Intratumoral and Peritumoral CT-Based Radiomics Strategy Reveals Distinct Subtypes of Nonsmall-Cell Lung Cancer

Xing Tang
Air Force Medical University Xijing Hospital: Xijing Hospital
Haolin Huang
Air Force Medical University School of Biomedical Engineering
Peng Du
Air Force Medical University School of Biomedical Engineering
Hong Yin
Air Force Medical University Xijing Hospital: Xijing Hospital
Xiaopan Xu (✉ alexander-001@163.com)
Air Force Medical University School of Biomedical Engineering https://orcid.org/0000-0003-3707-1104

Research Article

Keywords: Lung squamous cell carcinoma, Lung adenocarcinoma, CT radiomics, Intra- and peritumoral regions, Ensemble learning

Posted Date: November 2nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1029288/v1

Abstract

Purpose To evaluate a new radiomics strategy that incorporates peritumoral and intratumoral features extracted from lung CT images with ensemble learning for pretreatment prediction of lung squamous cell carcinoma (LUSC) and lung adenocarcinoma (LUAD).

Methods A total of 105 patients (47 LUSC and 58 LUAD) with pretherapy CT scans were involved in this retrospective study and were divided into training (n=73) and testing (n=32) cohorts. Seven categories of radiomics features involving 3078 metrics in total, were extracted from the intra- and peritumoral regions of each patient's CT data. Student's *t*-tests in combination with three feature selection methods were adopted for optimal features selection. An ensemble classifier that was generated with five machine learning classifiers and optimal features, was developed and the performance was quantitatively evaluated using both training and testing cohorts for the prediction task.

Results The classification models developed by using optimal feature subsets determined from intratumoral region and peritumoral region with the ensemble classifier achieved mean area under the curve (AUC) of 0.87, 0.83 in the training cohort and 0.66, 0.60 in the testing cohort, respectively. The model developed by using the optimal feature subset selected from both intra- and peritumoral regions with the ensemble classifier achieved great performance improvement, with AUC of 0.87 and 0.78 in both cohorts, respectively.

Conclusions The proposed new radiomics strategy that extracts image features from the intra- and peritumoral regions with ensemble learning, could greatly improve the diagnostic performance for the histological subtype stratification in patients with NSCLC.

1. Introduction

Lung cancer is the most frequently occurring cancer and the leading cause of cancer-related death in men globally [1]. In women, lung cancer is the third most commonly diagnosed cancer and the second most leading cause of cancer-related death [1]. Approximately 85% of primary lung malignancies are nonsmall-cell lung cancer (NSCLC), and the 5-year survival rate is less than 20% [2–6].

Lung squamous cell carcinoma (LUSC) and lung adenocarcinoma (LUAD) are two major histological subtypes of NSCLC that constitute approximately 35% and 60% of primary NSCLC cases, respectively [2, 3, 5, 7–9]. LUSC often shows keratinization, pearl formation, and intercellular bridges, whereas LUAD may exhibit lepidic, glandular, papillary or micropapillary, or solid architecture [2]. These two histological subtypes always present different anatomical sites and glucose metabolism levels, reflect the need for different optimal treatments to improve clinical outcomes [3, 6–8]. Therefore, accurately predicting LUSC and LUAD is of paramount importance prior to clinical interventions [6].

The first-line reference in preoperatively diagnosing LUSC and LUAD is lung biopsy [3, 6–8, 10], which is an invasive diagnostic approach with a high level of risks in clinical practice [11]. In addition, concerning the issue of tumor heterogeneity of NSCLC, lung biopsy examines only very limited proportions of the tumor tissue and is incapable of completely characterizing tumor properties [5, 7]. Developing a noninvasive strategy for the accurate prediction of LUSC and LUAD preoperatively is desirable.

Noninvasive imaging technologies, such as computed tomography (CT) and multiparametric magnetic resonance imaging (mpMRI), have recently been widely used for the pretherapy diagnosis of NSCLC [5, 7, 12–15]. Compared with mpMRI, CT offers considerably better imaging efficiency, higher resolution, and fewer motion artifacts caused by breathing and is thus recommended in the guidelines for NSCLC screening and diagnosis [2, 13]. However, it is very challenging for clinicians to visually predict the histological subtype of NSCLC directly from CT images to discriminate between LUSC and LUAD.

In recent years, radiomics strategies have been used for the prediction of LUSC and LUAD. In 2016, Wu *et al.* explored a CT-based radiomics strategy with 440 features extracted, and the Naïve Baye's classifier was used and achieved fair performance for the differentiation of LUSC and LUAD with an area under the curve (AUC) of the receiver operating characteristic (ROC) curve of 0.72 [16]. Bashir *et al.* extracted 115 radiomics features from CT data and developed a prediction model based on the optimal features and random forest (RF) classifier, achieving an AUC of 0.82 for discriminating between LUSC and LUAD [2]. Chaunzwa *et al.* introduced the convolutional neural network (CNN) to the prediction task and developed a prediction model based on the Visual Geometry Group-16 (VGG-16) network [17], obtaining an optimal AUC of 0.751.

In addition, some recent studies also integrated the radiomics strategy with positron emission tomography computed tomography (PET-CT) images, achieving favorable diagnostic performance in the differentiation of these two subtypes of NSCLC [18–20]. For instance, Koyasu *et al.* proposed a PET-CT-based radiomics strategy with an extreme gradient boosting (XGBoost) classifier for the prediction task [19], achieving good performance with an AUC of 0.843.

Although these previous studies have repeatedly demonstrated the feasibility of the radiomics strategy based on CT or PET-CT for the prediction of histological subtypes of NSCLC, all the features they extracted were from the intratumoral region of the image. We are not aware of any work that has attempted to evaluate the peritumoral area outside the tumor to distinguish LUSC from LUAD. According to a recent study [21], perinodular region-based radiomics features on lung CT images effectively reflect the difference between LUAD and granulomas and accurately distinguish these two types of lung nodules. Whether the radiomics features extracted from the peritumoral region of NSCLC can reflect the significant difference between LUSC and LUAD and further be used for the prediction task remains an open question to date.

Therefore, the first aim of this study was to investigate whether the radiomics features extracted from the peritumoral region of NSCLC could significantly reflect the difference between LUSC and LUAD. To achieve this goal, seven feature categories were employed in this study, including morphological features, histogram-based features (first-order features, hereafter), Haralick features of co-occurrence matrices (CM features, hereafter) [22], and features derived from the run length matrix (RLM features, hereafter) [23], the neighborhood gray-tone difference matrix (NGTDM features, hereafter) [24], the gray level size zone matrix (GLSZM features, hereafter) [25], and gray level dependence matrix (GLDM features, hereafter) [26] to fully characterize the global, local and regional differences of the tissue in the peritumoral region between LUSC and LUAD [27].

The second aim was to develop an accurate and consistent model for predicting LUSC and LUAD. To fulfil this aim, both intra- and peritumoral region-based radiomics features were utilized, and an ensemble classifier that combined multiple binary classifiers, such as support vector machine (SVM), RF and XGBoost, was used to form a more robust predictive model. The diagnostic performance of the model was then assessed with AUC for the differentiation of LUSC and LUAD.

2. Materials And Methods

This retrospective study was approved by the institutional ethics review board of Xijing Hospital, and informed content was waived. The overall methodological pipeline of this study is shown in Fig. 1.

2.1 Patients

A total of 146 archival patients with postoperatively confirmed NSCLC were collected from Xijing Hospital. The inclusion criteria were as follows: i) primary LUSC or LUAD was pathologically confirmed; ii) CT scan was performed prior to any therapies. Patients who met one of the following conditions were excluded: i) lack of postoperative pathological information to confirm the histopathological subtype of the patient as LUSC or LUAD (n=21); ii) missing preoperative CT scan (n=16); or iii) poor imaging quality makes accurate tumor annotations extremely difficult (n=4). Finally, 105 subjects were eligible for this study, including 47 patients with LUSC and 58 patients with LUAD. The patients were then randomly allocated into the training cohort (n=73) and testing cohort (n=32). The inclusion-exclusion process is illustrated in Fig. 2.

2.2 Image acquisition and region of interest annotation

All patients underwent thoracic CT imaging using a uCT 760 system (United Imaging Healthcare, China). The primary scanning parameters were as follows: 80 kV; 80 mAs; detector collimation: 64 × 0.6 mm; rotation time: 0.4 s; slice thickness: 5 mm; spacing between slices:5 mm; pixel spacing: 0.6 × 0.6 mm; and matrix size, 512 × 512. The entire lung region was scanned in each patient, and the image slice varied from 100 to 400.

