Radiomics combined with clinical features in distinguishing non-calcifying tuberculosis granuloma and lung adenocarcinoma in small pulmonary nodules

Qing Dong [1] , **Qingqing Wen** [2] , **Nan Li** [3] , **Jinlong Tong** [4] , **Zhaofu Li** [5] , **Xin Bao** [6] , **Jinzhi Xu** [1] , **Dandan Li** Corresp. [7]

[1] Department of Thoracic Surgery at No. 4 Affiliated Hospital, Harbin Medical University, Harbin, China

[2] Icahn School of Medicine at Mount Sinai, New York, NY, United States

[3] Department of Pathology at No. 4 Affiliated Hospital, Harbin Medical University, Harbin, China

[4] Department of Medical Imaging at No. 4 Affiliated Hospital, Harbin Medical University, Harbin, China

[5] Heilongjiang Institute of Automation, Harbin, China

[6] Harbin Medtech Innovative Company, Harbin, China

[7] Department of Radiology at Cancer Hospital, Harbin Medical University, Harbin, China

Corresponding Author: Dandan Li
Email address: hmu.cancer.hospital@gmail.com

Aim: To evaluate the performance of radiomics models with the combination of clinical features in distinguishing non-calcified tuberculosis granuloma (TBG) and lung adenocarcinoma (LAC) in small pulmonary nodules.

Methodology: We conducted a retrospective analysis of 280 patients with pulmonary nodules confirmed by surgical biopsy from January 2017 to December 2020. Samples were divided into LAC group (n=143) and TBG group (n=137). We assigned them to a training dataset (n=196) and a testing dataset (n=84). Clinical features including gender, age, smoking, CT appearance (size, location, spiculated sign, lobulated shape, vessel convergence, and pleural indentation) were extracted and included in the radiomics models. 3D slicer and FAE software were used to delineate the Region of Interest (ROI) and extract clinical features. The performance of the model was evaluated by the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC).

Results: Based on the model selection, clinical features gender, and age in the LAC group and TBG group showed a significant difference in both datasets (P<0.05). CT appearance lobulated shape was also significantly different in the LAC group and TBG group (Training dataset, P = 0.034; Testing dataset, P = 0.030). AUC were 0.8344 (95 % CI =0.7712- 0.8872) and 0.751 (95 % CI= 0.6382 – 0.8531) in training and testing dataset, respectively.

Conclusion: With the capacity to detect differences between TBG and LAC based on their clinical features, radiomics models with a combined of clinical features may function as the potential non-invasive tool for distinguishing TBG and LAC in small pulmonary nodules.

Manuscript Title

Radiomics combined with clinical features in distinguishing non-calcifying tuberculosis granuloma and lung adenocarcinoma in small pulmonary nodules

Qing Dong [1], Qingqing Wen [2], Nan Li [3], Jinlong Tong [4], Zhaofu Li [5], Xin Bao [6], Jinzhi Xu [1], Dandan Li [7]

1. Department of Thoracic Surgery at No. 4 Affiliated Hospital, Harbin Medical University, Harbin, Heilongjiang, China
2. Icahn School of Medicine, Mount Sinai, New York, United States
3. Department of Pathology at No. 4 Affiliated Hospital, Harbin Medical University, Harbin, Heilongjiang, China
4. Department of Medical Imaging at No. 4 Affiliated Hospital, Harbin Medical University, Harbin, Heilongjiang, China
5. Heilongjiang Institute of Automation, Harbin, Heilongjiang, China
6. Harbin Medtech Innovative Company, Harbin, Heilongjiang, China
7. Department of Radiology at Cancer Hospital, Harbin Medical University, Harbin, Heilongjiang, China

Dandan Li [7]
No. 150 Haping Road, Nangang Dist., Harbin, Heilongjiang, 150081, China
Email address: hmu.cancer.hospital@gmail.com

Abstract

Aim: To evaluate the performance of radiomics models with the combination of clinical features in distinguishing non-calcified tuberculosis granuloma (TBG) and lung adenocarcinoma (LAC) in small pulmonary nodules.

Methodology: We conducted a retrospective analysis of 280 patients with pulmonary nodules confirmed by surgical biopsy from January 2017 to December 2020. Samples were divided into the LAC group (n=143) and the TBG group (n=137). We assigned them to a training dataset (n=196) and a testing dataset (n=84). Clinical features including gender, age, smoking, CT appearance (size, location, spiculated sign, lobulated shape, vessel convergence, and pleural indentation) were extracted and included in the radiomics models. 3D slicer and FAE software were used to delineate the Region of Interest (ROI) and extract clinical features. The performance of the model was evaluated by the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC).

Results: Based on the model selection, clinical features gender, and age in the LAC group and TBG group showed a significant difference in both datasets ($P<0.05$). CT appearance lobulated shape was also significantly different in the LAC group and TBG group (Training dataset, $P = 0.034$; Testing dataset, $P = 0.030$). AUC were 0.8344 (95 % CI =0.7712- 0.8872) and 0.751 (95 % CI= 0.6382 – 0.8531) in the training and testing dataset, respectively.

Conclusion: With the capacity to detect differences between TBG and LAC based on their clinical features, radiomics models with a combined of clinical features may function as the potential non-invasive tool for distinguishing TBG and LAC in small pulmonary nodules.

