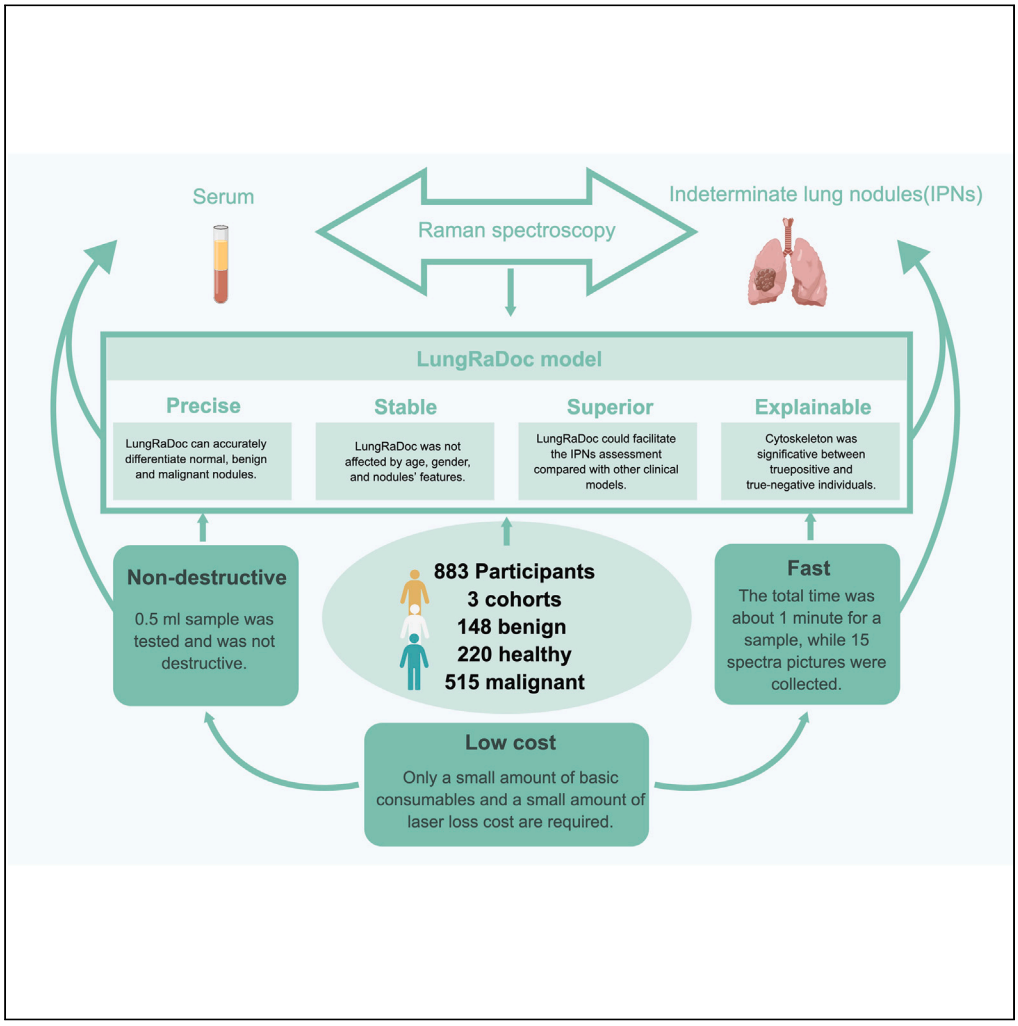
iScience

CellPress
OPEN ACCESS

Article

Serum laser Raman spectroscopy as a potential diagnostic tool to discriminate the benignancy or malignancy of pulmonary nodules



Huaichao Luo, Ruiling Zu, Lintao Li, ..., Gang Yin, Dezhong Yao, Dongsheng Wang

luo1987cc@163.com (H.L.)
wangdongsheng@scszlyy.org.cn (D.W.)

Highlights

Raman spectra differences were used to construct a classifier named LungRaDoc

LungRaDoc was not affected by age, gender, and nodules' features

LungRaDoc could facilitate the IPNs assessment compared with other clinical models

Cytoskeleton was significative between true-positive and true-negative individuals

Luo et al., iScience 26, 106693
May 19, 2023
https://doi.org/10.1016/j.isci.2023.106693

iScience

CellPress
OPEN ACCESS

Article

Serum laser Raman spectroscopy as a potential diagnostic tool to discriminate the benignancy or malignancy of pulmonary nodules

Huaichao Luo,[1,4,5,*] Ruiling Zu,[1,4] Lintao Li,[2,4] Yao Deng,[1,4] Shuya He,[1] Xing Yin,[1] Kaijiong Zhang,[1] Qiao He,[1] Yu Yin,[3] Gang Yin,[2] Dezhong Yao,[3] and Dongsheng Wang[1,*]

SUMMARY

It has been proved that Raman spectral intensities could be used to diagnose lung cancer patients. However, the application of Raman spectroscopy in identifying the patients with pulmonary nodules was barely studied. In this study, we revealed that Raman spectra of serum samples from healthy participants and patients with benign and malignant pulmonary nodules were significantly different. A support vector machine (SVM) model was developed for the classification of Raman spectra with wave points, according to ANOVA test results. It got a good performance with a median area under the curve (AUC) of 0.89, when the SVM model was applied in discriminating benign from malignant individuals. Compared with three common clinical models, the SVM model showed a better discriminative ability and added more net benefits to participants, which were also excellent in the small-size nodules. Thus, the Raman spectroscopy could be a less-invasive and low-costly liquid biopsy.

INTRODUCTION

As lung cancer has become a common health issue in China, the government was devoted to the implementation of lung cancer screening.[1] With the increasing screening by low dose computed tomography (LDCT), a million of patients have been identified indeterminate pulmonary nodules (IPNs), which mostly were false-positive nodules for malignancy.[2,3] At present, those IPNs are recommended to be on serial surveillance by LDCT according to expert guidelines, which may make patients suffer from physical, psychological, and financial harms.[4,5] It is a challenge to discriminate whether the pulmonary nodule (PN) is malignant or benign at the first time identified. Accordingly, methods ameliorating LDCT diagnosis in a short time could protect patients from those harms as soon as possible.

Recently, machine learning has become a method to improve accuracy of LDCT in estimating malignancy risk of IPNs. As reported, a deep learning algorithm with convolutional neural networks (CNNs) only based on the LDCT information reached a high accuracy.[6] Moreover, some algorithms integrating the LDCT information and clinical information also performed excellently, such as Mayo Clinic (MC), Brock University (BU), and Veterans Affairs (VA) model, with the area under the curve (AUC) over 0.8 in original research.[7–9] However, those algorithms did not perform stably in the other validation research, which might be due to the populations not matching with those in which they were developed.[7] Furthermore, to improve the stability in other independent population, a research integrated a biomarker in blood with MC model and significantly improved diagnostic accuracy of MC model, also validated in four independent cohorts.[10] Likewise, in our previous research, a diagnostic model integrating the LDCT information with platelet features outperformed MC, BU, and VA model in both internal and external cohorts.[11] It was obvious that the blood biomarkers could be noninvasive diagnosis of patients with IPNs, potentially reducing the unstable results caused by model-developed population. The blood biomarkers have always been characterized in liquid biopsies, which were less invasive and could be done serially.

Current liquid biopsies in cancer diagnostics often concerned the detection of biomarkers, which are nucleic acids, proteins, or other organic compounds. Surface-enhanced Raman spectroscopy (SERS), a kind of Raman spectroscopy (RS), could detect biomarkers like nucleic acid and proteins in serum by the substrate enhancing the produced Raman signal, which was also described as a liquid biopsy.[12,13] Not only focused

[1]Department of Clinical Laboratory, Sichuan Clinical Research Center for Cancer, Sichuan Cancer Hospital & Institute, Sichuan Cancer Center, Affiliated Cancer Hospital of University of Electronic Science and Technology of China, Chengdu, China

[2]Department of Radiation Oncology, Sichuan Clinical Research Center for Cancer, Sichuan Cancer Hospital & Institute, Sichuan Cancer Center, Affiliated Cancer Hospital of University of Electronic Science and Technology of China, Chengdu, China

[3]Sichuan Institute for Brain Science and Brain-Inspired Intelligence, MOE Key Lab for Neuroinformation, University of Electronic Science and Technology of China, Chengdu, China

[4]These authors contributed equally

[5]Lead contact

*Correspondence: luo1987cc@163.com (H.L.), wangdongsheng@scszlyy.org.cn (D.W.)

https://doi.org/10.1016/j.isci.2023.106693

on the detection of DNA, RNA, and microRNA (miRNA), a serum-based SERS test extracted unique peaks of cancerous exosomes using principal-component analysis (PCA) allowing for diagnosing non-small cell lung cancer.[13,14] Although the SERS analysis combined with machine learning methods for identifying lung cancer has demonstrated extreme strength and utility of the method, it still had potential drawbacks that the Raman signals acquired from the SERS nanoparticles were detected from known and predetermined depths.[15] That means, SERS was not suitable for the screening of IPNs either. Whereas spontaneous RS collecting multiple spectra from a heterogeneous sample could characterize the multi-component composition while no special substrate was in need, of which the lower cost and an easier manipulation might fit the IPNs screening.[12,16–18]

A Raman system based on spontaneous RS reported in a COVID-19 study could collect spectra with approximately 15 scans per serum sample within a 3 s accumulation for each scan.[19] Gathering spectra data from different population and analyzing by machine learning could reach the classification purpose, which has been already proved in the previous study.[19] Reasonably, it was assumed that the Raman system combining with appropriate machine learning could also be used in separating the healthiness from malignancy. And most RS research concentrated in the lung cancer diagnosis or staging; few focused on discriminating the IPNs. So, we screened the serum from patients with benign and malignant PNs and healthy individuals by RS. And machine learning was generated to access the commonalities of fingerprints in the same groups and the differences among the different groups. Then a rapid, less-invasive, easy-to-manipulate, and low-cost classification model was constructed based on the fingerprints, which could bring benefits at different levels to the patients with IPNs.