Two types of regions of interest (ROIs) including intra- and peritumoral regions, were annotated from the CT images, as shown in Fig. 3. Prior to the intratumoral region annotation of each CT dataset, the axial image slice was selected to obtain the largest area of the archived tumor with the maximal size in each patient’s lung region. Then, a manually depicted polygonal ROI was used to segment the intratumor region on the selected image slice. Two radiologists with 20 and 10 years of lung CT interpretation experience independently performed intratumoral region delineation using a custom-developed package. Then, divergence of their delineation results was carefully corrected by consensus.

After the intratumoral region mask was obtained, we adopted the morphological dilation operator to generate a new region mask that was approximately 10 mm larger in radial distance than the intratumoral region according to pixel size [21]. Then, the corresponding peritumoral region was the ring of the lung parenchyma around the tumor that was obtained by subtracting the intratumoral region mask from the new region mask after morphological expansion, as shown in Fig. 3. Finally, the peritumoral region was further divided into two rings including the first ring (0-5 mm) and the second ring (5-10 mm) for feature extraction and comparison [21].

2.3 Radiomics feature extraction

After intra- and peritumoral ROI segmentation, 10 filters including wavelet-HL, wavelet-LL, wavelet-LH, wavelet-HH, square, square root, logarithm, exponential, gradient, and local binary pattern (LBP), were utilized to the original image to magnify the tissue patterns and unearth important features. Then, six feature categories, including first-order features, GLCM features, GLRLM features, NGTDM features, GLSZM features and GLDM, were calculated from the original segmented image data and 10 generated images of the intratumoral and two rings of the peritumoral regions [28]. Given that the peritumoral region was dilated based on use of the intratumoral region, the shape 2D features were only calculated from the intratumoral region. Therefore, 1032, 1023, 1023 radiomics features were extracted from the intratumoral region and the first ring and the second ring of the peritumoral region, respectively, as shown in Table 1. Open source Pyradiomics (version 3.0.1) was used to perform this analysis [29]. All of the codes and results have been attached in the Appendix document.

Table 1
The demographics and clinical data of eligible patients

Characteristics	**Training cohort** **(n = 73)**	**Testing cohort** **(n = 32)**	***p*-value**
Age, years			0.87 [a]
Median (Range)	61 [35, 76]	59 [39, 83]	
Sex, No. (%)			0.91 [b]
Male	54 / 73 (73.97%)	24 / 32 (75.00%)	
Female	19 / 73 (26.03%)	8 / 32 (25.00%)	
Smoking, No. (%)			
Yes	49 / 73 (67.12%)	20 / 32 (62.50%)	0.65 [b]
No	24 / 73 (32.88%)	12 / 32 (37.50%)	
Side, No. (%)			0.90 [b]
Upper left lobe	22 / 73 (30.14%)	10 / 32 (31.25%)_	
Lower left lobe	12 / 73 (16.44%)	4 / 32 (12.50%)	
Upper right lobe	20 / 73 (27.40%)	7 / 32 (21.88%)	
Middle right lobe	2 / 73 (2.74%)	1 / 32 (3.13%)	
Lower right lobe	17 / 73 (23.29%)	10 / 32 (31.25%)	
Histopathological subtype, No. (%)			0.89 [b]
Squamous cell carcinoma (LUSC)	33 / 73(45.21%)	14 / 32(43.75%)	
Adenocarcinoma (LUAD)	40 / 73(54.79%)	18 /32(56.25%)	
a: Student's *t*-test			
b: Chi square test			

2.4 Feature selection

In this study, a two-step feature selection strategy was adopted to determine an optimal subset of features for model construction, as shown in Fig. 1. The first step was statistical analysis of all these features between LUSC and LUAD, which was performed with Scikit-learn. Student’s *t*-test with a significant *p*-value set as 0.05 was then performed with all radiomics features to select those with significant intergroup differences between LUSC and LUAD [30].

Then, all significant features were standardized to eradicate differences of the feature-value scales. The normalized feature z of each feature x for a specific patient is calculated as follows:

$$z = \frac{x - \overline{x}}{\sigma} \quad (1)$$

where $\overline{x}$ and σ are the mean and standard deviation, respectively, of each feature from the training cohort.

In the second step of feature selection, three widely-used feature selection algorithms, including the minimum redundancy maximum relevance method (mRMR) [31], the least absolute shrinkage and selection operator(LASSO) [32, 33], and the linear SVM-based recursive feature elimination (SVM-RFE) [34], were further implemented with these significant features to select an optimal feature subset from the training cohort for model development and external testing.

2.5 Model development based on ensemble learning and validation

With optimal features selected, the predictive model was developed using the training cohort and the ensemble learning strategy, which includes five commonly used binary classifiers, including the quadratic discriminant analysis(QDA) classifier, SVM with radial basis function(RBF) kernel, SVM with sigmoid/tanh kernel, RF, and XGBoost. QDA is the most commonly used binary classifier, which has no same-covariance assumption for each binary class [35, 36]. SVM is a classical machine learning classifier with several typical kernels, such as RBF and sigmoid/tanh, that is used to compute the decision boundary that separates two classes with the maximum marginal distance [37–39]. It has advantages in dealing with nonlinear features and is not easily overfit with even small datasets [40]. The RF classifier can build multiple random decision trees (100 trees of the default parameter in Scikit-learn to avoid overfitting) and integrate them to make an accurate diagnosis [40–42]. XGBoost offers many benefits in classification, including high precision and consistency and the prevention of overfitting [43, 44]; thus, it was also included in the ensemble learning strategy.

The ensemble classifier was finally developed by weighting the predictive value of these five classifiers in the model training process, which can be expressed as follows:

$$P(j) = \sum_{i=1}^{5} \omega_i p_i(j) \quad (2)$$

where $P(j)$ represents the final predictive value of the *j*-th patient; $p_i(j)$ denotes the predictive value of the *j*-th patient by using the *i*-th classifier; and ω_i is the weighting parameter of the *i*-th classifier in the ensemble learning process, which meets the following condition:

$$\sum_{i=1}^{5} \omega_i = 1 \qquad (3)$$

In this study, the optimal weight ω_i was determined based on minimizing the predictive error in the training process, and the cutoff $P(j)$ for assigning the patient to the LUAD group was set as 0.5. If $P(j)$ was greater than or equal to 0.5, the *j*-th patient was allocated to the LUAD group. The overall performance was evaluated using both the training cohort and the testing cohort with the quantitative metric of AUC [45–48]. The AUC value was widely used to comprehensively evaluate the general performance of the model developed for the prediction task [45–48].

2.6 Statistical analysis

Statistical analyses of the patient demographics were performed using IBM SPSS statistics (version 19.0, Armonk, NY), and **Python** software (version 3.6 DL-GPU) was used to perform statistical selection of features with significant differences between LUSC and LUAD. Chi square tests were performed to evaluate significant differences in primary clinical factors distributed between the training and testing cohorts, and Student's *t*-tests were used to select significant radiomics features between LUSC and LUAD. Two-sided *p*-values less than 0.05 were considered significant [27, 49, 50].

3. Results

3.1 Demographics of eligible patients

A total of 105 NSCLC patients were eligible for this study, including 47 patients with LUSC and 58 with LUAD. These patients were randomly allocated into the training cohort (n=73) and the testing cohort (n=32). The baseline demographics and clinical information of these patients was collected from the archival medical document, as shown in Table 1. Statistical analyses indicate no significant differences between both the training and testing cohorts in terms of all these primary factors.

3.2 Results of the two-step feature selection strategy

A total of 3078 standardized radiomics features, including 1032 features from the intratumoral region, 1023 from the first ring (0-5 mm) and 1023 from the second ring (5-10 mm) of peritumoral regions, were analyzed using Student's *t*-test (*p*-value < 0.05) to determine those with significant intergroup differences between LUSC and LUAD. Eventually, 500 significant features were selected from the intratumoral region, whereas, only 220 and 119 significant features were selected from the first ring and second ring of peritumoral regions, respectively, as shown in Fig. 4. These results indicate that i) a large number of radiomics features extracted from the peritumoral region can also reflect the significant differences in tissue distribution patterns between LUSC and LUAD; ii) the closer the peritumoral region is located to the intratumoral region, the more features with significant differences could be obtained to reflect the tumor property difference. Fig. 5 illustrates an example of the intra- and peritumoral tissue distribution differences of LUSC and LUAD determined using one of the significant radiomics features, energy, with 3×3 sliding patches on the CT image.

