistent with the final diagnosis. Otherwise, the WSI was excluded. The corresponding serum biological markers were also collected, including carcinoembryonic antigen (CEA), soluble fragment of cytokeratin 19 (CYFRA21-1), squamous cell carcinoma antigen (SCC), neuron-specific enolase (NSE), and carbohydrate antigen 125 (CA125). Flowchart of data collection and splitting was shown in Fig. 3.

Data preprocessing

The WSIs were tiled into patches of 224×224 pixels, and all background patches were discarded. To efficiently discard all background patches, we converted patches into binary format, with white areas denoting valid cells and black areas denoting invalid background. Patches with white areas in their binary formats were retained, while the other patches were considered as background and discarded. Although each patch may be treated as an input, the feature dimension of 224×224 is computationally too expensive as there may exist more than 150, 000 patches for one WSI. To improve computing efficiency, we derived low-dimensional features of size 1024 for each patch for further steps.

Deep convolutional neural network

Like CLAM [12], a method that aims to do data-efficient weakly-supervised learning for WSIs by embedding instances using a frozen encoder, we used the ResNet50 pre-trained on the ImageNet to extract low-dimensional features of patches. Then, for each WSI, we concatenated the extracted features of patches within a WSI as an input for the subsequent steps. This feature extraction procedure preserved both a local representation with low-dimensional features of each patch and a global representation with ensemble features of the whole WSI.

Attention module

The attention module aggregated the features of patches within a WSI to create WSI-level representations. Specifically, the attention module assigned one attention score to each patch to indicate its contribution to the WSI-level representation. We show such calculation in Eq. 1:

$$a_i = \frac{exp(f_{reg}\left(tanh(feat_i) \odot sigmoid\left(feat_i\right)\right))}{\sum_{j=1}^{N} exp(f_{reg}(tanh\left(feat_j\right) \odot sigmoid\left(feat_j\right)))} \quad (1)$$

where a_i represents the attention score for patch i, $feat_i$ stands for the feature extracted by the pre-trained ResNet50 from patch i, and N is the number of valid patches within the WSI. We mapped the $feat_i$ by a $tanh$ function and a $sigmoid$ function, respectively, and the mapped outputs were aggregated via dot product. We then applied a regression function on the result and normalized it to [0, 1] by a softmax function. To make the equation easy to follow, we only show the calculation of one attention score here.

Classifier

Each module, namely the B/M module, the M-sub module, and the NSCLC-sub module, was equipped with a dedicated fully connected layer serving as a classifier.

