BMC Medical Imaging

RESEARCH

Open Access

# Towards automatical tumor segmentation in radiomics: a comparative analysis of various methods and radiologists for both region extraction and downstream diagnosis

Ying Yu[1†], Gang-Feng Li[1†], Wei-Xiong Tan[2†], Xiao-Yan Qu[1†], Tao Zhang[3], Xing-Yi Hou[1], Yuan-Bo Zhu[1], Zhi-Ying Ma[1], Lu Yang[1], Ya Gao[1], Mei Yu[1], Cui Yue[1], Zhen Zhou[2], Yang Yang[1*], Lin-Feng Yan[1*] and Guang-Bin Cui[1*]

## Abstract

**Objective** By discussing the difference, stability and classification ability of tumor contour extracted by artificial intelligence and doctors, can a more stable method of tumor contour extraction be obtained?

**Methods** We propose a novel framework for the automatic segmentation of lung tumor contours and the differential diagnosis of downstream tasks. This framework integrates four key modules: tumor segmentation, extraction of radiomic features, feature selection, and the development of diagnostic models for clinical applications. Using this framework, we conducted a study involving a cohort of 1,429 patients suspected of lung cancer. Four automatic segmentation methods (RNN, UNET, WFCM, and SNAKE) were evaluated against manual segmentation performed by three radiologists with varying levels of expertise. We further studied the consistency of radiomic features extracted from these methods and evaluates their diagnostic performance across three downstream tasks: benign vs. malignant classification, lung adenocarcinoma infiltration, and lung nodule density classification.

**Results** The Dice coefficient of RNN is the highest among the four automatic segmentation methods ($0.803 > 0.751$, 0.576, 0.560), and all $P < 0.05$. In the consistency comparison of the seven contour-extracted radiomic features, that the features extracted by RNN and S1 (the senior radiologist) showed the highest similarity which was higher than the other automatic segmentation methods and doctors with low seniority. In all three downstream tasks, the radiomic features extracted from RNN segmentation contours showed the highest diagnostic discrimination. In the classification of benign and malignant nodules, the RNN method performed slightly better than the S1 method, with an AUC of $0.840 \pm 0.01$ and $0.824 \pm 0.015$, respectively, and significantly better than the other five methods.

†Ying Yu, Gang-Feng Li, Wei-Xiong Tan, Xiao-Yan Qu and contributed equally to this work.

*Correspondence:
Yang Yang
yyang507@126.com
Lin-Feng Yan
ylf8342@163.com
Guang-Bin Cui
cuigbtd@fmmu.edu.cn; cuigbtd@163.com
Full list of author information is available at the end of the article

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Similarly, the RNN method had an AUC value of 0.946 in lung adenocarcinoma infiltration, and a kappa value of 0.729 in lung nodule density classification, both of which were better than the other six methods.

**Conclusions** Our findings suggest that AI-driven tumor segmentation methods can enhance clinical decision-making by providing reliable and reproducible results, ultimately emphasizing the auxiliary role of automated tumor contouring in clinical practice. The findings will have important implications for the application of radiomics in clinical practice.

**Keywords** Tumor segmentation, Radiomics, Extraction, Diagnosis, Lung nodule

## Background

Studies have shown that radiomic features are associated with tumor histology, patient survival, metabolism, and clinical decision-making [1, 2]. Radiomics automatically extracts quantitative features from medical images that can be used to characterize the biology of a disease. Recently, experts from Cancer Research UK (CRUK) and the European Organization for Research and Treatment of Cancer (EORTC) made 14 key recommendations to accelerate the clinical use of radiology [3]. Two of these recommendations are the standardization and reproducibility of medical imaging biomarkers, that is, studying the reproducibility and stability of radiomic features for multimodal medical image data across institutions. The stability of tumor regions segmentation is key because the extraction and analysis of radiomic features depend on the results [4].

In clinical practice, tumor regions segmentation is most commonly performed manually. However, the manual segmentation process has significant inconsistencies due to several factors [5, 6], such as heterogeneity in tumor contour growth [7, 8], low resolution of medical imaging, volume effects at lesion edges [9], and inter-annotator variability. Especially for tumors with blurred borders or low contrast with the surrounding tissues, and there is even greater inconsistency among personnel involved in manual segmentation [6, 10, 11]. In addition, manual segmentation is time-consuming and subjective [12]. Studies have reported that the inherent variability of manual segmentation may have a significant impact on radiomic analysis and modeling [13]. It has been proven that computer-assisted semi-automatic and fully automatic segmentation can reduce the burden and variability of manual segmentation [14–17]. A previous study [18] validated that radiomic features from prostate middle lobe enlargement in MR images, extracted using seven semi-automatic segmentation algorithms, can assist physicians in radiomics analysis. However, semi-automatic methods still exhibit a certain degree of inter- and intra-observer variation [19, 20].

In previous studies on fully automated instead of manual segmentation of tumor regions, there were two main approaches: supervised methods of deep learning and unsupervised methods of traditional image processing. The current literature seems to focus on the stability between the radiomic features extracted from tumor regions segmented by manual and automated methods; however, the variability between multiple automatic segmentation methods and their diagnostic power has rarely been investigated [21–25]. Additionally, due to the homogeneous downstream tasks, it can only be concluded that the contours extracted by automatic segmentation methods are efficient for evaluating the diagnostic ability of radiomic features in the downstream process. However, it cannot prove that the automatic segmentation methods are better than the manual segmentation method in terms of diagnostic ability.

In our study, we used excised pulmonary nodules to evaluate the variability and diagnostic ability of these extracted lung nodule contours among multiple segmentation methods (two deep learning-based automatic methods (A tumor region segmentation method based on deep learning 3D recurrent neural network [26] in this study, hereinafter referred to as RNN. A tumor region segmentation method based on deep learning 3D-UNet [27] in this study, hereinafter referred to as UNET.), two traditional image-processing-based automatic methods(A tumor region segmentation method based on conventional machine learning methods weighted fuzzy C-means clustering [28] in this study, hereinafter referred to as WFCM. A tumor region segmentation method based on conventional machine learning methods snake [29] in this study, hereinafter referred to as SNAKE.), and three radiologists of different seniority (A senior radiologist with 18 years' experience in chest radiology, S1. A junior radiologist with 3 years' experience in chest radiology, R1. A junior radiologist with 5 years' experience in chest radiology, R2)) in terms of radiomic features. We then used five machine learning models to verify the discriminative ability and diagnostic value of the regions of interest (ROIs) extracted by each segmentation method for three downstream tasks (benign and malignant lung nodules, infiltrative lung adenocarcinoma, and lung nodule density type). Afterward, we demonstrated that the ROIs extracted by automatic segmentation methods can replace manual annotation to a certain extent.

## Materials and methods

The local Institutional Review Board approved this retrospective study and waived the requirement for informed consent.

### Patient population

From May 01, 2018 to March 31, 2021, patients who underwent surgery for lung nodules were retrospectively consecutively enrolled ($n=2,397$). Patients who met the following criteria were included: (1) The thickness of the chest CT plain scan image layer was $\leq 2.0$ mm within three months prior to surgery. (2) Underwent surgical resection and had a pathology report. (3) Patient age $\geq 18$ years. The exclusion criteria were applied to the remaining 1,867 cases, excluding patients with artifacts on CT ($n=5$), enhanced CT ($n=408$), and more than five pathology reports ($n=25$). The final cohort included 1,429 patients.

### Image acquisition protocols

The CT images were helical CT scans obtained using a 64-slice CT scanner (LightSpeed VCT, GE Medical Systems or Toshiba Aquilion, TOSHIBA Medical Systems). CT images were obtained using the regular-dose technique (120 kVp, 250–350 mA) and multiple convolution kernel reconstruction algorithms. The reconstruction thicknesses were 0.625 mm and 1.00 mm.

#### *Methods*

The overall methodological flow of the study is shown in eFigure 1. All pulmonary nodules were segmented using seven segmentation methods to obtain corresponding ROIs, and the radiomic features extracted from these seven ROIs and five classification methods were developed to accomplish three classification tasks. Finally, a series of performance metrics was used to evaluate the performance of the segmentation methods and the variability of radiomic features and diagnostic capabilities.

