



PAPER

A prognostic analysis method for non-small cell lung cancer based on the computed tomography radiomics

RECEIVED
31 July 2019REVISED
14 December 2019ACCEPTED FOR PUBLICATION
21 January 2020PUBLISHED
12 February 2020Xu Wang¹, Huihong Duan¹, Xiaobing Li¹, Xiaodan Ye^{2,4}, Gang Huang³ and Shengdong Nie^{1,4}¹ School of Medical Instrument and Food Engineering, University of Shanghai for Science and Technology, Shanghai 200093, People's Republic of China² Shanghai Chest Hospital, Shanghai Jiao Tong University, Shanghai 200030, People's Republic of China³ Shanghai University of Medicine and Health Science, Shanghai 201318, People's Republic of China⁴ Author to whom any correspondence should be addressed.E-mail: yuanxyd@163.com (X Ye) and nsd4647@163.com (S Nie)**Keywords:** non-small cell lung cancer, CT radiomics features, prognostic survival prediction model, prognostic factors**Abstract**

In order to assist doctors in arranging the postoperative treatments and re-examinations for non-small cell lung cancer (NSCLC) patients, this study was initiated to explore a prognostic analysis method for NSCLC based on computed tomography (CT) radiomics.

The data of 173 NSCLC patients were collected retrospectively and the clinically meaningful 3-year survival was used as the predictive limit to predict the patient's prognosis survival time range. Firstly, lung tumors were segmented and the radiomics features were extracted. Secondly, the feature weighting algorithm was used to screen and optimize the extracted original feature data. Then, the selected feature data combining with the prognosis survival of patients were used to train machine learning classification models. Finally, a prognostic survival prediction model and radiomics prognostic factors were obtained to predict the prognosis survival time range of NSCLC patients.

The classification accuracy rate under cross-validation was up to 88.7% in the prognosis survival analysis model. When verifying on an independent data set, the model also yielded a high prediction accuracy which is up to 79.6%. Inverse different moment, lobulation sign and angular second moment were NSCLC prognostic factors based on radiomics.

This study proved that CT radiomics features could effectively assist doctors to make more accurate prognosis survival prediction for NSCLC patients, so as to help doctors to optimize treatment and re-examination for NSCLC patients to extend their survival time.

Abbreviations

CT	Computed tomography
PET	Positron emission tomography
NSCLC	Non-small cell lung cancer
SCLC	Small cell lung cancer
TCIA	The Cancer Imaging Archive
CNN	Convolutional neural networks
LSP	Long survival group
SSP	Short survival group
SMOTE	Synthetic minority over-sampling technique
ML	Middle layer
GLCM	Gray level co-occurrence matrix
GLRLM	Gray level run length matrix
NGTDM	Neighborhood gray-tone difference matrix
GT	Gabor transform
SIFT	Scale-invariant feature transform

LBP	Local binary pattern
DT	Decision trees
DAC	Discriminant analysis classifiers
LR	Logistic regression
SVM	Support vector machine
KNN	K-nearest neighbor
EC	Ensemble classifiers
RF	Random forest
TP	True positive
FP	False positive
TN	True negative
FN	False negative
ACC	Accuracy
SE	Sensitivity
SP	Specificity
ROC	Receiver operating characteristic curve
AUC	Area under the curve
MIP	Maximal intensity projection
IDM	Inverse different moment
ASM	Angular second moment

1. Introduction

In 2018, the latest global cancer incidence and mortality data released by American Cancer Society showed that lung cancer has become a malignant tumor with the highest morbidity and mortality in the world now, which accounted for 11.6% of all cancers and 18.4% of total deaths due to cancer (Bray *et al* 2018). According to the classification of histology, lung cancer can be divided into non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC). Compared with SCLC, NSCLC has slower growth and division, late metastasis and less mortality. It accounts for 80%–85% of total number of lung cancer patients. Because the developmental status of NSCLC are usually various for patients as heterogeneity of tumors, the same type of tumors can show different therapeutic effect and prognosis in different individuals (Shepherd *et al* 2005). The epidemiological statistics have shown that a large number of NSCLC patients had not been treated properly due to inaccurate predictions for the disease developments which resulted in a mortality of 75%. Therefore, an effective method for predicting prognostic survival time of NSCLC patients is urgently needed to select the treatment and re-examinations to increase the cure rate and survival rate of NSCLC patients.

Recent researches have found that radiomics is highly related to NSCLC prognosis, it can be considered as a kind of prognostic factor for the prognosis prediction of patients. Wu *et al* retrospectively analyzed fluorine 18 Fluorodeoxyglucose positron emission tomography (PET) images of 101 NSCLC patients, and extracted 70 radiomics features of tumors. Through the analysis of Cox proportional hazard regression model, it was proved that PET radiomics features can be used for tumor distant metastasis prediction. When combined with histologic types, the prognostic power was further improved (Wu *et al* 2016a). Cook *et al* collected PET images of 53 NSCLC patients, and extracted four radiomics texture features of tumor region, namely, cosensence, contrast, busyness and complexity for survival analysis. In multi-factor analysis, cosensence was an independent prognostic factor of NSCLC (Cook *et al* 2013). Vaidya *et al* extracted and screened the pathological features and CT radiomics texture features of 93 NSCLC patients. Through statistical analysis and classification experiments, it is found that the pathological features and CT radiomics texture features can well predict the recurrence of early NSCLC (Vaidya *et al* 2018). Lee *et al* retrospectively analyzed multiple independent NSCLC cohorts and demonstrated that CT-derived pleural contact index and normalized inverse difference may serve as noninvasive prognostic markers in early stage NSCLC. When combined with known prognosticators, the two radiomics features can improve survival prediction (Lee *et al* 2018a, 2018b). Olya *et al* found that convexity and entropy ratio correlated with NSCLC prognosis. It was demonstrated that these two features were the independent prognosis factors of NSCLC through survival analysis (Olya *et al* 2015).

A large number of studies have shown that radiomics has great potential to evaluate cancer treatment at all stages, including auxiliary diagnosis, disease evaluation, efficacy monitoring and predictive analysis. On the basis of previous studies, this paper designed a new prognostic analysis model from the perspective of CT radiomics to predict the prognostic survival time range of NSCLC patients, and screened out the radiomics factors related to the prognostic survival.