Keywords: Radiomics, non-calcified tuberculosis granuloma, lung adenocarcinoma, pulmonary nodules, clinical features

Introduction

Tuberculosis (TB) is an infectious disease that is caused by a single source[1]. According to statistics, there are about 10 million new TB patients and 1.5 million deaths each year, more than any other infectious disease [2]. Among them, pulmonary TB is the most common, accounting for about 85 % of all tuberculosis cases [3]. Its pathological manifestation is chronic granulomatous inflammation [4]. In 2020, 1930 million new cancer cases and 10 million deaths were estimated worldwide, with approximately 2.2 million (11.4%) new lung cancer cases and 1.8 million (18%) deaths [5]. LAC is the most common malignant tumor, its prognosis is much worse than tuberculosis, so early diagnosis and treatment are very important. However, it is difficult to distinguish TBG and LAC in chest images, and even nuclear medicine is nonspecific [7, 8]. Because both diseases can be shown as solid nodules or masses on imaging studies and have similar radiological features. The confirmative diagnosis of pulmonary nodules is usually biopsy or surgery [9]. However, this invasive examination may lead to possible tissue damage [10]. Besides, unnecessary imaging studies may also delay treatment, or miss the best treatment time window [11]. Therefore, it is expected in clinical practice that a method can be used to monitor pulmonary nodules noninvasively, and may also provide effective support for the diagnosis and treatment of pulmonary nodules. Radiomics is used to extract features from radiological images and make these features in a quantifiable manner. Its purpose is to better or more consistently discover radiological features, and provide objective features that cannot be provided by standard visual image interpretation for quantitative and qualitative density and morphological characteristics of pulmonary nodules [12, 13]. Radiomics can be used for auxiliary diagnosis of pulmonary nodules and prognosis prediction of lung cancer [14,15]. Importantly, radiomics has been applied to evaluate the molecular and clinical features of lung cancer because of its capacity of detecting atypical features in tumor lesions. [16]. In this study, we hypothesized that radiomics analysis could distinguish TBG and LAC in small pulmonary nodules based on imaging and clinical features. To test this idea, we extracted the features of small nodules from lung CT using radiomics technology, obtained the radiological model through statistical analysis, and combined it with clinical features. Our goal is to develop a non-invasive method of distinguishing benign and malignant pulmonary nodules using radiomics models in a combination of clinical features.

Materials & Methods

Patients selection

Our research had been approved by the Ethics Review Committee of No.4th Affiliated Hospital of Harbin Medical University (Institutional Review Board that approved number: KY2020-04). Since it was a retrospective study, additional informed consent was waived. Samples that meet all the following criteria were included: (1) Pulmonary tuberculosis or primary LAC confirmed by biopsy or surgical pathology. (2) Enhanced chest CT images that were collected within 1 month before surgery. (3) Isolated non-calcified pulmonary nodules. (4) The maximum diameter was less than 30mm. Samples were excluded if they did not meet the above criteria. According to the above inclusion and exclusion criteria, we enrolled 280 patients (143 LAC, 137 TBG) who met the inclusion criteria from January 2017 to December 2020. Patients were randomly selected into training and testing data sets by FeAture Explorer (FAE) software based on the TBG or LAC group.

Evaluation of pathology

All specimens were fixed with formalin and stained with hematoxylin and eosin (HE). In order to judge the biopsy results separately, two pathologists with more than 10 years of working experience were blind to the clinical information. All lesions were classified according to the international standard [6]. Classification of Pulmonary Adenocarcinoma according to the latest IASLC/ATS/ERS criteria in previous study [27]: (1). Preinvasive lesions (2). Minimally invasive adenocarcinoma (≤3 cm lepidic predominant tumor with ≤5 mm invasion) (3) Invasive adenocarcinoma (4) Variants of invasive adenocarcinoma

CT data collection

Scanning parameters: The second generation gemstone spectral CT (Discovery CT750 HD) of the US General Electric Company was used to perform dual-phase enhanced CT examination of 280 patients. Patients were in the supine position, scan range was from chest entrance to the diaphragm, to ensure full coverage of all lung tissue. A total of 75 mL non-ionic iodine contrast agent Ioversol (350 mgI / ml) was injected with a double-tube high-pressure syringe at a flow rate of 3.5 mL / s. After injection into the elbow vein, the thoracic aorta at the level of tracheal protuberance was automatically selected as the starting point for monitoring. The intelligent tracking technology of the contrast agent was used to determine the starting time of scanning. When the threshold reached 130 Hu, the scanning was automatically triggered. A venous phase scan started at 80 seconds. Other parameters were as follows: layer thickness was 0.625 mm, frame rotation time was 0.6 s, pitch was 1.375, and tube current was 600 mA.

Image evaluation

The CT appearance including lesion size, location, burr, lobulation, vascular penetration, and pleural involvement was extracted by two radiologists with more than 10 years of imaging diagnosis experience. Other clinical features such as age, gender, and smoking history were obtained from the electronic health records. To keep a subjective clinical judgment, the two radiologists were blind to both baseline information and biopsy results. If there were conflicting opinions, an agreement would be achieved after discussion. For example, an average value of lesion size was taken after discussion if there were conflicting opinions between radiologists.

Tumor segmentation

We loaded CT images into 3D slicer software (version 4.10.0) for manual segmentation (Figure 1A). The region of interest (ROI) on CT was delineated by a thoracic surgeon with 10 years of lung surgery experience (Figure 1B). The ROI was then confirmed by another senior radiologist with chest radiograph experience for more than 10 years.