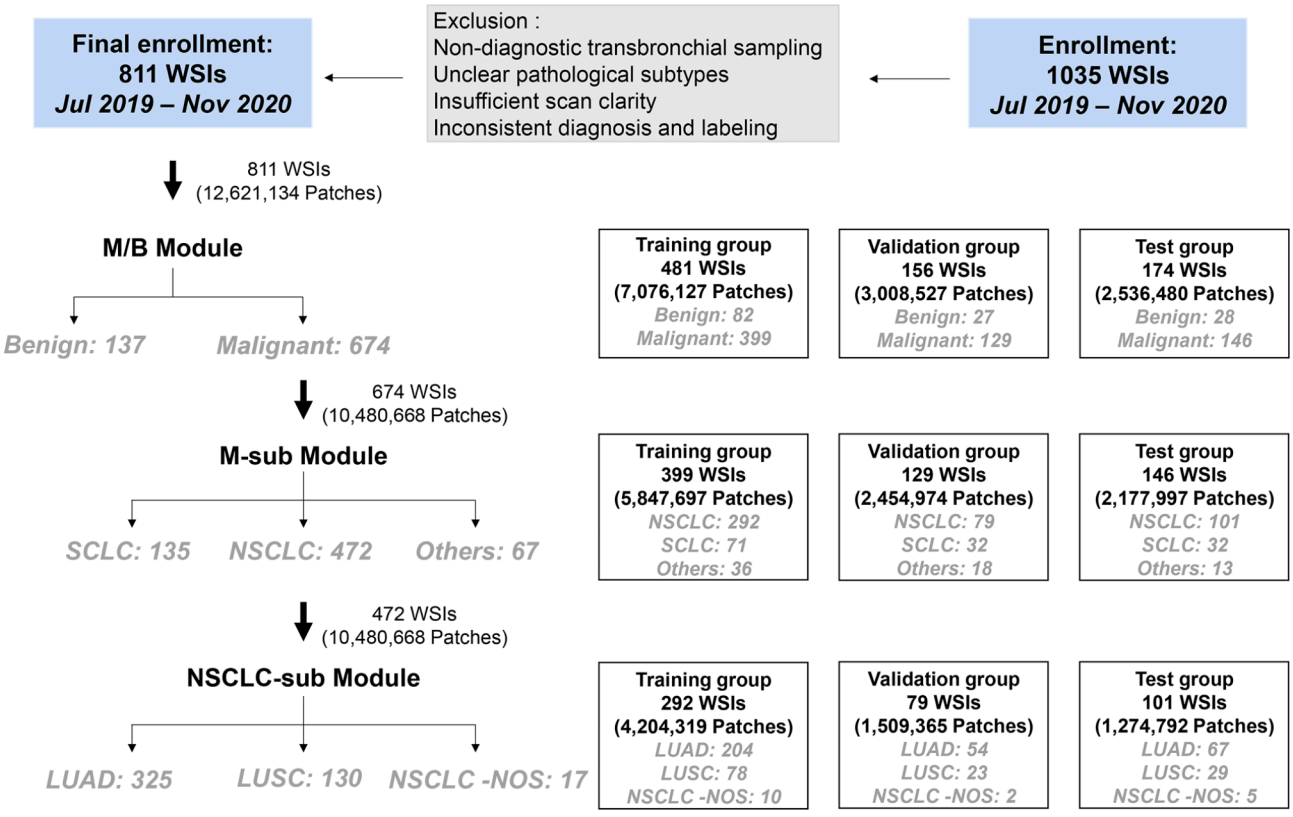

Fig. 3 Flowchart of data collection and splitting. LUAD: lung adenocarcinoma, LUSC: lung squamous cell carcinoma, NOS: not otherwise specified, NSCLC: non-small cell lung cancer, SCLC: small cell lung cancer, WSIs: whole slide images

To establish these classifiers, we initially aggregated the features of patches within a WSI by computing the weighted average based on the attention scores. It is noteworthy that, for each classifier, this feature fusion process was carried out independently to ensure that the attention score was learned in a focused manner specific to the respective classifier. In addition, we also concatenated five serum biological markers (CEA, CYFRA21-1, SCC, NSE, CA125) with the aggregated feature for the classifier to perform the classification. In summary, we employed a fully connected layer as the classifier for each module and trained the classifiers using the WSI-level labels.

Training protocol

The detailed training parameters of our model were as follows: a batch size of 4, a learning rate of $1\times e^{-5}$, an epoch of 100, an input size of 224×224×3, and a patch size of 8. Cross entropy loss was used as the loss function. The development of the model was implemented using Python version 3.6 and PyTorch version 1.3 based upon Ubuntu 16.04 on a workstation with two Xeon E5-2650 CPU (12C24T) at 2.2 GHz (Intel), 192GB RAM, and eight NVIDIA GTX 1080 Ti GPUs with 12GB GPU memory (NVidia).

Comparison between the HMLCS model and professional staff

The diagnostic performance of the HMLCS model was compared with professional staff, including two senior cytopathologists specializing in chest diseases, two bronchoscopists, and two technicians, on the same test set. The bronchoscopists and technicians were previously trained on ROSE regarding chest diseases. Professional staff independently made their predictions using the WSIs of ROSE slides and paired serum biological markers blinded to the final diagnoses.

Statistical analysis

The difference of clinical data from different datasets was compared by Mann-Whitney U test. The receiver operating characteristic (ROC) curve and the area under the curve (AUC) were introduced as graphical representations to showcase the balance between true positive and false positive rates. In addition, the accuracy and Cohen's kappa were calculated to assess the performance of the model and professional staff. We separately measured the agreement between professional staff and the annotation, as well as the agreement between the model and the annotation on the test set. The kappa value is denoted as κ, and the formula is as follows:

$$\kappa = \frac{P_A - P_e}{1 - P_e}$$

with

$$P_A = \frac{1}{N}\sum_{i=1}^{N}\frac{\left(\sum_{j=1}^{k} x_{ij}^2\right) - m}{m(m-1)}$$

And

$$P_e = \sum_{j=1}^{k}\left(\frac{\sum_{i=1}^{N} x_{ij}}{N \bullet m}\right)^2$$

, in which k is the number of categories, N is the number of datasets, and m is the number of experts.