After statistical analysis-based feature selection, three radiomics feature subsets were finally obtained, including i) 500 significant features from the intratumoral region, ii) 339 significant features from the entire peritumoral region, and iii) 839 significant features from both intratumoral and peritumoral regions. All these significant features in each feature subset were further selected using three commonly applied strategies: SVM-RFE, LASSO, and mRMR with the mutual information difference (MID), as shown in Figs. 6 - 8. Table 2 shows the results after the second-step feature selection procedure.

Table 2
Results after using the second-step feature selection strategy

Method	Optimal features selected from 500 significant features of the intratumoral region	Optimal features selected from 339 significant features of the peritumoral region	Optimal features selected from 839 significant features of both intra- and peritumoral regions
SVM-RFE	12	6	9
LASSO	6	6	8
mRMR with MID	12	12	12

3.3 Classification model development and performance evaluation

As these optimal feature subsets were determined, classification models were developed using five commonly used machine learning classifiers and the ensemble classifier with the training cohort, and the performance of each model was evaluated by using both training and testing cohorts for distinguishing LUSC from LUAD. The results are presented in Fig. 9. Three columns of subfigures in Fig. 9 exhibit the performance of predictive models developed by using optimal feature subsets determined from the intratumoral region, peritumoral region, and both intra- and peritumoral regions. These findings indicate that i) the classification model determined from the peritumoral region achieved comparable performance to that from the intratumoral region; ii) the classification model determined from intra- and peritumoral regions dramatically improved the overall performance for the prediction of LUSC and LUAD; and 3) the model developed by the ensemble classifier achieved more favorable and consistent performance with training and testing cohorts compared with those developed by five independent classifiers. Table 3 shows the performance of classification models developed by the ensemble classifier for the prediction task, indicating that the ensemble classification model developed by SVM-RFE-based optimal features determined from intra- and peritumoral regions achieved the best performance with AUC values of 0.87 and 0.78 in the training and testing cohorts, respectively.

Table 3
Performance of classification models developed by the ensemble classifier for the prediction of LUSC and LUAD with training and testing cohorts

Method	Classifier	From the intratumoral region		From the peritumoral region		From intra- & peritumoral regions	
		Training	Testing	Training	Testing	Training	Testing
SVM-RFE	Ensemble	**0.87**	**0.66**	**0.83**	0.60	**0.87**	**0.78**
LASSO	Ensemble	0.76	0.63	0.73	**0.63**	0.79	0.68
mRMR with MID	Ensemble	0.77	0.68	0.64	0.56	0.73	0.71
# The AUC value with **bold** ranks as the top place in each column.							

4. Discussion

In this study, we investigated the feasibility of CT-based radiomic features extracted from intra- and peritumoral regions of NSCLC to reflect the tissue distribution differences between LUSC and LUAD, and developed a CT-based radiomics strategy that incorporated high-throughput features with an ensemble classifier for the preoperative prediction of LUSC and LUAD. Three widely used methods, SVM-RFE, LASSO, and mRMR, were employed to select optimal features with significant intergroup differences between LUSC and LUAD for classification model development. Five independent classifiers, QDA, SVM with RBF kernel, SVM with sigmoid/tanh kernel, RF, and XGBoost, which were reported to have favorable classification performance and robustness for the diagnosis of cancer phenotypes with a small database, were utilized to form an ensemble classifier for classification model building. The results of the model that was developed using the ensemble classifier and optimal features selected by SVM-RFE from intra- and peritumoral regions demonstrate favorable discriminative power with both the training and testing cohorts.

In recent years, CT-/PET-CT/multimodal MRI-based radiomics strategies have been repeatedly demonstrated to have great capability for the prediction of LUSC and LUAD [2, 9, 16, 18–20]. The diagnostic performance ranged between 0.72 and 0.843. Nevertheless, all these previous studies only focused on how to extract an increasing number of features from the intratumoral region of the image, regardless of the peritumoral parenchyma, which might also contain substantial information and be of equal importance for the prediction task. Some studies have revealed that the interface of the tumor has a “rim” of densely packed tumor-infiltrating lymphocytes and tumor-associated macrophages in representative hematoxylin and eosin–stained images [8, 21, 51, 52]. At a macroscopic scale, the densely packed stromal tumor-infiltrating lymphocytes around LUAD represent fine and smooth textures on CT images and thus could be potential imaging biomarkers for the identification of LUAD from LUSC [21]. However, whether radiomics features extracted from the peritumoral parenchyma region effectively reflect the intergroup difference of the tissue and microenvironment between LUSC and LUAD, remains unknown to date.

In this study, we found that a large number of radiomics features extracted from the intratumoral region and peritumoral region were significantly different between LUSC and LUAD, and the total number of significant features extracted from the first ring (0-5 mm) peritumoral region was much greater than that of the significant features extracted from the second ring (5-10 mm) peritumoral region. These results demonstrate and verify for the first time the hypothesis that the peritumoral region on CT images also contains substantial information that can reflect the tissue texture difference between LUSC and LUAD. In addition, the closer the peritumoral region is to the intratumoral region, the more substantial the information it contains.

Most of the previous studies only focused on extracting features from the original image data, neglecting the image filters that not only reduce the noise but also enhance the quality and magnify the texture in the image [53, 54]. Therefore, in this study, 10 filters including wavelet-HL, wavelet-LL, wavelet-LH, wavelet-HH, square, square root, logarithm, exponential, gradient, and LBP were utilized to preprocess the image for feature extraction. Seven categories of radiomics features, including morphological features, first-order features, second-order features, and high-order texture features, were adopted in this study to fully characterize the shape properties and global, local and regional distribution patterns of the tissue, respectively. Student's *t*-tests integrated with three widely applied feature selection algorithms (SVM-RFE, LASSO and mRMR), were adopted for optimal feature selection and performance comparison. The results indicate that the optimal features selected using the SVM-RFE algorithm from all significant features of both intra- and peritumoral regions have the most powerful diagnostic ability for the discrimination between LUSC and LUAD.

Classification model development is the last but most crucial step in the proposed radiomics strategy for the prediction of LUSC and LUAD. In this step, the choice of an optimal decision classifier, for instance, SVM with RBF kernel or Sigmiod kernel, RF, QDA, or XGBoost represent the core influence of performance variation [40]. Hence, the determination of an optimal classifier is of critical importance. To fully integrate all the merits of these five independent classifiers, an ensemble classifier was generated using five independent classifiers, SVM with RBF kernel, or sigmoid kernel, RF, QDA, and XGBoost, and its diagnostic performance was compared with these independent classifiers. The results indicate that i) the classification model developed using the ensemble classifier achieves the most favorable, consistent and robust diagnostic performance compared with other independent classifiers, and ii) optimal features determined by SVM-RFE from both intra- and peritumoral regions with the ensemble classifier achieve the best diagnostic performance for the prediction of LUSC and LUAD with both training and testing cohorts. In addition, the classification results of all these models developed by each classifier with optimal features determined from intratumoral, peritumoral, or both of intratumoral and peritumoral regions using SVM-RFE, LASSO, and mRMR also revealed that although the model based on the ensemble classifier did not always obtain the best results, it always ranked as one of the top two models in terms of the AUC with both cohorts, suggesting remarkable consistency and robustness in the prediction of LUSC and LUAD.

The limitations of this study include the following aspects. First, inherent bias might exist given the retrospective nature of the present study with relatively small patient cohorts collected from a single clinical center. A larger number of participants from two or more clinical centers are further required to validate the performance of the model we developed. Moreover, other potential clinical factors, such as gene mutations and key molecular biomarkers, were not included in the current study given the incomplete data in the archival database, which should be further analyzed. In addition, deep radiomics features incorporating the current manual radiomics features might further improve current performance in the prediction of LUSC and LUAD.

In conclusion, the proposed CT-based radiomics strategy that extracts features from intra- and peritumoral regions, adopts SVM-RFE for optimal feature selection, and utilizes ensemble learning for classification model development is demonstrated with favorable predictive precision and stability for preoperatively prediction of LUSC and LUAD.