### Segmentation methods

Seven segmentation methods were used to extract the ROI for each pulmonary nodule. Two conventional machine learning methods (WFCM, and SNAKE) and two deep learning methods (UNET and RNN-based) were developed. The SNAKE model, developed by Bong et al. [29], combines Graph Cut and active contour methods for unsupervised lung nodule segmentation. It first enhances contrast and removes noise with a Median Filter, then segments the lungs using active contours and refines the results with a graph-cut method. The WFCM model, proposed by Liu et al. [28], extends traditional FCM by incorporating spatial and grayscale similarities. It uses prior knowledge from training samples to build a probabilistic matrix for weakly supervised segmentation of unlabeled lung CT images. The UNET and RNN models are supervised deep learning methods from the Dr. Wise system (Deepwise AI Lab). The RNN model uses a recurrent CNN to iteratively segment nodules, refining segmentation with each iteration by using attention maps. The UNET model is based on the 3D-UNet [27], a widely used network for medical image segmentation. The specific details of these four segmentation methods and the segmentation training are described in Supplementary Material eAppendix Supplemental Methods. In addition, to closely match the actual clinical process of lung nodule contouring, manual segmentation was implemented as follows: the locations of the pulmonary nodules were marked with boxes by two junior radiologists (G.Y., who had been a practicing chest radiologist for 2 years, and M.Z.Y., who had been a practicing chest radiologist for 3 years) based on the pathology report. It was then reviewed by a senior radiologist (Y. M., with 9 years' experience in chest radiography). After the location of the pulmonary nodule was determined, the pulmonary nodule was then outlined independently by three radiologists (L.G.F., a senior radiologist with 18 years' experience in chest radiology [S1], Z.Y.B., a junior radiologist with 3 years' experience in chest radiology [R1], and H.X.Y, a junior radiologist with 5 years' experience in chest radiology [R2]) utilizing the Deepwise labeling platform (https://label.deepwise.com/). The three readers who performed the outline were completely unaware of the pathological diagnosis, but had access to clinical information, such as age.

### Feature extraction and selection

Radiomic features were extracted from the ROIs by seven segmentation methods using PyRadiomics (version 3.0.1; https://pypi.org/project/pyradiomics). A total of 1,454-dimensional radiomic features were extracted from each ROI. These include first-order features, shape-based features, Gray Level Size Zone Matrix (GLSZM) features, Gray Level Co-occurence Matrix (GLCM) features, Neighbouring Gray Tone Difference Matrix (NLTDM) features,

Gray Level Dependence Matrix (GLDM) features, and Gray Level Run Length Matrix (GLRLM) features extracted from each ROI. Python (version 3.8) was used for the normalization of raw images and ROIs, as well as radiomic feature extraction. For a detailed description of each radiomic feature, see https://www.radiomics.io/pyradiomics.html.

Feature selection was performed on the features obtained by the seven methods. First, Mann Whitney U test [30] were performed for each feature of each method to select those with high predictive power. Features that

were statistically significant ($p < 0.05$) were retained. Subsequently, all retained features were matched in pairs, and if the Pearson correlation coefficient [31] between two features was > 0.85, the feature with the largest *p*-value in the significance test was excluded. Finally, a five-fold cross-validation-based least absolute shrinkage and selection operator (LASSO) was applied to select features with non-zero coefficients from all extracted radiomic features [32]. Finally, from the screened features, 30 features with the smallest *P*-value were selected for subsequent study.

#### *Classification methods*

We employed five conventional machine learning methods to classify three downstream tasks: (1) benign vs malignant lung nodules, (2) lung nodule density types (Solid vs Ground-Glass Opacity [GGO] vs Partial Solid), and (3) subtypes of lung adenocarcinoma (Atypical Adenomatous Hyperplasia [AAH], Adenocarcinoma In Situ [AIS], Minimally Invasive Adenocarcinoma [MIA] vs Invasive Adenocarcinoma [IAC]). These five classification methods were developed based on the Python scikit-learn ML package (version 0.20.4). The five machine learning approaches that we have chosen are as follows: logistic regression (LR), support vector machines (SVM), extreme gradient boosting (XGBoost), multilayer perception (MLP), and linear discriminant (LD). These models were chosen based on their ability to capture different patterns within the data, their computational efficiency, and their strong performance in classification tasks. LR is often chosen for its simplicity and interpretability, while SVM is well-known for handling high-dimensional feature spaces and providing robust generalization. XGBoost is a powerful ensemble method that excels in handling large datasets and complex relationships between features. MLP, as a neural network model, is capable of learning intricate non-linear relationships, making it suitable for complex tasks. Finally, LD is selected for its ability to perform well in problems where the data distribution is close to Gaussian and the classes are well-separated. To select the most suitable model and the most appropriate hyperparameters for each model, five-fold cross-validation was performed on the development set, where 80% of the data were randomly selected to train the model, and the remaining 20% of the data (tuned set) were used for model validation. The training and validation processes were repeated five times. In the model-testing phase, a five-fold cross-validated ensemble model was used to identify the final classification.

Of the three classification tasks, the gold standard for benign, malignant, and invasive pulmonary nodules was determined according to pathological reports. The density types of pulmonary nodules were determined by two junior radiologists (Y.L., a radiologist with 2 years' experience in chest radiology, Q.X.Y., a radiologist with 6 years' experience in chest radiology) based on pathology reports. They were then submitted to a senior radiologist (Y.C., a practicing chest radiologist for 13 years) for review.

#### *Statistical analysis*

SPSS 25.0 software (version 25.0; SPSS Inc., Chicago, IL, USA) was used for statistical analysis. Dice coefficient and Hausdorff were used to assess profile differences between S1 and the other six segmentation methods. ICC was used to quantify the consistency of the radiomic features, and Mann Whitney U test was used to determine whether the features had a significant discriminatory ability for classification tasks [33]. For categorical variables, Fisher's exact test and the chi-squared test were used to test for differences between the groups. The performance of the three classification tasks was evaluated based on accuracy, precision, sensitivity, specificity, and the kappa value. The area under the receiver operating characteristic curve (AUC) was used to evaluate the model, and a DeLong test [34] was used to determine whether there was a significant difference in diagnostic performance between the models. Models for density types were evaluated using Overall Statistics (Macro) and Kappa, and sampling tests were used to determine whether there was a significant difference in diagnostic performance at each ROI.

## Results

### Research population

From 1,429 patients, 1,626 pulmonary nodules were obtained. In the downstream task of benign-malignant classification, patients were randomly divided into a development set (number of patients: 759, malignant nodule: 749, benign nodule: 147) and a validation set (number of patients: 670, benign nodule: 641, malignant nodule: 89). In the downstream task of lung adenocarcinoma infiltration classification, patients were randomized into a development set (number of patients: 610, AAH: 26, AIS: 228, MIA: 119, IAC: 314) and a validation set (number of patients: 592, AAH: 23, AIS: 206, MIA: 129, IAC: 307). In the downstream task of pulmonary nodule density type classification, patients were randomized into a development set (number of patients: 686, solid nodule: 212, partial solid nodule: 227, GGO nodule: 366) and a validation set (number of patients: 688, solid nodule: 195, partial solid nodule: 238, ground glass nodule: 368). Details of the registered patients were shown in eTable 1.

### Segmentation performance

Among the 1,626 pulmonary nodules, the Dice value of RNN was the highest among the other automatic segmentation methods (0.803 > 0.751, 0.576, 0.560), which was less than 3% lower than that of junior radiologists (0.831, 0.821). eTable 2 and eTable 11 in the Supplementary Materials record the Dice and Hausdorff value statistics of the six methods in detail. In order to more intuitively present the differences between different methods in lung nodule segmentation, Fig. 1 shows some representative CT scan images and their corresponding segmentation results.

