Prediction of metastases in confusing mediastinal lymph nodes based on flourine-18 fluorodeoxyglucose (^{18}F-FDG) positron emission tomography/computed tomography (PET/CT) imaging using machine learning

Siqin Dong[1#], Ao Fu[2#], Jiacheng Liu[3^]

[1]Jiangsu Key Laboratory of Molecular and Functional Imaging, Medical School, Southeast University, Nanjing, China; [2]Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, Nanjing, China; [3]Department of Nuclear Medicine, Jiangsu Key Laboratory of Molecular and Functional Imaging, Zhongda Hospital, Medical School, Southeast University, Nanjing, China

Contributions: (I) Conception and design: J Liu; (II) Administrative support: J Liu; (III) Provision of study materials or patients: J Liu, S Dong; (IV) Collection and assembly of data: S Dong; (V) Data analysis and interpretation: S Dong, A Fu; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work as co-first authors.

Correspondence to: Jiacheng Liu, PhD. Department of Nuclear Medicine, Jiangsu Key Laboratory of Molecular and Functional Imaging, Zhongda Hospital, Medical School, Southeast University, 87 Dingjiaqiao Road, Nanjing 210009, China. Email: jiachengliu@seu.edu.cn.

Background: For patient management and prognosis, accurate assessment of mediastinal lymph node (LN) status is essential. This study aimed to use machine learning approaches to assess the status of confusing LNs in the mediastinum using positron emission tomography/computed tomography (PET/CT) images; the results were then compared with the diagnostic conclusions of nuclear medicine physicians.

Methods: A total of 509 confusing mediastinal LNs that had undergone pathological assessment or follow-up from 320 patients from three centres were retrospectively included in the study. LNs from centres I and II were randomised into a training cohort (N=324) and an internal validation cohort (N=81), while those from centre III patients formed an external validation cohort (N=104). Various parameters measured from PET and CT images and extracted radiomics and deep learning features were used to construct PET/CT-parameter, radiomics, and deep learning models, respectively. Model performance was compared with the diagnostic results of nuclear medicine physicians using the area under the curve (AUC), sensitivity, specificity, and decision curve analysis (DCA).

Results: The coupled model of gradient boosting decision tree-logistic regression (GBDT-LR) incorporating radiomic features showed AUCs of 92.2% [95% confidence interval (CI), 0.890–0.953], 84.6% (95% CI, 0.761–0.930) and 84.6% (95% CI, 0.770–0.922) across the three cohorts. It significantly outperformed the deep learning model, the parametric PET/CT model and the physician's diagnosis. DCA demonstrated the clinical usefulness of the GBDT-LR model.

Conclusions: The presented GBDT-LR model performed well in evaluating confusing mediastinal LNs in both internal and external validation sets. It not only crossed radiometric features but also avoided overfitting.

Introduction

Mediastinal lymph nodes (LNs) are the regional LNs that harbour metastases from thoracic tumours such as lung or oesophageal cancer, and while mediastinal metastases from nonthoracic tumours are less common than that from thoracic tumours, they can also be used as an indication of distant metastases (1). Chest computed tomography (CT) is the standard imaging modality used to assess mediastinal LNs, but it is of limited use in assessing LN status. CT assessment of LNs is based on LN size and morphology only and has low sensitivity and specificity (2,3). Endobronchial ultrasound-guided transbronchial needle aspiration biopsy (EBUS-TBNA) is commonly used for pathological confirmation of LN metastases, but not only is the process invasive, but the relatively limited amount of material that can be aspirated through the needle may limit its diagnostic ability in other mediastinal lesions, such as lymphoma and nodal disease (4). By combining the anatomical information from CT with the functional information from positron emission tomography (PET), PET/CT has emerged as a widely used modality in the diagnosis, staging, follow-up, treatment and prognosis of tumours (5,6). In previous studies, flourine-18 fluorodeoxyglucose (^{18}F-FDG) PET/CT has shown to be efficacious in the detection of mediastinal LN metastases in patients with lung, oesophageal and breast cancer (7-10). In clinical practice, a 2.5 maximum standardized uptake value (SUVmax) threshold is commonly used (11), while a short axis >1 cm is commonly used on CT (12). However, many infections and inflammatory or neoplastic conditions can cause enlarged mediastinal LNs or FDG uptake (13), leading to false positives. PET/CT should not be used for mediastinal LN staging in areas where sarcoidosis is endemic or in patients with pneumoconiosis and lung cancer because of the high rate of false-positive results (14,15).

Recently, artificial intelligence (AI) has seen extensive applications in various fields, particularly in the field of medical imaging. Several studies have shown that machine learning is effective in distinguishing between benign and malignant mediastinal LNs (12,16,17). Recently, deep learning methods have shown positive results in tumour segmentation, histological subtype classification, diagnosis and prognosis (18-21). A recent study showed that a deep learning method based on enhanced CT imaging performs well in predicting mediastinal LN metastasis in lung cancer patients (22). However, it is widely recognised that deep learning techniques offer significant advantages for large datasets but are prone to overfitting in smaller datasets (23). For medical images with small datasets, radiomics is more suitable. Recently, some researchers have used radiomics based on CT or PET/CT images for the diagnosis of benign and malignant LNs (12,24,25). Most studies of LN benignity and malignancy are single-centre studies, and so reproducibility and generalisability are not guaranteed. Furthermore, most studies focus on the predictive power of radiological features and often use linear models to fit features. This approach may neglect to consider how different features interact and merge. However, some machine learning algorithms that can directly handle feature interactions may overfit. To address the above limitations, our study included patients from three centres and used multiple machine learning algorithms to build diagnostic models, culminating in the gradient boosting decision tree-logistic regression (GBDT-LR) algorithm, which addresses both feature crossing and overfitting. Finally, in assessing the benign and malignant nature of confusing mediastinal LNs, we also compared the performance of deep learning models, radiomics models, parametric PET/CT models and physician diagnosis. We present this article in accordance with the TRIPOD reporting checklist (available at https://qims.amegroups.com/article/view/10.21037/qims-24-100/rc).