Abbreviations

AUC, area under the curve; CM, co-occurrence matrices; CNN, convolutional neural network; CT, computed tomography; FN, false negative; FP, false positive; GLCM, gray level co-occurrence matrix; GLDM, gray level dependence matrix; GLRLM, gray level run length matrix; GLSZM, gray level size zone matrix; LASSO, least absolute shrinkage and selection operator; LBP, local binary pattern; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; MID, mutual information difference; mpMRI, multiparametric magnetic resonance imaging; mRMR, minimum redundancy maximum relevance; NGTDM, neighbouring gray tone difference matrix; NSCLC, nonsmall-cell lung cancer; PET-CT, positron emission tomography computed tomography; QDA, quadratic discriminant analysis; RBF, radial basis function; RF, random forest; RLM, run length matrix; ROC, receiver operating characteristic curve; SVM, support vector machine; SVM-RFE, support vector machine -based recursive feature elimination; TN, true negative; TP, true positive; VGG, visual geometry group network; XGBoost, extreme gradient boosting.

Declarations

Funding

This work was funded by the National Natural Science Foundation of China (No. 81901698), and Young Eagle plane of High Ambition Project (No. 2020CYJHXXP).

Conflict of Interest

The authors declare that they have no conflict of interest.

Ethics Approval

This study was approved by the institutional ethics review board of Xijing Hospital, and informed content was waived.

Data Availability Statement

The raw/processed data of this study cannot be publicly shared at present as it forms part of an ongoing study, but it could be available under reasonable request from the corresponding author with the permission of the Institutional Review Board. Results and code package in each step of this study have been arranged in a document named as "**Appendix**". The code package has also been uploaded to Gitee for publicly sharing and further perfection (https://gitee.com/yang-tianran-01/radiomics_-ensemble_learning.git)

Author contributions

Xiaopan Xu, Xing Tang and Haolin Huang contributed to the study concept, design, and data interpretation. Xing Tang contributed to the CT and clinical data collection. Xingtang and Hong Yin contributed to the intratumoral region annotation. Haolin Huang, Xiaopan Xu and Peng Du performed the peritumoral region extraction and radiomics feature calculcation; Xiaopan Xu, Haolin Huang and Xing Tang contributed to the model construction and data analysis. Xiaopan Xu, Xing Tang and Haolin Huang contributed to the manuscript drafting, editing and revision. All authors approve the final version of the manuscript for submission.

References

1. Sung H et al (2021) Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. A Cancer Journal for Clinicians, CA, p caac.21660
2. Bashir U et al (2019) Non-invasive classifcation of non-small cell lung cancer: a comparison between random forest models utilising radiomic and semantic features. Br J Radiol 92(20190159):1–8
3. Herbst RS, Heymach JV, Lippman SM (2008) Lung Cancer. The New England Journal of Medicine 359(13):1367–1380
4. Bray F et al., *Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.* CA: A Cancer Journal for Clinicians, 2018. **doi:10.3322/caac.21492**
5. Su R et al., *Identification of Expression Signatures For Non-Small-Cell Lung Carcinoma Subtype Classification.* Bioinformatics, 2019
6. Ma Y et al (2018) Intra-tumoural heterogeneity characterization through texture and colour analysis for differentiation of non-small cell lung carcinoma subtypes. Phys Med Biol 63(16):165018
7. Zhu X et al (2018) Radiomic signature as a diagnostic factor for histologic subtype classification of non-small cell lung cancer. Eur Radiol 28(7):1–7
8. Hoffman PC, Mauer AM, Vokes EE (2000) Lung cancer. Lancet 355(9202):479–485

9. Tang X et al., *Elaboration of a multimodal MRI-based radiomics signature for the preoperative prediction of the histological subtype in patients with non-small-cell lung cancer*. BioMedical Engineering OnLine, 2020. 19(1)
10. Mahon RN, Hugo GD, Weiss E, *Repeatability of texture features derived from magnetic resonance and computed tomography imaging and use in predictive models for non-small cell lung cancer outcome*. Phys Med Biol, 2019
11. Ebrahimi M et al (2016) Diagnostic concordance of non–small cell lung carcinoma subtypes between biopsy and cytology specimens obtained during the same procedure. Cancer Cytopathol 124(10):737–743
12. Sun W et al (2018) Effect of machine learning methods on predicting NSCLC overall survival time based on Radiomics analysis. Radiat Oncol 13(1):197
13. Starkov P et al (2018) The use of texture-based radiomics CT analysis to predict outcomes in early-stage non-small cell lung cancer treated with stereotactic ablative radiotherapy. Br J Radiol 91(20180228):1–7
14. Sollini M et al (2017) PET Radiomics in NSCLC: state of the art and a proposal for harmonization of methodology. Sci Rep 7(1):358
15. Shen C et al (2017) 2D and 3D CT Radiomics Features Prognostic Performance Comparison in Non-Small Cell Lung Cancer. Transl Oncol 10(6):886–894
16. Wu W et al (2016) Exploratory Study to Identify Radiomics Classifiers for Lung Cancer Histology, 6. Frontiers in Oncology
17. Chaunzwa TL et al (2018) Using deep-learning radiomics to predict lung cancer histology. J Clin Oncol 36(15_suppl):8545–8545
18. Ma Y et al (2018) Intra-tumoural heterogeneity characterization through texture and colour analysis for differentiation of non-small cell lung carcinoma subtypes. Phys Med Biol 63(16):165018
19. Koyasu S et al., *Usefulness of gradient tree boosting for predicting histological subtype and EGFR mutation status of non-small cell lung cancer on 18F FDG-PET/CT.* Annals of Nuclear Medicine, 2020. 34(1): p. 49–57
20. Ren C et al (2020) Machine learning based on clinico-biological features integrated 18F-FDG PET/CT radiomics for distinguishing squamous cell carcinoma from adenocarcinoma of lung. European Journal of Nuclear Medicine and Molecular Imaging
21. Beig N et al., *Perinodular and Intranodular Radiomic Features on Lung CT Images Distinguish Adenocarcinomas from Granulomas* Radiology, 2019. 290(3): p. 783–792
22. Haralick RM, Shanmugam K, Dinstein IH (1973) Textural Features for Image Classification. IEEE Transactions on Systems Man Cybernetics SMC-3(6):610–621
23. Galloway MM (1975) Texture Analysis Using Gray Level Run Lengths. Computer Graphics Image Processing 4:172–179

24. Amadasun M, King R (1989) Texural Features Corresponding to Texural Properties. IEEE Transactions on Systems Man Cybernetics 19(5):1264–1274
25. Thibault G, Angulo J, Meyer F (2014) Advanced statistical matrices for texture characterization: application to cell classification. IEEE Trans Biomed Eng 61(3):630–637
26. Sun C, Wee WG, *Neighboring Gray Level Dependence Matrix for Texture Classification.* Compute Vision, Graphics, and Image Processing (1983) **23**: p. 341-352
27. Xu X et al., *A predictive nomogram for individualized recurrence stratification of bladder cancer using multiparametric MRI and clinical risk factors*. Journal of Magnetic Resonance Imaging, 2019 Apr 13. **0**(0)
28. Zwanenburg A et al., *Image biomarker standardisation initiative.* arXiv:1612.07003 [cs, eess], 2019
29. van Griethuysen JJM et al (2017) Computational Radiomics System to Decode the Radiographic Phenotype. Can Res 77(21):e104–e107
30. Kotz S, Johnson NL, Editors (1992) The Probable Error of a Mean. In: Breakthroughs in Statistics: Methodology and Distribution. Springer, New York, pp 33–57
31. Peng H, Long F, Ding C (2005) Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27(8):1226–1238
32. Tibshirani R (1996) Regression Shrinkage and Selection Via the Lasso. J Roy Stat Soc B 58(1):267–288
33. Sauerbrei W, Royston P, Binder H (2007) Selection of important variables and determination of functional form for continuous predictors in multivariable model building. Stat Med 26(30):5512–5528
34. Fehr D et al., *Automatic classification of prostate cancer Gleason scores from multiparametric magnetic resonance images.* Proceedings of the National Academy of Sciences, 2015. **112**(46): p. E6265-E6273
35. *Linear & Quadratic Discriminant Analysis · UC Business Analytics R Programming Guide*
36. Tharwat A (2016) Linear vs. quadratic discriminant analysis classifier: a tutorial. International Journal of Applied Pattern Recognition 3(2):145
37. Hastie T, Tibshirani R, Friedman J, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition* (2009) Springer Science & Business Media. 757
38. Lam H-K et al (2012) Computational Intelligence and Its Applications: Evolutionary Computation, Fuzzy Logic, Neural Network and Support Vector Machine Techniques. World Scientific, p 318
39. Stenzinger A et al., *Artificial intelligence and pathology: From principles to practice and future applications in histomorphology and molecular profiling*. Seminars in Cancer Biology, 2021: p. S1044579X21000341
40. Liang C et al (2018) A computer-aided diagnosis scheme of breast lesion classification using GLGLM and shape features: Combined-view and multi-classifiers. Phys Med 55:61–72