Radiomics feature extraction and model building

We selected 196 cases as the training dataset (96 / 100 = TBG / LAC) and 84 cases as the testing dataset (41 / 43 = TBG / LAC). 851 radiomics features were extracted from each ROI and divided into three main categories : (1) First-order features. (2) Shape characteristics. (3) Texture features, including gray level co-occurrence matrix (GLCM) features, grey-level run-length matrix (GLRLM) features, gray level size zone matrix (GLSZM) features, neighborhood grey tone difference matrix (NGTDM) features, and grey level dependence matrix (GLDM) features. Figure 2 showed how Grey Level Histogram worked. FAE applied uniformization automatically to the feature matrix when preprocessing CT data, where each feature vector subtracted its average value and then divided by its length. Since the dimensional feature space was very high, the similarity of each feature pair was compared. If the Pearson Correlation Coefficient (PCC) of one feature pair was greater than 0.99, one of them from the pair was removed. After this preprocessing procedure, the size of the feature space was reduced, and each feature was independent of another. Kruskal Wallis was utilized to explore the important features corresponding to labels. In the FAE software, Pearson and Kruskal Wallis methods were automatically selected in the FAE software and we applied them to the training dataset. To evaluate the relationship between features and labels, we calculated the F value. Afterward, we ranked the top 14 features according to the corresponding F value. These 14 features were chosen by the FAE software based on the highest F value. Eventually, Random Forest Model with the highest AUC value was chosen automatically by FAE software as a classifier from all existing models including Support Vector Machine (SVM), Latent Dirichlet Allocation (LDA), Autoencoder (AE), Random Forest, Logistic Regression-Lasso, Adaboost, Decision Tree, Gaussian Process, Naive Bayes. To determine the hyperparameters of the model (eg. The number of features), we applied 10 times cross-validation on the training dataset. Therefore, hyperparameters were set according to the model performance on the validation dataset. (Figure 1.).

Statistical analysis

We used the Statistical Program for Social Science (SPSS, version 16.0) to test statistical differences in clinical features between LAC and TBG groups. The independence of categorical variables was examined by the Chi-square test and Fisher exact test. To test the continuous variables with normal distribution, a t-test was conducted (P < 0.05 indicates statistical significance). We used the Chi-square test for categorical variables such as location, smoking, and other clinical features. The performance of the model and quantitative analysis were evaluated by the ROC curve and AUC (Figure 1D), respectively. Sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated when the Youden index was maximized to its cut-point value. We estimated 95 % confidence intervals for 1000 samples by bootstrapping. All the above processes were operated via FeAture Explorer Pro (FAEPro, V0.3.5, Figure 1E) on Python (3.7.6) according to the software operation reference related literature [17].

Results

Clinical features

Table 1 listed the statistical test results in the training dataset and testing dataset. There were 196 patients in the training dataset, including 96 males (age range: 40 – 79 years old, average age: 64.53 ± 9.21 years old) and 100 females (age range: 33-72 years old, mean age: 56.06 ± 10.98 years). The testing dataset included 84 patients with 42 males (age range: 41 – 79 years old, average age: 63.71 ± 10.22 years old) and 42 females (age range: 33-73 years, mean age: 58.65 ± 10.71 years). Patients' gender and age were significantly different in the LAC group and TBG group in both datasets (Training dataset, Gender: P=0.001, Age: P=0.006; Testing dataset, Gender: P=0.016, Age: P=0.005). However, TB and LAC were indistinguishable by some clinical features such as smoking status. For example, there was no statistical difference between smoking history and patients' LAC or TB status (Training dataset, P = 0.15; Testing dataset, P = 0.536). In CT appearance, the lobulated shape was found to show a significant difference in the LAC group and TBG group in the training dataset (P = 0.03) and the testing dataset (P = 0.030). The rest CT features did not show any statistical difference in two groups, including size (Training dataset, P = 0.60; Testing dataset, P = 0.67), location (Training dataset, P = 0.910; Testing dataset, P = 0.43), spiculated sign (Training dataset, P = 0.97; Testing dataset, P = 0.79), vessel convergence (Training dataset, P=0.40; Testing dataset, P=0.43), and pleural indentation (Training dataset, P=0.34; Testing dataset, P=0.85. These CT features were not distinguishable between LAC and TB in the model.

Feature Selection and Radiological Model Construction

Table 2 illustrated the prediction performance of the training dataset and testing dataset. The accuracy of the training data set was 0.781, AUC was 0.834 (95 % Confidence Interval=0.7712 – 0.887), NPV was 0.782, PPV was 0.779, sensitivity was 0.771, and specificity was 0.790. Accuracy of the testing dataset was 0.726, AUC was 0.751 (95 % confidence interval=0.6382 – 0.853), NPV was 0.794, PPV was 0.680, sensitivity was 0.829, and specificity was 0.628. Table 3 showed features with the 14 highest AUC values on the testing dataset (Table 3 and Figure 3). In addition, the ROC curve was shown in Figure 4 (Training dataset AUC=0.834; Testing dataset AUC=0.751).