2. Material and methods

2.1. Experimental data set

The data used in the experiments were derived from two data sets, which are lung 1 data set (called data set 1 below) of TCIA NSCLC-Radiomics project⁵ and lung 3 data set (called data set 2 below) of TCIA NSCLC Radiogenomics project⁶. The data inclusion-exclusion criteria is as follows: INCLUDE single tumor (In order to control the variables and ensure the influence of a single lesion on the prognosis of patients, we only selected the data of patients with one primary lesion) and EXCLUDE patients who had a cut-off survival time of less than three years and were still alive at that time (Such patients cannot be given a proper label so that they do not meet our experimental requirements).

Based on the above requirements, we screened a total of 173 eligible patient data. Among them, 124 patients from data set 1 are used for model training, and 49 patients from data set 2 are used for model testing. Besides, the database also included many clinical information such as age, gender and TNM stage. The full information of the eligible data is shown in table 1.

2.2. Experimental design

The clinical experiments indicated that the chance of recurrence will be greatly reduced when lung cancer patients have survived for more than 3 years. If the patient's disease-free survival (DFS) was more than 5 years, it was considered to be probably recovered. These time points were important for patient treatment design and review options in clinical practice (Zeng *et al* 2018). Therefore, according to the postoperative survival time of patients in the data set, the survival period of 3 years was selected as grouping threshold clinically. The experimental data from data set 1 was divided into 'long survival group' (LSP) and 'short survival group' (SSP), which correspond to survival time more than 1100 d (from 1141 to 1926 d with the median time of 1479 d) and less than 900 d (from 10 to 896 d with the median time of 418 d) respectively. After grouping, lung tumors were segmented and the features were extracted and optimized at first. Then, various classifiers were applied to classify the feature optimized data, and the prediction model of NSCLC prognosis survival was constructed based on CT radiomics. At the same time, the survival-related radiomics prognostic factors were obtained. Finally, the effectiveness of radiomics prognostic factors was verified and evaluated for performance of the classification model. The process of experiments is shown in figure 1.

2.3. Tumor segmentation

Tumor segmentation is the basis for tumor radiomics feature extraction. In order to extract accurately the radiomics features from tumor regions, the tumor regions should be firstly segmented accurately from CT image sequence to prevent interference of surrounding background (Shakibapour *et al* 2019).

For the requirements of segmentation, a semi-automatic segmentation method was used in this paper. Firstly, tumor area sequence images were selected that contain the entire tumor through the interactive medical imaging software RadiAntViewer. Secondly, a bicubic interpolation algorithm along the axial direction was applied to generate isotropic scans in order to solve the problem of anisotropic 3D spatial resolution CT scanning. Then, the semi-automatic segmentation method was used to segment the tumor, and different segmentation schemes were adopted for tumors with different types: (1) For the solitary tumors, the gray threshold segmentation algorithm and three-dimensional region growing algorithm were used to segment; (2) For the pleural tumors, the edge of lung area was repaired based on concave hulls algorithm (Soltaninejad *et al* 2016) and chain code algorithm (Sun and Wang 2016). On the basis of a complete repair of lung boundary, the same segmentation method as solitary pulmonary tumor segmentation method was used; (3) For the vascular tumors, the above-described solitary pulmonary tumor segmentation method is used for coarse segmentation at first. Moreover, the method of Geodesic distance histogram (Sun *et al* 2014) and fuzzy C-means clustering were adopted to achieve fine segmentation, and the segmentation outcome was the intersection of the above two segments. The final results were selected from the average of several times of segmentations, which are shown in figure 2.

2.4. Feature extraction

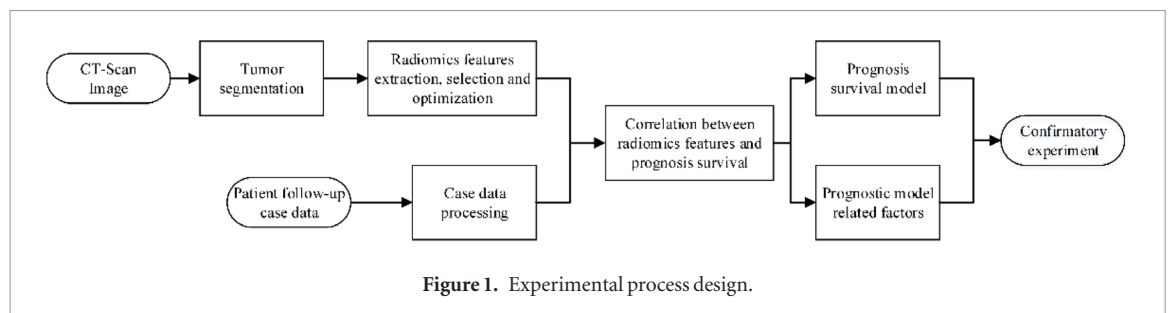
The extraction of radiomics features is an important method for calculating and quantifying tumor image information. In order to reveal the correlation between CT radiomics features and NSCLC prognosis, quantitative indicators of tumors must be extracted, as many as possible, to avoid the loss of information and increase the accuracy of predictions (Way *et al* 2006). In this paper, a total of 258 radiomics features based on gray, shape, texture and other information were extracted (Han *et al* 2015, Dhara *et al* 2016, de Carvalho Filho *et al* 2017, Ferreira *et al* 2017). In addition to traditional gray features, shape features, texture features, we also used mathematical modeling methods to accurately quantify medical signs commonly used in clinical practice. The features extracted by the experiment are shown in table 2.

⁵TCIA NSCLC-Radiomics: <https://wiki.cancerimagingarchive.net/display/Public/NSCLC-Radiomics>

⁶TCIA NSCLC Radiogenomics: <https://wiki.cancerimagingarchive.net/display/Public/NSCLC+Radiogenomics>

Table 1. Data basic information.