41. Khalilia M, Chakraborty S, Popescu M (2011) Predicting disease risks from highly imbalanced data using random forest. BMC Med Inform Decis Mak 11(1):51
42. Seera M, Lim CP (2014) A hybrid intelligent system for medical data classification. Expert Syst Appl 41(5):2239–2249
43. Chen T, Guestrin C (2016) XGBoost: A Scalable Tree Boosting System. Association for Computing Machinery
44. Colen RR et al (2021) Radiomics analysis for predicting pembrolizumab response in patients with advanced rare cancers. Journal for ImmunoTherapy of Cancer 9(4):e001752
45. Gupta V, Mittal M, *R-Peak Detection in ECG Signal Using Yule–Walker and Principal Component Analysis.* IETE Journal of Research: p. 1-14
46. Gupta V, Mittal M (2019) A Comparison of ECG Signal Pre-Processing Using FrFT, FrWT and IPCA for Improved Analysis. IRBM
47. Gupta V, Mittal M, *QRS Complex Detection Using STFT, Chaos Analysis, and PCA in Standard and Real-Time ECG Databases*. Journal of The Institution of Engineers (India): Series B
48. Kora P, Krishna KSR. *Myocardial infarction detection using magnitude squared coherence and Support Vector Machine*. in *Medical Imaging, m-Health and Emerging Communication Systems (MedCom)*. 2014
49. Wu S et al (2017) A Radiomics Nomogram for the Preoperative Prediction of Lymph Node Metastasis in Bladder Cancer. Clin Cancer Res 23(22):6904–6911
50. Wu S et al (2018) Development and Validation of an MRI-Based Radiomics Signature for the Preoperative Prediction of Lymph Node Metastasis in Bladder Cancer. EBioMedicine 34:76–84
51. Kirienko M et al (2018) Prediction of disease-free survival by the PET/CT radiomic signature in non-small cell lung cancer patients undergoing surgery. Eur J Nucl Med Mol Imaging 45(2):207–217
52. de Jong EEC et al (2018) Applicability of a prognostic CT-based radiomic signature model trained on stage I-III non-small cell lung cancer in stage IV non-small cell lung cancer. Lung Cancer 124:6–11
53. Xu X et al (2017) Preoperative prediction of muscular invasiveness of bladder cancer with radiomic features on conventional MRI and its high-order derivative maps. Abdom Radiol (NY) 42(7):1896–1905
54. Xu X et al (2017) Three-dimensional texture features from intensity and high-order derivative maps for the discrimination between bladder tumors and wall tissues via MRI. Int J Comput Assist Radiol Surg 12(4):645–656

Figures

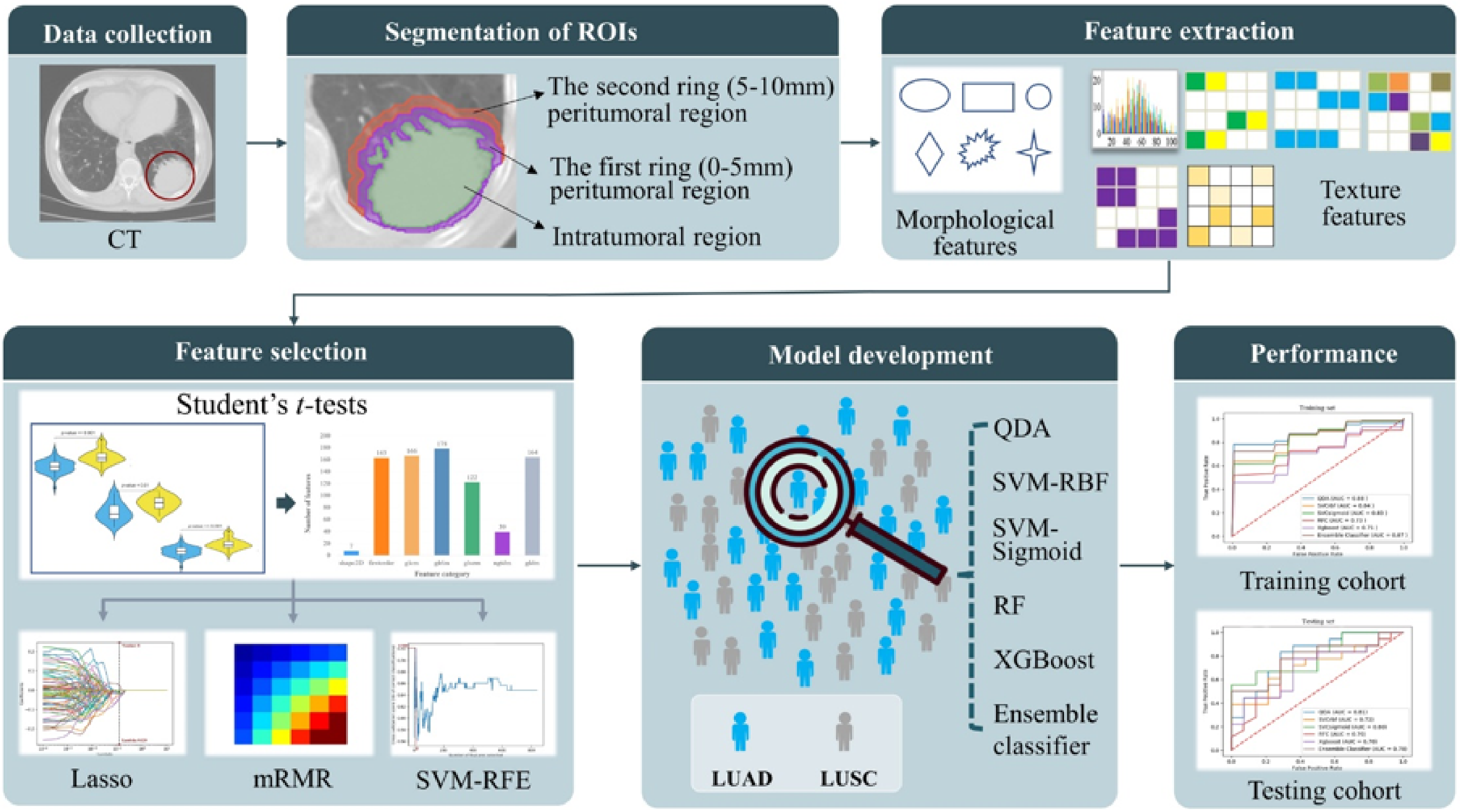


Figure 1

The schematic pipeline of the proposed strategy for the prediction of lung squamous cell carcinoma (LUSC) and lung adenocarcinoma (LUAD) via intra- and peritumoral CT radiomics features and ensemble learning