RESULTS

Clinical characteristics of participants

A total of 883 participants were enrolled in this research, including 148 benign, 220 healthy, and 515 malignant participants (Figure 6). This research was performed in two centers; one was in Sichuan Cancer Hospital (SCH), and another was Sichuan Provincial People's Hospital (SPH). The participants from Sichuan Cancer hospital were divided into SCH batch 1 and SCH batch 2 according to the samples' saving time. There were 94 benign, 105 healthy, and 317 malignant participants in SCH batch 1, whose samples were stored for more than 1 year. And other 19 benign, 67 healthy, and 120 malignant participants were included in SCH batch 2 with saving time less than 1 year. The components of each group were presented in Table 1. In both discover and validation group, the gender and age were not significantly different among benign, healthy, and malignant subsets. In the discover group, the patients diagnosed as lung adenocarcinoma (LACC) predominantly consisted of the malignant subset (80.4%). And the nodules were mostly small size, distributing from 0 to 3 cm in both benign (67.5%) and malignant subsets (57.6%). The common LDCT symptom of malignant patients was ground-glass nodule (GGN) (21.6%), which was also a significant symptom compared with benign subset ($p = 0.035$).

Raman spectra comparison

Figure 1 represented the Raman spectra of three subsets in the discover group. The average of each subsets ' preprocessed spectra was displayed in Figure 1A. Among the three subsets, healthy subset represented an obvious peak intensity at 1437.6 cm^{-1}, while malignant subset showed a least peak at the same shift. There were still other different peaks that were not distinguished sufficiently by human eyes, for which heatmaps were used to reflect the subtle differences among different subsets (Figure 1B). Overall, the heatmaps illustrated the RS differences in the three subsets according to ANOVA test ($p < 0.05$), while the peaks appeared at range of 500–1800 cm^{-1}. As it is shown, more differences were observed between malignant and healthy subset, where the peaks concentrated around at 821.9 cm^{-1}, 1041.5 cm^{-1}, and 1246.7 cm^{-1}. Moreover, the heatmap constructs further confirmed the marked different intensities at 1437.6 cm^{-1} in malignant patients as compared to benign and healthy individuals. The results also showed other spectra displayed difference of malignant subset as compared to benign subset.

Model construction

The support vector machine (SVM) models were developed for the classification of Raman spectra with wave points, which were significantly different in ANOVA test results. In order to reduce selection bias, all benign and malignant participants from discover group were split into 10 disjunctive parts, each time combining a benign and malignant part to a combination with roughly the same proportion of benign

Table 1. The clinical characteristics of participants

	Discover group				Validation group			
	benign (N = 108)	health (N = 180)	malignant (N = 475)	*P*	benign (N = 40)	health (N = 40)	malignant (N = 40)	*P*
Center:								
SCH_batch1	69 (63.9%)	86 (47.8%)	294 (61.9%)		25 (62.5%)	19 (47.5%)	25 (62.5%)	
SCH_batch2	14 (13.0%)	55 (30.6%)	111 (23.4%)		5 (12.5%)	12 (30.0%)	9 (22.5%)	
SPH	25 (23.1%)	39 (21.7%)	70 (14.7%)		10 (25.0%)	9 (22.5%)	6 (15.0%)	
Gender (male)	62 (57.4%)	94 (52.2%)	233 (49.1%)	0.13	28 (70.0%)	23 (57.5%)	24 (60.0%)	0.47
Pathology:				–				–
LACC	–	–	337 (80.4%)		–	–	25 (62.5%)	
LSCC	–	–	16 (3.82%)		–	–	2 (5.00%)	
NSCLC	–	–	18 (4.30%)		–	–	3 (7.50%)	
SCC	–	–	48 (11.5%)		–	–	7 (17.5%)	
Benign	108 (100%)	–	–		40 (100%)	–	–	
Up (Y)	58 (53.7%)	–	278 (58.8%)	0.39	22 (55.0%)	–	23 (57.5%)	1.0
GGN (Y)	13 (12.0%)	–	102 (21.6%)	**0.035**	8 (20.0%)	–	14 (35.0%)	0.21
Spiculation (Y)	19 (17.6%)	–	115 (24.3%)	0.17	8 (20.0%)	–	6 (15.0%)	0.77
age	57.5 (12.2)	54.8 (9.63)	59.3 (10.0)	0.14	58.0 (14.3)	55.2 (7.19)	60.1 (8.71)	0.43
Stage:								
I	–	–	178 (37.6%)		1 (2.50%)	–	11 (27.5%)	
I	–	–	49 (10.4%)		–	–	4 (10.0%)	
III	–	–	88 (18.6%)		–	–	9 (22.5%)	
IV	–	–	85 (18.0%)		–	–	10 (25.0%)	
smoking:	39 (36.1%)	–	167 (35.3%)	0.96	15 (37.5%)	–	16 (40.0%)	
Size(cm):		–		0.25		–		**<0.001**
0-1	16 (18.0%)	–	76 (18.9%)		10 (35.7%)	–	7 (19.4%)	
1-2	24 (27.0%)	–	91 (22.6%)		12 (42.9%)	–	5 (13.9%)	
2-3	20 (22.5%)	–	65 (16.1%)		1 (3.57%)	–	6 (16.7%)	
≥3	29 (32.5%)	–	171 (42.4%)		5 (17.9%)	–	18 (50.0%)	

Abbreviation: LACC, lung adenocarcinoma; LSCC, lung small cell cancer; NSCLC, non-small cell lung cancer; SCC, squamous cell carcinoma; up, the nodule located in upper lobes; GGN, the nodule was a ground-glass nodule; Spiculation, margin of the nodule was spiculation; smoking, the participants has smoked; size, the diameter of the largest nodule.

and malignant instances as in the original datasets. For each combination, the receiver operating characteristic (ROC) curves were used to evaluate the performance of the SVM model, which repeated 5 times in every combination. The process was also applied in healthy-benign combination and healthy-malignant combination (Figure 2A) (See also Table S1). When the SVM model was used to classify the healthy and benign individuals, the AUCs were in the range from 0.81 to 0.9 in the 10 combinations, and the median AUC was 0.86 in the validation group (Figure 2B). The AUCs were in the range from 0.83 to 1 in the classification of healthy and malignant individuals, and the median AUC was 0.87 in the validation group (Figure 2B). Delightedly, the AUCs were in the range from 0.87 to 0.95 in the discover group and the median AUC was 0.89 in the validation group, when the SVM model was applied to discriminate benign from malignant individuals (Figure 2B).

Model estimation

The SVM model named LungRaDoc could provide a value of each sample, which was used to access the diagnostic value of LungRaDoc. As performed in the reserved validation group, the predicted values showed gradually rising levels in gradually advanced stage (Figure 3A, $p < 0.001$). Figure 3A also showed the predicted values with significant increases in the patients with LACC compared to the benignancy ($p < 0.001$). As ground

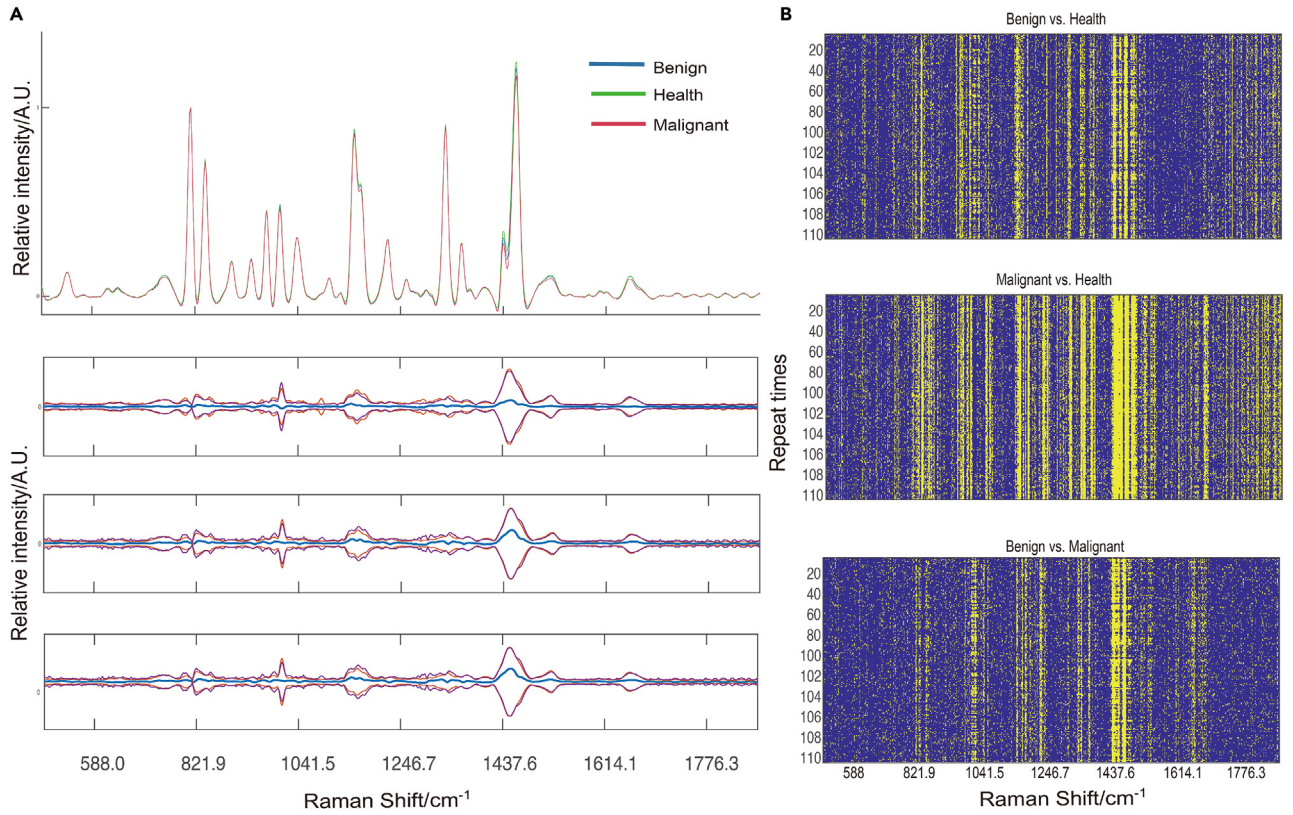


Figure 1. The average Raman spectra of the three groups and the difference among three groups

(A) The total average of spectra in three subsets; the color band represents the standard deviation. The difference of spectra signal between two groups (blue) and the spectra signal of two groups between ±2 standard deviations (red and purple).