Methods

Patients

Between June 2016 and July 2023, a total of 320 patients from three centres (Zhongda Hospital, Shengjing
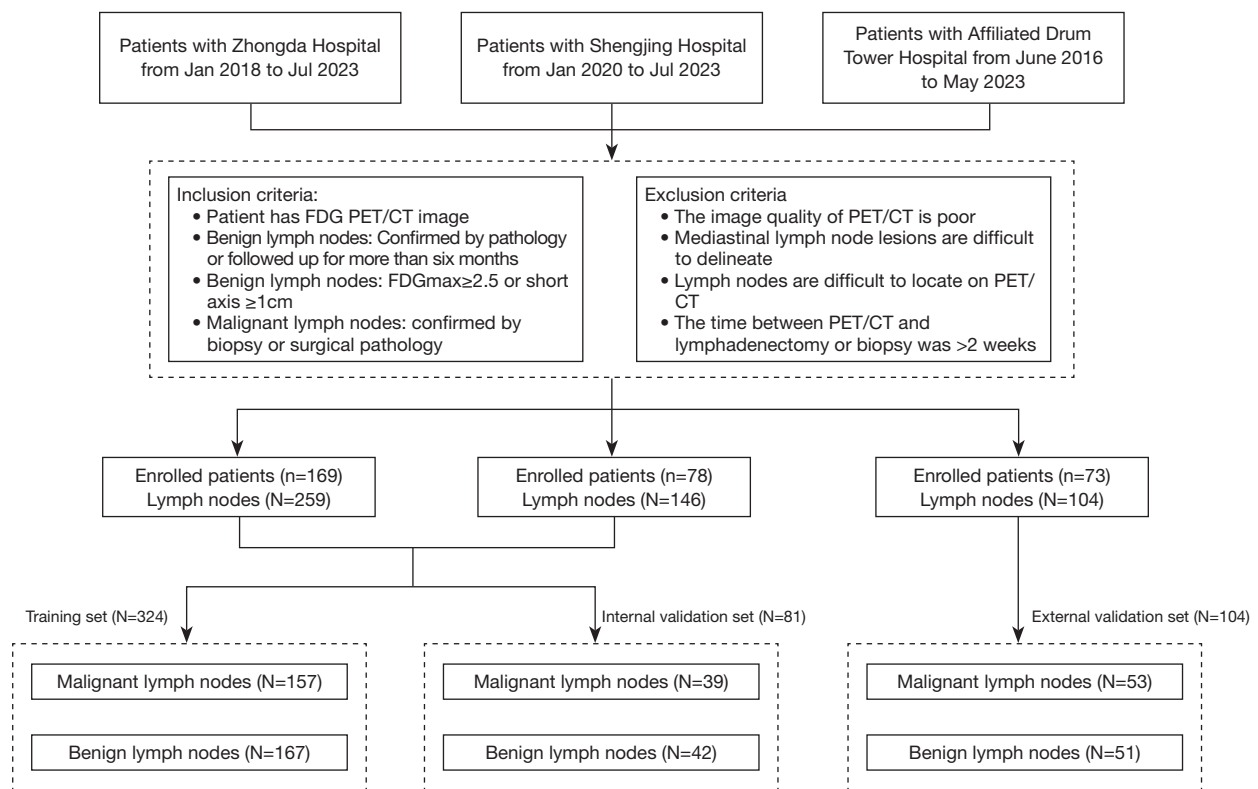
Figure 1 The workflow diagram of patient recruitment. In the figure, n is the number of patients and N is the number of confusing mediastinal lymph nodes. FDG, fluorodeoxyglucose; PET/CT, positron emission tomography/computed tomography.

Hospital and Affiliated Drum Tower Hospital) were retrospectively enrolled. This retrospective study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by the Institutional Review Board (IRB) of Zhongda Hospital (No. 2021ZDSYLL239-P01). The Affiliated Drum Tower Hospital and Shengjing Hospital were informed and agreed with the study. The requirement of individual consent for this retrospective analysis was waived. We included (I) surgery- or biopsy-confirmed malignant mediastinal LNs; (II) benign mediastinal LNs confirmed pathologically or followed up for >6 months (III) benign mediastinal LNs with an SUVmax ≥2.5 or short axis ≥1 cm; and (IV) PET/CT scans less than 2 weeks from the time of surgery or biopsy. The exclusion criteria were as follows: (I) poor PET/CT quality. (II) Difficulty locating LNs on PET/ CT or delineating LN lesions. (III) Patients with a primary malignant tumour and a negative pathological diagnosis of EBUS-TBNA in the mediastinal LNs. (IV) Patients who had enlarged LNs or who had a primary malignancy at the 6-month follow-up. We selected 1–3 LNs for each patient based on the pathology results related to the LNs. If a patient had metastatic LNs, no more benign LNs were included for that patient. In previous studies, the criterion for assessing LNs by CT was that a short axis diameter (SAD) of ≥1 cm was considered malignant, whereas the clinical criterion for assessing LNs by PET imaging was that an SUVmax ≥2.5 was considered positive for PET and an SUVmax <2.5 was considered negative for PET. However, the size and SUVmax of normal, inflammatory proliferative, and metastatic LNs partly overlapped and could be easily confused. Therefore, in this paper we defined LNs with a SAD greater than 1 cm or an SUVmax greater than 2.5 as confusing LNs. The classification of LNs as benign or malignant was based on pathological diagnosis and follow-up findings, which were considered the gold standard. Ultimately, 509 confusing LNs were included in the study. To train the model, LNs from Centre I and Centre II patients were randomly allocated 4:1 to training and internal validation sets. Using data from Centre III, we validate the model. *Figure 1* showed the patient recruitment.

4726

Dong et al. Using machine learning to evaluate LNs status in PET/CT

^{18}F-FDG PET/CT analysis

In total, three PET/CT scanners acquired PET/CT images at the three centres, with acquisition parameters reported in the Appendix 1.