### Consistency performance of radiomic features

The radiomic features of all pulmonary nodules were extracted. First, the differences in radiomic features between S1 and the other six segmentation methods were compared. Figure 2a showed differences in radiomic features with S1 and six other segmentation methods were analyzed through seven subgroups (benign, malignant, AAH/AIS/MIA, IAS, solid, partial solid, GGO). The feature difference values of the RNN were the smallest among the six methods; the feature difference values were less than 0.12 for all seven subgroups in the seven major categories of features. Therefore, the features of the RNN and S1 methods were remarkably similar. The WFCM has the highest differences (the difference values of features in most subgroups were greater than 0.2). Intraclass correlation coefficient (ICC) [35] was calculated by S1 and the other six methods, and the radiomic feature consistency of each method was evaluated. As showed in Fig. 2b, the number of features with consistency (ICC > 0.75) of RNN, UNET, R1, and R2 was similar, and they were all greater than 90, of which the RNN was the best. WFCM and SNAKE had a feature stability percentage of less than 70%. The variability of the radiomic features of the seven segmentation methods was examined. For each class of radiomic features, the number of features with high consistency between two of the seven segmentation methods was shown in Fig. 2d. Between the RNN and S1, the number of highly consistent features exceeded 80% for all seven classes of radiomic features. R2 (Gray Level Size Zone Matrix (GLSZM), Gray Level Co-occurence Matrix (GLCM), Neighbouring Gray Tone Difference Matrix (NLTDM)) and S1 had three major classes of radiomic features with a number of highly consistent features between 40 and 80%, and the other classes of radiomic features exceeded 80%. While R1 (NLTDM) and UNET (GLSZM) had only one major class with a number of highly consistent features between 40 and 80%, the other classes had more than 80% of radiomic features. For WFCM and SNAKE, the number of highly consistent features between them and S1, R1, R2, UNET, and the RNN for each class of radiomic features was mostly less than 40%, especially for WFCM. By calculating the p-value of each radiomic feature for each of the seven segmentation methods with three downstream tasks, it was found that the RNN with S1 had the largest number of features with

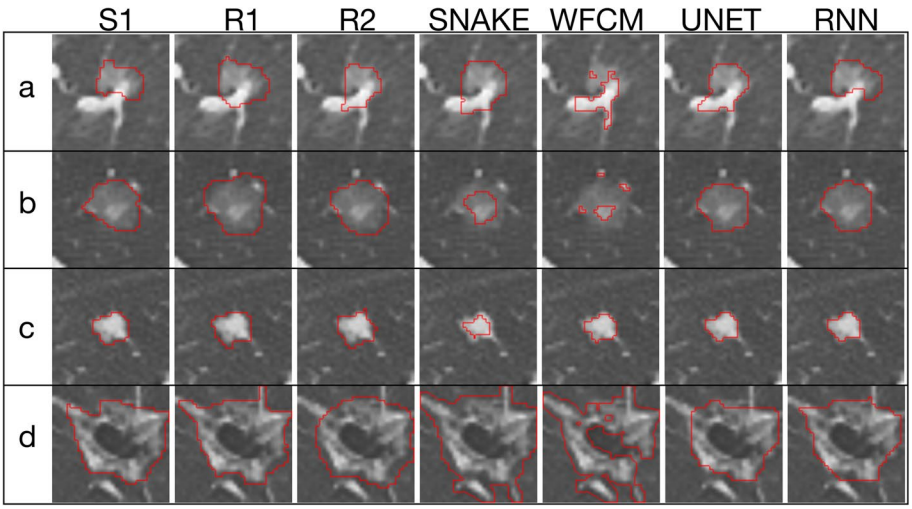


**Fig. 1** Segmentation results of 4 tumors in CT images using 7 segmentation methods. **a** This case is a benign pure ground glass nodule, pathological subtype atypical adenomatous hyperplasia, age 56 years old, gender is female. **b** This case is a malignant partially solid nodule, pathological subtype carcinoma in situ, age 61 years old, gender is male. **c** This case is a malignant solid nodule, pathological subtype microinvasion, age 41 years old, gender is male. **d** This case is a malignant pure ground glass nodule, pathological subtype invading, age 66 years old, gender is female

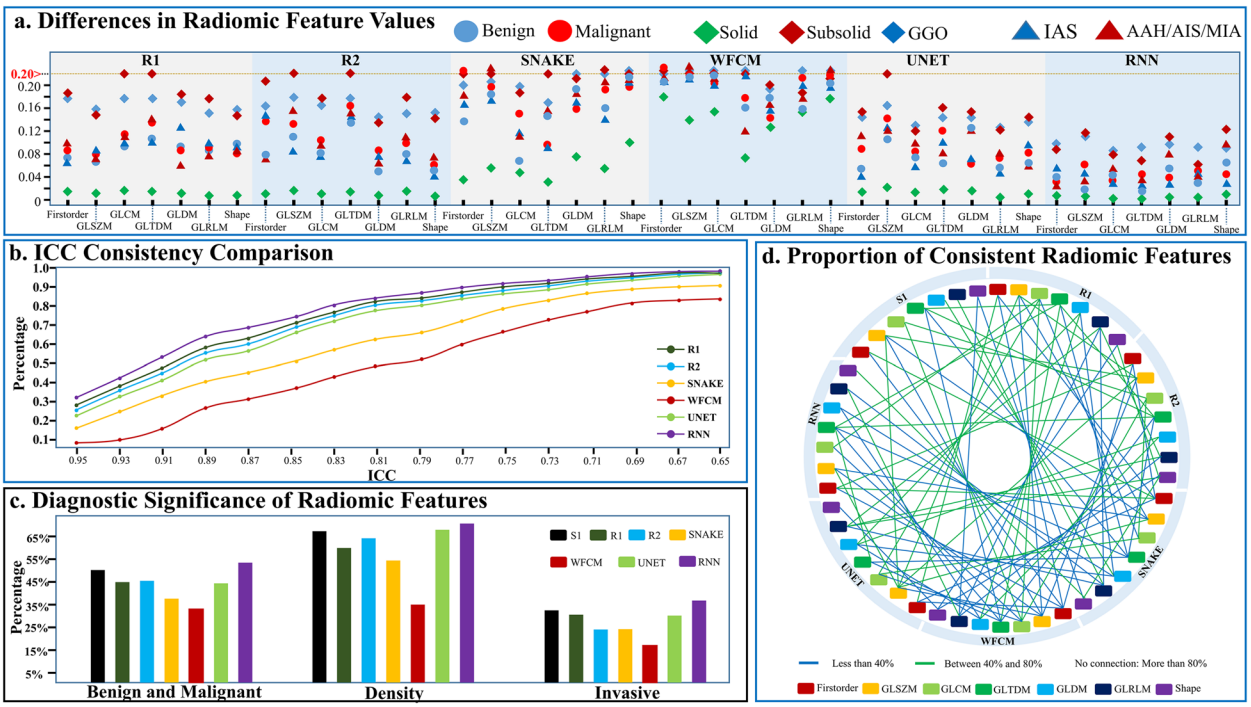


**Fig. 2** Differences between the radiomic features of the lung nodule region extracted by seven segmentation methods, and the diagnosis of features to three downstream classification tasks. **a** Difference in radiomics feature values: This illustrates the difference in radiomics features between the S1 method and the other six segmentation methods. The horizontal axis represents the seven major radiomics feature categories, while the vertical axis shows the differences between the features of S1 and those of the other segmentation methods within each category. For each category, the features of S1 are subtracted from the corresponding features of the other methods, and the absolute differences are calculated and summed to obtain the average. **b** ICC Consistency Comparison: Percentage of ICC values of the radiomic features between S1 and the other 6 segmentation methods. The horizontal axis represents the ICC value; the vertical axis represents the percentage of the number of ICC features under the maximum value to the total number. **c** Diagnostic Significance of Radiomic Features: The percentage of the number of features with significant diagnosis in the three downstream tasks from the radiomic features of lung nodules extracted by seven segmentation methods. **d** Proportion of Consistent Radiomic Features: Among the radiomic features in each category, the proportion of features with highly consistency (ICC > 0.75) between the seven segmentation methods. Abbreviations: AAH, atypical adenomatous hyperplasia; AIS, adenocarcinoma in situ; MIA, minimally invasive adenocarcinoma; IAC, invasive adenocarcinoma; GGO, ground glass nodule; GLCM, Gray Level Co-occurence Matrix; GLDM,Gray Level Dependence Matrix; GLRLM, Gray Level Run Length Matrix; GLSZM, Gray Level Size Zone Matrix; NGTDM: Neighbouring Gray Tone Difference Matrix

discrimination ability ($p < 0.05$) in each downstream task, followed by R1, R2, and UNET, while WFCM had the least (Fig. 2c).