(B) The results of the ANOVA test. The spectra range with significant differences in the ANOVA test ($p < 0.05$) was indicated in yellow, while no significance was indicated in blue.

glass was commonly discovered in a malignant nodule, the validation group was divided into ground-glass subsets and non-ground-glass subsets to evaluate whether the performance of LungRaDoc was limited by ground glass. As expected, in both ground-glass subsets and non-ground-glass subsets, the values were significantly increased in malignant patients ($p < 0.001$) (Figure 3A). The effect of nodule size was next sought to determine. The malignant patients developed higher-grade values than the benignancy no matter in small-size subsets, whose nodule size was smaller than 3 cm ($p < 0.001$), and lager-size subsets, whose nodule size was larger than 3 cm ($p < 0.001$) (Figure 3A). Using logistic regression analysis, we examined which, if any, of the predicted values were infected by age, gender, characteristics of nodules, and saving time of samples. It could be reflected in Figure 3B that there was an observed correlation of LungRaDoc with the saving time. Following that, we looked at the effect of saving time on the malignant prediction using logit regression adjustment (See also Table S2). As shown in Figure 3C, only LungRaDoc and nodule size performed for reliability of malignant prediction; the saving time did not affect the prediction ($p > 0.05$).

Model comparison

The LungRaDoc showed a better discriminative ability than other common clinical models with the AUC of 0.89 (95% confidence interval [CI] 0.82–0.96) (Figure 4A). The ROC curves of LungRaDoc and other three models across validation group were shown in Figure 4A. As displayed, the MC model, VA model, and BU model did not achieve good classification performance as LungRaDoc, with AUC values of 0.46 (95% CI 0.33–0.59) for MC model, 0.61 (95% CI 0.48–0.73) for VA model, and 0.74 (95% CI 0.62–0.85) for BU model. The decision curve analysis (DCA) indicated that the threshold probabilities of LungRaDoc were 1–80%, while the thresholds of other three models were at smaller range. LungRaDoc could better predict malignancy as it added more net benefits (NBs) compared with the other three clinical models for almost all threshold values (Figure 4B). The improvement in reclassification was indicated by net reclassification improvement (NRI) and integrated

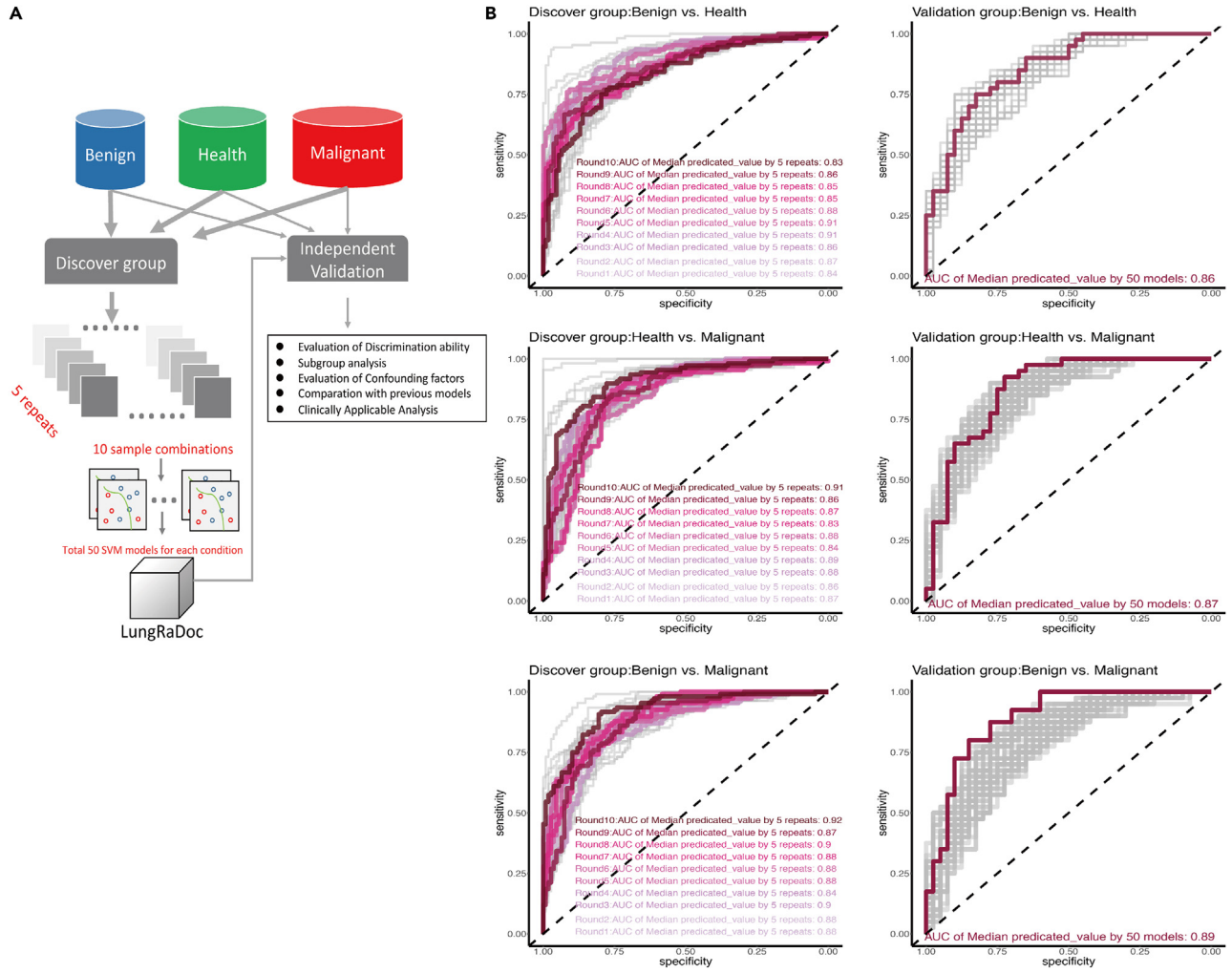


Figure 2. LungRaDoc construction

(A)The steps of LungRaDoc construction and validation.

(B) The ROC curves of LungRaDoc used in both discover and validation group to discriminate the healthy from benign individuals and malignant individuals and to discriminate the benign from malignant individuals.

discrimination improvement (IDI). Compared with MC, BU, and VA model, the NRI was 21% (95% CI = 12%–29%), 21% (95% CI = 12%–29%), and 19% (95% CI = 10%–29%), respectively, and the IDI was 13% (95% CI = 9.1%–16%), 12% (95% CI = 9%–16%), and 14% (95% CI = 10%–18%) (Figure 4C), respectively. These results were performed in the validation group, indicating that LungRaDoc discriminated malignancy from benignancy with greater accuracy than the other three clinical models.

Model application

In the LungRaDoc, a predictive value cutoff of 0.0773 had the highest AUC (0.89, 95% CI 0.82–0.96) in distinguishing between benign and malignant patients. To some extent, the differences between cutoff value and actual value could reflect the diagnostic accuracy. As shown in Figure 5A, the false positive and negative individuals were all called wrong discrimination displayed in yellow. In 27 patients whose nodule size was less than or equal to 1 cm, only 2 were false negatives and 3 were false positives. There were 4 false negatives and 6 false positives in 49 patients whose nodule size were less than or equal to 2 cm. In the validation group, 40 benign and 40 malignant patients were further selected to validate diagnostic performance. At the pathology's specificity, LungRaDoc achieved a specificity of 85%, whereas, at pathology's sensitivity, LungRaDoc achieved a sensitivity of 80% (Figure 5B). With the high positive predicted value of 0.93, LungRaDoc had a potential application in IPNs screening. Correlation analysis was used to estimate