Two nuclear medicine specialists, who were unaware of the clinicopathological information, evaluated the PET/CT images. Parameters analysed subsequently included LN SAD, density, CT values, SUVmax, minimal SUV (SUVmin), average SUV (SUVavg), peak SUV (SUVpeak), metabolic tumour volume (MTV) and total lesion glycolysis (TLG). MedEx and LIFEx postprocessing software were used to perform parametric measurements of the LNs. Two nuclear medicine physicians evaluated the images according to the following criteria throughout the study period (3,5,7,26). LNs with increased glucose uptake and a distinct margin were considered malignant. If the LNs had increased glucose uptake, it was assumed that they had a higher level of ^{18}F-FDG uptake than the tissue in the mediastinum. Even if ^{18}F-FDG uptake was high (higher than background activity), calcified LNs, LNs with higher attenuation than the surrounding macrovessels, or HUmax >120 on CT images with integrated PET/CT were considered benign. If disagreements arose, a consensus was reached by discussion.

Radiomics signature development

Image segmentation was performed by a trained nuclear medicine specialist using the open-source software 3D Slicer (version 5.0.3) for CT images and LIFEx (version v7.3.6) for PET images, as detailed in the Appendix 1. PyRadiomics (https://pyradiomics.readthedocs.io/en/latest/index.html) software was used to extract radiomics features from the outlined mediastinal LNs. The Image Biomarker Standardisation Initiative (IBSI) was used as a reference. It was accounted for in the extraction and selection of radiomics features (27). To enhance the robustness of the results, we employed 10-fold cross-validation prior to modelling to obtain the most effective predictive features. We also ensured the reliability and repeatability of the results through a rigorous process of feature extraction and screening throughout the study. See Appendix 1 for more details.

Deep learning signature development

Our mediastinal data on confusing LNs contained an annotation at the pixel level for each LN, for which a corresponding square bounding box was generated. Specifically, we cropped and resampled to 256×256 pixels using bilinear interpolation by creating a minimum square box centred around the centre of the LN annotation to ensure it wrapped around the entire node. We had also linked PET, CT and labelling in the dimension of the channel, as well as the conditions that were used to increase the input.

We used an end-to-end deep learning model and augmentation to increase the amount of data. The ResNet18 network pretrained on ImageNet in torchvision was used for training. To prevent model overfitting, we also added a dropout layer to the model. The predicted probability that the LN was malignant was output from the 3-channel images. Details can be found in the Appendix 1.

Harmonisation

Radiomics features were affected by differences in the scanning machine, acquisition parameters, reconstruction algorithms, number of iterations and voxel size (28). The variability of radiomics features results in a central effect in multicentre radiomics studies. The ComBat harmonisation approach was prevalent in genomics, while more recent studies had standardised radiomic signatures in order to aid with multicentre studies (29-31). After feature extraction, we harmonised the radiomics features of the three centres using the ComBat method.

PET/CT parameter model

PET/CT parameters that were significant in univariate analysis were selected for the training cohort. To avoid multicollinearity, we excluded variables with a correlation coefficient of more than 0.7. The PET/CT parameter model was constructed by incorporating the remaining PET/CT parameters into a support vector machine (SVM) model.

Statistical analysis

Statistical analyses were carried out using SPSS (IBM, version 22.0, NY, USA) and R studio software (version 4.2.2). Depending on the data distribution, continuous variables were compared using independent samples t tests, one-way ANOVA or Mann-Whitney U tests. Comparisons of categorical variables were made using the Chi-squared

test or Fisher's exact test, as appropriate. Spearman's correlation coefficient was used to assess the correlations between the variables. Receiver operating characteristic (ROC) curves were plotted, and the area under the curve (AUC) as well as the sensitivity and specificity were calculated to assess the performance of the physicians' assessment, the parametric PET/CT model, the radiomics model and the deep learning model. AUCs were compared using DeLong's test. Decision curve analysis (DCA) was also performed as an assessment of the potential for clinical application of the diagnostic models.

Results

Clinical characteristics

A total of 509 confusing LNs were identified in 320 patients from three centres as part of this study. Institution I and II LNs were allocated to a training cohort (N=324) and an internal validation cohort (N=81), and Institution III LNs were used as an external validation cohort (N=104) using a 4:1 stratified sampling method. The differences in the clinicopathological characteristics between the three cohorts were not statistically significant. *Table 1* showed the pathological findings and clinical information of the confusing mediastinal LNs. There were no significant differences in the benign or malignant distribution of LNs (P=0.896), the percentage distribution of LNs larger than 1 cm (P=0.327), or the distribution of LN stations (P=0.164) among the three cohorts. There was a statistically significant difference in SAD (P<0.001) between malignant and benign LNs in the three cohorts, while no significant difference was observed for age (P=0.104) or sex (P=0.327).

Development of the radiomics signature and deep learning signature

From the PET and CT images, a total of 2,632 radiomics features were extracted (Figure S1). We removed features with an intraclass correlation coefficient (ICC) below 0.75, leaving more stable radiomics features. The distribution of radiomics features differed between the three facilities, but after ComBat harmonisation, the radiomics feature distributions were approximately similar. The minimum redundancy-maximum relevance (mRMR) (32) algorithm and the least absolute shrinkage and selection operator (LASSO) method were used to further refine the harmonised features. Five machine learning techniques,

linear regression (LR), decision tree (DT), SVM, Naïve Bayes (NB) and GBDT-LR, were used for radiomics-based modelling (see Appendix 1). We obtained the predicted probability that the LN was malignant from the last layer of the deep learning model.

PET/CT parameter model construction

LNs with SAD >1 cm were more likely to be malignant (P<0.001), but no correlation was observed between age, sex, LN location distribution and SUVmin and LN metastasis, as shown in *Table 1*. The SUVmax, SUVavg, SUVpeak, MTV and TLG were more likely to be higher in malignant LNs than in benign LNs (P<0.001). Calcified LNs were more likely to be benign LNs (P=0.033). Therefore, a SAD greater than 1 or less than or equal to 1, SUVmax, SUVavg, SUVpeak, MTV, TLG and the presence or absence of calcifications were candidates for the construction of the PET/CT parameter model (PA model). However, as SUVmax, SUVpeak and TLG were highly correlated with all the other parameters, they were excluded from the analysis (see Table S1). SAD, SUVavg, MTV and calcification were included as PET/CT parameters in the SVM model to construct the PET/CT parameter model.