### Benign and malignant classification

In the downstream task of benign and malignant lung cancer classification, as showed in Fig. 3a, radiomic features under seven segmentation methods (RNN, UNET, S1, R1, R2, WFCM, SNAKE) were feature selection and modeling, respectively, the percentage of the same feature intersection was over 63.3% between RNN, UNET S1, R1, and R2, the RNN and S1 was the highest, reaching 96.6%, only one feature was different. However, WFCM and SNAKE had less intersections with the same features than other methods, especially WFCM, whose highest intersection ratio was only 43.3%. Specific selecting features were shown in Supplementary Material eTable 3.

Figure 4a1 shows the area under ROC (receiver-operating characteristic) curve (AUC) values of the seven segmentation methods among the five machine-learning methods. The AUC values of the RNN methods were found to be the highest, except for the p-value of S1 in the logistic regression (LR) model ($p > 0.05$), while the others were significantly different ($p < 0.05$, Fig. 4b1). Among the average AUC values of the five machine learning models, the AUC value of S1 (0.824) was second only to that of the RNN (0.840). UNET was found to be slightly higher than the low R1 and R2 values, and the traditional automatic segmentation method had the worst performance, where the AUC of WFCM was only 0.757 (Fig. 4c1). Table 1 showed the performance of the seven segmentation methods for benign and malignant classification in the linear discriminant (LD) model. Where the RNN method was optimal in accuracy, precision, sensitivity, and kappa,

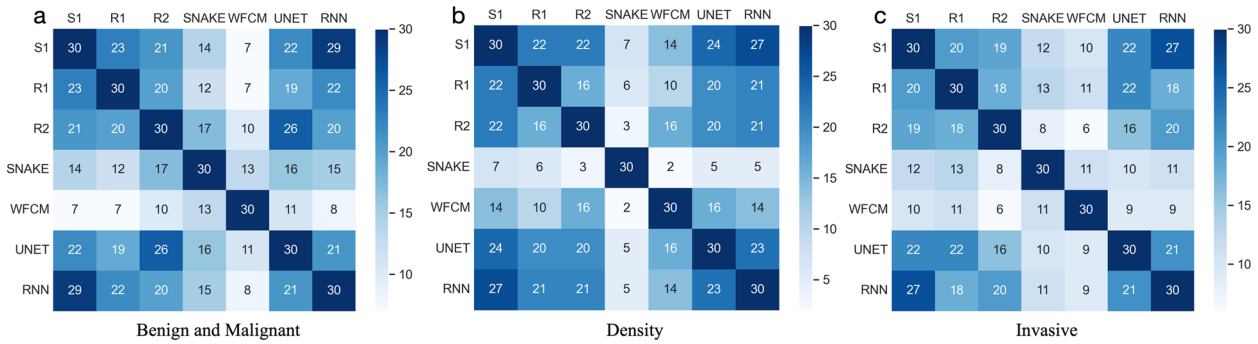


**Fig. 3** In the three downstream classification tasks, radiomic features under seven segmentation methods were feature selection, respectively, heat map of the same feature intersection between pairwise. For example, the number 23 in the second column of the first row of Figure a indicates that, in the benign and malignant classification task, the omics features selected by the S1 method and those selected by the R1 method share 23 common features

**Table 1** Baseline characteristics of the patients and pulmonary nodules

| Characteristic | Total (benign-malignant/infiltration/density) | Development (benign-malignant/infiltration/density) | Validation (benign-malignant/infiltration/density) | *P* value |
|---|---|---|---|---|
| **Patients** | **1429/1202/1374** | **759/610/686** | **670/592/688** | |
| Age in years, mean ± SD | 56.6 ± 10.7/56.4 ± 10.7/56.6 ± 10.7 | 56.4 ± 10.9/55.9 ± 10.6/56.1 ± 10.6 | 56.8 ± 10.6/56.9 ± 10.7/57.1 ± 10.9 | 0.509/0.208/0.197 |
| SEX | | | | 0.615/0.453/0.583 |
| Male | 605(42%)/456(38%)/552(40%) | 337(44%)/235(38%)/279(34%) | 268(40%)/217(37%)/273(40%) | |
| Female | 824(58%)/746(62%)/822(60%) | 422(56%)/375(62%)/407(66%) | 402(60%)/375(63%)/415(60%) | |
| **Nodule** | **1626/1352/1606** | **896/687/805** | **730/665/801** | |
| Location | | | | 0.735/0.989/0.201 |
| RUL | 316(19%)/261(19%)/378(24%) | 169(19%)/137(20%)/178(22%) | 147(20%)/124(19%)/197(25%) | |
| RML | 334(21%)/262(19%)/304(19%) | 181(20%)/132(19%)/152(19%) | 153(21%)/130(20%)/154(19%) | |
| RLL | 310(19%)/257(19%)/309(19%) | 174(20%)/130(19%)/150(19%) | 136(19%)/127(19%)/160(20%) | |
| LUL | 339(21%)/293(22%)/309(19%) | 192(21%)/136(20%)/167(21%) | 147(20%)/157(23%)/141(18%) | |
| LLL | 327(20%)/279(21%)/306(19%) | 180(20%)/152(22%)/158(19%) | 147(20%)/127(19%)/149(18%) | |
| benign- malignant | | | | 0.139 |
| Benign | 1390(85%) | 749(84%) | 641(87%) | |
| Malignant | 236(15%) | 147(16%) | 89(13%) | |
| infiltration | | | | 0.521 |
| AAH | 49(4%) | 26(4%) | 23(4%) | |
| AIS | 434(32%) | 228(33%) | 206(31%) | |
| MIA | 248(19%) | 119(17%) | 129(19%) | |
| IAS | 621(45%) | 314(46%) | 307(46%) | |
| density | | | | 0.390 |
| Solid | 407 | 212 | 195 | |
| Subsolid | 465 | 227 | 238 | |
| GGO | 734 | 366 | 368 | |

Mean data are ± standard deviation; data in parentheses are percentages; / is to separate the three classification tasks of benign and malignant, invasive and density type

*Abbreviations: AAH* atypical adenomatous hyperplasia, *AIS* adenocarcinoma in situ, *MIA* minimally invasive adenocarcinoma, *IAC* invasive adenocarcinoma, *RUL* right upper lobe, *RML* right middle lobe, *RLL* right lower lobe, *LUL* left upper lobe, *LLL* left lower lobe, *GGO* ground glass nodule

the specificity was slightly lower than S1 (0.828 vs. 0.83). The performance of the other four methods was detailed in eTable 6 of the Supplementary Material.

#### Lung nodule density type classification

In the downstream task of lung nodule density type classification, radiomic features under seven segmentation methods were feature selection and modeling, respectively, the percentage of the same feature intersection was similar to the results for benign and malignant lung cancer (Fig. 3b, specific selecting features were shown in Supplementary Material eTable 4). Figure 4a2 shows the kappa values for each of the seven segmentation methods across the five machine learning methods, with the RNN methods possessing the highest values with p-values less than 0.05 (Fig. 4b2). Among the five machine learning methods, the overall kappa values of the seven segmentation methods were shown in Fig. 4c2, including WFCM (0.642), SNAKE (0.681), R1 (0.701), R2 (0.705), UNET (0.717), S1 (0.723), and the RNN (0.729). Table 2 showed the performance of the seven segmentation methods for lung nodule density classification in the LD model. In the overall macroscopic performance index, the RNN showed better performance, with sensitivity, specificity, and kappa being the highest among the seven methods. This was followed by S1, UNET, R1, R2, WFCM, and SNAKE. The performance of the other four methods was detailed in eTable 7 of the Supplementary Material.

#### Invasive adenocarcinoma classification

In the downstream task of invasive lung adenocarcinoma classification, the intersection of features screened by the seven segmentation methods supported similar results to the intersection of features from the first two classifications (Fig. 3c, and the specific filtered features were listed in Supplementary Material eTable 5). Figure 4a3 showed the AUC values of each of the seven segmentation methods among the five machine learning methods. The RNN methods had the highest values, and all the *p*-values were less than 0.05 (Fig. 4b3). Among the average AUC values of the five machine learning methods, the magnitudes of the AUC values were WFCM (0.879), SNAKE (0.913), R1 (0.926), R2 (0.927), UNET (0.934), S1 (0.940), and the RNN (0.946) (Fig. 4c3). Table 3 showed the performance of the seven segmentation methods for infiltrative lung adenocarcinoma classification in the LD model. The RNN showed better performance in the overall performance index, where accuracy, precision, specificity, and kappa were the highest among the seven methods, followed by S1, UNET, R1, and R2. WFCM and SNAKE performed the worst again. The performances of the other four methods were detailed in Supplementary Material eTable 8.