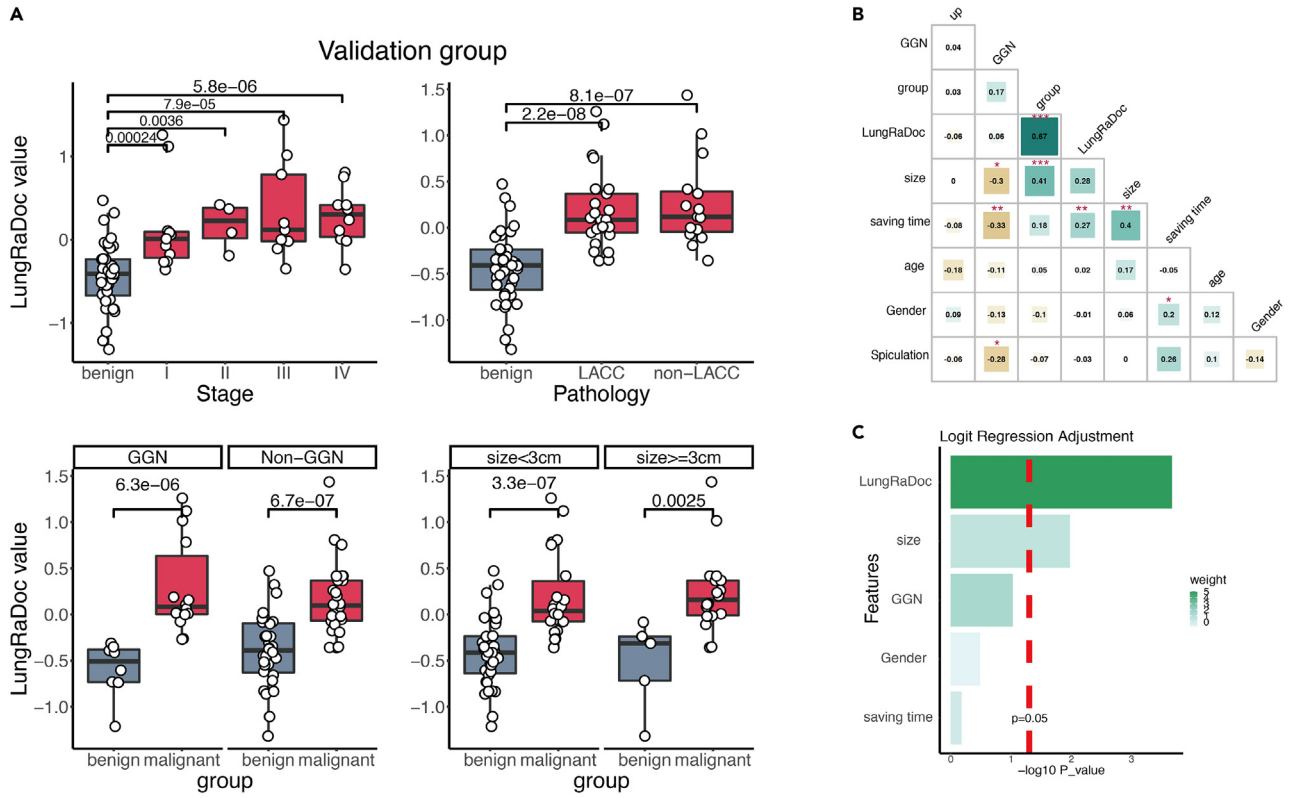


Figure 3. Estimation of LungRaDoc

(A) The predicted values calculated by LungRaDoc in validation group stratified by the stage and pathology of lung cancer; the predicted values between benignancy with malignancy in ground-glass subsets and non-ground-glass subsets; the predicted values between benignancy with malignancy in smaller-size subsets and lager-size subsets. Boxplots are defined as median (center), with the bounds of the box representing the interquartile range (upper and lower bounds), the whiskers representing upper and lower extremes, and points indicating individual patients. Statistical analysis was conducted using Wilcoxon rank-sum test (to test for specific intergroup differences).

(B) Heatmap showing correlation of LungRaDoc with age, gender, characteristics of nodules, and saving time of samples. The characteristics and P-value were displayed in the cross square.

(C) The weight of LungRaDoc, gender, characteristics of nodules, and saving time of samples in the malignant prediction, displayed as a bar graph on the right-hand side. A significance level of <0.05 was adopted.

the factors influencing LungRaDoc, as depicted in Figure 5C, integrating the histogram and heatmap. The diagnostic performance of LungRaDoc was not affected by age, gender, smoking status, nodule location, and symptoms. And the patients with lower predicted values were concentrated in SPH batch, while the patients with close predicted values to cutoff were concentrated in SCH batch 2 (Figure 5C), which might also indicate the influence of saving time.

Possible origin exploration

To illuminate the molecular modifications causing the Raman signal changes, proteome sequencing was used to seek out the possible origins of the changes. The proteome sequencing analysis identified 36 differentially expression proteins (DEPs) between malignancy and benignancy, in which 35 proteins were upregulated and 1 protein was downregulated (Figure 5D). After Gene Ontology (GO) (see Figure S2), Cluster of Orthologous Groups of proteins (KOG) (Figure 5E), Kyoto Encyclopedia of Genes and Genomes (KEGG), protein domain, subcellular localization (see Figure S2), and signal peptide analysis, the DEPs were mainly from cytoskeleton, also involved in cell process vesicle and focal adhesion, even related to intermediate filament protein and keratin. While keratin was reported visible at the peaks of 1030 cm^{-1} on the Raman spectrum of the serum, it might explain the significant difference between benignancy and malignancy around the peaks of 1030 cm^{-1}.[20] And the reflection of intermediate filament protein on Raman spectrum and the other significant different spectra would be analyzed in the next work.

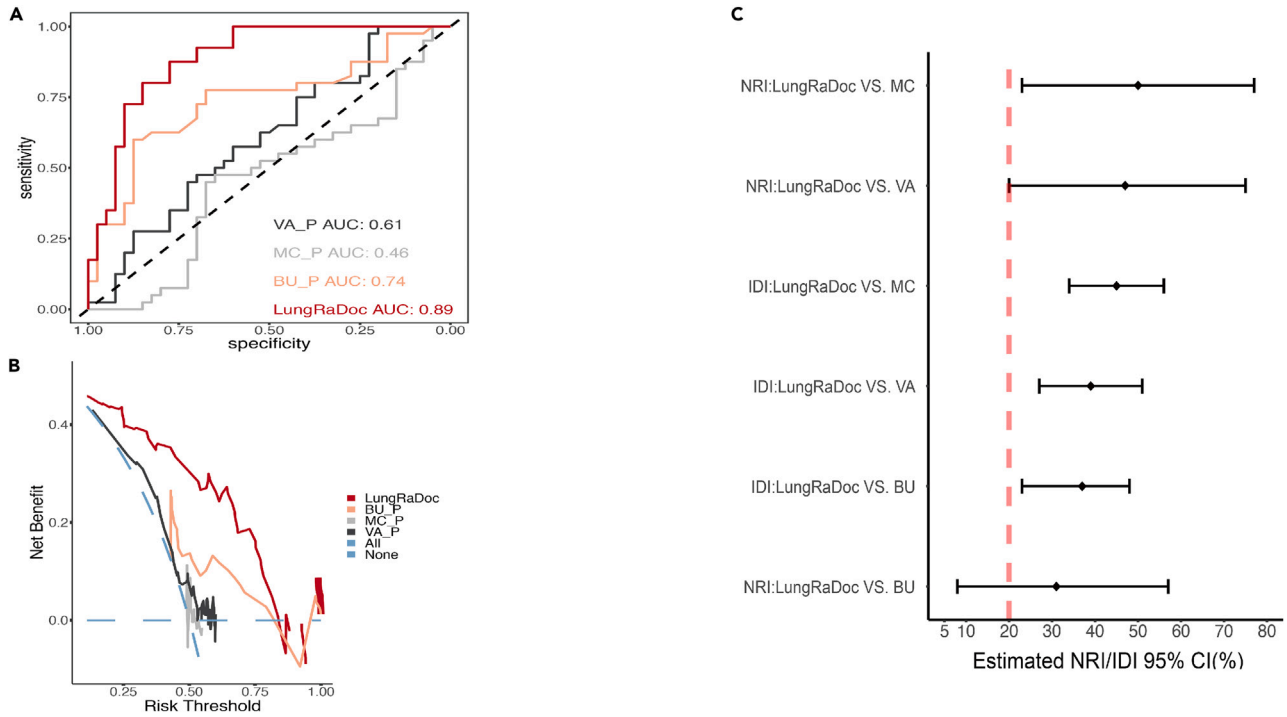


Figure 4. Comparison of the LungRaDoc with other common clinical models

(A) The ROC curves of LungRaDoc and three common clinical models used in discriminating the malignant from benign individuals. Black, gray, yellow, and red ROC curves indicated VA, MC, BU, and LungRaDoc, respectively. AUCs were included.

(B) Black, gray, yellow, and red DCA curves indicated VA, MC, BU, and LungRaDoc at whole risk threshold, respectively.

(C) The NRI and IDI of LungRaDoc compared with MC, BU, and VA model, respectively. The comparison pairs were displayed on the left-hand side, and associated 95% confidence intervals of each NRI and IDI were included.

DISCUSSION

With the increasing PNs discovered by LDCT, a part of suspicious patients receipted unnecessary surgery whose nodules were difficult to discriminate as benignancy. In this research, an RS system developed previously by our team was performed to screen the serum from patients with PNs, which operated easily, detected fast, and cost low. After combing with the SVM and ANOVA, the Raman spectra differences were used to construct a classifier named LungRaDoc, discriminating the IPNs, exceeding than the other clinical models. For LungRaDoc was not affected by age, gender, and nodules' features, it could facilitate the IPNs assessment. And LungRaDoc achieved a high positive predicted value indicating a helpful application as a screen tool for the identification of IPNs.

PNs were always defined as single and less than or equal to 3 cm in diameter, always discovered by LDCT. The possibilities of malignancy increased with the nodule size; LDCT could accurately discriminate the lesions with diameter over 3 cm. However, the nodules below 3 cm in size had more difficulties for LDCT in discrimination of malignancy. Thus, other invasive methods providing pathological results were always used to define the indistinct nodules, such as *trans*-thoracic needle aspiration (TTNA) and *trans*-bronchial forceps biopsy (TBB). For the common shortages of invasive methods, TTNA could only be attempted if the nodules were located within the periphery of the lung, which might also carry a high risk of periprocedural pneumothorax.[21] TBB was also limited by the nodule size with the sensitivity below 40% when the nodule size was smaller than 2.5 cm, which might also carry an invasive damage for patients.[22] Furthermore, surgical resection might be the final diagnosis and treatment of choice for those indistinct nodules; but more than 23% in those indistinct nodules were finally proved benign after surgery.[3] As suggested in the guideline, the IPNs at small size would be rather on LDCT surveillance than unclear resection. So, a surveillance method ameliorating LDCT for the IPNs screening with no or less-invasive damage was increasingly urgent.