Discussion

The paper discussed a non-invasive diagnostic method for distinguishing non-calcifying tuberculosis granuloma from lung adenocarcinoma. The results of this study showed that age, gender, and lobulation were important predictors for distinguishing the LAC group and TB group [18]. On the one hand, the average age of patients in the TB group was lower than that in the LAC group, which may be explained by the fact that LAC is a malignant tumor, which is common in elderly patients. On the other hand, the number of female patients in the LAC group was more than that in the TB group, whereas the number of male patients in the LAC group was more than that in the TB group. The gender imbalance in the two groups may lead to statistical differences. It could be explained by the fact that females are prone to LAC compared to males, and males are more susceptible to TB compared to females[24]. Radiomics is a process that transforms the subjective evaluation of images into objective quantitative data. Many studies have shown that it can be used as a non-invasive method to predict the benign and malignant effects of pulmonary nodules [19, 25,26]. These objective data cannot be identified visually but can be determined in a computer-aided manner. The CT appearance ‘lobulated shape’ in this study was statistically different in both groups. This feature can reflect the heterogeneity within pulmonary nodules and help to identify benign and malignant nodules [20]. In this study, 196 cases were selected as the training dataset and 84 cases were chosen as the testing dataset. 851 radiomics features were extracted from each ROI randomly and automatically by software, including 18 first-order features, 14 shape features, 24 gray level co-occurrence matrices, 16 gray area size matrices, 16 gray level travel matrices, 5 domain gray difference matrices, 14 gray level correlation matrices, and 744 wavelet features. The features were sorted according to the corresponding F value, and the first 14 features are selected according to the verification performance. There are three firstorder features, including original _ firstorder _ 90Percentile, original _ firstorder _ Energy and original _ firstorder _ Mean. The first-order features stand for the difference in the distribution of individual prime parameter values, which reflects the difference in the density of lesions. This is the density difference in internal space between lung adenocarcinoma and non-calcified granuloma, which is difficult to identify from the eyes since it is a high-dimensional spatial feature. These features are related to gray matrix parameters. This indicates that the change of gray level in CT images of lung lesions may potentially contribute to the differential diagnosis of lung adenocarcinoma and non-calcified granuloma [18]. Random forest was used as a classifier in the model because of its highest AUC value among all models. Lung cancer and granuloma were commonly found in the upper lobe in this study. This may be due to changes in lobulation caused by lung cancer infiltration. However, chronic inflammation may also have similar characteristics. This could explain the reason for the relatively low AUC in the results. The AUC of the training dataset and the testing dataset were 0.834 and 0.751, respectively. The AUC of the training dataset is 0.834 compared to 0.751 in the AUC of the testing dataset. The NPV, PPV, sensitivity, and specificity have high similarities when compared to previous studies of its kind [21, 22]. It may not be appropriate to observe lung cancer for a long time without providing treatment, but suspected nodules that grow slowly are not easily identifiable with imaging studies without a sufficient waiting period. In addition, lung cancer and granuloma cannot be accurately distinguished in PET scans as well [23]. Although the gold standard for lung cancer diagnosis is the surgical biopsy, it is considered overtreatment if the nodule is a granuloma. On the contrary, conservative treatment may delay the timely treatment for lung cancer. Overall, it is difficult to distinguish benign and malignant pulmonary nodules merely using a lung CT scan. Physicians have been seeking a non-invasive examination to solve this problem. Radiomics, in combination with clinical features, shows its potential to be used as an effective tool to assist radiologists to distinguish benign and malignant pulmonary nodules. However, we have several limitations in this study. Firstly, it was a retrospective analysis. The sample size was relatively small and selection bias could be a potential issue. More high-quality samples are needed to prove the validity of the study in the future. Secondly, selected patients who had surgeries were more likely to be patients diagnosed with malignant tumors. Future research should maintain a relatively equal number of pathology results in both LAC and TB groups. Thirdly, different CT scans may affect the quality of image parameters. Therefore, thin-layer CT scanning (with a value of 0.625 mm) was adopted, and Radiomics normalization preprocessing was used to improve the quality of the data.

Conclusions

In summary, radiomics combined with clinical features is a possible non-invasive tool to distinguish non-calcifying tuberculosis granuloma and lung adenocarcinoma in small pulmonary nodules. The application of this combination has a great potential to decrease overdiagnosis and overtreatment in the future.

References

1. MacNeil, A., Glaziou, P., Sismanidis, C., Date, A., Maloney, S., & Floyd, K. (2020). Global Epidemiology of Tuberculosis and Progress Toward Meeting Global Targets - Worldwide, 2018. *MMWR. Morbidity and mortality weekly report*, *69*(11), 281–285. https://doi.org/10.15585/mmwr.mm6911a2
2. Thwaites, G., & Nahid, P. (2020). Triumph and Tragedy of 21st Century Tuberculosis Drug Development. *The New England journal of medicine*, *382*(10), 959–960. https://doi.org/10.1056/NEJMe2000860
3. Reid, M., Arinaminpathy, N., Bloom, A., Bloom, B. R., Boehme, C., Chaisson, R., Chin, D. P., Churchyard, G., Cox, H., Ditiu, L., Dybul, M., Farrar, J., Fauci, A. S., Fekadu, E., Fujiwara, P. I., Hallett, T. B., Hanson, C. L., Harrington, M., Herbert, N., Hopewell, P. C., … Goosby, E. P. (2019). Building a tuberculosis-free world: The Lancet Commission on tuberculosis. *Lancet (London, England)*, *393*(10178), 1331–1384. https://doi.org/10.1016/S0140-6736(19)30024-8
4. Yuan, T., & Sampson, N. S. (2018). Hit Generation in TB Drug Discovery: From Genome to Granuloma. *Chemical reviews*, *118*(4), 1887–1916. https://doi.org/10.1021/acs.chemrev.7b00602
5. Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: a cancer journal for clinicians*, *71*(3), 209–249. https://doi.org/10.3322/caac.21660
6. Rami-Porta, R., Asamura, H., Travis, W. D., & Rusch, V. W. (2017). Lung cancer - major changes in the American Joint Committee on Cancer eighth edition

cancer staging manual. *CA: a cancer journal for clinicians*, *67*(2), 138–155. https://doi.org/10.3322/caac.21390