**Table 2** Results of the benign and malignant performance metrics for the 7 methods in LD

| | Accuracy | Precision | Sensitivity | Specificity | Kappa |
|---|---|---|---|---|---|
| S1 | 0.826 | 0.394 | 0.798 | **0.83** | 0.435 |
| R1 | 0.784 | 0.332 | 0.764 | 0.786 | 0.352 |
| R2 | 0.8 | 0.352 | 0.764 | 0.805 | 0.378 |
| SNAKE | 0.778 | 0.327 | 0.775 | 0.778 | 0.348 |
| WFCM | 0.764 | 0.305 | 0.73 | 0.769 | 0.312 |
| UNET | 0.795 | 0.35 | 0.798 | 0.794 | 0.381 |
| RNN | **0.826** | **0.396** | **0.809** | 0.828 | **0.439** |

## Discussion

In this study, we developed a general framework for the automatic segmentation of tumors and the differential diagnosis of downstream tasks. It integrated four modules: automatic segmentation of tumors, extraction of radiomic features, radiomic feature selection, and building diagnostic models for downstream tasks. Under this framework, we first investigated the variability of tumor contours segmented by four automatic segmentation methods and three radiologists, as well as the consistency of their radiomic features. We further investigated the differential diagnostic ability of the radiomic features extracted from each contour using seven segmentation methods for three downstream tasks. The main objective was to verify whether automatic segmentation of tumor contours could replace manual segmentation. In this study, it was found that the contours segmented by the RNN were the closest to those manually outlined by S1 among the four automatic segmentation methods. The radiomic features extracted by the RNN were the most similar to those extracted by S1. This was based on a comparison of the consistency of the radiomic features for each of the seven methods. In the model performance analysis of the three downstream tasks, the features extracted by the RNN had the most discriminative diagnostic power.

Tumor segmentation is a fundamental step in radiomics analysis [36–38]. However, because there is no "gold standard" for tumor segmentation, it is challenging to use manual or semi-automated segmentation methods. Automatic CT-based tumor segmentation methods to replace manual segmentation have a high degree of consistency, and a strong correlation with the macroscopic diameter is considered the "gold standard" [39]. However, not all automatic segmentation methods are suitable for specific tumors [40]. In this study, we selected two traditional methods and two deep learning-based methods for comparison. Among the various

**Table 3** Results of the density performance metrics for the 7 methods in the LD

| | | Class Statistics | | | | Overall Statistics (Macro) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | **Accuracy** | **Precision** | **Sensitivity** | **Specificity** | **Accuracy** | **Precision** | **Sensitivity** | **Specificity** | **Kappa** |
| S1 | Solid | 0.949 | 0.914 | 0.872 | 0.974 | 0.885 | **0.829** | 0.824 | 0.91 | 0.731 |
| | Subsolid | 0.833 | 0.711 | 0.735 | 0.874 | | | | | |
| | GGO | 0.874 | 0.862 | 0.864 | 0.882 | | | | | |
| R1 | Solid | 0.948 | 0.909 | 0.872 | 0.972 | 0.88 | 0.821 | 0.813 | 0.905 | 0.718 |
| | Subsolid | 0.824 | 0.71 | 0.689 | 0.881 | | | | | |
| | GGO | 0.869 | 0.843 | 0.878 | 0.861 | | | | | |
| R2 | Solid | 0.945 | 0.876 | 0.903 | 0.959 | 0.874 | 0.809 | 0.815 | 0.903 | 0.708 |
| | Subsolid | 0.819 | 0.687 | 0.718 | 0.861 | | | | | |
| | GGO | 0.859 | 0.863 | 0.823 | 0.889 | | | | | |
| SNAKE | Solid | 0.941 | 0.889 | 0.867 | 0.965 | 0.868 | 0.801 | 0.795 | 0.895 | 0.689 |
| | Subsolid | 0.809 | 0.687 | 0.655 | 0.874 | | | | | |
| | GGO | 0.853 | 0.826 | 0.861 | 0.845 | | | | | |
| WFCM | Solid | 0.939 | 0.892 | 0.851 | 0.967 | 0.843 | 0.772 | 0.743 | 0.868 | 0.622 |
| | Subsolid | 0.778 | 0.685 | 0.466 | 0.909 | | | | | |
| | GGO | 0.811 | 0.74 | 0.91 | 0.727 | | | | | |
| UNET | Solid | 0.954 | 0.911 | 0.897 | 0.972 | 0.882 | 0.823 | 0.822 | 0.907 | 0.724 |
| | Subsolid | 0.826 | 0.704 | 0.718 | 0.872 | | | | | |
| | GGO | 0.865 | 0.855 | 0.851 | 0.878 | | | | | |
| RNN | Solid | 0.946 | 0.884 | 0.897 | 0.962 | **0.888** | 0.827 | **0.829** | **0.913** | **0.738** |
| | Subsolid | 0.836 | 0.724 | 0.727 | 0.883 | | | | | |
| | GGO | 0.88 | 0.874 | 0.864 | 0.894 | | | | | |

traditional baseline segmentation algorithms [41–43] (such as Otsu, MCET, SNAKE, and FCM), we chose SNAKE and WFCM due to their superior performance in handling complex tumor contours. For deep learning methods, we selected U-Net and RNN, which are highly representative. These four automatic segmentation methods were compared to manual segmentation performed by radiologists of varying qualifications across three aspects: segmentation shape, feature similarity after segmentation, and the discriminative ability of the extracted features. The results of the study showed that the deep learning-based segmentation method was closer to manual segmentation by radiologists than the segmentation method based on traditional image processing in terms of segmentation contours. In addition, the radiomic features extracted by the RNN segmented contours performed best in the three downstream tasks, with AUC for benign and malignant classification, AUC on invasive adenocarcinoma classification, and kappa values for density type were found to perform significantly better than those of the manual segmentation method. Therefore, we believe that the radiomic features extracted by the high-precision automatic segmentation method are sufficient to support efficient analysis of the diagnosis of downstream tasks.

In this study, we found that the Dice values of the two junior radiologists were 2–3 percentage points higher than that of the RNN, but the RNN was more consistent than the junior radiologists when comparing the consistency of imaging characteristics with S1. In view of this phenomenon, we performed a subgroup analysis by interval of Dice values of single pulmonary nodules, computing ICC in two groups, which were divided by Dice values (greater than or equal to 0.6 and less than 0.6). The detailed procedure is described in the Supplementary Material eTable 9. It was found that in the group with Dice values greater than 0.6, R1, R2, and the RNN had similar numbers of highly consistent features of 94.6%, 95.2%, and 94.9%, respectively. However, in the group with Dice values less than 0.6, R1 and R2 had a much lower number of highly consistent features than RNN (71.2%, 72.1%, and 81.7%, respectively). Therefore, when the tumor contour edge was complex and the consistency of the contours segmented by different radiologists is poor, the RNN segmentation of contours has a higher number of highly consistent features than the less senior radiologists. Thus, it can be concluded that the complexity of the tumor contour has less influence on the high consistency feature of the RNN. In addition, we counted the respective contour sizes of the radiologists and