Liquid biopsies as the less-invasive methods have become the research focuses currently. Most liquid biopsies have depended on the identification of tumor-derived nucleic acids and antibodies or proteins

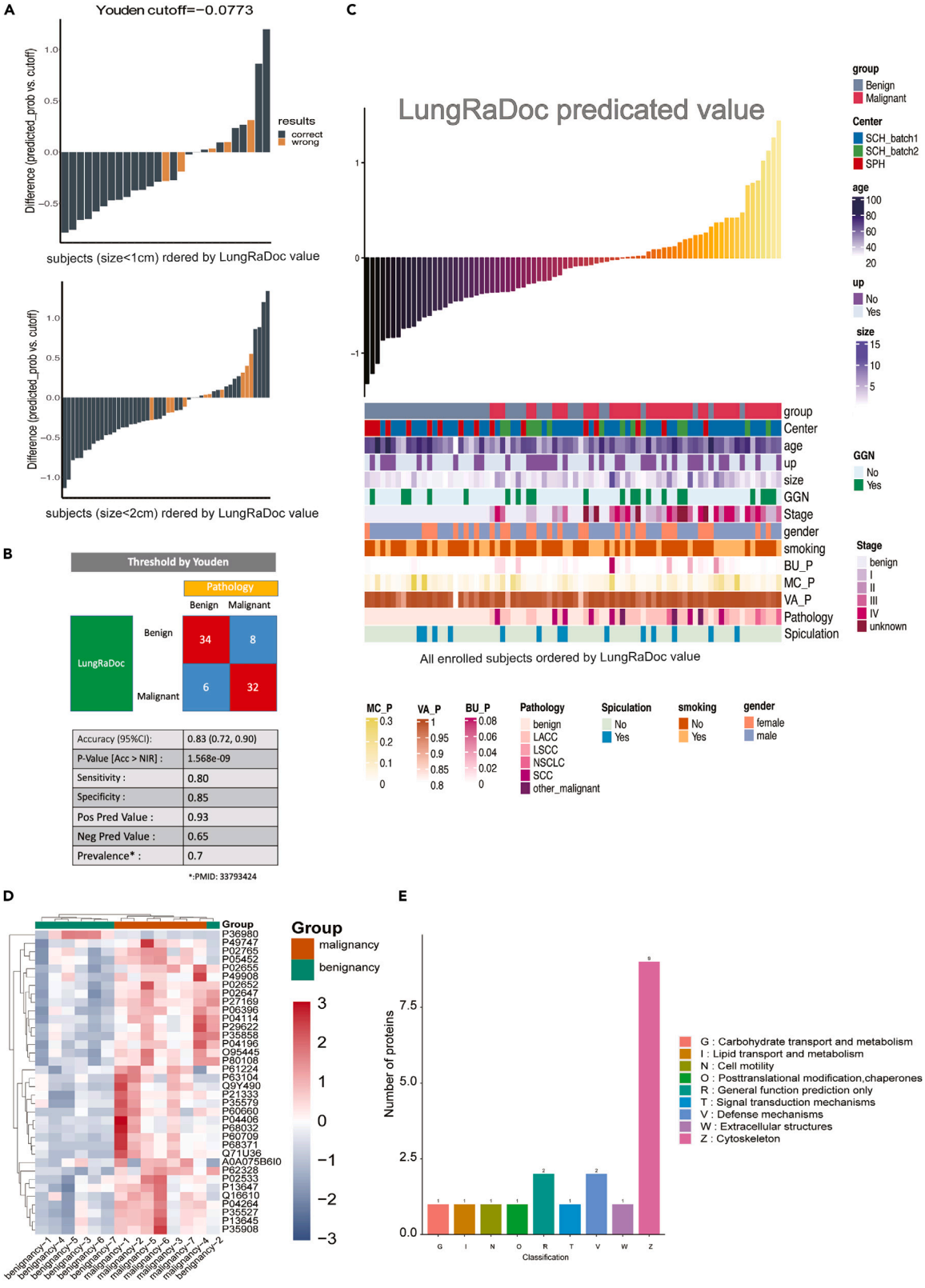

A
Youden cutoff=−0.0773
Difference (predicted_prob vs. cutoff)
results
correct
wrong
subjects (size<1cm) rdered by LungRaDoc value
subjects (size<2cm) rdered by LungRaDoc value
B
Threshold by Youden
Pathology
Benign
Malignant
LungRaDoc
34
8
6
32
Accuracy (95%CI): 0.83 (0.72, 0.90)
P-Value [Acc > NIR] : 1.568e-09
Sensitivity : 0.80
Specificity : 0.85
Pos Pred Value : 0.93
Neg Pred Value : 0.65
Prevalence* : 0.7
*:PMID: 33793424
C
LungRaDoc predicated value
group
Benign
Malignant
Center
SCH_batch1
SCH_batch2
SPH
age
up
No
Yes
size
GGN
Stage
benign
I
II
III
IV
unknown
group
Center
age
up
size
GGN
Stage
gender
smoking
BU_P
MC_P
VA_P
Pathology
Spiculation
All enrolled subjects ordered by LungRaDoc value
MC_P
VA_P
BU_P
Pathology
benign
LACC
LSCC
NSCLC
SCC
other_malignant
Spiculation
smoking
gender
female
male
D
Group
malignancy
benignancy
E
Number of proteins
Classification
G : Carbohydrate transport and metabolism
I : Lipid transport and metabolism
N : Cell motility
O : Posttranslational modification,chaperones
R : General function prediction only
T : Signal transduction mechanisms
V : Defense mechanisms
W : Extracellular structures
Z : Cytoskeleton

Figure 5. The application of LungRaDoc

(A) The recognition ability of LungRaDoc at the cutoff in the group whose nodule size was less or equal to 1 cm (left) and size was less or equal to 2 cm (right). Data are presented as differences (differences = predictive value – cutoff, cutoff = 0.0773) of the patients whose nodule size was less than or equal to 1 or 2 cm, with bars indicating individual patients. The yellow bars reflected the false positive and negative individuals recognized by the LungRaDoc, whose predictive values were close to cutoff with 0.5 differences.

(B) Compared with the pathology, the performance of LungRaDoc at Youden threshold in validation group was depicted. The 4-fold contingency table represents counts of benignancy and malignancy classified by pathology and LungRaDoc, respectively. The accuracy, sensitivity, specificity, positive predicted value, and negative predicted value were displayed in a table.

(C) The discrimination ability of LungRaDoc in validation group, and the correlation of LungRaDoc with LDCT information and other three models. The histogram indicated the individual levels of predicted values, presented as the differences. Heatmap constructs of predicted values with group, center, age, gender, smoking status, nodule characteristics, pathology, and probabilities calculated by three clinical models. For each attribute and further generation of heatmap constructs, the differences scaled from −1 to 1 were illustrated as defined in the color key gradients provided at the right of the Figure.

(D) Heatmap of differentially expressed proteins (DEPs) between true positive and negative individuals.

(E) KOG enrichment analysis of DEPs.

present in blood, plasma, serum, or sputum.[23] Combining with machine learning, the diagnostic values of those molecules mostly increased due to the construction of diagnostic classifiers,[24–26] including mRNA, DNA methylation, and circulating tumor cells (CTCs) detection. These classifiers all performed well in the discrimination of PNs with AUC higher than 0.8. However, the plentiful gene detection, DNA methylation detection, and CTC detection were all high costly and needed professional operation, which might not suit for massive screening. Some researchers raised RS might be a satisfactory liquid biopsy for screening as it is less invasive, low cost, and easy to operate and assess.[12]

RS was widely applied in the diagnosis in different kinds of diseases, for the reason that it was consisted of light generated by different vibration modes of various biomolecules including nucleic acids, proteins, lipids, and carbohydrates,[18,27–29] and this molecular "fingerprint" could be used for the sensitive and specific detection of cancer.[18] Until now, RS in cancer diagnostics has investigated a multitude of different cancer types including lung, cervix, breast, prostate, lymph nodes, esophagus, colon, larynx, bladder, and brain.[30] And in the past research, the SERS was widely used in the lung cancer diagnosis. SERS has advantages that characterize the interaction of biomolecules with metallic surfaces, to quantify biomolecules in complex matrices and to investigate cells such as CTCs, attributing to the optical properties of plasmonic nanostructures.[31] In our past work, SiO2@Au structure was applied to the SERS measurement, and with the aid of SVM an accuracy of over 90% was shown in predicting the lung cancers.[32] However, since the noble metallic nanostructure was essential in the SERS, a number of potential drawbacks appeared due to the structure and optical properties of the nanostructure.[15] Although SERS application is possible for disease detection due to the lower price of the device, the consumables used for SERS were more expensive, caused by the signal amplification generated through the use of plasmonic nanostructures.[33] And the high-quality requirements of the nanostructure might bring a high cost, which made SERS not appropriate for the IPNs screening.

Compared to SERS, spontaneous RS could collect multiple spectra from a heterogeneous sample to characterize the multi-component composition, making it useful to integrate multiple potential biomarkers into one spectroscopic signature.[12] In the serum of malignant patients, there are numbers of biomolecules changed, which could be reflected in the RS spectra. When the spontaneous RS was combined with machine learning to analyze the spectra signature, the sensitivity and specificity were highly improved in cancer diagnosis.[12] As reported previously, the Raman spectral intensity was significantly different among healthy individuals and lung cancer patients with stage I, stage II, or stage III/IV, which indicated that laser Raman spectroscopy could be used in the diagnosis of lung cancer.[16] So, we suspected that, compared with benign serum, the composition and content of biomolecules in the serum of lung cancer patients may have subtle changes. The differences of Raman spectra could be useful to analyze the metabolic changes between benignancy and malignancy. As reported previously, the peak intensities of lung cancer patients were different at 1446 cm^{-1} and 1658 cm^{-1}.[16] Similarly, in our results, the differences of peak intensities between benignancy and malignancy were around at 1437 cm^{-1}, which suggested that protein and phospholipids in the sera of lung cancer patients varied from benign patients. The RS spectra derive the morphological and biochemical composition of serum. Since some previous RS research constructed diagnostic model using components directly verified by mass spectrometry (MS), the molecular modifications causing the Raman signal changes could be more clearly stated. The proteome sequencing was also used to seek out the possible origins of the changes; then 36 DEPs between malignancy and benignancy were identified, which were mainly from cytoskeleton, also involved in cell process vesicle and focal adhesion, even related to intermediate filament protein and keratin. As previously reported, the peaks of