7. Fischer, B. M., Lassen, U., & Højgaard, L. (2011). PET-CT in preoperative staging of lung cancer. *The New England journal of medicine*, *364*(10), 980–981. https://doi.org/10.1056/NEJMc1012974
8. McWilliams, A., Tammemagi, M. C., Mayo, J. R., Roberts, H., Liu, G., Soghrati, K., Yasufuku, K., Martel, S., Laberge, F., Gingras, M., Atkar-Khattra, S., Berg, C. D., Evans, K., Finley, R., Yee, J., English, J., Nasute, P., Goffin, J., Puksa, S., Stewart, L., … Lam, S. (2013). Probability of cancer in pulmonary nodules detected on first screening CT. *The New England journal of medicine*, *369*(10), 910–919. https://doi.org/10.1056/NEJMoa1214726
9. Siegel, R. L., Miller, K. D., & Jemal, A. (2019). Cancer statistics, 2019. *CA: a cancer journal for clinicians*, *69*(1), 7–34. https://doi.org/10.3322/caac.21551
10. Pisano, C., O'Connor, J., Krick, S., & Russell, D. W. (2020). A Fatal Case of Pneumocephalus during Computed Tomography-guided Lung Biopsy. *American journal of respiratory and critical care medicine*, *201*(12), e83–e84. https://doi.org/10.1164/rccm.201902-0280IM
11. Huo, J., Xu, Y., Sheu, T., Volk, R. J., & Shih, Y. T. (2019). Complication Rates and Downstream Medical Costs Associated With Invasive Diagnostic Procedures for Lung Abnormalities in the Community Setting. *JAMA internal medicine*, *179*(3), 324–332. https://doi.org/10.1001/jamainternmed.2018.6277
12. Bi, W. L., Hosny, A., Schabath, M. B., Giger, M. L., Birkbak, N. J., Mehrtash, A., Allison, T., Arnaout, O., Abbosh, C., Dunn, I. F., Mak, R. H., Tamimi, R. M., Tempany, C. M., Swanton, C., Hoffmann, U., Schwartz, L. H., Gillies, R. J., Huang, R. Y., & Aerts, H. (2019). Artificial intelligence in cancer imaging: Clinical challenges and applications. *CA: a cancer journal for clinicians*, *69*(2), 127–157. https://doi.org/10.3322/caac.21552
13. Peikert, T., Bartholmai, B. J., & Maldonado, F. (2020). Radiomics-based Management of Indeterminate Lung Nodules? Are We There Yet?. *American journal of respiratory and critical care medicine*, *202*(2), 165–167. https://doi.org/10.1164/rccm.202004-1279ED
14. Mu, W., Jiang, L., Zhang, J., Shi, Y., Gray, J. E., Tunali, I., Gao, C., Sun, Y., Tian, J., Zhao, X., Sun, X., Gillies, R. J., & Schabath, M. B. (2020). Non-invasive decision support for NSCLC treatment using PET/CT radiomics. *Nature communications*, *11*(1), 5228. https://doi.org/10.1038/s41467-020-19116-x
15. Hosny, A., Parmar, C., Coroller, T. P., Grossmann, P., Zeleznik, R., Kumar, A., Bussink, J., Gillies, R. J., Mak, R. H., & Aerts, H. (2018). Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. *PLoS medicine*, *15*(11), e1002711. https://doi.org/10.1371/journal.pmed.1002711
16. Grossmann, P., Stringfield, O., El-Hachem, N., Bui, M. M., Rios Velazquez, E., Parmar, C., Leijenaar, R. T., Haibe-Kains, B., Lambin, P., Gillies, R. J., & Aerts, H. J. (2017). Defining the biological basis of radiomic phenotypes in lung cancer. *eLife*, *6*, e23421. https://doi.org/10.7554/eLife.23421
17. Song, Y., Zhang, J., Zhang, Y. D., Hou, Y., Yan, X., Wang, Y., Zhou, M., Yao, Y. F., & Yang, G. (2020). FeAture Explorer (FAE): A tool for developing and

comparing radiomics models. *PloS one*, *15*(8), e0237587. https://doi.org/10.1371/journal.pone.0237587