Parameter	Result	
	Data set 1	Data set 2
Image size	512 Pixel \times 512 Pixel	512 Pixel \times 512 Pixel
Pixel size		
X direction	0.977 mm	0.604–0.865 mm
Y direction	0.977 mm	0.604–0.865 mm
Slice thickness	3 mm	0.8–3 mm
Gender		
Male	82	41
Female	42	8
Histology		
Adenocarcinoma	16	35
Squamous cell carcinoma	44	12
Large cell carcinoma	41	0
Other	23	2
T stage		
T1	36	18
T2	54	22
T3	9	3
T4	25	6
N stage		
N0	57	29
N1	6	7
N2	40	13
N3	21	0
M stage		
M0	123	43
M1	1	6
Tumor type		
Solitary	8	5
Vascular adhesions	43	19
Pleural adhesions	73	25



A large number of studies have shown that the parameters of imaging equipment affect radiomics feature values. Since the data used in this experiment is from public data sets, we cannot obtain the complete imaging protocol, but we try to consider all factors that can be obtained. In the process of feature extraction, we comprehensively considered these parameters which are easy to affect the robustness and reproducibility of radiomics features, including pixel size, layer thickness, layer spacing. We obtained this information in the Dicom file. In the process of analyzing the Dicom file information of the experimental data, we found that the device information of the CT image sequence of the data set 1 is consistent, and the type of imaging device is SIEMENS CT VA1 DUMMY. So during the training of the model, we can ensure that the CT imaging device parameters have the same impact on the imaging feature extraction for each patient. In the process of model verification using data set 2, we also considered these parameters, and tried to ensure that the radiomics feature extraction of each sample data is performed under the same operation. Finally, the feature data is normalized, and value interval of all features is scaled to the range of [0, 1].

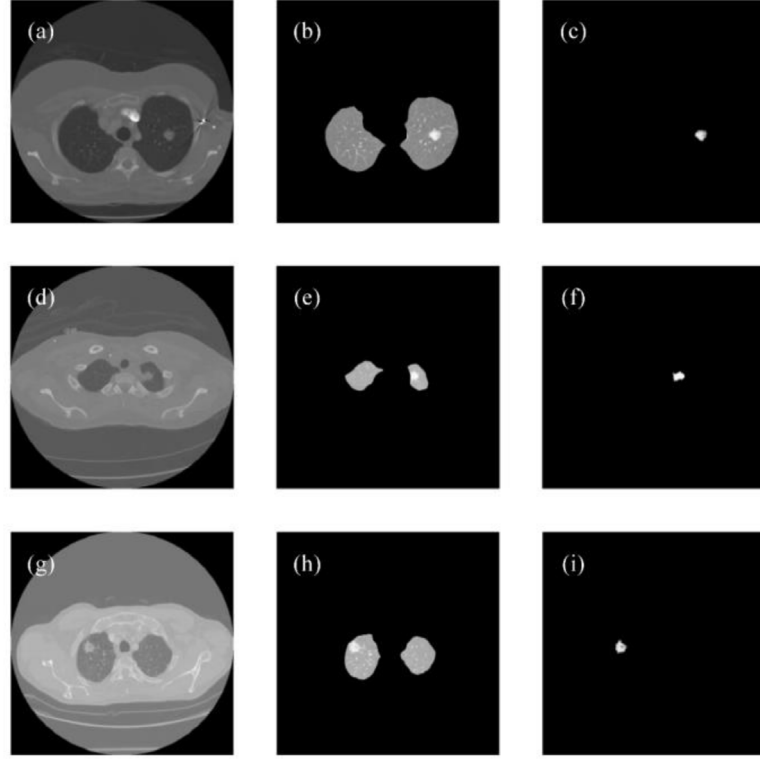


Figure 2. Segmentation results: (a), (d) and (g) are original images. (b), (e) and (h) are the segmentation results of parenchyma. (c), (f) and (i) are segmentation results of solitary, pleural and vascular tumors.

2.5. Feature selection

In order to maximize the collection of information on all aspects of the tumor, a large number of radiomics features describing the characteristics of the tumor need to be extracted (Lohrmann *et al* 2018). However, too high dimension of features often tend to make features matching too complicated and affect operation speed. At the same time, some of the extracted features will bring irrelevant information or error information to the prediction. Therefore, in order to obtain a good performance prediction model, it is necessary to filter redundant features data and retain effective features (Lee *et al* 2008 Jaffar and Al Eisa 2015). So the Relief feature weighting algorithm (Hancer *et al* 2018) was used to retain features which are highly relevant to categories to reduce redundancy features and improve performance of the prediction model. The pseudo-code for the Relief algorithm is shown as follows.

Algorithm: Pseudo-code for the Relief algorithm

INPUT: Sample set S , Sampling number m , Feature weight threshold τ .

OUTPUT: the vector w of estimations of the qualities of attributes and the Selected feature subset.

Begin:

1. set all weights $W(i) = 0$,
2. for $i = 1$ to m
 - Randomly select a sample R from S ,
 - Find the k -nearest neighbors H_j ($j = 1, 2, L, k$) of R from the same sample set of R ,
 - Find k -nearest neighbors from each different class of sample set $M_j(C)$,
3. for $i = 1$ to N

$$W(A) = W(A) - \sum_{j=1}^k \text{diff}(A, R, M_j(C)) / (mk) + \sum_{C \notin \text{class}(R)} \left[\frac{p(C)}{1 - p(\text{class}(R))} \sum_{j=1}^k \text{diff}(A, R, M_j(C)) / (mk) \right] \quad (1)$$

4. Sort W and sort its corresponding features.

End

$M_j(C)$ represents the j th nearest neighbor sample in class C .

$\text{diff}(A, R_1, R_2)$ represents the difference between the sample R_1 and the sample R_2 in the feature A , as shown below:

$$\text{diff}(A, R_1, R_2) = \begin{cases} \frac{|R_1[A] - R_2[A]|}{\max(A) - \min(A)}, & \text{if } A \text{ is continuous} \\ 0, & \text{if } A \text{ is discrete and } R_1[A] = R_2[A] \\ 1, & \text{if } A \text{ is discrete and } R_1[A] \neq R_2[A] \end{cases} \quad (2)$$

Table 2. The features extracted in experiment.