Model performance evaluation

Among the radiomics models generated by the five machine learning algorithms, the AUCs of the DT, SVM and GBDT-LR models in the training set were all greater than 90%, as shown in Table S2 in the Appendix 1. However, the DT and SVM models performed much worse than the GBDT-LR algorithm and the logistic regression algorithm in the internal and external validation sets, and neither models were stable. Of the five models, the GBDT-LR model not only had good performance in the training set but also had an AUC of more than 80% in both the internal validation set and the external validation set.

Due to the high AUC values of the GBDT-LR model in the three cohorts, its diagnostic efficacy for metastatic LNs was compared with the physician's diagnosis and the PET/CT parameter model and the deep learning models, as shown in *Figure 2* and *Table 2*. Encouragingly, the AUC of the GBDT-LR model was significantly higher than that of the other three models in all three cohorts, and the sensitivity and specificity of the model ranged from 84.6% to 87.3% and from 73.8% to 89.8%, respectively.

The DCA curves in *Figure 3* and Figure S2; showed

Table 1 Summary of characteristics in training cohorts, internal validation cohort, and external validation cohort

Characteristics	Training set			Internal test set			External test set			P[a]
	LN (+)	LN (−)	P[b]	LN (+)	LN (−)	P[b]	LN (+)	LN (−)	P[b]	
Sex, n (%)			0.086			0.093			0.086	0.327
From male patients	103 (65.6)	94 (56.3)		24 (61.5)	18 (42.9)		38 (71.7)	25 (49.0)		
From female patients	54 (34.4)	73 (43.7)		15 (38.5)	24 (57.1)		15 (28.3)	26 (51.0)		
Age (years), mean ± SD	63±8	65±13	0.056	61±10	64±14	0.323	64±12	63±10	0.644	0.104
Clinical diagnosis, n (%)			N/A			N/A			N/A	N/A
Sarcoidosis	–	28 (16.8)		–	11 (26.2)		–	11 (21.6)		
Tuberculosis	–	3 (1.8)		–	3 (7.1)		–	–		
Esophageal cancer	–	2 (1.2)		1 (2.6)	1 (2.4)		1 (1.9)	1 (2.0)		
Breast cancer	–	3 (1.8)		–	1 (2.4)		–	–		
Lung biopsy	–	5 (2.3)		–	3 (7.1)		–	2 (3.9)		
Lung cancer	138 (87.9)	71 (42.5)		35 (90.0)	13 (31.0)		47 (88.7)	32 (62.7)		
Gastric cancer	1 (0.6)	7 (4.2)		–	3 (7.1)		3 (5.7)	–		
Others	18 (11.5)	48 (28.7)		3 (7.7)	7 (16.7)		2 (3.8)	5 (9.8)		
LN station, n (%)			0.728			0.450			<0.001	0.164
1–4	90 (57.3)	90 (53.9)		16 (41.0)	23 (54.8)		28 (52.8)	21 (41.2)		
5–6	15 (9.6)	20 (12.0)		3 (7.7)	2 (4.8)		4 (7.5)	6 (11.8)		
7–9	52 (35.7)	57 (34.1)		20 (51.3)	17 (40.5)		21 (39.6)	24 (47.1)		
SAD (cm), n (%)			<0.001			0.001			<0.001	0.212
≥1	120 (76.4)	83 (49.7)		34 (87.2)	22 (52.4)		49 (92.5)	25 (49.0)		
<1	37 (23.6)	84 (50.3)		5 (12.8)	20 (47.6)		4 (7.5)	26 (51.0)		

[a], P value indicates the significance of differences between the characteristics in training cohorts, internal validation cohorts and external validation cohorts; [b], P value indicates the significance of differences between the LN (+) group and LN (−) group. LN, lymph node; SD, standard deviation; N/A, not available; SAD, short-axis diameter.

that the GBDT-LR model had greater benefit than the 'all treatment' or 'no treatment' programmes when the threshold probability for the patient or clinician was greater than 20% and less than 80%. Compared to other methods, the GBDT-LR model had higher net gains.

Discussion

LNs are common in the mediastinum. The majority of benign LNs are reactive proliferative, inflammatory, tuberculous and granulomatous lesions, while the majority of malignant LNs are metastatic diseases or lymphomas (33,34). These LNs are sensitive to pathological and physiological changes. As a result, they often show increased

FDG uptake or enlargement on PET/CT imaging. This results in a substantial reduction in the accuracy of PET. For example, Onal and colleagues used PET/CT for evaluating isolated mediastinal LNs in cervical cancer patients and found a 75% false-positive rate (35). Thus, when encountering LNs with increased FDG uptake or enlargement, pathological confirmation is needed, which is not only invasive but also not possible for LNs in all locations (36). In this study, we selected benign LNs with SADs greater than 1 cm or SUVmax values greater than 2.5 and aimed to better evaluate these LNs using deep learning, radiomics and parametric PET/CT approaches.

The deep learning models had good performance on the training dataset but poor performance on the two
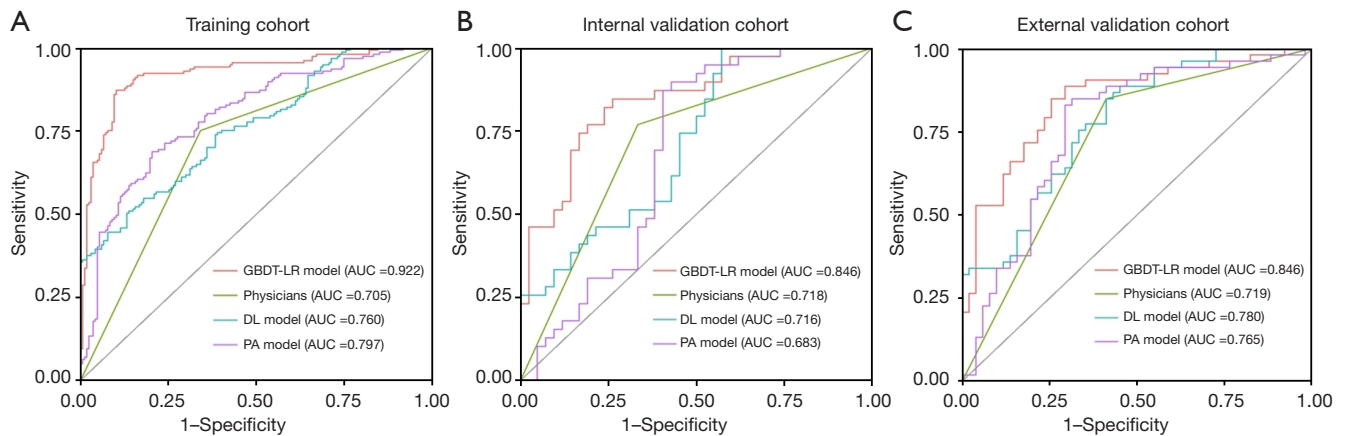
Figure 2 The receiver operating characteristic curves of GBDT-LR model, DL model, Physicians and PA model in training cohort (A), internal validation cohort (B), and external validation cohort (C). Number in parenthesis is the area under receiver operating characteristic curve. GBDT-LR, Gradient Boosting Decision Tree-Logistic Regression; AUC, area under the curve; DL, deep learning; PA, parametric.