18. Cui, E. N., Yu, T., Shang, S. J., Wang, X. Y., Jin, Y. L., Dong, Y., Zhao, H., Luo, Y. H., & Jiang, X. R. (2020). Radiomics model for distinguishing tuberculosis and lung cancer on computed tomography scans. *World journal of clinical cases*, *8*(21), 5203–5212. https://doi.org/10.12998/wjcc.v8.i21.5203
19. Feng, B., Chen, X., Chen, Y., Lu, S., Liu, K., Li, K., Liu, Z., Hao, Y., Li, Z., Zhu, Z., Yao, N., Liang, G., Zhang, J., Long, W., & Liu, X. (2020). Solitary solid pulmonary nodules: a CT-based deep learning nomogram helps differentiate tuberculosis granulomas from lung adenocarcinomas. *European radiology*, *30*(12), 6497–6507. https://doi.org/10.1007/s00330-020-07024-z
20. Jiang, Y., Che, S., Ma, S., Liu, X., Guo, Y., Liu, A., Li, G., & Li, Z. (2021). Radiomic signature based on CT imaging to distinguish invasive adenocarcinoma from minimally invasive adenocarcinoma in pure ground-glass nodules with pleural contact. *Cancer imaging : the official publication of the International Cancer Imaging Society*, *21*(1), 1. https://doi.org/10.1186/s40644-020-00376-1
21. .Feng, B., Chen, X., Chen, Y., Liu, K., Li, K., Liu, X., Yao, N., Li, Z., Li, R., Zhang, C., Ji, J., & Long, W. (2020). Radiomics nomogram for preoperative differentiation of lung tuberculoma from adenocarcinoma in solitary pulmonary solid nodule. *European journal of radiology*, *128*, 109022. https://doi.org/10.1016/j.ejrad.2020.109022
22. Chen, X., Feng, B., Chen, Y., Liu, K., Li, K., Duan, X., Hao, Y., Cui, E., Liu, Z., Zhang, C., Long, W., & Liu, X. (2020). A CT-based radiomics nomogram for prediction of lung adenocarcinomas and granulomatous lesions in patient with solitary sub-centimeter solid nodules. *Cancer imaging : the official publication of the International Cancer Imaging Society*, *20*(1), 45. https://doi.org/10.1186/s40644-020-00320-3
23. Du, D., Gu, J., Chen, X., Lv, W., Feng, Q., Rahmim, A., Wu, H., & Lu, L. (2021). Integration of PET/CT Radiomics and Semantic Features for Differentiation between Active Pulmonary Tuberculosis and Lung Cancer. *Molecular imaging and biology*, *23*(2), 287–298. https://doi.org/10.1007/s11307-020-01550-4
24. Marçôa, R., Ribeiro, A. I., Zão, I., & Duarte, R. (2018). Tuberculosis and gender - Factors influencing the risk of tuberculosis among men and women by age group. *Pulmonology*, *24*(3), 199–202. https://doi.org/10.1016/j.pulmoe.2018.03.004
25. Xu, Y., Lu, L., E, L. N., Lian, W., Yang, H., Schwartz, L. H., Yang, Z. H., & Zhao, B. (2019). Application of Radiomics in Predicting the Malignancy of Pulmonary Nodules in Different Sizes. *AJR. American journal of roentgenology*, *213*(6), 1213–1220. https://doi.org/10.2214/AJR.19.21490
26. Wilson, R., & Devaraj, A. (2017). Radiomics of pulmonary nodules and lung cancer. *Translational lung cancer research*, *6*(1), 86–91. https://doi.org/10.21037/tlcr.2017.01.04
27. Eguchi, T., Kadota, K., Park, B. J., Travis, W. D., Jones, D. R., & Adusumilli, P. S. (2014). The new IASLC-ATS-ERS lung adenocarcinoma classification: what

the surgeon should know. *Seminars in thoracic and cardiovascular surgery*, *26*(3), 210–222. https://doi.org/10.1053/j.semtcvs.2014.09.002

Figure 1

Research method

Overview of research methods : A. Collection of chest CT data B. ROI delineation C. Feature extraction. The image is the gray scale histogram of the lesion D. Data analysis E. Operation using FAE software

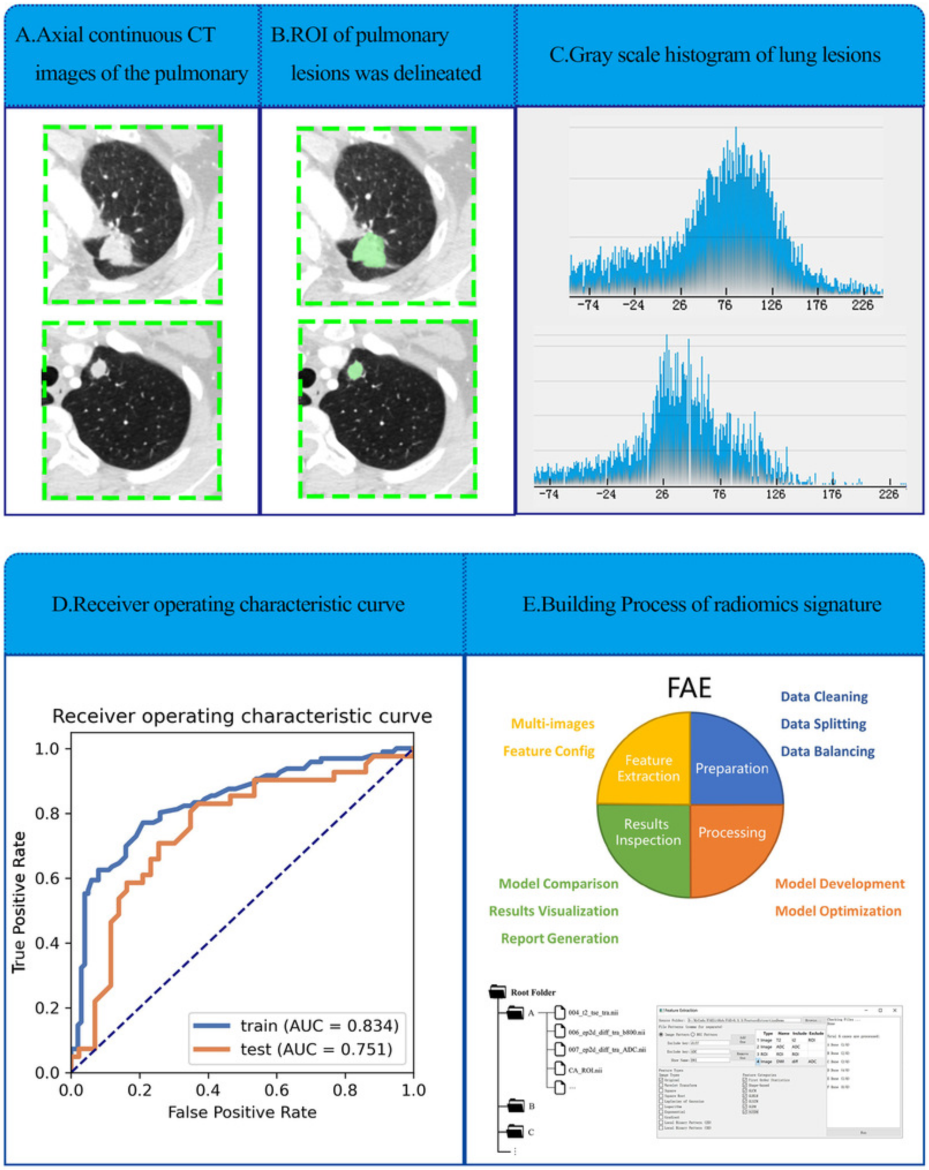
A.Axial continuous CT images of the pulmonary
B.ROI of pulmonary lesions was delineated
C.Gray scale histogram of lung lesions
−74
−24
26
76
126
176
226
D.Receiver operating characteristic curve
E.Building Process of radiomics signature
Receiver operating characteristic curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
train (AUC = 0.834)
test (AUC = 0.751)
FAE
Multi-images
Feature Config
Feature Extraction
Preparation
Data Cleaning
Data Splitting
Data Balancing
Results Inspection
Processing
Model Comparison
Results Visualization
Report Generation
Model Development
Model Optimization
Root Folder
A
004_t2_tse_tra.nii
006_ep2d_diff_tra_b800.nii
007_ep2d_diff_tra_ADC.nii
CA_ROI.nii
...
B
C