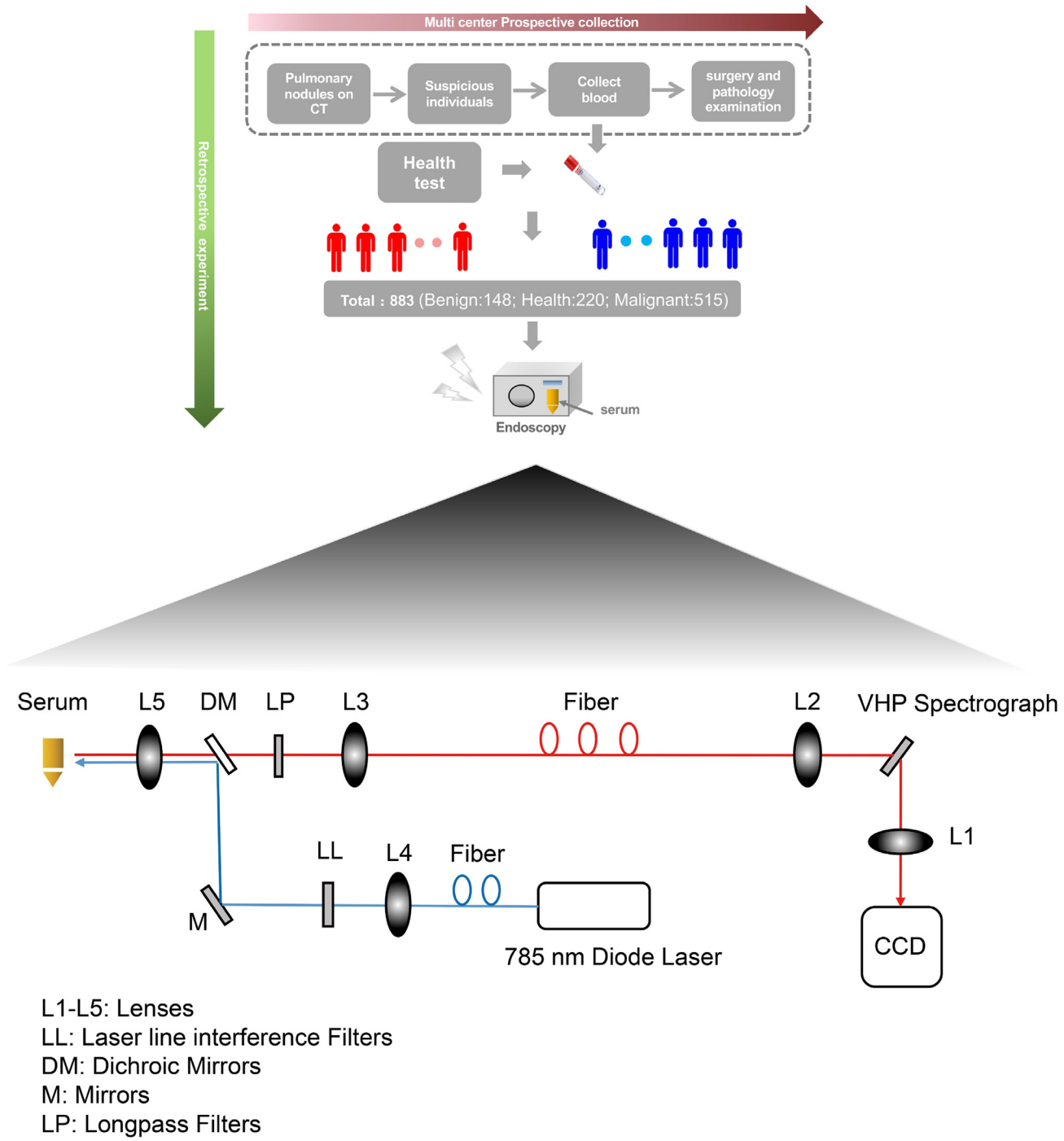


Figure 6. Workflow of this research

1030 cm^{-1} were also visible on the Raman spectrum of the serum, representing the expression of the keratin in the patient's serum.[20] Same in our research, multiple proteins changed in serum of malignant patients reflecting on the spectra peaks, while more nucleic acids, amino acid, and lipid might also change the spectra peaks. As a result, the diagnostic model could be constructed using spectra peaks without general pictures of the biophysical origins, as generality in same set and specificity in different set of the spectra figures could be analyzed using machine learning.

Moreover, the RS system used in our research was a system that needed only 0.5 mL serum and 1 min in each test. Since blood-based biomarkers might have a role in refining selection criteria for lung cancer screening, miRNA, cell-free nucleic acids, proteins, DNA methylation, and other promising biomarkers were established for early detection of lung cancer.[34] Even those molecular biology methods also outperformed in required time—a factor to consider in developing lung cancer screening consultation processes; several hours were also needed, which were much slower than RS. Therefore, the few experimental consumables indicated less strict requirements of samples, and the short scanning time indicated a rapid testing rate, which meant the RS system was much more applicable for the IPNs screening.

In order to improve the sensitivity and specificity of the RS system and make it extensive in clinical application, ANOVA and SVM were used to obtain qualified results from the spectrum data analyzing. And in our previous research, combining the Raman spectra with SVM performed excellent in the discrimination of COVID-19 from suspected patients.[19] Same in this research, the analyzing method had a discernible effect in the discrimination of IPNs. According to the information of Raman spectra, the LungRaDoc model could provide a value. After estimated by ROC, the threshold of LungRaDoc value was set at 0.0773 with the highest AUC. In the patients with suspicious nodules, the Raman spectrum could be used to evaluate the malignancy if the SVM value was higher than 0.0773. The MC model, VA model, and BU model were used to compare with our LungRaDoc, which were classical clinical models and widely cited in a lot of other similar research.[35,36] These three models calculated the probabilities of malignancy in patients with PNs using the patients' clinical information and LDCT information. Since the three models were set up with retrospective data based on specific population, which might contain racial and other differences, the three models showed limited values in our validation. The results of NRI, IDI, and NB also reflected that the LungRaDoc performed better in classification of the suspicious patients, which could bring more benefit to those patients and further reduce the unnecessary surgery. The nodules' characteristics including location and whether GGN or speculation was necessary for the three clinical models needed assessment by imaging experts. While shown in our results, the LungRaDoc was not correlated with those features, indicating an easier assessment than the other models. And the LungRaDoc was also more convenient to manipulate, while it still retained a higher diagnostic value with the ROC-AUC of 0.89 in our results. Thus, the LungRaDoc could be a useful method for the screening of IPNs. And in the future work, a large external validation would be accomplished; the application in the prognosis and treatment monitoring would also be researched.

In conclusion, our results indicated that there were significant differences in Raman spectra among patients with malignant nodules, the patients with benign nodules, and the healthy individuals. In addition, LungRaDoc was constructed based on the Raman spectra data, which could distinguish the patients with benign or malignant nodules in a high accuracy. The LungRaDoc also performed excellently and stably in the small-size nodules and exceeded the other clinical models. While the LungRaDoc was convenient to assess, less invasive, and low costly, it would be more of benefit to the patients with IPNs.

Limitations of the study

During cross-validation procedure, spectra from one sample would appear both in test set and training. Since sample size was unbalanced, the benign samples were much less than malignant samples. Spectra results from malignant samples were divided into 10 groups randomly to match with the benign groups, which might cause some overlap. However, the samples in validation group were absolutely independent, which could avoid the spectra from one sample being both in test set and in validation set.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Human subjects
- METHOD DETAILS
  - Sample preparation and storage

- RS detection
- Feature extraction and model construction
- Model estimation
- Sample-selection for the possible origins'exploration
- Proteome sequencing analysis
- Packages used

- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.106693.

ACKNOWLEDGMENTS

There are sincerely thanks to all participants in this study. This study was funded by Sichuan Natural Science Foundation (No. 2022NSFSC0654, China), UESTC Sichuan Cancer Hospital 2021 Medical-engineering Oncology Innovation Fund (No. ZYGX2021YGCX013, China), Sichuan Medical Association Research project (S20087), and Sichuan Cancer Hospital Outstanding Youth Science Fund (YB2021033).

This multi-center and prospective research started from June 2017, which was approved by the medical ethical committee of Sichuan Cancer Hospital (SCCHEC-02-2021-073) and Sichuan Provincial People's Hospital (2021-NO.404).

AUTHOR CONTRIBUTIONS

Huaichao Luo, Dongsheng Wang, and Yu Yin designated the study. Lintao Li and Gang Yin convened the participants in this research. Ruiling Zu completed the ethical works. Yao Deng, Shuya He, Xing Yin, Kaijiong Zhang, and Qiao He prepared samples and performed RS analyses. Dezhong Yao performed ANOVA analysis and SVM model construction. Huaichao Luo performed the ROC, DCA, NRI, IDI, NB analysis, Logistic regression, and heatmap. Ruiling Zu wrote the first draft. Ruiling Zu and Huaichao Luo completed the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: November 21, 2022
Revised: February 23, 2023
Accepted: April 13, 2023
Published: April 23, 2023

REFERENCES

1. Wu, F., Wang, L., and Zhou, C. (2021). Lung cancer in China: current and prospect. Curr. Opin. Oncol. *33*, 40–46. https://doi.org/10.1097/CCO.0000000000000703.

2. Slatore, C.G., and Wiener, R.S. (2018). Pulmonary nodules: a small problem for many, severe distress for some, and how to communicate about it. Chest *153*, 1004–1015. https://doi.org/10.1016/j.chest.2017.10.013.

3. Harzheim, D., Eberhardt, R., Hoffmann, H., and Herth, F.J.F. (2015). The solitary pulmonary nodule. Respiration *90*, 160–172. https://doi.org/10.1159/000430996.

4. Callister, M.E.J., Baldwin, D.R., Akram, A.R., Barnard, S., Cane, P., Draffan, J., Franks, K., Gleeson, F., Graham, R., Malhotra, P., et al. (2015). British Thoracic Society guidelines for the investigation and management of pulmonary nodules. Thorax *70*, ii1–ii54. https://doi.org/10.1136/thoraxjnl-2015-207168.

5. González Maldonado, S., Delorme, S., Hüsing, A., Motsch, E., Kauczor, H.U., Heussel, C.P., and Kaaks, R. (2020). Evaluation of prediction models for identifying malignancy in pulmonary nodules detected via low-dose computed tomography. JAMA Netw. Open *3*, e1921221. https://doi.org/10.1001/jamanetworkopen.2019.21221.