Table 2 The model performances in the training cohort, internal validation cohort and external validation cohort

Model	AUC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Accuracy (95% CI)	PPV (95% CI)	NPV (95% CI)
Training set						
Physicians	70.5% (65.6–75.5%)	75.2% (68.4–81.9%)	65.9% (58.7–73.1%)	70.4% (70.2–70.5%)	67.4% (60.5–74.4%)	73.8% (66.8–80.9%)
PA model	79.7% (74.8–84.5%)	68.8% (61.5–76.0%)	79.6% (73.5–85.7%)	74.4% (74.3–74.5%)	76.1% (69.0–83.1%)	73.1% (66.6–79.5%)
GBDT-LR model	92.2% (89.0–95.3%)	87.3% (82.0–92.5%)	89.8% (85.2–94.4%)	88.6% (88.5–88.6%)	89.0% (84.0–93.9%)	88.2% (83.4–93.1%)
DL model	76.0% (70.9–81.1%)	51.0% (43.1–58.8%)	86.2% (81.0–91.5%)	69.1% (69.0–69.3%)	77.7% (69.6–85.7%)	65.2% (58.9–71.4%)
Internal test set						
Physicians	71.8% (62.0–81.6%)	76.9% (63.7–90.1%)	66.7% (52.4–80.9%)	71.6% (71.1–72.1%)	68.2% (54.4–81.9%)	75.7% (61.9–89.5%)
PA model	68.3% (56.2–80.5%)	89.7% (80.2–99.3%)	57.1% (42.2–72.1%)	72.8% (72.4–73.3%)	66.0% (53.3–78.8%)	85.7% (72.8–98.7%)
GBDT-LR model	84.6% (76.1–93.0%)	84.6% (73.3–95.9%)	73.8% (60.5–87.1%)	79.0% (78.6–79.4%)	75.0% (62.2–87.8%)	83.8% (71.9–95.7%)
DL model	71.6% (60.5–82.7%)	100% (100–100%)	42.9% (27.9–57.8%)	70.4% (69.9–70.9%)	61.9% (49.9–73.9%)	100% (100–100%)
External test set						
Physicians	71.9% (63.5–80.2%)	84.9% (75.3–94.5%)	58.8% (45.3–72.3%)	72.1% (71.7–72.5%)	68.2% (56.9–79.4%)	78.9% (66.0–91.9%)
PA model	76.5% (66.9–86.0%)	83.0% (72.9–93.1%)	70.6% (58.1–83.1%)	76.9% (76.6–77.3%)	74.6% (63.5–85.7%)	80.0% (68.3–91.7%)
GBDT-LR model	84.6% (77.0–92.2%)	84.9% (75.3–94.5%)	74.5% (62.5–86.5%)	79.8% (79.5–80.1%)	77.6% (66.9–88.3%)	82.6% (71.7–93.6%)
DL model	78.0% (69.2–86.7%)	84.9% (75.3–94.5%)	58.8% (45.3–72.3%)	72.1% (71.7–72.5%)	68.2% (56.9–79.4%)	78.9% (66.0–91.9%)

AUC, area under the curve; CI, confidence interval; PPV, positive predictive value; NPV, negative predictive value; PA, parametric; GBDT-LR, Gradient Boosting Decision Tree-Logistic Regression; DL, deep learning.

validation sets, indicating overfitting. To avoid overfitting, the model was given an additional drop layer, but the model still underperformed. This can be attributed to the need for deep learning models to have a large number of high-quality images, and medical images, especially PET/CT images, being more difficult to obtain, more expensive to annotate, and more time consuming than other imaging disciplines (37,38). Despite attempts to increase the sample size and improve deep learning model classification accuracy through migration learning (39) and data augmentation (40), the results remain unsatisfactory. However, neural networks are usually seen as powerful "black boxes" with weak interpretability (41).

Radiomics is an emerging and promising field.

4730

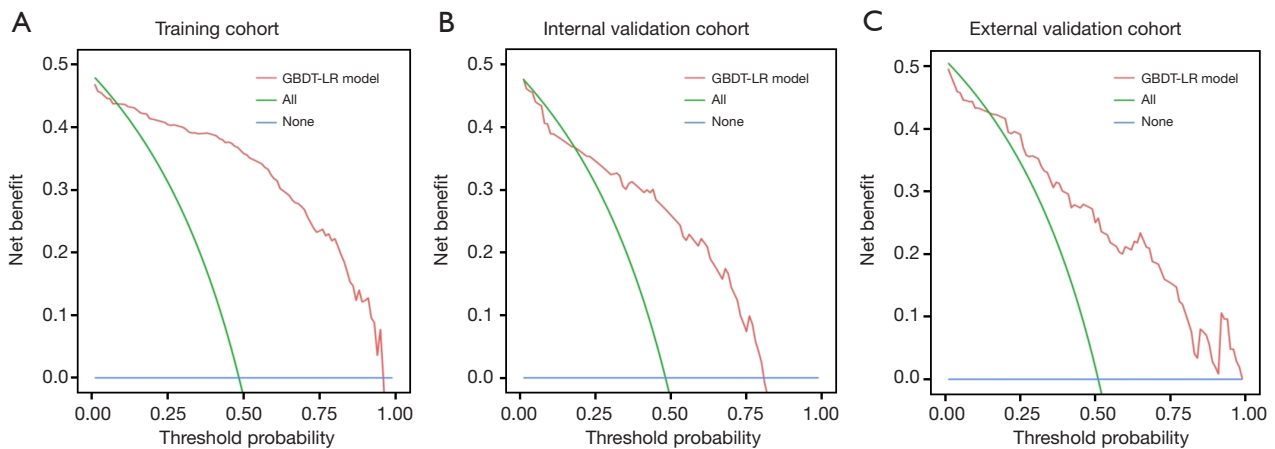Dong et al. Using machine learning to evaluate LNs status in PET/CT