Feature type	Feature subtype	Feature name
Grayscale features	3D	Mean, standard deviation, absolute maximum, absolute minimum, local maximum, local minimum, median, information entropy, kurtosis, slope, contrast, energy, density
Shape features	2D or 3D	Volume(3D), surface area(3D), sphericity(3D), three-dimensional longest diameter(3D), surface area to volume ratio(3D), middle layer area circularity(2D), middle layer area rectangle(2D), middle layer area elongation(2D), middle layer area compactness(2D), middle layer area outline size ratio(2D), middle layer area perimeter(2D), middle layer area concavity(2D), middle layer area 2, 3, 4th order invariant moment(2D), middle layer area boundary irregularity(2D), middle layer area boundary fourier descriptor(2D)
	Gray level co-occurrence matrix (MIP, each feature computed in 4 orientations: (0°,45°,90°,135°))	Sum average, variance, sum variance, differential variance, inverse different moment, contrast, dissimilarity, entropy, sum entropy, difference entropy, correlation, angular second moment, related information metric
	Gray level run length matrix (MIP, each feature computed in 4 orientations: (0°,45°,90°,135°))	Short run emphasis, long run emphasis, gray scale inhomogeneity, run length non-uniformity, run length distribution, run percentages, low gray level run emphasis, high gray level run emphasis, short run low gray level emphasis, short run high gray level emphasis, long run low gray level emphasis, long run high gray level emphasis, mean, mean square error, energy, entropy
Texture features	Neighborhood gray-tone difference matrix (3D)	Coarseness, busyness, complexity, contrast, texture strength
	Gabor transform (3D)	Gabor texture mean, gabor texture variance
	Scale-invariant feature transform (ML, 2D)	72 feature statistics
	Local binary pattern (3D)	Texture mean, texture variance
Medical signs	3D	Lobulation sign, spiculation sign, uniformity of tumor density, vacuole sign, stelliform sign, patchy shadow, alveolar ectasia, notch sign

Abbreviations: MIP, maximum intensity projection; ML, middle layer.

Table 3. Radiomics features of the top 10 weight values.

Class	Feature name	Importance
Texture feature	GLCM-inverse different moment (MIP, orientations = 135°)	0.3137
Medical signs	Lobulation sign	0.2564
Texture feature	GLCM-angular second moment (MIP, orientations = 90°)	0.2189
Texture feature	NGTDM -busyness	0.2020
Shape features	Compactness	0.1852
Texture feature	GLCM-related information metric1 (MIP, orientations = 135°)	0.1791
Texture feature	GLCM-related information metric2 (MIP, orientations = 0°)	0.1586
Medical signs	Spiculation sign	0.1544
Gray feature	Contrast	0.1319
Texture feature	SIFT-#20	0.1203

The feature screening was introduced into every cross-validation process to avoid errors in feature screening. According to the quantitative relationship between samples and features, the top ten features with the highest weighted value were selected for the training and testing by using the Relief feature weighting algorithm. The features sorted by weight value from high to low is shown in table 3.

2.6. Data balance

In 124 case samples, the huge quantity difference between ‘short survival group’ and ‘long survival group’ results in data imbalance (sample ratio: $94/30 = 3.13$). Data imbalance will cause errors in classifier training and affect classifier performance. Therefore, synthetic minority over-sampling technique (SMOTE) resampling method (Last *et al* 2017) was used to resample the feature data of ‘long survival group’ samples in the data set to balance the two types of samples.

SMOTE algorithm is a data synthesis algorithm based on interpolation to synthesize a small number of new samples. The implementation of the algorithm was shown as follows:

- (1) For each sample x in a few classes, distance to all samples in a small sample set were calculated by Euclidean distance and the k -nearest neighbor should be obtained.
- (2) A resampling magnification N was reset according to the sample imbalance ratio. For each minority sample x , several samples from its k nearest neighbors were randomly selected assuming the selected neighbors as x_n .
- (3) For each randomly selected neighbor x_n , a new sample x_{new} was constructed by the primitive sample according to equation (3).

$$x_{new} = x + \text{rand}(0, 1) \times (x - x_n). \quad (3)$$

Where: $\text{rand}(0, 1)$ indicates a random number between 0 and 1.

By using SMOTE data resampling method, the number of 'long survival group' has increased from 30 to 90. Experimental data has been contained 94 samples of 'short survival group' and 90 samples of 'long survival group'. During the training and testing, SMOTE resampling method was also involved into verification process. Since the added 'long survival group' sample from resampling technique is not the real data, the data of this part only were used for classifier training but have not been used to test. In the process of classifier training and testing with all data sample, the classification results of only 124 real samples were obtained in the original data set. This method is widely used in the design of computer aided diagnosis (CAD) of pulmonary nodules (Aghaei *et al* 2016, Yan *et al* 2016).

2.7. Classification

The classification model of lung cancer patients' survival time range was trained by the feature data with strong characterization ability of lung cancer regions obtained from feature selection, and model predictive performance was tested on all 124 cases by cross-validation. In order to construct a performance-optimized classification model as much as possible, seven representative classifiers with strong generalization ability and often used for small sample data set training have been selected (Messay and Rogers 2010, Cascio *et al* 2012, Wang *et al* 2013). They are decision trees (DT), discriminant analysis classifiers (DAC), logistic regression (LR), support vector machine (SVM), k -nearest neighbor (KNN), ensemble classifiers (EC), random forest (RF). In this paper, all the above biomedical data classification methods had been used to select the classification model with best performance for the experimental results, and the optimal performance model was used to predict survival of patients.

2.8. Correlation analysis

In order to explore the degree of correlation between the features obtained by the above feature selection algorithm and prognosis survival, and to select statistically significant correlation features as factors of prognosis survival analysis model, Pearson correlation analysis (Grkovski *et al* 2016, Bailey *et al* 2018) was applied to calculate the correlation between the top ten features obtained by the Relief algorithm and the prognostic survival labels. Correlation coefficient $|r|$ and P-value were used to evaluate the 'significant degree' of correlation degree.

2.9. Survival analysis

According to the prediction results obtained by the above performance optimal prediction model, the Kaplan-Meier survival analysis method was used to perform survival analysis on the predicted two groups of cases to verify the validity of the experiment (Wu *et al* 2016b).