The κ values were interpreted as follows: 0.00–0.20, poor agreement; 0.21–0.40, fair agreement; 0.41–0.60, moderate agreement; 0.61–0.80, substantial agreement; and 0.81–1.00, almost perfect agreement. The test statistics were approximated by a normal distribution to calculate the *p*-value and the 95% confidence interval (CI). $p<0.05$ was considered statistically significant. Statistical analysis was carried out with MedCalc (Version 22.009, MedCalc Software Ltd, Belgium) and IBM SPSS Statistics for Windows, Version 25.0 (IBM Corp., Armonk, NY, USA).

Results

Clinical characteristics

A total of 811 WSIs and corresponding serum biological markers were included in this study, including benign lesions ($n=137$), LUAD ($n=325$), LUSC ($n=130$), NSCLC-NOS ($n=17$), SCLC ($n=135$), and other malignancies ($n=67$). The detailed information is presented in Table 1. The results of biomarker values for different diseases are shown in Table S1 in the Supplementary Materials.

P values were calculated based on training and test groups. LUAD: lung adenocarcinoma, LUSC: lung squamous cell carcinoma, SCLC: small cell lung cancer, NSCLC: non-small cell lung cancer, NSCLC-NOS: non-small cell lung cancer not otherwise specified, TBB: transbronchial biopsy, TBLB: transbronchial lung biopsy, TBNA: transbronchial needle aspiration, IQR: interquartile range, CEA: carcinoembryonic antigen, CYFRA21-1: soluble fragment of cytokeratin 19, SCC: squamous cell carcinoma antigen, NSE: neuron-specific enolase, CA125: carbohydrate antigen 125.

Table 1 Clinical characteristics

Characteristics			Total ($n=811$)	Training ($n=481$)	Validation ($n=156$)	Test ($n=174$)	*P*
Final diagnoses, n(%)							0.773
Malignant/Benign	M-subtyping	NSCLC-subtyping					
Malignant	NSCLC	LUAD	325 (40.07%)	204 (42.41%)	54 (34.62%)	67 (38.51%)	
		LUSC	130 (16.03%)	78 (16.22%)	23 (14.74%)	29 (16.67%)	
		NSCLC-NOS	17 (2.10%)	10 (2.08%)	2 (1.28%)	5 (2.87%)	
	SCLC		135 (16.65%)	71 (14.76%)	32 (20.51%)	32 (18.39%)	
	Others		67 (8.26%)	36 (7.48%)	18 (11.54%)	13 (7.47%)	
Benign			137 (16.89%)	82 (17.05%)	27 (17.31%)	28 (16.09%)	
Transbronchial sampling techniques, n (%)							*0.501*
TBB			146 (18.00%)	94 (19.54%)	28 (17.95%)	24 (13.79%)	
TBNA			258 (31.81%)	138 (28.69%)	58 (37.18%)	62 (35.63%)	
TBLB			407 (50.19%)	249 (51.77%)	70 (44.87%)	88 (50.58%)	
Lesions, n (%)							*0.681*
Lung lesions			596 (73.49%)	366 (76.09%)	108 (69.23%)	122 (70.11%)	
Intrathoracic lymphadenopathy			215 (26.51%)	115 (23.91%)	48 (30.77%)	52 (29.89%)	
Serum biological markers, median (IQR)							
CEA			4.09 (2.46–8.55)	4.63 (2.69–9.10)	3.47 (2.01–6.14)	3.68 (2.40–8.49)	*0.086*
CYFRA21-1			2.73 (1.92–4.05)	2.84 (1.93–4.43)	2.52 (1.75–3.15)	2.79 (1.95–4.53)	*0.506*
SCC			0.79 (0.55–1.10)	0.79 (0.57–1.10)	0.71 (0.50–1.08)	0.80 (0.59–1.33)	*0.643*
NSE			18.95 (13.92–24.62)	19.32 (13.77–24.21)	18.22 (14.21–25.23)	18.66 (15.13–28.62)	*0.231*
CA125			16.08 (10.31–33.67)	16.36 (10.97–33.48)	16.08 (9.67–33.50)	14.43 (10.40-36.24)	*0.074*

Diagnostic performance of the HMLCS model

To evaluate the diagnostic performance of each classification module, the confusion matrix and AUC of each module on the test set were calculated as shown in Fig. 4. For the confusion matrix, each column represents the number of WSIs in each predicted class according to the model, while each row represents the actual number of WSIs in each class according to the annotation. The B/M module achieved an AUC of 0.9540 (95% CI: 0.9257–0.9823) in classifying malignant and benign cases (Fig. 4A). The M-sub module exhibited an AUC of 0.9126 (95% CI: 0.8756–0.9365) in distinguishing subtypes of malignancies (Fig. 4B). The NSCLC-sub module obtained an AUC of 0.9297 (95% CI: 0.9026–0.9603) in classifying subtypes of NSCLC (Fig. 4C). In total, the HMLCS model achieved an AUC of 0.8721 (95% CI: 0.7714–0.9258) and an accuracy of 0.7184 in the six-class classification (benign, LUAD, LUSC, NSCLC-NOS, SCLC, or other malignancies) (Fig. 4D). Additionally, the predictions of the HMLCS model demonstrated substantial agreement with the annotated labels, with a κ value of 0.6183 ($p<0.0001$).