Figure 3 Decision curve analyses of GBDT-LR model in training cohort (A), internal validation cohort (B), and external validation cohort (C). GBDT-LR, Gradient Boosting Decision Tree-Logistic Regression.

Previously, Xie (12) and others thus developed a PET/CT nomogram by combining SUVmax and CT radiomics for preoperative LN staging of non-small cell lung cancer. Bayanati *et al.* (42) used a combination of three textural features and three shape-based features for the prediction of LN metastasis in primary lung cancer. However, these were single-centre studies with a small number of LNs and did not place much emphasis on the comparison of predictive modelling methods. Most studies use linear models, such as logistic regression, in the modelling phase, which have the advantages of low computational complexity and high parallelism, but they ignore the ability of the features to interact with each other and to fuse information. Feature crossing refers to the combination or interaction of different features to create new features. Feature crosses allow the model to learn nonlinear relationships and interactions between features, improving the ability to model complex relationships. Although several machine learning methods, such as SVMs, DT and random forest, are capable of learning nonlinear relationships, these algorithms have been shown in previous studies to have an overfitting bias (43). Here, both SVM and DT models were found to overfit the data, indicating that these models were not stable when applied. To reduce overfitting and account for the correlations between features, we introduce a combined model combining gradient boosted DTs and logistic regression, that is, GBDT-LR. This combinatorial model was first proposed on Facebook, where GBDT is used to automatically filter and combine features to generate new discrete feature vectors, which are then fed into an LR model along with the original vectors to

produce the final prediction (44). The GBDT algorithm consists of several weak classification algorithms and is a typical example of an integrated learning approach (45). It can build combined features and automatically perform feature filtering. However, the data become sparse, and the feature dimensions can become too large when the GBDT model is used to construct new training features from the original features. This is addressed by using the LR algorithm, and then L1 regularisation is employed to reduce the risk of overfitting. The GBDT-LR radiomics model used in this study comprises 6 PET and 11 CT radiomics features, mostly belonging to 17 radiomics features related to shape and texture, consistent with previous studies. Previous scholars (42) have predicted LN metastasis in primary lung cancer by combining six shape and texture CT radiomics features with 71% accuracy. The experimental results demonstrate that the combined model yields high AUC, specificity, and sensitivity in all three cohorts, outperforming other single machine learning algorithms.

The physicians' analyses and the PET/CT parameter models were compared with the GBDT-LR model. The GBDT-LR model had higher AUC, ACC, PPV and specificity in all three cohorts. We hypothesise that this may be because the benign LNs we included would all be considered ambiguous in clinical practice, resulting in poorer PET/CT parameter modelling and physician assessment. In contrast, radiomics provides a more objective assessment of the status of mediastinal LNs by rapidly extracting many quantitative features from images in a high-throughput manner. There are several limitations to our study. First, this was a retrospective study, and the

LNs dissected at the time of surgery may be different from the LNs found in the same area on PET/CT images, so prospective studies are needed to address this issue. Second, we should include more data and more centres in future studies because although this was a multicentre study, the amount of data is still insufficient to cause overfitting in deep learning methods and some machine learning methods. Third, although we used radiomics, deep learning methods, and PET/CT parametric analysis, we included only images and data from the mediastinal LNs themselves. We did not include images or parameters related to the primary lesion. To improve diagnostic efficacy, it may be beneficial to include relevant information about the primary lesion. In future studies, we will incorporate additional relevant information about the lesion and follow-up with patients for an extended period to better evaluate the predictive performance of the machine learning model in terms of patient prognosis, including recurrence-free survival and overall survival.

Conclusions

In conclusion, in this study, we introduced a combined model called GBDT-LR to assess the status of confusing mediastinal LNs. This model outperformed deep learning models, physicians' analyses and PET/CT parameter models. Therefore, the status of confusing mediastinal LNs can be effectively assessed using radiological features from ^{18}F-FDG PET/CT images. Combining radiomics and machine learning approaches can improve the accuracy of diagnosing disease and can be applied to clinical practice by providing a safe and noninvasive approach for managing patients.

Acknowledgments

Footnote

Reporting Checklist: The authors have completed the TRIPOD reporting checklist. Available at https://qims.amegroups.com/article/view/10.21037/qims-24-100/rc

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-24-100/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by the Institutional Review Board (IRB) of Zhongda Hospital (No. 2021ZDSYLL239-P01) and the requirement of individual consent for this retrospective analysis was waived.

References

1. Kang WJ, Chung JK, So Y, Jeong JM, Lee DS, Lee MC. Differentiation of mediastinal FDG uptake observed in patients with non-thoracic tumours. Eur J Nucl Med Mol Imaging 2004;31:202-7.
2. McLoud TC, Bourgouin PM, Greenberg RW, Kosiuk JP, Templeton PA, Shepard JA, Moore EH, Wain JC, Mathisen DJ, Grillo HC. Bronchogenic carcinoma: analysis of staging in the mediastinum with CT by correlative lymph node mapping and sampling. Radiology 1992;182:319-23.
3. Kim YK, Lee KS, Kim BT, Choi JY, Kim H, Kwon OJ, Shim YM, Yi CA, Kim HY, Chung MJ. Mediastinal nodal staging of nonsmall cell lung cancer using integrated 18F-FDG PET/CT in a tuberculosis-endemic country: diagnostic efficacy in 674 patients. Cancer 2007;109:1068-77.
4. Fan Y, Zhang AM, Wu XL, Huang ZS, Kontogianni K, Sun K, Fu WL, Wu N, Kuebler WM, Herth FJF. Transbronchial needle aspiration combined with cryobiopsy in the diagnosis of mediastinal diseases: a multicentre, open-label, randomised trial. Lancet Respir Med 2023;11:256-64.