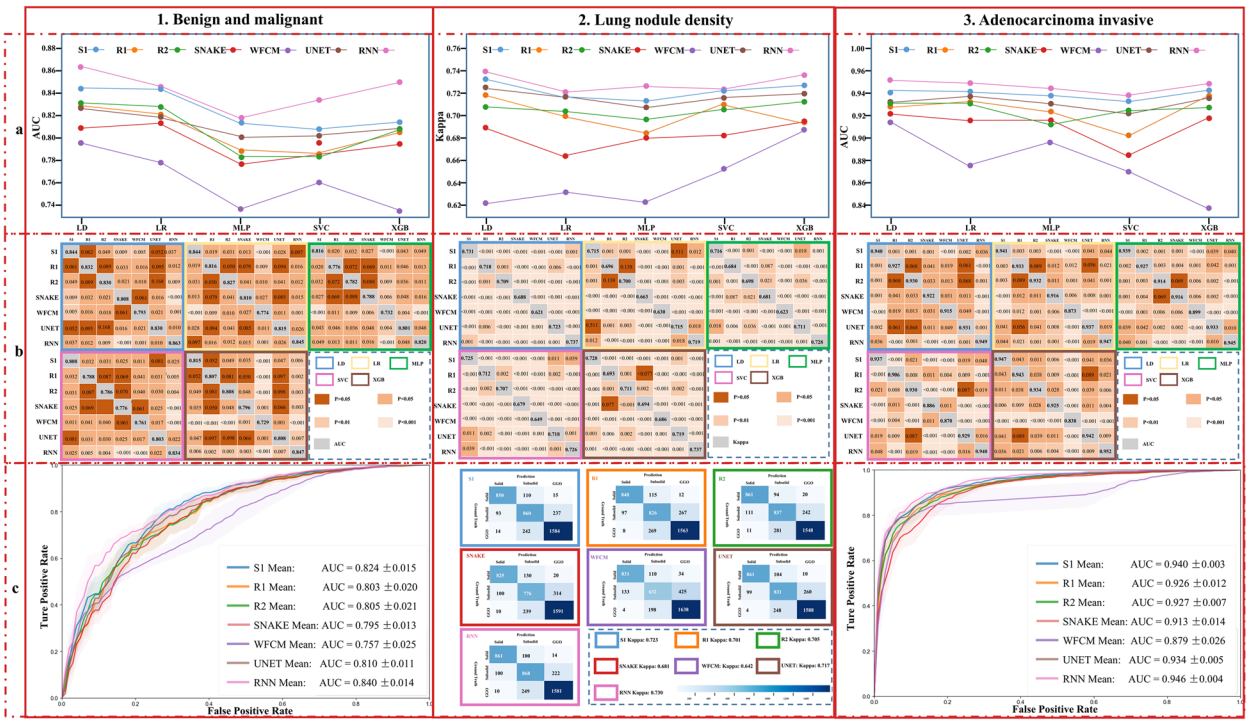


**Fig. 4** In the three downstream tasks, the diagnostic performance under seven kinds of segmentation and the comparison between them are different. The figure in the first row shows the performance comparison line chart of three downstream tasks in five machine learning classification models and seven segmentation methods; The graph in the second row shows the statistical difference (*P* value) between the segmentation methods in the first row. The graph in the third row shows a comparison graph of the tie performance of seven segmentation methods for three downstream tasks among five machine learning methods. a1: Comparing AUC values between segmentation methods in benign and malignant classification tasks. a2: Kappa values between segmentation methods in the density type classification task. a3: Comparing AUC values among segmentation methods in the invasive classification task of lung adenocarcinoma. b1: The statistical difference (*P* value) of AUC values between two segmentation methods in each machine learning method in the benign and malignant classification task. b2: The statistical difference (*P* value) of Kappa value between two segmentation methods in the density type classification task. b3: The statistical difference (*P* value) of AUC value between two segmentation methods in the invasive classification task. c1: ROC curve of seven segmentation methods in the benign and malignant classification task. c2: The confusion matrix diagram of seven segmentation methods in the density type classification task. c3: ROC curve of seven segmentation methods in the invasive classification task of lung adenocarcinoma

automatic segmentation methods. The contours outlined by radiologists were larger than those segmented by the RNN, and the contours were more outward for the radiologists with less seniority. It is probable that when radiologists outline a tumor, in an effort to try to include the whole tumor, they will expand outward; thus, the contour will not be as close to the tumor edge. In contrast, automatic segmentation methods, such as the RNN, stick to the tumor edge when performing contour segmentation. Most of the texture information was contained within the tumor; therefore, even though the Dice values were higher for the junior radiologists than for the RNN, the consistency is lower.

Among the seven segmentation methods, the features extracted in the RNN were the most similar to S1, and the number of features with ICC > 0.75 is up to 92%. Although the consistency of the features they obtained was high, there were still some differences in their diagnostic abilities for downstream tasks. It was found that in the lung nodule density type classification, the RNN and S1 had three different feature choices in the final 30 feature filters for modeling.The three different feature choices in S1 were not included in the last 30 features in RNN, but these from RNN were diagnostic for density types ($P < 0.05$). On the contrary, the different feature choices (GLCM_Gray_Level_Variance, GLRLM_Gray_Level_Non-Uniformity, and Firstorder_Root_Mean_Squared) in RNN come from S1 but had no diagnostic ability. This was mainly because, in clinical practice, the mean variance of the HU values of CT was one of the discriminatory features when distinguishing the type of density in pulmonary nodules [44, 45]. Thus, the accuracy of contour delineation of the lung nodules was critical for modeling radiomic features. These three features were based on the variance of the gray value of the lesion area and the mean of the HU value to count; in lung nodule contour

segmentation, the RNN was closer to the contour than S1, so these three features make the RNN performed better than S1 in density-type discrimination. Furthermore, it can also be verified from the final classification results that the RNN had more discriminative power than the S1 features (kappa: $0.729 > 0.723$; $p < 0.05$). To further verify the more discriminative ability of the RNN than the features selected by S1, we removed the different features selected from the RNN and S1 in the other two downstream tasks and kept only the same features for classification. In the benign-malignant classification task, the AUC results of S1 and the RNN were 0.816 and 0.819, respectively. In the infiltrative classification task, the AUC results of S1 and the RNN were 0.931 and 0.933, respectively. Therefore, the RNN automatic segmentation method can extract additional valuable features compared to radiologists.

There are noticeable differences between the various segmentation methods, particularly when tumors have complex backgrounds, such as the presence of blood vessels and trachea passing through them, or when volume effects come into play. In these situations, less experienced doctors may inadvertently include blood vessels in the tumor delineation, as seen in Fig. 1, where doctor R1 outlined some blood vessels as part of the tumor. In case (d), doctor R2 missed part of the tumor due to the interference from blood vessels. For tumors with complex backgrounds, the WFCM method tends to focus solely on areas with high brightness, resulting in suboptimal segmentation, particularly in regions with low tumor density, such as the center of case (d). Similarly, the SNAKE method faces the same issue. However, more experienced radiologists, like senior doctor S1, are able to handle these complexities effectively. The RNN method, on the other hand, demonstrates relatively consistent performance and similarity to senior doctor S1. The variability in segmentation results directly influences the radiomic features extracted, which in turn affects the diagnostic performance of the model. Different downstream tasks prioritize features most relevant to their specific objectives. We observed that when significant discrepancies occur in the tumor boundaries between different segmentation methods, the extraction of key radiomic features—such as those related to tumor texture and shape—becomes inconsistent, ultimately affecting diagnostic accuracy. As shown in Fig. 4b, the downstream task least affected by segmentation variability was the benign vs. malignant classification, where most *P*-values were greater than 0.05, indicating little impact on model performance. In contrast, the lung nodule density type classification task was most affected, with very few *P*-values above 0.05 and most below 0.001. This aligns with our earlier analysis, as the classification of lung nodule density is particularly sensitive to segmentation discrepancies, especially those related to boundaries and vascular structures.

This study has several limitations. First, the data were homogeneous, consisting of patients from a single institution and focused on lung cancer, which limits the generalizability of the findings to other cancers or institutions. Multicenter data and the inclusion of various tumor types have been shown in the literature to be crucial for studying the relationship between segmentation methods and radiomic features [46, 47]. Future studies should incorporate multicenter data with diverse patient populations to validate the framework's robustness. Second, the performance of segmentation methods may be influenced by factors such as tumor complexity and image quality, which were not sufficiently addressed in our analysis. Further research is needed to explore the impact of imaging parameters, scanner types, and image modalities (e.g., CT, pathology, ultrasound) on segmentation accuracy. Given the critical role of radiation dose reduction and improved predictive classification across multiple data types (such as pathology images and ultrasound) [48, 49], this should be prioritized. Lastly, while we provide a comparative analysis of segmentation methods, the absence of external validation with datasets from other institutions or different patient populations remains a limitation. We plan to include such validation in future studies to confirm the broader applicability of our findings.