The Kaplan-Meier survival analysis method used the principle of conditional probability and probability multiplication to calculate the survival rate by combining the cut-off survival time and survival status in the case information. In SPSS software, we inputted patients' cut-off survival time, patients' cut-off survival status, original grouping label and prognosis prediction results of patients into Kaplan Meier survival analysis model to draw the survival curve of NSCLC patients.

3. Results

The results of the above experiments are shown in this section. Table 4 shows the classification results and evaluation indicators for each classifier subtype in the experiment. The ROC curves of the best classification subtype under each classifier model are shown in figure 3. Experimental results showed that the highest cross-

Table 4. Classification experiment results.

Classifier	Subtype	ACC(%)	AUC	SE	SP
DT	Deep tree	72.6	0.75	0.667	0.745
	medium tree	72.6	0.75	0.667	0.745
	shallow tree	63.7	0.71	0.700	0.617
DAC	Linear discriminant	71.0	0.79	0.800	0.681
	Quadratic discriminant	83.1	0.90	0.867	0.819
LR	—	69.4	0.77	0.733	0.681
SVM	Linear SVM	69.4	0.79	0.800	0.660
	Quadratic SVM	78.2	0.89	0.933	0.734
	Cubic SVM	73.4	0.85	0.900	0.681
	Fine Gaussian SVM	88.7	0.92	0.833	0.904
	Medium Gaussian SVM	68.5	0.82	0.867	0.628
KNN	Coarse Gaussian SVM	75.8	0.80	0	1
	Fine KNN	77.4	0.82	0.933	0.723
	Medium KNN	53.2	0.74	0.933	0.404
	Coarse KNN	24.2	0.72	1	0
	Cosine KNN	76.6	0.84	0.800	0.755
EC	Cubic KNN	55.6	0.77	0.933	0.436
	Weighted KNN	58.9	0.87	0.967	0.468
	AdaBoosted trees	79.0	0.84	0.633	0.840
	RusBoosted trees	74.2	0.78	0.667	0.766
	Bagged trees	82.3	0.88	0.867	0.809
RF	Subspace KNN	75.8	0.90	0.967	0.702
	Subspace discriminant	70.1	0.79	0.767	0.681
RF	—	80.6	0.86	0.800	0.613

Abbreviations: DT, decision trees; DAC, discriminant analysis classifiers; LR, logistic regression; SVM, support vector machine; KNN, K-nearest neighbor; EC, ensemble classifiers; RF, random forest; ACC, accuracy; SE, sensitivity; SP, specificity; AUC, area under the curves.

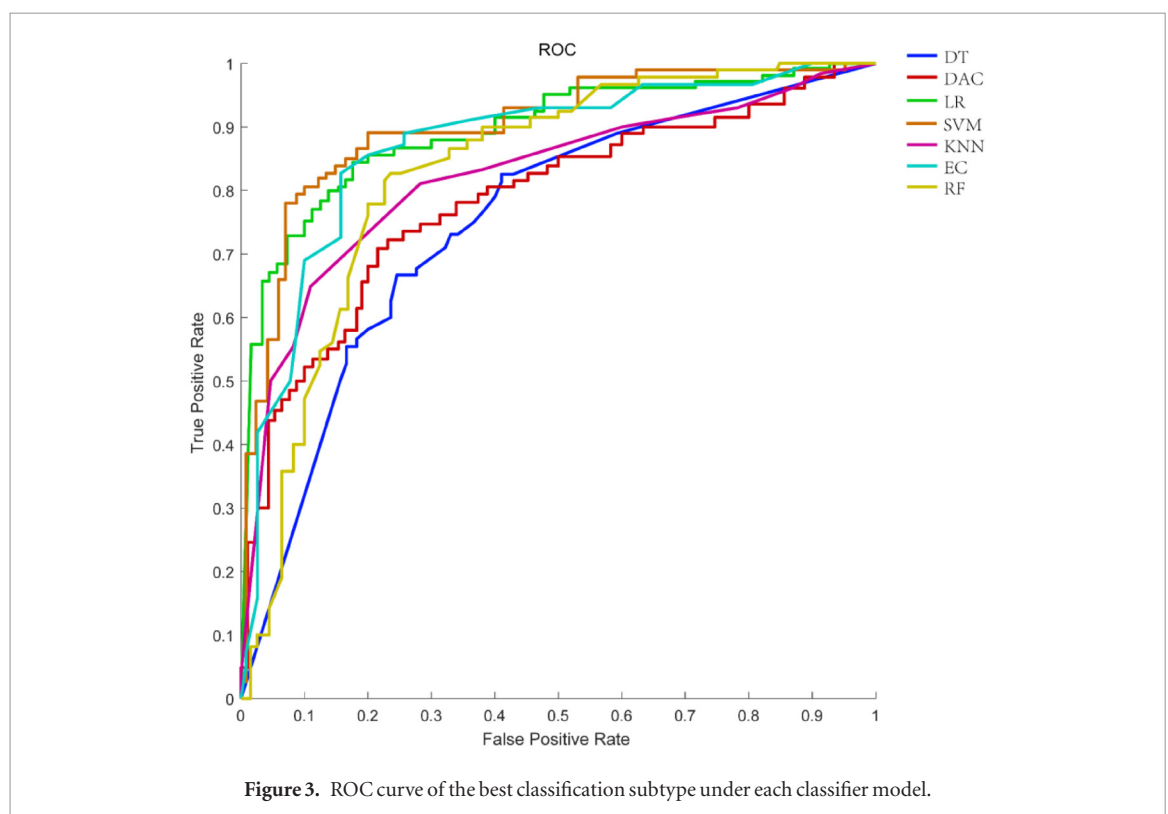
**Figure 3.** ROC curve of the best classification subtype under each classifier model.

Table 5. Independent dataset verification results.

Confusion matrix		Predictive value		ACC(%)	AUC	SE	SP
		SSP	LSP				
Actual value	SSP	29	6	79.6	0.84	0.714	0.829
	LSP	4	10				

Abbreviations: LSP, long survival group; SSP, short survival group; ACC, accuracy; SE, sensitivity; SP, specificity; AUC, area under the curves.