We analyzed the performance of our model on lymph node and lung lesion data respectively. The HMLCS model achieved an AUC of 0.8347 (95% CI: 0.7882–0.8606) and an accuracy of 0.6923 in the six-class classification on lymph node data, and achieved an AUC of 0.8961 (95% CI: 0.8616–0.9317) and an accuracy of 0.7295 on lung lesion data. Although there were differences in imaging characteristics between lung and lymph node samples, such as the differences in background cell composition, the performance of our model on lung and lymph node data was similar.

To further analyze the significance of each component in our algorithm, we conducted comprehensive studies to assess the effect of individual elements. Initially, we trained hierarchical models using solely WSIs. The AUC for classifying malignant and benign was 0.9792 (95% CI: 0.9127–0.9838), the AUC for classifying SCLC, NSCLC, and other malignancies was 0.8907 (95% CI: 0.8628–0.9236), and the AUC for classifying LUAD, LUSC, and NSCLC-NOS was 0.8464 (95% CI: 0.8012–0.8829). Overall, the hierarchical model using solely WSIs achieved an AUC of 0.8067 (95% CI: 0.7489–0.8256) and an accuracy of 0.6494 in the six-class classification (Fig. S1 in the Supplementary Materials displays the ROC curve and the confusion matrix of the hierarchical model using solely WSIs). At most levels of classification, our method (using both WSIs and serum biological markers as the input) outperformed the model using solely WSIs, demonstrating the additional predictive power gained through the inclusion of serum biological markers. Additionally, to analyze the significance of our hierarchical classification approach, we attempted using a flat classification model instead of the hierarchical approach, where the six leaf classification labels were equally represented in the output. The flat model achieved an AUC of 0.7967 (95% CI: 0.7601–0.8333) and an accuracy of 0.4770 in the six-class classification (Fig. S2 in the Supplementary Materials displays the ROC curve and the confusion matrix of the flat classification model). Our method outperformed the flat