4732

Dong et al. Using machine learning to evaluate LNs status in PET/CT

5. Kim BT, Lee KS, Shim SS, Choi JY, Kwon OJ, Kim H, Shim YM, Kim J, Kim S. Stage T1 non-small cell lung cancer: preoperative mediastinal nodal staging with integrated FDG PET/CT--a prospective study. Radiology 2006;241:501-9.

6. Lv YL, Yuan DM, Wang K, Miao XH, Qian Q, Wei SZ, Zhu XX, Song Y. Diagnostic performance of integrated positron emission tomography/computed tomography for mediastinal lymph node staging in non-small cell lung cancer: a bivariate systematic review and meta-analysis. J Thorac Oncol 2011;6:1350-8.

7. Lu P, Sun Y, Sun Y, Yu L. The role of (18)F-FDG PET/CT for evaluation of metastatic mediastinal lymph nodes in patients with lung squamous-cell carcinoma or adenocarcinoma. Lung Cancer 2014;85:53-8.

8. Eubank WB, Mankoff DA, Takasugi J, Vesselle H, Eary JF, Shanley TJ, Gralow JR, Charlop A, Ellis GK, Lindsley KL, Austin-Seymour MM, Funkhouser CP, Livingston RB. 18fluorodeoxyglucose positron emission tomography to detect mediastinal or internal mammary metastases in breast cancer. J Clin Oncol 2001;19:3516-23.

9. De Leyn P, Stroobants S, De Wever W, Lerut T, Coosemans W, Decker G, Nafteux P, Van Raemdonck D, Mortelmans L, Nackaerts K, Vansteenkiste J. Prospective comparative study of integrated positron emission tomography-computed tomography scan compared with remediastinoscopy in the assessment of residual mediastinal lymph node disease after induction chemotherapy for mediastinoscopy-proven stage IIIA-N2 Non-small-cell lung cancer: a Leuven Lung Cancer Group Study. J Clin Oncol 2006;24:3333-9.

10. Flamen P, Lerut A, Van Cutsem E, De Wever W, Peeters M, Stroobants S, Dupont P, Bormans G, Hiele M, De Leyn P, Van Raemdonck D, Coosemans W, Ectors N, Haustermans K, Mortelmans L. Utility of positron emission tomography for the staging of patients with potentially operable esophageal carcinoma. J Clin Oncol 2000;18:3202-10.

11. Antoch G, Vogt FM, Freudenberg LS, Nazaradeh F, Goehde SC, Barkhausen J, Dahmen G, Bockisch A, Debatin JF, Ruehm SG. Whole-body dual-modality PET/CT and whole-body MRI for tumor staging in oncology. JAMA 2003;290:3199-206.

12. Xie Y, Zhao H, Guo Y, Meng F, Liu X, Zhang Y, Huai X, Wong Q, Fu Y, Zhang H. A PET/CT nomogram incorporating SUVmax and CT radiomics for preoperative nodal staging in non-small cell lung cancer. Eur Radiol 2021;31:6030-8.

13. Fiterman N, Berkman N, Kuint R. Predictors of malignant lymph node involvement in patients with mediastinal lymphadenopathy and previous cancer: A cohort study. Thorac Cancer 2022;13:631-6.

14. Lee JW, Kim BS, Lee DS, Chung JK, Lee MC, Kim S, Kang WJ. 18F-FDG PET/CT in mediastinal lymph node staging of non-small-cell lung cancer in a tuberculosis-endemic country: consideration of lymph node calcification and distribution pattern to improve specificity. Eur J Nucl Med Mol Imaging 2009;36:1794-802.

15. Choi EK, Park HL, Yoo IR, Kim SJ, Kim YK. The clinical value of F-18 FDG PET/CT in differentiating malignant from benign lesions in pneumoconiosis patients. Eur Radiol 2020;30:442-51.

16. Rogasch JMM, Michaels L, Baumgärtner GL, Frost N, Rückert JC, Neudecker J, Ochsenreither S, Gerhold M, Schmidt B, Schneider P, Amthauer H, Furth C, Penzkofer T. A machine learning tool to improve prediction of mediastinal lymph node metastases in non-small cell lung cancer using routinely obtainable [18F]FDG-PET/CT parameters. Eur J Nucl Med Mol Imaging 2023;50:2140-51.

17. Zhong Y, Yuan M, Zhang T, Zhang YD, Li H, Yu TF. Radiomics Approach to Prediction of Occult Mediastinal Lymph Node Metastasis of Lung Adenocarcinoma. AJR Am J Roentgenol 2018;211:109-13.

18. Hosny A, Bitterman DS, Guthier CV, Qian JM, Roberts H, Perni S, Saraf A, Peng LC, Pashtan I, Ye Z, Kann BH, Kozono DE, Christiani D, Catalano PJ, Aerts HJWL, Mak RH. Clinical validation of deep learning algorithms for radiotherapy targeting of non-small-cell lung cancer: an observational study. Lancet Digit Health 2022;4:e657-66.

19. Yang H, Chen L, Cheng Z, Yang M, Wang J, Lin C, Wang Y, Huang L, Chen Y, Peng S, Ke Z, Li W. Deep learning-based six-type classifier for lung cancer and mimics from histopathological whole slide images: a retrospective study. BMC Med 2021;19:80.

20. Courtiol P, Maussion C, Moarii M, Pronier E, Pilcer S, Sefta M, Manceron P, Toldo S, Zaslavskiy M, Le Stang N, Girard N, Elemento O, Nicholson AG, Blay JY, Galateau-Sallé F, Wainrib G, Clozel T. Deep learning-based classification of mesothelioma improves prediction of patient outcome. Nat Med 2019;25:1519-25.