6. Venkadesh, K.V., Setio, A.A.A., Schreuder, A., Scholten, E.T., Chung, K., W Wille, M.M., Saghir, Z., van Ginneken, B., Prokop, M., and Jacobs, C. (2021). Deep learning for malignancy risk estimation of pulmonary nodules detected at low-dose screening CT. Radiology *300*, 438–447. https://doi.org/10.1148/radiol.2021204433.

7. Choi, H.K., Ghobrial, M., and Mazzone, P.J. (2018). Models to estimate the probability of malignancy in patients with pulmonary nodules. Ann. Am. Thorac. Soc. *15*, 1117–1126. https://doi.org/10.1513/AnnalsATS.201803-173CME.

8. Gould, M.K., Ananth, L., and Barnett, P.G.; Veterans Affairs SNAP Cooperative Study Group (2007). A clinical model to estimate the pretest probability of lung cancer in patients with solitary pulmonary nodules. Chest *131*, 383–388. https://doi.org/10.1378/chest.06-1261.

9. Chung, K., Mets, O.M., Gerke, P.K., Jacobs, C., den Harder, A.M., Scholten, E.T., Prokop, M., de Jong, P.A., van Ginneken, B., and Schaefer-Prokop, C.M. (2018). Brock malignancy risk calculator for pulmonary nodules: validation outside a lung cancer screening population. Thorax *73*, 857–863. https://doi.org/10.1136/thoraxjnl-2017-211372.

10. Kammer, M.N., Lakhani, D.A., Balar, A.B., Antic, S.L., Kussrow, A.K., Webster, R.L., Mahapatra, S., Barad, U., Shah, C., Atwater, T., et al. (2021). Integrated biomarkers for the management of indeterminate pulmonary nodules. Am. J. Respir. Crit. Care Med. *204*, 1306–1316. https://doi.org/10.1164/rccm.202012-4438OC.

11. Zu, R., Wu, L., Zhou, R., Wen, X., Cao, B., Liu, S., Yang, G., Leng, P., Li, Y., Zhang, L., et al. (2022). A new classifier constructed with platelet features for malignant and benign pulmonary nodules based on prospective real-world data. J. Cancer *13*, 2515–2527. https://doi.org/10.7150/jca.67428.

12. Ralbovsky, N.M., and Lednev, I.K. (2020). Towards development of a novel universal medical diagnostic method: Raman spectroscopy and machine learning. Chem. Soc. Rev. *49*, 7428–7453. https://doi.org/10.1039/d0cs01019g.

13. Wu, L., Dias, A., and Diéguez, L. (2022). Surface enhanced Raman spectroscopy for tumor nucleic acid: towards cancer diagnosis and precision medicine. Biosens. Bioelectron. *204*, 114075. https://doi.org/10.1016/j.bios.2022.114075.

14. Shin, H., Jeong, H., Park, J., Hong, S., and Choi, Y. (2018). Correlation between cancerous exosomes and protein markers based on surface-enhanced Raman spectroscopy (SERS) and principal component analysis (PCA). ACS Sens. *3*, 2637–2643. https://doi.org/10.1021/acssensors.8b01047.

15. Cialla-May, D., Zheng, X.S., Weber, K., and Popp, J. (2017). Recent progress in surface-enhanced Raman spectroscopy for biological and biomedical applications: from cells to clinics. Chem. Soc. Rev. *46*, 3945–3961. https://doi.org/10.1039/c7cs00172j.

16. Wang, H., Zhang, S., Wan, L., Sun, H., Tan, J., and Su, Q. (2018). Screening and staging for non-small cell lung cancer by serum laser Raman spectroscopy. Spectrochim. Acta Mol. Biomol. Spectrosc. *201*, 34–38. https://doi.org/10.1016/j.saa.2018.04.002.

17. Sinica, A., Brožáková, K., Brůha, T., and Votruba, J. (2019). Raman spectroscopic discrimination of normal and cancerous lung tissues. Spectrochim. Acta Mol. Biomol. Spectrosc. *219*, 257–266. https://doi.org/10.1016/j.saa.2019.04.055.

18. Zheng, Q., Li, J., Yang, L., Zheng, B., Wang, J., Lv, N., Luo, J., Martin, F.L., Liu, D., and He, J. (2020). Raman spectroscopy as a potential diagnostic tool to analyse biochemical alterations in lung cancer. Analyst *145*, 385–392. https://doi.org/10.1039/c9an02175b.

19. Yin, G., Li, L., Lu, S., Yin, Y., Su, Y., Zeng, Y., Luo, M., Ma, M., Zhou, H., Orlandini, L., et al. (2021). An efficient primary screening of COVID-19 by serum Raman spectroscopy. J. Raman Spectrosc. *52*, 949–958. https://doi.org/10.1002/jrs.6080.

20. Wang, S.S., Ye, D.X., Wang, B., and Xie, C. (2020). The expressions of keratins and P63 in primary squamous cell carcinoma of the thyroid gland: an application of Raman spectroscopy. OncoTargets Ther. *13*, 585–591. https://doi.org/10.2147/OTT.S229436.

21. Yang, W., Jiang, H., Khan, A.N., Allen, C., Bertolaccini, L., Lv, T., and Song, Y.; written on behalf of the (2017). Transthoracic needle aspiration in solitary pulmonary nodule. Transl. Lung Cancer Res. *6*, 76–85. AME Lung Cancer Collaborative Group. https://doi.org/10.21037/tlcr.2017.02.03.

22. Gould, M.K., Donington, J., Lynch, W.R., Mazzone, P.J., Midthun, D.E., Naidich, D.P., and Wiener, R.S. (2013). Evaluation of individuals with pulmonary nodules: when is it lung cancer? Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. Chest *143*, e93S–e120S. https://doi.org/10.1378/chest.12-2351.

23. Kathuria, H., Gesthalter, Y., Spira, A., Brody, J.S., and Steiling, K. (2014). Updates and controversies in the rapidly evolving field of lung cancer screening, early detection, and chemoprevention. Cancers *6*, 1157–1179. https://doi.org/10.3390/cancers6021157.

24. Kossenkov, A.V., Qureshi, R., Dawany, N.B., Wickramasinghe, J., Liu, Q., Majumdar, R.S., Chang, C., Widura, S., Kumar, T., Horng, W.H., et al. (2019). A gene expression classifier from whole blood distinguishes benign from malignant lung nodules detected by low-dose CT. Cancer Res. *79*, 263–273. https://doi.org/10.1158/0008-5472.CAN-18-2032.

25. Liang, W., Chen, Z., Li, C., Liu, J., Tao, J., Liu, X., Zhao, D., Yin, W., Chen, H., Cheng, C., et al. (2021). Accurate diagnosis of pulmonary nodules using a noninvasive DNA methylation test. J. Clin. Invest. *131*, e145973. https://doi.org/10.1172/JCI145973.

26. Zhang, W., Duan, X., Zhang, Z., Yang, Z., Zhao, C., Liang, C., Liu, Z., Cheng, S., and Zhang, K. (2021). Combination of CT and telomerase+ circulating tumor cells improves diagnosis of small pulmonary nodules. JCI Insight *6*, e148182. https://doi.org/10.1172/jci.insight.148182.

27. Cui, S., Zhang, S., and Yue, S. (2018). Raman spectroscopy and imaging for cancer diagnosis. J. Healthc. Eng. *2018*, 8619342. https://doi.org/10.1155/2018/8619342.

28. Ralbovsky, N.M., and Lednev, I.K. (2019). Raman spectroscopy and chemometrics: a potential universal method for diagnosing cancer. Spectrochim. Acta Mol. Biomol. Spectrosc. *219*, 463–487. https://doi.org/10.1016/j.saa.2019.04.067.

29. Santos, I.P., Barroso, E.M., Bakker Schut, T.C., Caspers, P.J., van Lanschot, C.G.F., Choi, D.H., van der Kamp, M.F., Smits, R.W.H., van Doorn, R., Verdijk, R.M., et al. (2017). Raman spectroscopy for cancer detection and cancer surgery guidance: translation to the clinics. Analyst *142*, 3025–3047. https://doi.org/10.1039/c7an00957g.

30. Moisoiu, V., Stefancu, A., Gulei, D., Boitor, R., Magdo, L., Raduly, L., Pasca, S., Kubelac, P., Mehterov, N., Chiș, V., et al. (2019). SERS-based differential diagnosis between multiple solid malignancies: breast, colorectal, lung, ovarian and oral cancer. Int. J. Nanomed. *14*, 6165–6178. https://doi.org/10.2147/IJN.S198684.

31. Xu, X., Lin, J., Guo, Y., Wu, X., Xu, Y., Zhang, D., Zhang, X., Yujiao, X., Wang, J., Yao, C., et al. (2022). TiO(2)-based Surface-Enhanced Raman Scattering bio-probe for efficient circulating tumor cell detection on microfilter. Biosens. Bioelectron. *210*, 114305. https://doi.org/10.1016/j.bios.2022.114305.

32. Wang, Z., Hong, Y., Yan, H., Luo, H., Zhang, Y., Li, L., Lu, S., Chen, Y., Wang, D., Su, Y., and Yin, G. (2022). Fabrication of optoplasmonic particles through electroless deposition and the application in SERS-based screening of nodule-involved lung cancer. Spectrochim. Acta Mol. Biomol. Spectrosc. *279*, 121483. https://doi.org/10.1016/j.saa.2022.121483.

33. Bratchenko, L.A., Al-Sammarraie, S.Z., Tupikova, E.N., Konovalova, D.Y., Lebedev, P.A., Zakharov, V.P., and Bratchenko, I.A. (2022). Analyzing the serum of hemodialysis patients with end-stage chronic kidney disease by means of the combination of SERS and machine learning. Biomed. Opt Express *13*, 4926–4938. https://doi.org/10.1364/BOE.455549.

34. Adams, S.J., Stone, E., Baldwin, D.R., Vliegenthart, R., Lee, P., and Fintelmann, F.J. (2023). Lung cancer screening. Lancet *401*, 390–408. https://doi.org/10.1016/S0140-6736(22)01694-4.