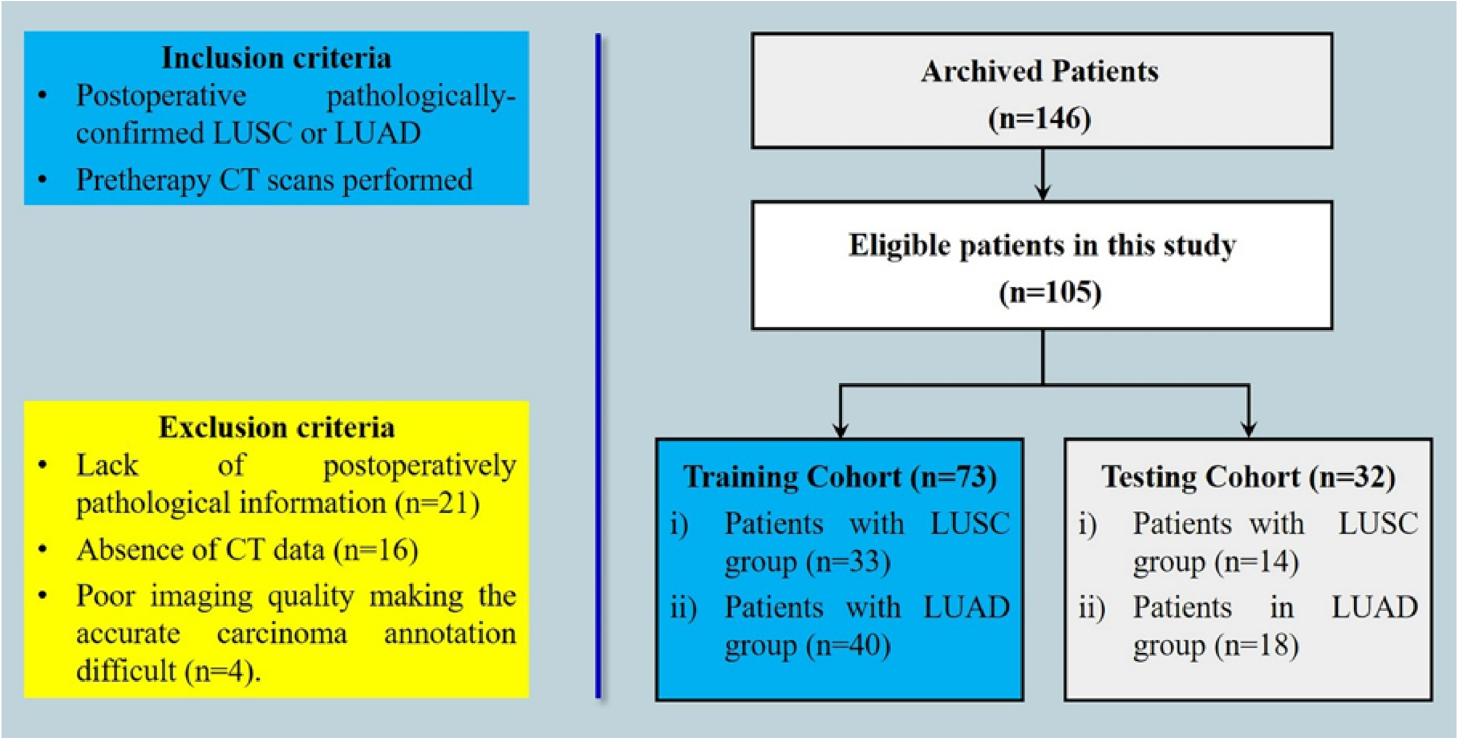

Inclusion criteria
• Postoperative pathologically-confirmed LUSC or LUAD
• Pretherapy CT scans performed
Exclusion criteria
• Lack of postoperatively pathological information (n=21)
• Absence of CT data (n=16)
• Poor imaging quality making the accurate carcinoma annotation difficult (n=4).
Archived Patients (n=146)
Eligible patients in this study (n=105)
Training Cohort (n=73)
i) Patients with LUSC group (n=33)
ii) Patients with LUAD group (n=40)
Testing Cohort (n=32)
i) Patients with LUSC group (n=14)
ii) Patients in LUAD group (n=18)

Figure 2

Inclusion-exclusion criteria of this study to obtain 105 eligible subjects including 47 ones with lung squamous cell carcinoma (LUSC) and 58 with lung adenocarcinoma (LUAD)

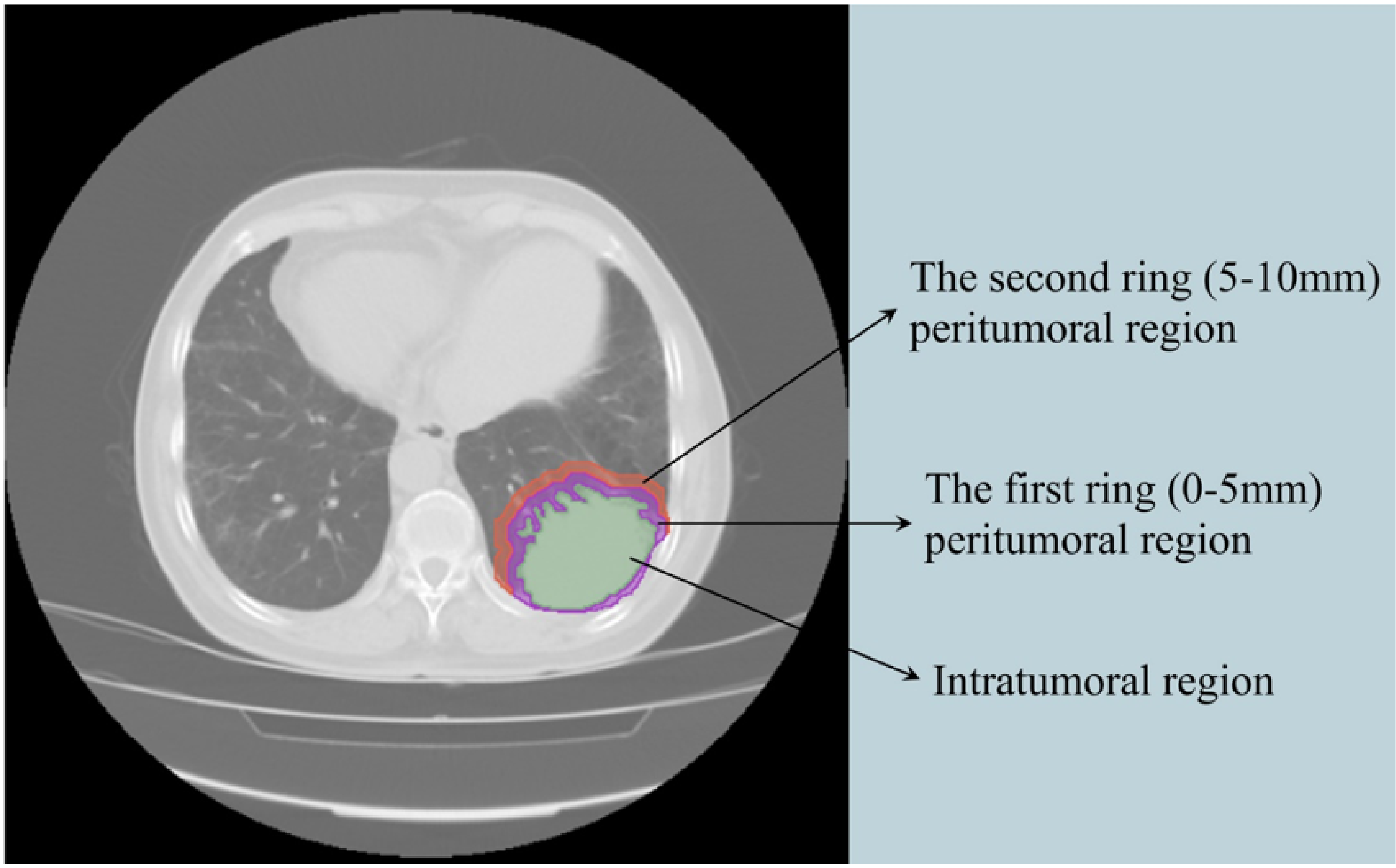

The second ring (5-10mm) peritumoral region
The first ring (0-5mm) peritumoral region
Intratumoral region

Figure 3

Illustration of the intratumoral region (light green) manually delineated and the first ring (0 – 5 mm, light purple) and second ring (5 – 10 mm, red) of the peritumoral regions generated by morphologically expanding the segmented intratumoral region mask

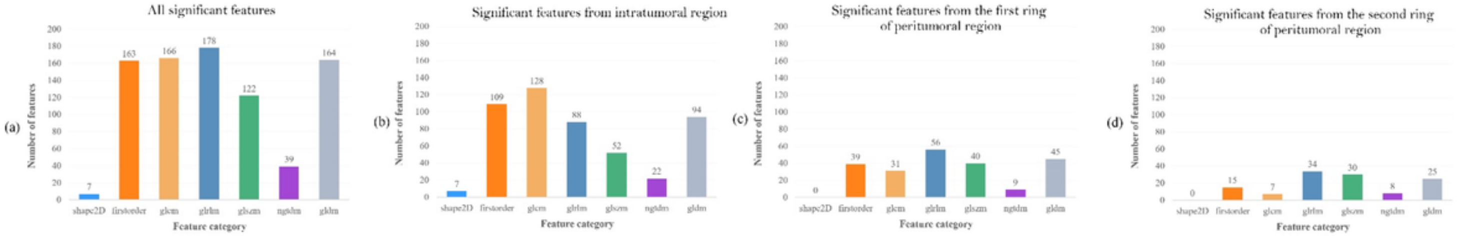


Figure 4

Statistical analysis-based feature selection results: (a) all 839 significant features from intra- and peritumoral regions; (b) 500 significant features from the intratumoral region; (c) 220 significant features from the first ring (0 – 5 mm) of peritumoral region; (d) 119 significant features from the second ring (5 – 10 mm) of peritumoral region