21. Nam JG, Park S, Park CM, Jeon YK, Chung DH, Goo JM, Kim YT, Kim H. Histopathologic Basis for a Chest CT Deep Learning Survival Prediction Model in Patients with Lung Adenocarcinoma. Radiology 2022;305:441-51.

22. Ma X, Xia L, Chen J, Wan W, Zhou W. Development and validation of a deep learning signature for predicting lymph node metastasis in lung adenocarcinoma: comparison with radiomics signature and clinical-semantic model. Eur Radiol 2023;33:1949-62.

23. Wang H, Wang L, Lee EH, Zheng J, Zhang W, Halabi S, Liu C, Deng K, Song J, Yeom KW. Decoding COVID-19 pneumonia: comparison of deep learning and radiomics CT image signatures. Eur J Nucl Med Mol Imaging 2021;48:1478-86.

24. Dong M, Hou G, Li S, Li N, Zhang L, Xu K. Preoperatively Estimating the Malignant Potential of Mediastinal Lymph Nodes: A Pilot Study Toward Establishing a Robust Radiomics Model Based on Contrast-Enhanced CT Imaging. Front Oncol 2020;10:558428.

25. Chen X, Wang X, Zhang K, Fung KM, Thai TC, Moore K, Mannel RS, Liu H, Zheng B, Qiu Y. Recent advances and clinical applications of deep learning in medical image analysis. Med Image Anal 2022;79:102444.

26. Yin G, Song Y, Li X, Zhu L, Su Q, Dai D, Xu W. Prediction of mediastinal lymph node metastasis based on (18)F-FDG PET/CT imaging using support vector machine in non-small cell lung cancer. Eur Radiol 2021;31:3983-92.

27. Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. Radiology 2020;295:328-38.

28. Galavis PE, Hollensen C, Jallow N, Paliwal B, Jeraj R. Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters. Acta Oncol 2010;49:1012-6.

29. Orlhac F, Boughdad S, Philippe C, Stalla-Bourdillon H, Nioche C, Champion L, Soussan M, Frouin F, Frouin V, Buvat I. A Postreconstruction Harmonization Method for Multicenter Radiomic Studies in PET. J Nucl Med 2018;59:1321-8.

30. Orlhac F, Lecler A, Savatovski J, Goya-Outi J, Nioche C, Charbonneau F, Ayache N, Frouin F, Duron L, Buvat I. How can we combat multicenter variability in MR radiomics? Validation of a correction procedure. Eur Radiol 2021;31:2272-80.

31. Hu Y, Xie C, Yang H, Ho JWK, Wen J, Han L, Lam KO, Wong IYH, Law SYK, Chiu KWH, Vardhanabhuti V, Fu J. Computed tomography-based deep-learning prediction of neoadjuvant chemoradiotherapy treatment response in esophageal squamous cell carcinoma. Radiother Oncol 2021;154:6-13.

32. Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. J Bioinform Comput Biol 2005;3:185-205.

33. Vassallo P, Wernecke K, Roos N, Peters PE. Differentiation of benign from malignant superficial lymphadenopathy: the role of high-resolution US. Radiology 1992;183:215-20.

34. Tomlinson GS, Thomas N, Chain BM, Best K, Simpson N, Hardavella G, Brown J, Bhowmik A, Navani N, Janes SM, Miller RF, Noursadeghi M. Transcriptional Profiling of Endobronchial Ultrasound-Guided Lymph Node Samples Aids Diagnosis of Mediastinal Lymphadenopathy. Chest 2016;149:535-44.

35. Onal C, Oymak E, Findikcioglu A, Reyhan M. Isolated mediastinal lymph node false positivity of [18F]-fluorodeoxyglucose-positron emission tomography/computed tomography in patients with cervical cancer. Int J Gynecol Cancer 2013;23:337-42.

36. Vincent BD, El-Bayoumi E, Hoffman B, Doelken P, DeRosimo J, Reed C, Silvestri GA. Real-time endobronchial ultrasound-guided transbronchial lymph node aspiration. Ann Thorac Surg 2008;85:224-30.

37. Chen X, Xu X, Chrysikos S, Zhao M, Zhou Y. Value of 18-fluorodeoxyglucose positron emission tomography/computed tomography (18F-FDG PET/CT) in the differential diagnosis of sarcoidosis and lung cancer with lymph node metastasis: a retrospective study. Transl Lung Cancer Res 2022;11:1926-35.

38. Jiang X, Hu Z, Wang S, Zhang Y. Deep Learning for Medical Image-Based Cancer Diagnosis. Cancers (Basel) 2023;15:3608.

39. Stahlschmidt SR, Ulfenborg B, Synnergren J. Multimodal deep learning for biomedical data fusion: a review. Brief Bioinform 2022;23:bbab569.

40. Athalye C, Arnaout R. Domain-guided data augmentation for deep learning on medical imaging. PLoS One 2023;18:e0282532.

41. Singh A, Sengupta S, Lakshminarayanan V. Explainable Deep Learning Models in Medical Image Analysis. J Imaging 2020;6:52.

42. Bayanati H, E Thornhill R, Souza CA, Sethi-Virmani V, Gupta A, Maziak D, Amjadi K, Dennie C. Quantitative CT texture and shape analysis: can it differentiate benign and malignant mediastinal lymph nodes in patients with primary lung cancer? Eur Radiol 2015;25:480-7.

43. Zhou Y, Ma XL, Zhang T, Wang J, Zhang T, Tian R. Use

of radiomics based on (18)F-FDG PET/CT and machine learning methods to aid clinical decision-making in the classification of solitary pulmonary lesions: an innovative approach. Eur J Nucl Med Mol Imaging 2021;48:2904-13.

44. He X, Bowers S, Candela JQO, Pan J, Jin O, Xu T, Liu B, Xu T, Shi Y, Atallah A, editors. Practical Lessons from Predicting Clicks on Ads at Facebook. Eighth International Workshop on Data Mining for Online Advertising, 2014.

45. Zhou S, Wang S, Wu Q, Azim R, Li W. Predicting potential miRNA-disease associations by combining gradient boosting decision tree with logistic regression. Comput Biol Chem 2020;85:107200.