35. Li, M., He, H., Huang, G., Lin, B., Tian, H., Xia, K., Yuan, C., Zhan, X., Zhang, Y., and Fu, W. (2021). A novel and rapid serum detection technology for non-invasive screening of gastric cancer based on Raman spectroscopy combined with different machine learning methods. Front. Oncol. *11*, 665176. https://doi.org/10.3389/fonc.2021.665176.

36. Wu, J., Zhang, H., Li, L., Hu, M., Chen, L., Xu, B., and Song, Q. (2020). A nomogram for predicting overall survival in patients with low-grade endometrial stromal sarcoma: a population-based analysis. Cancer Commun. *40*, 301–312. https://doi.org/10.1002/cac2.12067.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
blood samples	Sichuan Cancer hospital, and Sichuan Provincial People's Hospital	N/A
Software and algorithms		
R	R Foundation for Statistical Computing	http://www.R-project.org
FragPipe	github	https://github.com/Nesvilab/FragPipe
Diamond	github	https://github.com/bbuchfink/diamond/releases
InterProscan	EMBL-EBI	https://www.ebi.ac.uk/interpro/
WoLF PSORT	NAKAI Lab	https://wolfpsort.hgc.jp
SignalP	DTU Health Tech	https://services.healthtech.dtu.dk/service.php?SignalP-4.1
Raman spectrum analysis software	Sichuan Institute for Brain Science and Brain-Inspired Intelligence	Andor camera
Other		
silicone-coated serum tubes	C.D.RICH Co., Ltd, Chengdu, China	N/A
LC-MS	Bruker Daltonics, Germany	N/A
caret package	https://github.com/topepo/caret/	https://github.com/topepo/caret/
ggplot2 package	https://ggplot2.tidyverse.org	https://ggplot2.tidyverse.org
pROC packages	https://web.expasy.org/pROC/	https://web.expasy.org/pROC/
ggDCA packages	https://cran.rstudio.com/web/packages/ggDCA/index.html	https://cran.rstudio.com/web/packages/ggDCA/index.html
PredictABEL package	https://cran.r-project.org/web/packages/PredictABEL/index.html	https://cran.r-project.org/web/packages/PredictABEL/index.html
glmnet package	https://glmnet.stanford.edu	https://glmnet.stanford.edu
ComplexHeatmap package	https://github.com/stemangiola/tidyHeatmap	https://github.com/stemangiola/tidyHeatmap
clusterProfiler package	https://www.bioconductor.org/packages/clusterProfiler/	https://www.bioconductor.org/packages/clusterProfiler/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to the lead contact, Dr. Huaichao Luo (luohuaichao@scszlyy.org.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- All data reported in this paper will be shared by the lead contact upon any reasonable request.
- This paper does not report original code.
- Any reasonable request for additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human subjects

This multi-center and prospective research was approved by the medical ethical committee of Sichuan Cancer Hospital (SCCHEC-02-2021-073), and Sichuan Provincial People's Hospital (2021-NO.404). And informed consent was obtained from all subjects. As displayed in Figure 6, when the pulmonary nodules were first observed in Sichuan Cancer Hospital and Sichuan Provincial People's Hospital, the suspicious patients were enrolled in this study. After the surgery or biopsy, the enrolled patients were separated into benign and malignant group according to the pathological results. The guidelines of Chinese society of clinical oncology (2019 version) were diagnostic basis of histological type and TNM staging. The basic information, anamnesis, family history, personal history, LDCT information, and pathology information were captured and summarized. And the healthy individuals were recruited from the health examination population with the blood samples collected. So, the exclusion criteria were as followed:1) patients who were not confirmed by the surgery or biopsy; 2) patients with acute inflammation; 3) patients who were pregnant; 4) patients who received plasma or blood transfusion in half of a month; 5) patients without complete clinical and LDCT data. A total of 883 participants were enrolled in this research, including 148 benign, 220 healthy, and 515 malignant participants, respectively. The single subject details including gender and age information are reported in Table S3.

METHOD DETAILS

Sample preparation and storage

The blood sample of all the suspicious patients were collected at the same time they admitted to hospital. The blood samples were taken from all participants using silicone-coated serum tubes (C.D.RICH Co., Ltd, Chengdu, China). After one-hour reposed, the serum was isolated from blood samples by centrifuging at 3000 rpm for 10 min. The serum samples were transferred into a new EP tube, and stored at 4 °C. All the samples were needed to measure within 36h after the collection. 2 ml cryopreservation tubes made up with polypropylene were prepared in advance. Approximately 0.5 ml of the serum samples were strictly sealed in the cryopreservation tubes for the Raman scan. Finished with Raman scan, each sample was transferred back to EP tube and stored at -80°C.

RS detection

As previously, the Raman system (Ando_iVac_316) used in this research was designed by the Sichuan Institute for Brain Science and Brain-Inspired Intelligence.[19] Before manipulation, the device was pre-cooled to -60°C. After the wave-number was calibrated using ethanol spectrum, the samples were excited with laser power around 760mA, while the spectra were recorded in the range of 600–1800 cm−1 and collected 15 times per sample, with 1s exposure time for each collection. The total time was about 1 minute for a sample, while 15 spectra pictures were collected. Collecting multiple spectra per sample ensures an accurate representation of the heterogeneous composition of a sample.

Feature extraction and model construction

In this research, a total of 883 individuals were final involved, where 120 individuals were random preselected into validation group including 40 malignant patients, 40 benign patients and 40 healthy participants. And then the rest of 761 individuals were enrolled into discover group, including 180 healthy individuals, 108 patients with benign nodules, and 475 patients with malignant nodules. Wave-number data points from all the participants were considered to extract features. After the data smoothed by Savitzky–Golay digital–moving average filter, and the baseline corrected by the Improved Modified Multi-Polynomial Fitting algorithm, ANOVA analysis was used for the statistical test. Random sampling 70% of the data was repeated one hundred times. Once the distributions of random samples satisfied Gaussian distribution, variances test was selected for ANOVA test. Points showing statistical significance of ANOVA test for more than 70 times out of 100 between two comparison groups, were selected as the features to input in support vector machine (SVM) model, and built the classification model. The data set was trained using slide-wise cross-validation where discover group was used for training and validation group was used for cross-validation. The whole training procedure was subjected to 10-fold cross-validation to obtain malignant probabilities for the training samples. The probabilities were also used to calculate cross-validation performance. For each cross-validation, the ROC curves were used to evaluate the performance of the SVM model, which was repeated 5 times in every cross-validation. Performance was also determined by applying SVM model to the validation group.

Model estimation

The SVM model could output probabilities for each particular group that could be evaluated using ROC curves. The predicted values calculated by SVM model were compared in validation group stratified by the stage and pathology of lung cancer. After the validation group was separated into ground glass subsets and non-ground glass subsets, the predicted values were compared in the two subsets. While the validation group was separated into smaller size subsets and lager size subsets, the predicted values were also compared in the two subsets. Logistic regression was used to investigate the correlation of SVM model with age, gender, characteristics of nodules and saving time of samples. In order to compare the classifying performance of SVM model with other clinical models, ROC curves reflecting the sensitivity (Sens), specificity (Spec), accuracy (ACC), the area under the curve (AUC), Net reclassification improvement (NRI), integrated discrimination improvement (IDI) and net benefit (NB) were used. The recognition ability of SVM at the cutoff in the individuals with small size nodules was in contrast with pathology, and displayed as four-fold table.

Sample-selection for the possible origins'exploration

A total of 14 samples were selected based on the predicted values of SVM model, including 7 true positive individuals from group of which predictive values were over than cutoff value and 7 true negative individuals from group of which predictive values lower than cutoff value. At same time, the 14 samples were randomly from SCH batch 1, SCH batch 2 and SPH batch with age and gender matched, and all samples were stored at -80°C after RS test previously.

Proteome sequencing analysis

The total proteins were extracted (Bio-Rad) and analyzed by LC-MS (Bruker Daltonics, Germany). MS raw data were analyzed with FragPipe (v17.1) which relies on MSFragger for qualitative analysis and uses Phosopher for validation and filtering. Spectra files were searched against the homo sapiens SwissProt database (20425 entries). IonQuant mode and TMT-Integrator was used to perform isobaric labeling-based quantification (TMT/iTRAQ). Proteins denoted as contaminants were removed, the remaining identifications were used for further quantification analysis. The conditions use to filter the deferentially expressed proteins were as follows: $|\log2\text{-fold change}| \geq 1.5$ and adjusted P-value < 0.05. Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway, Cluster of Orthologous Groups of proteins (KOG) and Gene Ontology (GO) enrichment analyses were used to identify the significant pathways, InterPro (v5.59-91.0) was used to provide functional analysis of protein sequences by classifying them into families and predicting the presence of domains and important sites, and WoLF PSORT was used to reveal the subcellular location. $P < 0.05$ was set as the cutoff criterion for significant enrichment.

Packages used

Statistical analysis was conducted with caret package (version 6.0) (https://github.com/topepo/caret/). Boxplot visualization was carried out using the ggplot2 package (version 3.3.5) (https://ggplot2.tidyverse.org). ROC depiction and cut-off selection used the pROC R packages (version 1.18) (https://web.expasy.org/pROC/). DCA analysis is conducted by ggDCA R packages (version 1.1) (https://cran.rstudio.com/web/packages/ggDCA/index.html). NRI, IDI, and NB was performed using PredictABEL package (version 1.2) (https://cran.r-project.org/web/packages/PredictABEL/index.html). Logistic regression and heatmap was carried out using the glmnet package (version 4.1) (https://glmnet.stanford.edu) and ComplexHeatmap package (version 2.7.11) (https://github.com/stemangiola/tidyHeatmap). Enrichment analyses were finished with clusterProfiler package (v3.10.1) (https://www.bioconductor.org/packages/clusterProfiler/).

QUANTIFICATION AND STATISTICAL ANALYSIS

All analyses were performed by using R 4.0.4 (Version 1.74) (R Foundation for Statistical Computing, http://www.R-project.org). The data were compared among the three groups using the Nonparametric test. Thus, $p<0.05$ were considered significant with two sided, and values were indicated in medians. The construction, performance and validation of the model were implemented under R either.