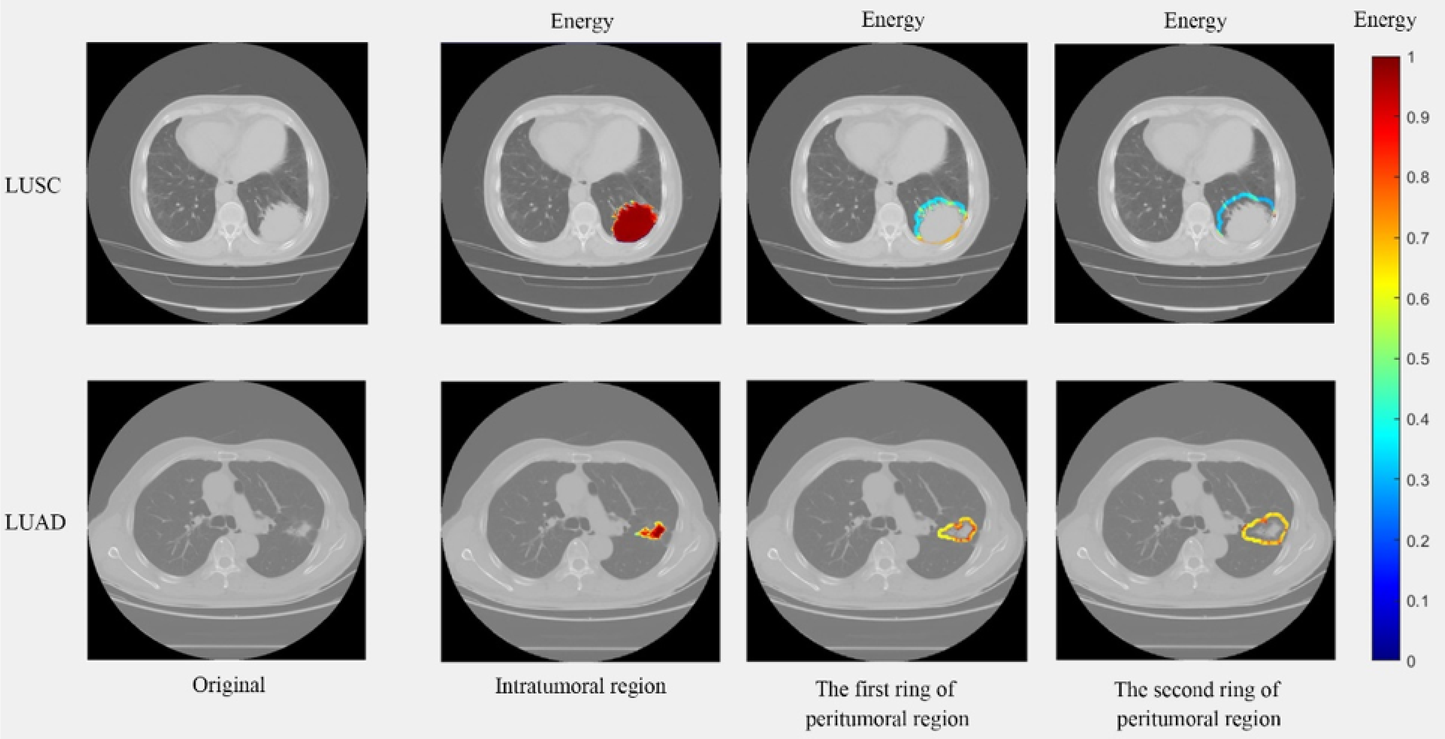
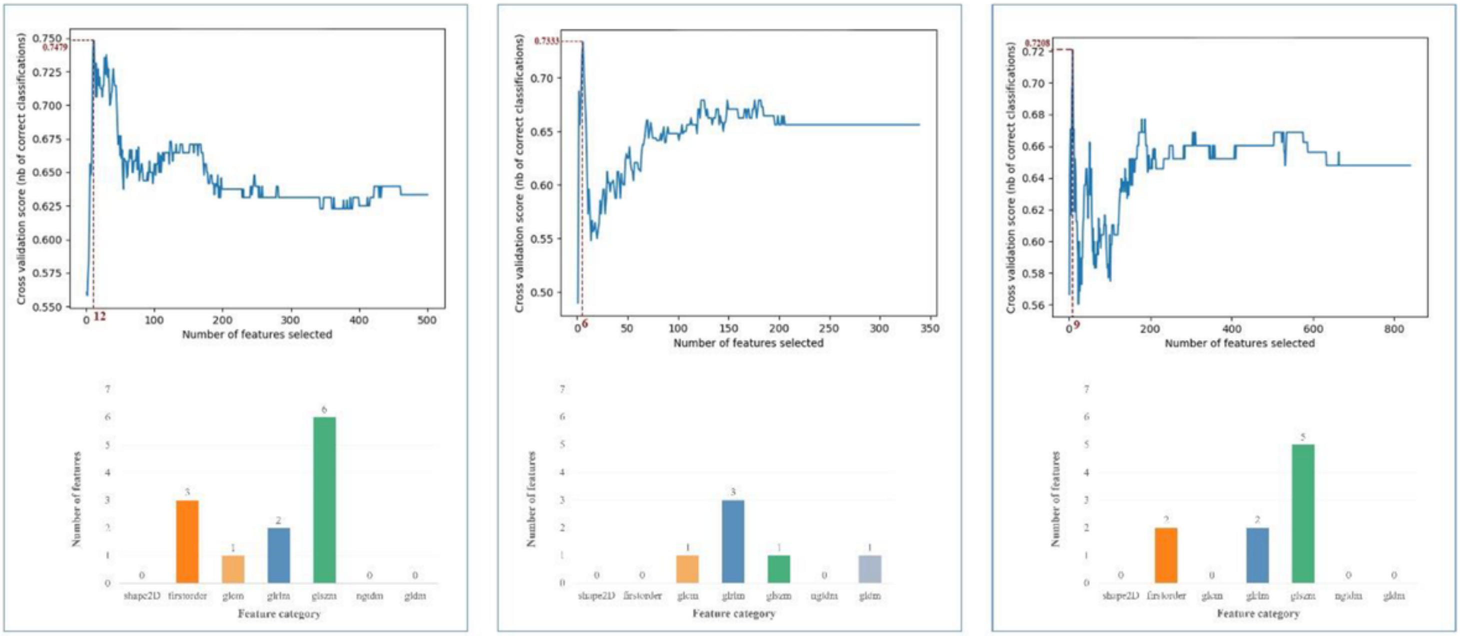


Figure 5

Intra- and peritumoral tissue distribution differences between LUSC and LUAD characterized by the significant radiomics feature Energy on CT images with the unit normalized as “1” on the color bar



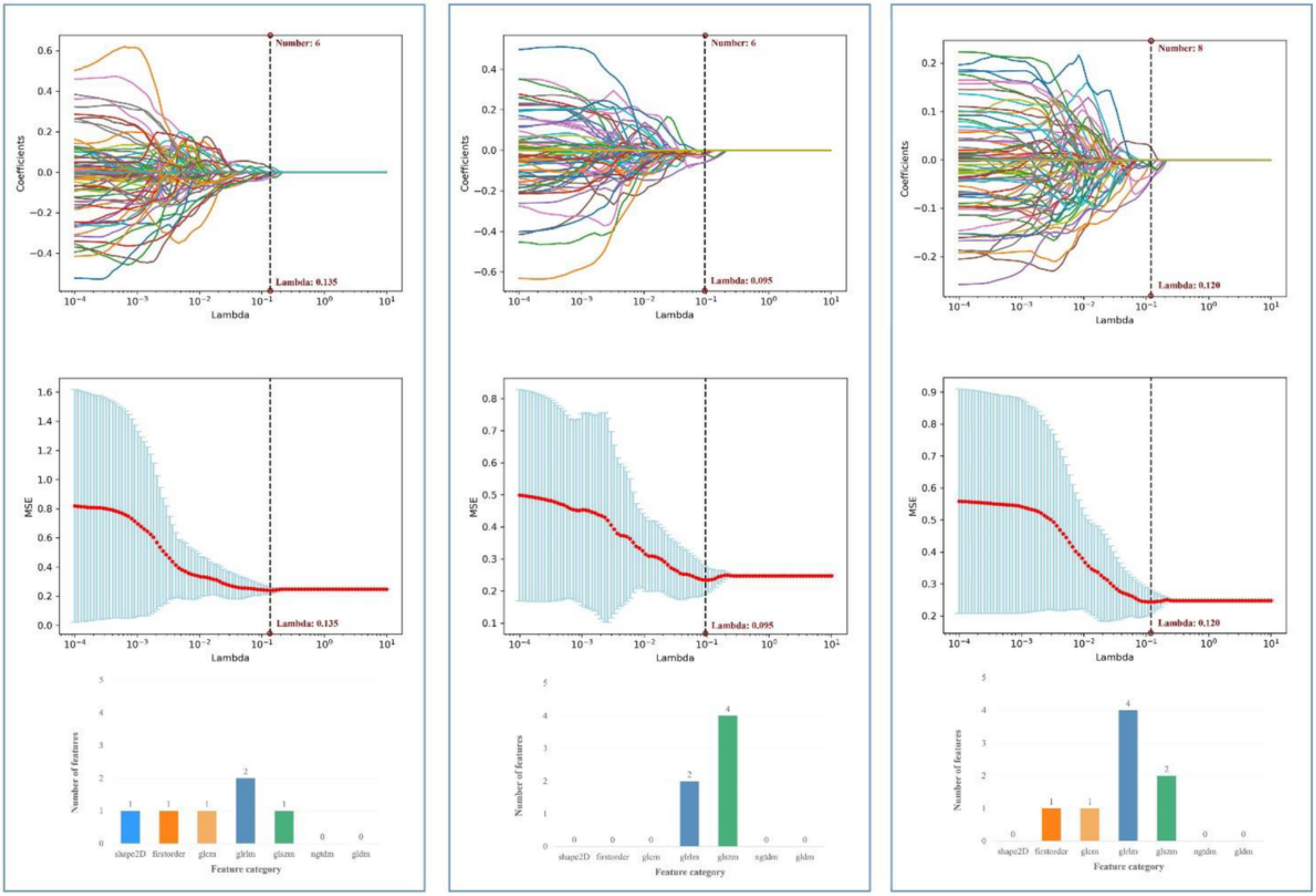
(a) Optimal features selected from the intratumoral region