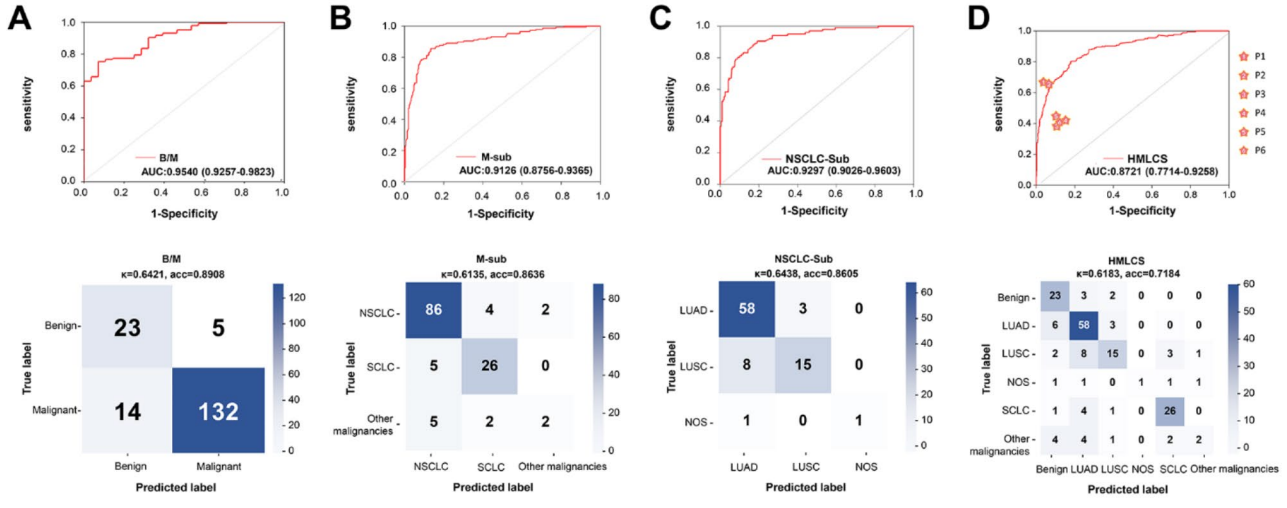

Fig. 4 The diagnostic performance of the HMLCS model on the test set. The diagnostic performance of each module and the total model is presented. The first row displays the ROC curve with the AUC value, while the second row shows the confusion matrix with the accuracy and κ value. (**A**) The diagnostic performance of the B/M module. (**B**) The diagnostic performance of the M-sub module. (**C**) The diagnostic performance of the NSCLC-sub module. (**D**) The diagnostic performance of the total HMLCS model. P1 to P6 indicate the performance of two cytopathologists, two bronchoscopists, and two technicians, respectively. acc: accuracy, AUC: aera under the curve, LUAD: lung adenocarcinoma, LUSC: lung squamous cell carcinoma, NOS: not otherwise specified, NSCLC: non-small cell lung cancer, ROC: receiver operating characteristic, SCLC: small cell lung cancer

model, underscoring the significance of our hierarchical classification approach.

Comparison between the HMLCS model and professional staff

To further evaluate the performance of the HMLCS model, we conducted a comprehensive comparison with professional staff, including two cytopathologists, two bronchoscopists and two technicians. Fig. 5 illustrates the Cohen's kappa and accuracy analysis for the cytopathologists (P1 and P2), bronchoscopists (P3 and P4) and technicians (P5 and P6) in relation to the annotated labels. In this analysis, we treated the problem as a simple six-class classification problem to calculate the κ value and the accuracy. Higher κ values of 0.7074 and 0.7422 were found in P1 and P2, demonstrating substantial agreement between cytopathologists and the annotated labels. Moderate agreement was found in P5 (κ=0.4483). Only weak agreement was shown in P3, P4, and P6, with κ values of 0.3443, 0.3514, and 0.3362, respectively. The accuracy analysis showed similar trends as depicted in Fig. 5C. In comparison, the κ value of our HMLCS model was 0.6183, which was comparable to cytopathologists and superior to trained bronchoscopists and technicians.

Discussion

In this study, a three-step model named HMLCS model was proposed to discriminate between benign and malignant lesions and recognize different subtypes of lung cancer. A large-scale dataset containing 811. ROSE slides and paired serum biological markers from patients with lung lesions and intrathoracic lymphadenopathy was constructed to train and evaluate the model. The HMLCS model demonstrated promising performance, achieving AUC values of 0.9540, 0.9126, and 0.9297 in benign/malignant classification, malignancy subtyping (NSCLC, SCLC, or other malignancies), and NSCLC subtyping (LUAD, LUSC, or NSCLC-NOS), respectively. In total, the HMLCS model achieved an AUC of 0.8721 and an accuracy of 0.7184 in the six-class classification task (benign, LUAD, LUSC, NSCLC-NOS, SCLC, or other malignancies). In addition, the κ value of the HMLCS model in relation to the annotated labels was 0.6183, which was comparable to cytopathologists and superior to trained bronchoscopists and technicians.

Previous studies have only focused on the malignant and benign discrimination on ROSE slides using deep learning [16–19]. Accuracies of 84.6-95.5% were obtained in these studies. The accuracy of our model in malignant and benign discrimination was 89.08%, which was similar to previous studies. To our knowledge, this is the first