Figure 2

CT imagines

CT images showed lung adenocarcinoma (LAC) and non-calcified tuberculous granuloma (TB); 1a and 2a : CT scan showed irregular solid nodules (red area) in the left upper lobe; 1b and 2b : gray scale histogram of the nodule; 1c : LAC with hematoxylin and eosin (H&E) stain, x 400; 2c : TB with hematoxylin and eosin (H&E) stain, x 400

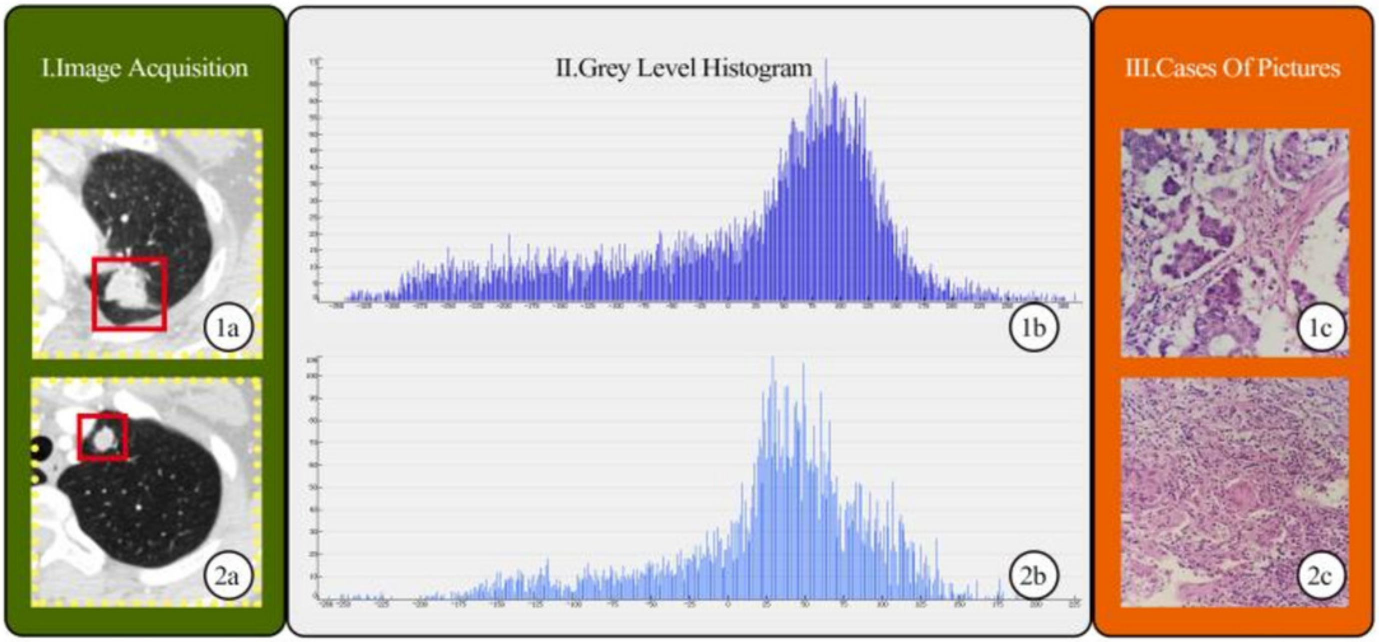

Figure 3

features selection

14 features selection (above the yellow line)