## Conclusions

We believe that the main contributions of this paper are as follows:

(1) A general framework for the automatic segmentation of tumor contours and the differential diagnosis of downstream tasks was established. The framework integrates four modules: automatic segmentation of tumor contours, extraction of image histological features corresponding to the ROI, dimensionality reduction of image histological features, and establishment of diagnostic models for downstream tasks. The framework can easily be migrated and extended and can be applied to other diseases.
(2) Seven segmentation methods (two automatic methods based on deep learning, two automatic methods based on conventional image processing, and three doctors of different seniority) were used to analyze the variability and diagnostic performance of their corresponding histological features.
(3) The diagnostic ability and clinical value of each segmentation method were validated using three downstream tasks (lung nodule benignity and malignancy, lung adenocarcinoma infiltration, and lung nodule density type).

To conclude, in this study, seven ROIs were extracted using automatic segmentation based on deep learning and traditional methods and radiologists of different seniority levels, from which the radiomic features were extracted and then modeled in three downstream tasks of diagnosis using five machine learning methods. It has been demonstrated that the contours extracted by deep learning segmentation methods are efficient for evaluating the diagnostic ability of radiomic features in the downstream process. Our results further emphasized the feasibility of applying deep learning segmentation methods to assist in clinical diagnosis.

**Abbreviations**

| | |
|---|---|
| AAH | Atypical adenomatous hyperplasia |
| AIS | Adenocarcinoma in situ |
| AUC | Area under receiver-operating characteristic curve |
| CRUK | Cancer research UK |
| EORTC | European Organization for Research and Treatment of Cancer |
| FCM | Fuzzy C-means |
| GGO | Ground glass nodule |
| GLCM | Gray Level Co-occurence Matrix |
| GLDM | Gray Level Dependence Matrix |
| GLRLM | Gray Level Run Length Matrix |
| GLSZM | Gray Level Size Zone Matrix |
| IAC | Invasive adenocarcinoma |
| ICC | Intraclass correlation coefficient |
| LASSO | Least absolute shrinkage and selection operator |
| LD | Linear discriminant |
| LR | Logistic regression |
| MIA | Invasive adenocarcinoma |
| MLP | Multilayer perception |
| NLTDM | Neighbouring Gray Tone Difference Matrix |
| R1 | A junior radiologist with 3 years' experience in chest radiology |
| R2 | A junior radiologist with 5 years' experience in chest radiology |
| RNN | Recurrent neural networks algorithm |
| ROIs | Regions of interest |
| ROC | Receiver-operating characteristic |
| S1 | A senior radiologist with 18 years' experience in chest radiology |
| SVM | Support vector machines |
| UNET | U-net algorithm |
| WFCM | Weighted fuzzy C-means |
| XGBoost | Extreme gradient boosting |

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12880-025-01596-2.

Supplementary Material 1.

**Acknowledgements**
We gratefully acknowledge all the support from Department of Radiology & Functional and Molecular Imaging Key Lab of Shaanxi Province, Tangdu Hospital, Fourth Military Medical University (Air Force Medical University) and Deepwise Artificial Intelligence (AI) Lab, Deepwise Inc.

**Authors' contributions**
G.B.C., L.F.Y. and Y.Yang conceived the overall idea of this study, designed the research plan in detail, and revised the manuscript. Y.Yu, G.F.L. and X.Y.Q. sorted all relevant work and wrote the manuscript. T.Z., X.Y.H., Y.B.Z. and Z.Y.M. collected and organized the imaging data. Z.Z. and W.X.T., undertook the MATLAB program for lung parenchyma segmentation, nodule classification and diagnosis, and 3-D reconstruction. Y.L., M.Y., C.Y. and Y.G. designed the experiments to validate the effectiveness and revised the manuscript. All authors approved and reviewed the final version of the manuscript.

**Funding**
This study was supported by the Special subject of health care in China military logistics research project [No. 23BJZ14, GBC], the Air Force Medical University clinical research program of China [No. 2021LC2217, GBC], the Air Force Medical University health special research project of China [No. 22KYBJ03, GBC], and the major clinical research project of Tangdu Hospital of China [No. 2021LCYJ013, GBC].

**Data availability**
The datasets analysed during the current study available from the corresponding author on reasonable request.

## Declarations

**Ethics approval and consent to participate**
All the procedures were implemented based on the principles of the Declaration of Helsinki. Ethical approval was obtained from the Ethics Committee of Tangdu Hospital (No. K202108-18).

**Consent for publication**
Not applicable.

**Competing interests**
The authors declare no competing interests.

**Author details**
[1]Department of Radiology & Functional and Molecular Imaging Key Lab of Shaanxi Province, Tangdu Hospital, Fourth Military Medical University (Air Force Medical University), 569 Xinsi Road, Xi'an , Shaanxi 710038, China. [2]Deepwise Artificial Intelligence (AI), Deepwise Inc, 8 Haidian Street, Beijing 100080, China. [3]Department of Pulmonary and Critical Care Medicine, Tangdu Hospital, Fourth Military Medical University (Air Force Medical University), 569 Xinsi Road, Xi'an, Shaanxi 710038, China.

Received: 7 October 2024 Accepted: 14 February 2025
Published online: 26 February 2025

**References**

1. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, Zegers CM, Gillies R, Boellard R, Dekker A, Aerts HJ. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer. 2012;48(4):441–6.
2. Tomaszewski MR, Gillies RJ. The biological meaning of radiomic features. Radiology. 2021;299(2):E256.
3. O'Connor JP, Aboagye EO, Adams JE, Aerts HJ, Barrington SF, Beer AJ, Boellaard R, Bohndiek SE, Brady M, Brown G, et al. Imaging biomarker roadmap for cancer studies. Nat Rev Clin Oncol. 2017;14(3):169–86.
4. Shi C, Cheng Y, Wang J, Wang Y, Mori K, Tamura S. Low-rank and sparse decomposition based shape model and probabilistic atlas for automatic pathological organ segmentation. Med Image Anal. 2017;38:30–49.
5. Avanzo M, Stancanello J, El Naqa I. Beyond imaging: the promise of radiomics. Phys Med. 2017;38:122–39.
6. Liu Z, Wang S, Dong D, Wei J, Fang C, Zhou X, Sun K, Li L, Li B, Wang M, Tian J. The applications of radiomics in precision diagnosis and treatment of oncology: opportunities and challenges. Theranostics. 2019;9(5):1303–22.
7. Padmanaban V, Tsehay Y, Cheung KJ, Ewald AJ, Bader JS. Between-tumor and within-tumor heterogeneity in invasive potential. PLoS Comput Biol. 2020;16(1): e1007464.
8. van Timmeren JE, Leijenaar RTH, van Elmpt W, Wang J, Zhang Z, Dekker A, Lambin P. Test-retest data for radiomics feature stability analysis: generalizable or study-specific? Tomography. 2016;2(4):361–5.
9. Soret M, Bacharach SL, Buvat I. Partial-volume effect in PET tumor imaging. J Nucl Med. 2007;48(6):932–45.