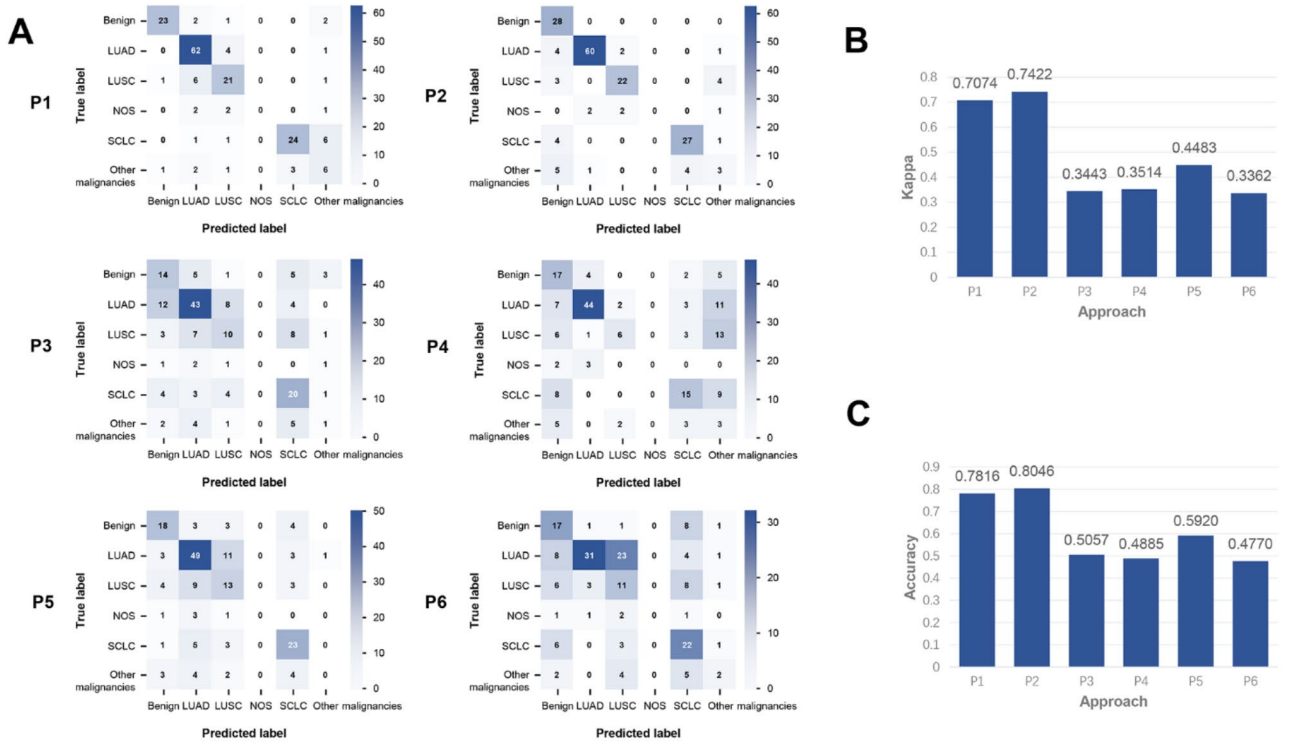

Fig. 5 The diagnostic performance of professional staff on the test set. (**A**) Confusion matrixes of professional staff. (**B**) κ values of professional staff. (**C**) Accuracies of professional staff. P1 and P2: two cytopathologists, P3 and P4: two bronchoscopists, P5 and P6: two technicians, LUAD: lung adenocarcinoma, LUSC: lung squamous cell carcinoma, NOS: not otherwise specified, SCLC: small cell lung cancer

study to develop a deep learning model that can identify lung cancer subtypes, and promising performance in lung cancer subtyping was obtained. A large-scale dataset containing both lung lesions and intrathoracic lymphadenopathy was collected to train and evaluate the model.

The proposed HMLCS model addressed the lung cancer subtyping problem by a novel hierarchical framework with three modular classification levels: (1) a B/M module that distinguishes benign and malignant lesions, (2) a M-sub module that subtypes malignancies into SCLC, NSCLC, and other malignancies, and (3) a NSCLC-sub module that further classifies NSCLC subtypes. To formalize hierarchical learning, the HMLCS model calculated independent attention scores at each level, ensuring the model allocated representational capacity based on each decision. According to our results, this cascading structure provided advantages over a single multi-class classifier. Dividing the problem into successive sub-modules allows focused learning on narrowly defined tasks, potentially capturing subtle differences obscured in a holistic approach.

The HMLCS model has the potential to provide real-time feedback regarding preliminary diagnoses of specimens during transbronchial sampling. If the lesion is likely to be malignant according to the patient's history and preoperative examinations but the specimen is predicted to be benign by the HMLCS model, the specimen obtained may be unqualified, and the biopsy site should be adjusted and new specimens should be obtained. Otherwise, if the lesion is likely to be malignant and the specimen is also predicted to be malignant by the HMLCS model, the specimens are satisfactory. AI-assisted ROSE evaluation could help compensate for the shortage of experienced cytopathologists in the field and address disparities in expertise among medical institutions. Furthermore, diagnostic results of the HMLCS model may be useful to facilitate an early clinical decision, especially for malignant results.