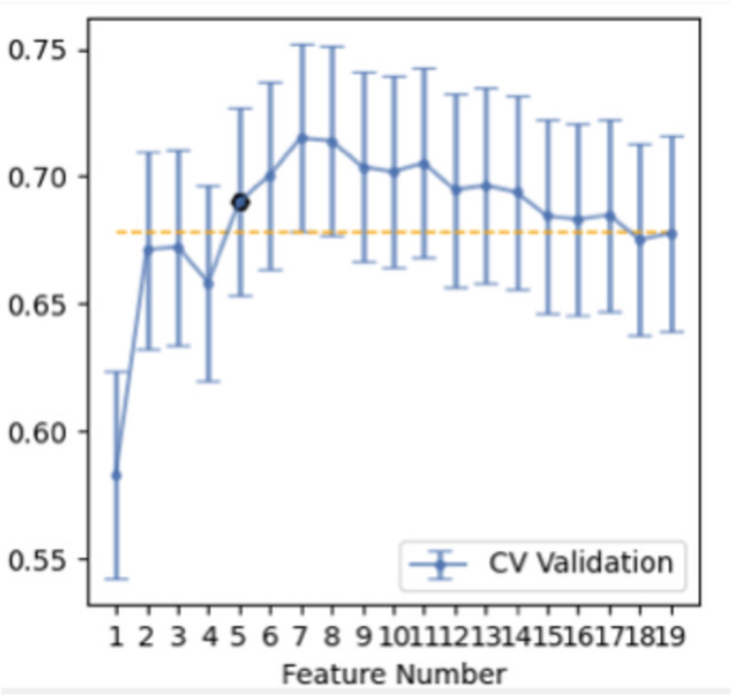
0.75
0.70
0.65
0.60
0.55
CV Validation
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
Feature Number

Figure 4

ROC curve selection

ROC curve of the model

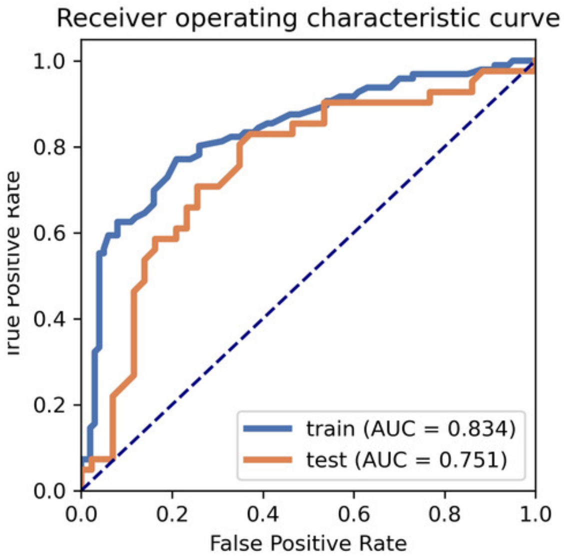
Receiver operating characteristic curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
train (AUC = 0.834)
test (AUC = 0.751)

Table 1(on next page)

Clinical characteristics and CT findings in LAC and TB

Note: The differences were assessed with the Wilcoxon rank sum test or Pearson chi-squared test CT:computed tomography, LAC:lung adenocarcinoma, TB:pulmonary tuberculosis, SD:standard deviation *P < 0.05

Characteristic	Training data set (n-=196)		P	Test data set(n=84)		P
	LAC(100)	TB(96)		LAC(43)	TB(41)	
Gender			*0.001			*0.016
Male	37	59		16	26	
Female	63	37		27	15	
Age (mean ± SD, years)	64.53±9.21	56.06±10.98	*0.006	63.71±10.22	58.65±10.71	*0.005
Smoking history			0.148			0.536
Absence	69	75		29	25	
Presence	31	21		14	16	
Size (mean ± SD, mm)	19.81±7.47	18.69±5.44	0.595	20.71±7.62	19.03±9.01	0.667
Location			0.910			0.425
Upper and middle	68	66		28	30	
Lower	32	30		15	11	
Spiculated sign			0.967			0.791
Absence	57	55		25	25	
Presence	43	41		18	16	
Lobulated shape			*0.034			*0.030
Absence	36	49		14	23	
Presence	64	47		29	18	
Vessel convergence			0.400			0.425
Absence	44	48		22	21	

Presence	56	48		21	20	
Pleural indentation			0.337			0.884
Absence	30	35		13	13	
Presence	70	61		30	28	

Table 1. Clinical characteristics and CT findings in LAC and TB

Note: The differences were assessed with the Wilcoxon rank sum test or Pearson chi-squared test

CT:computed tomography, LAC:lung adenocarcinoma, TB:pulmonary tuberculosis, SD:standard deviation

*P < 0.05

Table 2(on next page)

Clinical statistics in the diagnosis

	Accuracy	AUC	AUC 95% CIs	NPV	PPV	Sensitivity	Specificity
Training data set	0.7806	0.8344	0.7712-0.8872	0.7822	0.7789	0.7708	0.7900
Test data set	0.7262	0.751	0.6382-0.8531	0.7941	0.68	0.8293	0.6279

Table 2. Clinical statistics in the diagnosis

Table 3(on next page)

The rank of selected features

Features	Rank
original_firstorder_90Percentile	1
original_firstorder_Energy	2
original_firstorder_Mean	3
wavelet-HHL_firstorder_Median	4
wavelet-HHL_glcm_ClusterProminence	5
wavelet-HHL_glcm_Imc1	6
wavelet-HHL_glcm_Imc2	7
wavelet-HHL_gldm_DependenceEntropy	8
wavelet-HHL_glrlm_RunEntropy	9
wavelet-HHL_glszm_GrayLevelNonUniformityNormalized	10
wavelet-HHL_glszm_SizeZoneNonUniformityNormalized	11
wavelet-HHL_ngtdm_Busyness	12
wavelet-HHL_ngtdm_Strength	13
wavelet-LLH_glcm_MCC	14

Table 3. The rank of selected features