10. Pavic M, Bogowicz M, Würms X, Glatz S, Finazzi T, Riesterer O, Roesch J, Rudofsky L, Friess M, Veit-Haibach P, et al. Influence of inter-observer delineation variability on radiomics stability in different tumor sites. Acta Oncol. 2018;57(8):1070–4.
11. Zhao B, Tan Y, Bell DJ, Marley SE, Guo P, Mann H, Scott ML, Schwartz LH, Ghiorghiu DC. Exploring intra- and inter-reader variability in uni-dimensional, bi-dimensional, and volumetric measurements of solid tumors on CT scans reconstructed at different slice intervals. Eur J Radiol. 2013;82(6):959–68.
12. Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA, Schabath MB, Forster K, Aerts HJ, Dekker A, Fenstermacher D, et al. Radiomics: the process and the challenges. Magn Reson Imaging. 2012;30(9):1234–48.
13. Yamashita R, Perrin T, Chakraborty J, Chou JF, Horvat N, Koszalka MA, Midya A, Gonen M, Allen P, Jarnagin WR, et al. Radiomic feature reproducibility in contrast-enhanced CT of the pancreas is affected by variabilities in scan parameters and manual segmentation. Eur Radiol. 2020;30(1):195–205.
14. Jin J, Zhu H, Zhang J, Ai Y, Zhang J, Teng Y, Xie C, Jin X. Multiple U-Net-based automatic segmentations and radiomics feature stability on ultrasound images for patients with ovarian cancer. Front Oncol. 2020;10: 614201.
15. Mottola M, Ursprung S, Rundo L, Sanchez LE, Klatte T, Mendichovszky I, Stewart GD, Sala E, Bevilacqua A. Reproducibility of CT-based radiomic features against image resampling and perturbations for tumour and healthy kidney in renal cancer patients. Sci Rep. 2021;11(1):11542.
16. Teng Y, Ai Y, Liang T, Yu B, Jin J, Xie C, Jin X. The effects of automatic segmentations on preoperative lymph node status prediction models with ultrasound radiomics for patients with early stage cervical cancer. Technol Cancer Res Treat. 2022;21: 15330338221099396.
17. Stefano A, Comelli A, Bravatà V, Barone S, Daskalovski I, Savoca G, Sabini MG, Ippolito M, Russo G. A preliminary PET radiomics study of brain metastases using a fully automatic segmentation method. BMC Bioinformatics. 2020;21(Suppl 8):325.
18. Tixier F, Um H, Young RJ, Veeraraghavan H. Reliability of tumor segmentation in glioblastoma: Impact on the robustness of MRI-radiomic features. Med Phys. 2019;46(8):3582–91.
19. Saha A, Grimm LJ, Harowicz M, Ghate SV, Kim C, Walsh R, Mazurowski MA. Interobserver variability in identification of breast tumors in MRI and its implications for prognostic biomarkers and radiogenomics. Med Phys. 2016;43(8):4558.
20. Beresford MJ, Padhani AR, Taylor NJ, Ah-See ML, Stirling JJ, Makris A, d'Arcy JA, Collins DJ. Inter- and intraobserver variability in the evaluation of dynamic breast cancer MRI. J Magn Reson Imaging. 2006;24(6):1316–25.
21. Bianconi F, Fravolini ML, Pizzoli S, Palumbo I, Minestrini M, Rondini M, Nuvoli S, Spanu A, Palumbo B. Comparative evaluation of conventional and deep learning methods for semi-automated segmentation of pulmonary nodules on CT. Quant Imaging Med Surg. 2021;11(7):3286–305.
22. Huang B, Lin X, Shen J, Chen X, Chen J, Li ZP, Wang M, Yuan C, Diao XF, Luo Y, Feng ST. Accurate and feasible deep learning based semi-automatic segmentation in CT for radiomics analysis in pancreatic neuroendocrine neoplasms. IEEE J Biomed Health Inform. 2021;25(9):3498–506.
23. Bleker J, Kwee TC, Rouw D, Roest C, Borstlap J, de Jong IJ, Dierckx R, Huisman H, Yakar D. A deep learning masked segmentation alternative to manual segmentation in biparametric MRI prostate cancer radiomics. Eur Radiol. 2022;32(9):6526–35.
24. Caballo M, Pangallo DR, Mann RM, Sechopoulos I. Deep learning-based segmentation of breast masses in dedicated breast CT imaging: radiomic feature stability between radiologists and artificial intelligence. Comput Biol Med. 2020;118: 103629.
25. Shboul ZA, Alam M, Vidyaratne L, Pei L, Elbakary MI, Iftekharuddin KM. Feature-guided deep radiomics for glioblastoma patient survival prediction. Front Neurosci. 2019;13: 966.
26. Visin F, Ciccone M, Romero A, Kastner K, Cho K, Bengio Y, Matteucci M, Courville A. ReSeg: a recurrent neural network-based model for semantic segmentation. In: Computer vision & pattern recognition workshops: 2016. 2016.
27. Iek Z, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. Cham: Springer; 2016.
28. Zhang ZC. A fast weak-supervised pulmonary nodule segmentation method based on modified self-adaptive FCM algorithm. Soft computing. 2018;22(12):3983.
29. Mirderikvand N, Naderan M, Jamshidnezhad A. Accurate automatic localisation of lung nodules using graph cut and snakes algorithms. In: International conference on computer & knowledge engineering: 2017. 2017.
30. Mcknight PE, Najab J. Mann-Whitney U Test. In: The Corsini encyclopedia of psychology. 2010.
31. Gooch JW. Pearson correlation coefficient. In: Encyclopedic dictionary of polymers. 2011.
32. Kukreja SL, Lofberg J, Brenner MJ. A least absolute shrinkage and selection operator (LASSO) for nonlinear system identification. In: IFAC symposium on system identification: 2009. 2009.
33. Kotz S, Johnson NL. [Springer series in statistics] Breakthroughs in statistics || Introduction to Huber (1964) Robust estimation of a location parameter, vol. 10.1007/978-1-4612-4380-9. 1992.
34. Demler OV, Pencina MJ, D'Agostino RB Sr. Misuse of DeLong test to compare AUCs for nested models. Stat Med. 2012;31(23):2577–87.
35. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979;86(2):420–8.
36. Guiot J, Vaidyanathan A, Deprez L, Zerka F, Danthine D, Frix AN, Lambin P, Bottari F, Tsoutzidis N, Miraglio B, et al. A review in radiomics: making personalized medicine a reality via routine imaging. Med Res Rev. 2022;42(1):426–40.
37. D'Arnese E, Donato GWD, Sozzo ED, Sollini M, Sciuto D, Santambrogio MD. On the automation of radiomics-based identification and characterization of NSCLC. IEEE J Biomed Health Inform. 2022;26(6):2670–9.
38. Zhang L, Luo Z, Chai R, Arefan D, Sumkin J, Wu S. Deep-learning method for tumor segmentation in breast DCE-MRI. In: Imaging informatics for healthcare, research, and applications: 2019. 2019.
39. Rios Velazquez E, Aerts HJ, Gu Y, Goldgof DB, De Ruysscher D, Dekker A, Korn R, Gillies RJ, Lambin P. A semiautomatic CT-based ensemble segmentation of lung tumors: comparison with oncologists' delineations and with the surgical specimen. Radiother Oncol. 2012;105(2):167–73.
40. Steger S, Sakas G. FIST: fast interactive segmentation of tumors. In: International conference on abdominal imaging: computational & clinical applications: 2011. 2011.
41. Kittaneh OA. The variance entropy multi-level thresholding method. Multimedia Tools Applications. 2023;82(28):43075.
42. Jumiawi WAH, El-Zaart A. Otsu thresholding model using heterogeneous mean filters for precise images segmentation. In: 2022 International Conference of Advanced Technology in Electronic and Electrical Engineering (ICATEEE). 2022. p. 1–6.
43. Li CH, Lee CK. Minimum cross entropy thresholding. Pattern Recogn. 1993;26(4):617–25.
44. Tan Y, Schwartz LH, Zhao B. Segmentation of lung lesions on CT scans using watershed, active contours, and Markov random field. Med Phys. 2013;40(4): 043502.
45. Kubota T, Jerebko AK, Dewan M, Salganicoff M, Krishnan A. Segmentation of pulmonary nodules of various densities with morphological approaches and convexity models. Med Image Anal. 2011;15(1):133–54.
46. Wang L, Zhou H, Xu N, Liu Y, Jiang X, Li S, Feng C, Xu H, Deng K, Song J. A general approach for automatic segmentation of pneumonia, pulmonary nodule, and tuberculosis in ct images. iScience. 2023;26(7):107005.
47. Astley JR, Biancardi AM, Hughes PJC, Marshall H, Collier GJ, Chan HF, Saunders LC, Smith LJ, Brook ML, Thompson R, et al. Implementable deep learning for multi-sequence proton MRI lung segmentation: a multi-center, multi-vendor, and multi-disease study. J Magn Reson Imaging. 2023;58(4):1030–44.
48. Comelli A. Artificial intelligence and statistical models for the prediction of radiotherapy toxicity in prostate cancer: a systematic review. Applied Sciences. 2024;14:10947.
49. Choi W, Oh JH, Riyahi S, Liu CJ, Jiang F, Chen W, White C, Rimner A, Mechalakos JG, Deasy JO. Radiomics analysis of pulmonary nodules in low-dose CT for early detection of lung cancer. Med Phys. 2018;45(4):1537–49.


## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.