Table 6. Prognostic prediction results of traditional prognostic markers.

Confusion matrix		Predictive value		ACC(%)	AUC	SE	SP
		SSP	LSP				
Actual value	SSP	61	33	63.7	0.69	0.600	0.649
	LSP	12	18				

Abbreviations: LSP, long survival group; SSP, short survival group; ACC, accuracy; SE, sensitivity; SP, specificity; AUC, area under the curves.

Table 7. Pearson correlation analysis.

Feature weight ranking	Feature name	r	P-value
#1	GLCM-inverse different moment (MIP, orientations = 135°)	0.637	0.000 013(≤0.05)
#2	Lobulation sign	0.487	0.005 721(≤0.05)
#3	GLCM-angular second moment (MIP, orientations = 90°)	0.256	0.034 519(<0.05)
#4	NGTDM-busyness	0.237	0.271 931
#5	Shape-compactness	0.013	0.858 854
#6	GLCM-related information metric1 (MIP, orientations = 135°)	0.033	0.652 019
#7	GLCM-related information metric2 (MIP, orientations = 0°)	0.093	0.202 318
#8	Spiculation sign	0.281	0.183 632
#9	Gray-contrast	0.146	0.157 295
#10	SIFT-#20	0.261	0.478 626

Abbreviations: GLCM, gray level co-occurrence matrix; NGTDM, neighborhood gray-tone difference matrix; SIFT, scale-invariant feature transform; MIP, maximal intensity projection.

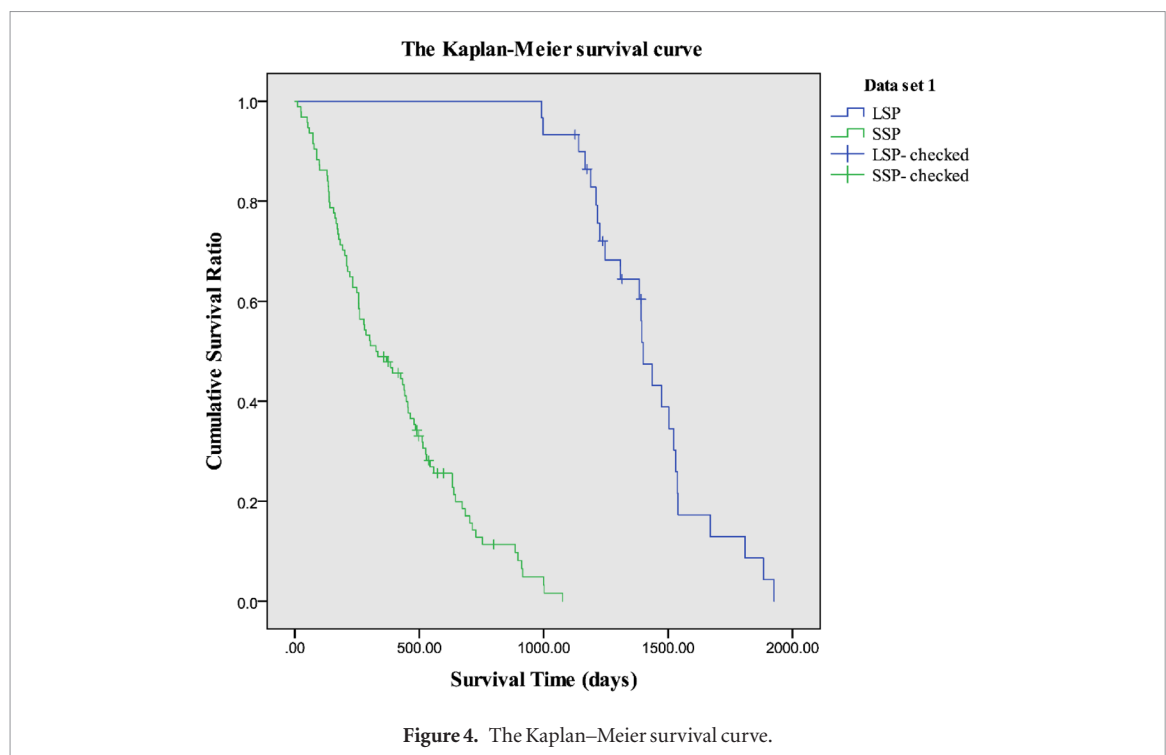


Table 8. Prognostic analysis model performance comparison results.

Algorithm	S.H. Hawkins's	R. Paul's	Ours
Sample size	81	81	124
Feature Type	Traditional imaging features	Deep learning features	Mixed (more traditional imaging features + medical signs) features
Classifier Used	Decision trees	K-nearest neighbor	Support vector machine
Total number of features	219	4096	258
Feature selector used	Relief-f	Symmetric uncertainty	Relief
Number of features selected	5	5	10
Verification of method	Leave-one-out cross validation	Leave-one-out cross validation	<i>k</i> -fold cross validation
ACC(%)	77.5	82.5	88.7
AUC	0.712	—	0.92

Abbreviations: ACC, accuracy; AUC, area under the curves.

validation accuracy of classification was 88.7% and was obtained with the support vector machine classifier. After that, we used all 124 patient data from data set 1 to train the model, and adopted the parameters, average error rate and classifier subtype of the above optimal model to train the data to get the new classification model. Then, we tested all 49 patient data sets of data from data set 2 to verify the accuracy and robustness of the model. The results of the test set are shown in table 5. The 79.6% prediction accuracy of other independent cohort also proved that experimental method and feature extraction for predicting the prognostic survival of patients are effective and robust.

In order to verify that the radiomics features can bring better prediction capabilities to traditional markers, we also took the currently recognized clinical prognostic factors of NSCLC (volume, TNM staging, histological classification) as training input, and then used them to train the patient prognosis survival model to predict the prognosis survival time range of 124 NSCLC patients. The best experimental result is shown in the table 6. The experimental results showed that the radiomics features can significantly improve the prognostic ability, which once again proved that CT radiomics has great potential in the prognosis prediction of NSCLC.