(b) Optimal features selected from the peritumoral region

(c) Optimal features selected from intra- and peritumoral regions

Figure 6

Optimal features selected using SVM-RFE approach: (a) 12 optimal features selected from the intratumoral region; (b) six optimal features selected from the peritumoral region; and (c) nine optimal features selected from intra- and peritumoral regions.



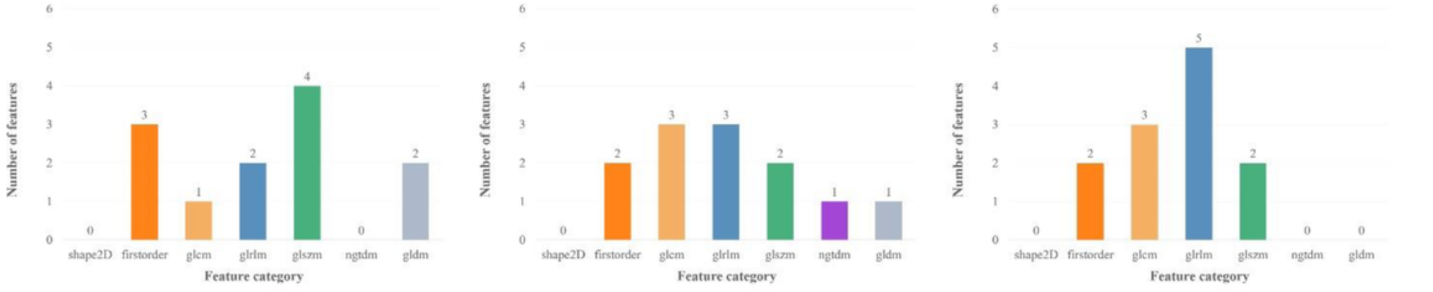
(a) Optimal features selected from the intratumoral region

(b) Optimal features selected from the peritumoral region

(c) Optimal features selected from intra- and peritumoral regions

Figure 7

Optimal features selected using LASSO approach: (a) six optimal features selected from the intratumoral region; (b) six optimal features selected from the peritumoral region; and (c) eight optimal features selected from intra- and peritumoral regions.



(a) Optimal features selected from the intratumoral region

(b) Optimal features selected from the peritumoral region

(c) Optimal features selected from intra- and peritumoral regions

Figure 8

Optimal features selected using mRMR with MID: (a) 12 optimal features selected from the intratumoral region; (b) 12 optimal features selected from the peritumoral region; and (c) 12 optimal features selected from intra- and peritumoral regions.

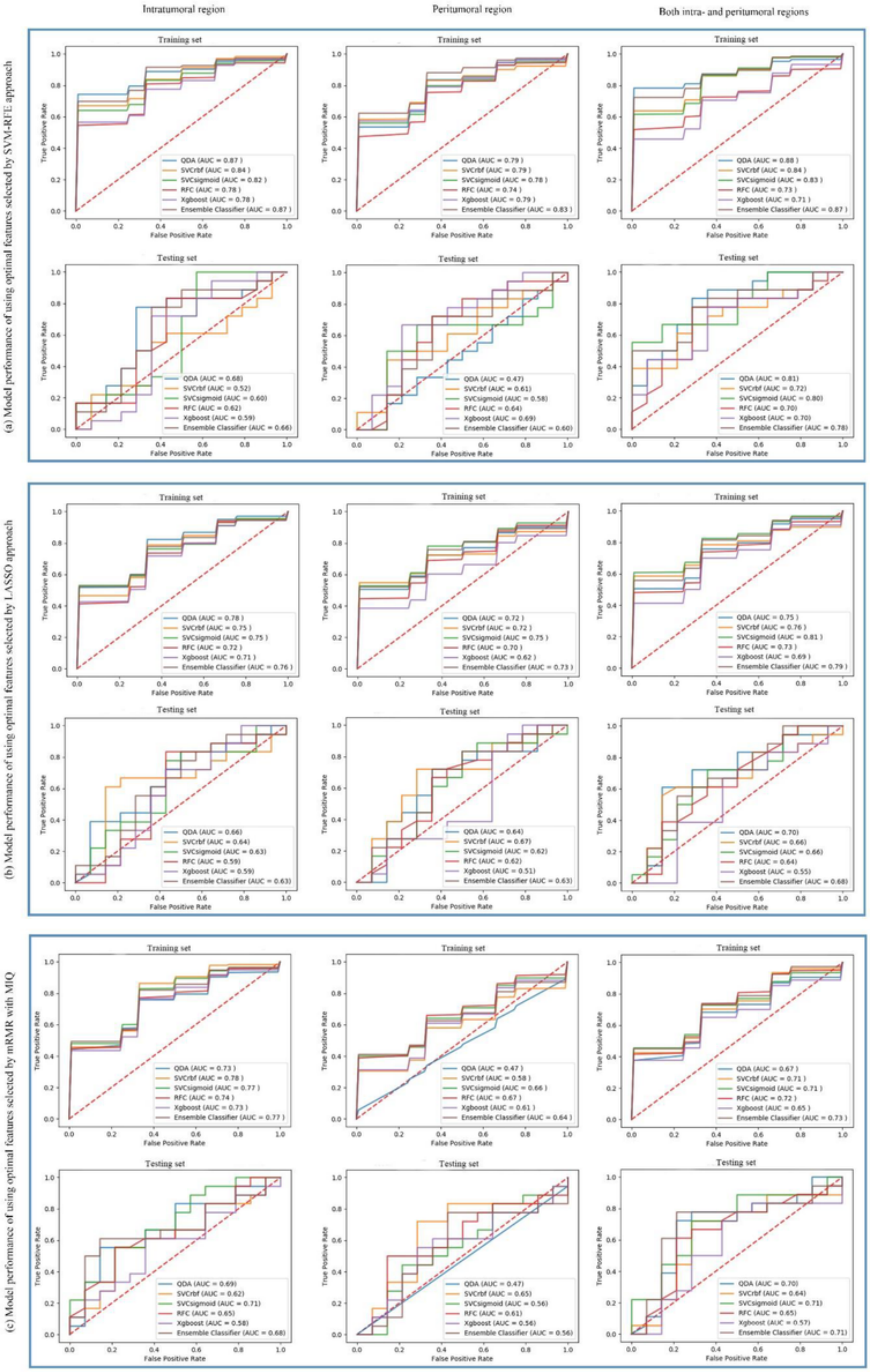


Figure 9

Classification models developed by using five independent classifiers and the ensemble classifier with optimal features determined by three different feature selection methods: (a) performance of classification models developed by using different classifiers and optimal features selected by SVM-RFE approach; (b) performance of classification models developed by using different classifiers and optimal features selected by LASSO approach; (c) performance of classification models developed by using different classifiers and optimal features selected by mRMR with MID

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Appendix20211008.docx