This study acknowledges certain limitations that warrant further investigation. First, this is a single-center retrospective study. To evaluate the generalizability of our model, it is necessary to carry out a multicenter and prospective research. Second, the dataset used in this study came from a specialized hospital, resulting in a high proportion of lung cancer, especially LUAD. However, this phenomenon was reasonable in lesions scheduled for an invasive examination. In a future study, larger dataset including sizeable cases of different diseases should be collected to optimize our model. Third, although the performance of our model was similar on lymph node and lung lesion data, the number of lymph node samples is not large enough (26.51% of the total dataset). The generalizability of our model on lymph node and lung lesion samples needs to be further validated. Finally, although the diagnostic performance of our model was promising, its clinical value needs to be assessed prospectively in real clinical scenarios.

Conclusion

In summary, by utilizing ROSE slides and serum biological markers as the input, our HMLCS model demonstrated cytopathologist-level diagnostic performance in benign and malignant discrimination and lung cancer subtype classification for lung lesions and intrathoracic lymphadenopathy. The HMLCS model could be a useful tool during transbronchial sampling procedures, providing real-time feedback regarding preliminary diagnoses of specimens to the bronchoscopist.

Abbreviations

ROSE	rapid on-site evaluation
HMLCS	hierarchical multi-label lung cancer subtyping
AUC	area under the curve
CI	confidence interval
NSCLC	non-small cell lung cancer
SCLC	small cell lung cancer
LUAD	lung adenocarcinoma
LUSC	lung squamous cell carcinoma
NOS	not otherwise specified
AI	artificial intelligence
WSI	whole slide image
TBNA	transbronchial needle aspiration
TBLB	transbronchial lung biopsy
TBB	transbronchial biopsy
CEA	carcinoembryonic antigen
CYFRA21-1	soluble fragment of cytokeratin 19
SCC	squamous cell carcinoma antigen
NSE	neuron-specific enolase
CA125	carbohydrate antigen 125
ROC	receiver operating characteristic

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12931-024-03021-8.

Supplementary Material 1

Acknowledgements

The authors are grateful to all participating patients. In addition, we would like to thank our colleagues in Shanghai Chest Hospital for their helpful contributions in data collection.

Author contributions

J. Chen and C. Zhang: data collection, methodology, investigation, and writing – original draft. J. Xie: methodology, algorithm support, and writing – review & editing. X. Zheng and P. Gu: methodology and algorithm support. S. Liu and Y. Zhou: investigation and data collection. J. Wu, Y. Chen, and Y. Wang: investigation and data collection. C. He: project administration and resources. J. Sun: methodology, project administration, resources, supervision, and writing – review & editing. All authors contributed to this work and have read and approved this work.

Funding

This study was supported by projects from National Multidisciplinary Treatment Project for Major Diseases (2020NMDTP), Science and Technology Commission of Shanghai Municipality (21XD1434400), SJTU Trans-med Awards Research (20210101), Joint Clinical Research Center of Institute

of Medical Robotics-Chest Hospital, Shanghai Jiao Tong University (IMR-XKH202102), Shanghai Municipal Health Commission (20214Y0417).

Data availability
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
This study was approved by the Ethics Committee of Shanghai Chest Hospital (No. KS2023), and the requirement for informed consent was waived.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Respiratory Endoscopy, Shanghai Chest Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China
[2]Department of Respiratory and Critical Care Medicine, Shanghai Chest Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China
[3]Shanghai Engineering Research Center of Respiratory Endoscopy, Shanghai, China
[4]Shanghai Aitrox Technology Corporation Limited, Shanghai, China
[5]Department of Pathology, Jiahui International Hospital, Shanghai, China
[6]Department of Pathology, Fudan University Shanghai Cancer Center, Shanghai, China

Received: 9 July 2024 / Accepted: 21 October 2024
Published online: 29 October 2024