The result of the Pearson correlation analysis is shown in table 7. Experimental results showed that the *p*-value of the Pearson correlation analysis of GLCM-inverse different moment (IDM), Lobulation sign and GLCM-angular second moment (ASN) was less than 0.05 which was statistically significant. Only the first three of the ten features with the highest weighted value were significantly correlated with the survival of the model.

The survival curve is shown in figure 4. It can be seen from the figure that there is a large difference between the predicted survival rates of the two groups. Chi-square test was further carried out on the prediction results. With a *p*-value of 0.0173, we rejected the null hypothesis that the groups are the same. Thus, the predicted classes were distinct from one another when predicting survival groups, which also verified the feasibility and effectiveness of this experimental method.

4. Discussion

This study was based on CT radiomics features to construct a prognostic survival prediction model for NSCLC for predicting the survival of 173 patients with NSCLC, and screening for radiomics factors associated with prognostic survival prediction. We conducted accurate tumor segmentation and multi-scale feature extraction to carry out in-depth exploration of tumor radiomics information and used feature optimization method and various types of machine learning methods to constructed a comprehensive and effective prognostic survival prediction model to predict the prognosis survival of patients with NSCLC.

The accuracy of 88.7% is a high accuracy for this problem, which again demonstrated the feasibility and effectiveness of radiomics features in the prognosis of NSCLC. The algorithm of this study is compared with existing method (Hawkins *et al* 2014, Paul *et al* 2017), and the prediction accuracy of the model (ACC) and the area under the ROC curve (AUC) were used as the evaluation indexes of the algorithm. The comparison results are shown in table 8, which indicated that the accuracy and AUC of our method were significantly improved in the case of the same cross-validation. We not only used the traditional cross-validation method, but also used other independent dataset verification methods not used in the literature, which showed that our method is more universal and the prediction effect is more accurate and stable.

After proving the effectiveness of radiomics features in predicting the prognosis of NSCLC, we also explored the specific relationship between the imaging information of these features and the clinical background. Three conclusions can be drawn as below:

- (1) GLCM-inverse different moment reflects the homogeneity of image texture and measures the local variation of image texture. The small value indicates that the different regions of the image texture change violently and the local area is very inhomogeneous.
- (2) GLCM-angular second moment reflects the uniformity of gray distribution of images. The larger ASN value, the more inhomogeneous the gray distribution of the image.
The roughness of tumor surface can be revealed by the above two radiomics shape texture features, which is an important reference index for the discrimination of tumor benign and malignant, even the determination of tumor malignancy. The lower the IDM and ASM values of tumor indicate the rough of the tumor surface and the higher the degree of malignancy of the tumor clinically.
- (3) Lobulation sign is considered to be an important indicator for diagnosing benign and malignant tumors clinically. In general, lobulation sign reflects the infiltration of the tumor. The depth the lobulation makes the invasiveness intensive, the low the degree of differentiation and the worse the prognosis. Besides, lobulation sign also reflects the tumor heterogeneity. Deep lobulation reveals a high level of heterogeneity to tumors which will lead to the bad prognosis.

All of the above three features can reflect the degree of malignancy of the tumor at the level of radiomics. The degree of malignancy of the tumor usually determines the prognosis of the patient.

Although experimental results verified the effectiveness of the proposed method, there were still some shortcomings in this study. In spite of the experimental data came from two data sets, there were a large number of data in the data set that cannot meet experimental requirements. Through detailed screening, CT data of only 173 NSCLC patients were enrolled. The experimental data might not completely capture the variability of NSCLC at imaging due to the limitation of small sample size. In the future research, it is necessary to use a large size sample data to further verify and test the effectiveness of the algorithm. At the same time, the future work will be focused on the combination of radiomics and other omics information such as genetic test results, tumor marker, histopathological grade and other important prognosis related information in order to improve the accuracy and robustness of prognostic survival prediction model.

5. Conclusion

In conclusion, CT radiomics features of tumors can be used to quantitatively evaluate the subtle differences of tumors of the same type, which can solve the problem that tumor heterogeneity is difficult to quantitatively estimate, and can assist doctors in diagnosis, analysis and prediction of prognosis survival of patients. The method is an important application of radiomics in the diagnosis, analysis and prediction of NSCLC, and has important clinical value and significance.

Acknowledgments

The study was supported by research grants from The National Natural Science Foundation of China (Grant No.81830052,81530053,81571629), Shanghai Key Laboratory of Molecular Imaging (Grant No.18DZ2260400), and the Project of Science and Technology Commission of Shanghai Municipality (Grant No.19411965200).

References

- Aghaei F, Tan M, Hollingsworth A B and Zheng B 2016 Applying a new quantitative global breast MRI feature analysis scheme to assess tumor response to chemotherapy *J. Magn. Reson. Imaging* **44** 1099–106
- Bailey M H, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, Colaprico A, Wendl M C, Kim J and Reardon B 2018 Comprehensive characterization of cancer driver genes and mutations *Cell* **173** 371–85.e18
- Bray F, Ferlay J, Soerjomataram I, Siegel R L, Torre L A and Jemal A 2018 Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries *CA: Cancer J. Clin.* **68** 394–424
- Cascio D, Magro R, Fauci F, Iacomini M and Raso G 2012 Automatic detection of lung nodules in CT datasets based on stable 3D mass-spring models *Comput. Biol. Med.* **42** 1098–109
- Cook G J R, Connie Y, Muhammad S, Vicky G, Sugama C, Arunabha R, Paul M, Shahreen A and David L 2013 Are pretreatment ¹⁸F-FDG PET tumor textural features in non-small cell lung cancer associated with response and survival after chemoradiotherapy? *J. Nucl. Med.* **54** 19–26
- de Carvalho Filho A O, Silva A C, de Paiva A C, Nunes R A and Gattass M 2017 Computer-aided diagnosis system for lung nodules based on computed tomography using shape analysis, a genetic algorithm, and SVM *Med. Biol. Eng. Comput.* **55** 1129–46

- Dhara A K, Mukhopadhyay S, Dutta A, Garg M and Khandelwal N 2016 A combination of shape and texture features for classification of pulmonary nodules in lung CT images *J. Digit. Imaging* **29** 466–75
- Ferreira J R, de Azevedo-Marques P M and Oliveira M C 2017 Selecting relevant 3D image features of margin sharpness and texture for lung nodule retrieval *Int. J. Comput. Assist. Radiol. Surg.* **12** 509–17
- Grkovski M, Schwartz J, Rimner A, Schöder H, Carlin S D, Zanzonico P B, Humm J L and Nehmeh S A 2016 Reproducibility of ^{18}F -fluoromisonidazole intratumour distribution in non-small cell lung cancer *EJNMMI Res.* **6** 79
- Grove O, Berglund A E, Schabath M B, Aerts H J W L, Andre D, Hua W, Emmanuel Rios V, Philippe L, Yuhua G and Yoganand B 2015 Quantitative computed tomographic descriptors associate tumor shape complexity and intratumor heterogeneity with prognosis in lung adenocarcinoma *PLoS One* **10** e0118261
- Han F, Wang H, Zhang G, Han H, Song B, Li L, Moore W, Lu H, Zhao H and Liang Z 2015 Texture feature analysis for computer-aided diagnosis on pulmonary nodules *J. Digit. Imaging* **28** 99–115
- Hancer E, Xue B and Zhang M 2018 Differential evolution for filter feature selection based on information theory and feature ranking *Knowl.-Based Syst.* **140** 103–19
- Hawkins S H, Korecki J N, Balagurunathan Y, Gu Y, Kumar V, Basu S, Hall L O, Goldgof D B, Gatenby R A and Gillies R J 2014 Predicting outcomes of nonsmall cell lung cancer using CT image features *IEEE Access* **2** 1418–26
- Jaffar M A and Al Eisa E 2015 *Progress in Systems Engineering* (Berlin: Springer) pp 645–51
- Last F, Douzas G and Bacao F 2017 Oversampling for imbalanced learning based on k-means and smote (arXiv:1711.00837)
- Lee J, Cui Y, Sun X, Li B, Wu J, Li D, Gensheimer M F, Loo B W, Diehn M and Li R 2018a Prognostic value and molecular correlates of a CT image-based quantitative pleural contact index in early stage NSCLC *Eur. Radiol.* **28** 736–46
- Lee J, Li B, Cui Y, Sun X, Wu J, Zhu H, Yu J, Gensheimer M F, Loo B W Jr and Diehn M 2018b A quantitative CT imaging signature predicts survival and complements established prognosticators in stage I non-small cell lung cancer *Int. J. Radiat. Oncol. Biol. Phys.* **102** 1098–106
- Lee M C, Boroczky L, Sungur-Stasik K, Cann A D, Borczuk A C, Kawut S M and Powell C A 2008 *21st IEEE Int. Symp. on Computer-Based Medical Systems* (IEEE) pp 548–53
- Lohrmann C, Luukka P, Jablonska-Sabuka M and Kauranne T 2018 A combination of fuzzy similarity measures and fuzzy entropy measures for supervised feature selection *Expert Syst. Appl.* **110** 216–36
- Messay T and Rogers H S K 2010 A new computationally efficient CAD system for pulmonary nodule detection in CT imagery *Med. Image Anal.* **14** 390–406
- Paul R, Hawkins S H, Hall L O, Goldgof D B and Gillies R J 2017 *IEEE Int. Conf. on Systems*
- Shakibapour E, Cunha A, Aresta G, Mendonça A M and Campilho A 2019 An unsupervised metaheuristic search approach for segmentation and volume measurement of pulmonary nodules in lung CT scans *Expert Syst. Appl.* **119** 415–28
- Shepherd F A, Rodrigues Pereira J, Ciuleanu T, Tan E H, Hirsh V, Thongprasert S, Campos D, Maoleekoonpiroj S, Smylie M and Martins R 2005 Erlotinib in previously treated non-small-cell lung cancer *New Engl. J. Med.* **353** 123–32
- Soltaninejad S, Cheng I and Basu A 2016 *IEEE Int. Conf. on Systems, Man, and Cybernetics* (IEEE) pp 004775–80
- Sun S, Guo Y, Guan Y, Ren H, Fan L and Kang Y 2014 Juxta-vascular nodule segmentation based on flow entropy and geodesic distance *IEEE J. Biomed. Health Inf.* **18** 1355–62
- Sun Y and Wang J 2016 *9th Int. Congress on Image and Signal Processing, BioMedical Engineering and Informatics* (IEEE) pp 330–5
- Vaidya P, Wang X, Bera K, Khunger A, Choi H, Patil P, Velcheti V and Madabhushi A 2018 RaPtomics: integrating radiomic and pathomic features for predicting recurrence in early stage lung cancer *Proc. SPIE* **10581** 105810M
- Wang Q, Kang W, Wu C and Wang B 2013 Computer-aided detection of lung nodules by SVM based on 3D matrix patterns *Clin. Imaging* **37** 62–9
- Way T W, Hadjiiski L M, Sahiner B, Chan H P, Cascade P N, Kazerooni E A, Bogot N and Zhou C 2006 Computer-aided diagnosis of pulmonary nodules on CT scans: Segmentation and classification using 3D active contours *Med. Phys.* **33** 2323–37
- Wu J, Aguilera T, Shultz D, Gudur M, Rubin D L, Loo B W Jr, Diehn M and Li R 2016a Early-stage non-small cell lung cancer: quantitative imaging characteristics of ^{18}F fluorodeoxyglucose PET/CT allow prediction of distant metastasis *Radiology* **281** 270–8
- Wu J, Gensheimer M F, Dong X, Rubin D L, Napel S, Diehn M, Loo B W Jr and Li R 2016b Robust intratumor partitioning to identify high-risk subregions in lung cancer: a pilot study *Int. J. Radiat. Oncol. Biol. Phys.* **95** 1504–12
- Yan S, Qian W, Guan Y and Zheng B 2016 Improving lung cancer prognosis assessment by incorporating synthetic minority oversampling technique and score fusion method *Med. Phys.* **43** 2694–703
- Zeng H, Chen W, Zheng R, Zhang S, Ji J S, Zou X, Xia C, Sun K, Yang Z and Li H 2018 Changing cancer survival in China during 2003–15: a pooled analysis of 17 population-based cancer registries *Lancet Global Health* **6** e555–67