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer statistics 2020: GLOBOCAN estimates of incidence and Mortality Worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71(3):209–49.
2. Travis WD, Brambilla E, Noguchi M, Nicholson AG, Geisinger KR, Yatabe Y, et al. International association for the study of lung cancer/american thoracic society/european respiratory society international multidisciplinary classification of lung adenocarcinoma. J Thorac Oncol. 2011;6(2):244–85.
3. Zhong CH, Su ZQ, Luo WZ, Rao WY, Feng JX, Tang CL, et al. Hierarchical clock-scale hand-drawn mapping as a simple method for bronchoscopic navigation in peripheral pulmonary nodule. Respir Res. 2022;23:245.
4. Nadig TR, Thomas N, Nietert PJ, Lozier J, Tanner NT, Wang Memoli JS, et al. Guided bronchoscopy for the evaluation of Pulmonary lesions: an updated Meta-analysis. Chest. 2023;163(6):1589–98.
5. Kuijvenhoven JC, Leoncini F, Crombag LC, Spijker R, Bonta PI, Korevaar DA, et al. Endobronchial Ultrasound for the diagnosis of centrally located lung tumors: a systematic review and Meta-analysis. Respiration. 2020;99(5):441–50.
6. Varela-Lema L, Fernández-Villar A, Ruano-Ravina A. Effectiveness and safety of endobronchial ultrasound-transbronchial needle aspiration: a systematic review. Eur Respir J. 2009;33(5):1156–64.
7. Mondoni M, Carlucci P, Di Marco F, Rossi S, Santus P, D'Adda A, et al. Rapid on-site evaluation improves needle aspiration sensitivity in the diagnosis of central lung cancers: a randomized trial. Respiration. 2013;86(1):52–8.
8. Mondoni M, Sotgiu G, Bonifazi M, Dore S, Parazzini EM, Carlucci P, et al. Transbronchial needle aspiration in peripheral pulmonary lesions: a systematic review and meta-analysis. Eur Respir J. 2016;48(1):196–204.
9. Chen CH, Cheng WC, Wu BR, Chen CY, Chen WC, Hsia TC, et al. Improved diagnostic yield of bronchoscopy in peripheral pulmonary lesions: combination of radial probe endobronchial ultrasound and rapid on-site evaluation. J Thorac Dis. 2015;7(Suppl 4):S418–425.
10. Sehgal IS, Dhooria S, Aggarwal AN, Agarwal R. Impact of Rapid On-Site cytological evaluation (ROSE) on the Diagnostic yield of Transbronchial Needle Aspiration during Mediastinal Lymph Node Sampling: systematic review and Meta-analysis. Chest. 2018;153(4):929–38.
11. Lu L, Xu H. An update on the classification of lung and pleural tumors. J Clin Transl Pathol. 2023;3(2):106–13.
12. Lu MY, Williamson DFK, Chen TY, Chen RJ, Barbieri M, Mahmood F. Data-efficient and weakly supervised computational pathology on whole-slide images. Nat Biomed Eng. 2021;5(6):555–70.
13. Chen CL, Chen CC, Yu WH, Chen SH, Chang YC, Hsu TI, et al. An annotation-free whole-slide training approach to pathological classification of lung cancer types using deep learning. Nat Commun. 2021;12(1):1193.
14. Teramoto A, Kiriyama Y, Tsukamoto T, Sakurai E, Michiba A, Imaizumi K, et al. Weakly supervised learning for classification of lung cytological images using attention-based multiple instance learning. Sci Rep. 2021;11(1):20317.
15. Zhang S, Zhou Y, Tang D, Ni M, Zheng J, Xu G, et al. A deep learning-based segmentation system for rapid onsite cytologic pathology evaluation of pancreatic masses: a retrospective, multicenter, diagnostic study. EBioMedicine. 2022;80:104022.
16. Lin CK, Chang J, Huang CC, Wen YF, Ho CC, Cheng YC. Effectiveness of convolutional neural networks in the interpretation of pulmonary cytologic images in endobronchial ultrasound procedures. Cancer Med. 2021;10(24):9047–57.
17. Wang CW, Khalil MA, Lin YJ, Lee YC, Huang TW, Chao TK. Deep learning using endobronchial-ultrasound-guided transbronchial needle aspiration image to improve the overall diagnostic yield of sampling Mediastinal Lymphadenopathy. Diagnostics (Basel). 2022;12(9):2234.
18. Ai D, Hu Q, Chao Y-C, Fu C-C, Yuan W, Lv L, et al. Artificial intelligence-based rapid on-site cytopathological evaluation for bronchoscopy examinations. Intelligence-Based Med. 2022;6:100069.
19. Yan S, Li Y, Pan L, Jiang H, Gong L, Jin F. The application of artificial intelligence for Rapid On-Site evaluation during flexible bronchoscopy. Front Oncol. 2024;14:1360831.
20. Ren C, Zhang J, Qi M, Zhang J, Zhang Y, Song S, et al. Machine learning based on clinico-biological features integrated 18F-FDG PET/CT radiomics for distinguishing squamous cell carcinoma from adenocarcinoma of lung. Eur J Nucl Med Mol Imaging. 2021;48(5):1538–49.
21. Wang L, Zhang M, Pan X, Zhao M, Huang L, Hu X, et al. Integrative serum metabolic fingerprints based multi-modal platforms for Lung Adenocarcinoma Early Detection and Pulmonary Nodule classification. Adv Sci (Weinh). 2022;9(34):e2203786.

Publisher's note