Reproducibility and Non-Redundancy Feature Selection Methods in Radiomics

Bingyan Wei [1, a], Jianlin Song [2, b] and Jinli Sun [1, c]

[1]Hebei University, Baoding 071000, China

[2]Hebei eye hospital, Xingtai 054000, China

[a]wei_bingyan@163.com, [b]289982964@qq.com, [c]1259915579@qq.com

Abstract

"Radiomics" is a process of extracting a great quantity of descriptive features from biomedical images that can be used for prognosis .One of the major challenges of radiomics is to acquire the features that reproducibility and non-redundancy. The reproducibility of features depends on the segmentation algorithm, which should provide accurate and reproducible results. In this study, a semi-automated segmentation method based on CV model with shape constraint ,which has an good effect for gray scale inhomogeneous of tumors was used to get tumor region for computed tomographic(CT) images of 35 non–small cell lung cancer (NSCLC) patients. A set of features (125 3D and 92 2D) was computed for each tumor region in the test/retest data set. In terms of comparing the intra-class correlation coefficient (ICC) with manual segmentation method in feature extracting, it is indicated that the features obtained by the method of on CV model with shape constraint has a better representatives. By utilizing concordance correlation coefficient (CCC) and dynamic range (DR), series quantitative features of better reproducibility could be obtained. However, these features were redundant. A feature selection method based on sparse representation coefficient (SRC) was used to filter these redundant features. It is indicated that the features obtained by SRC have better non-redundancy through comparison the selection methods based on Pearson correlation coefficient (PCC) and symmetrical uncertainty (SU).Thus quantitative image features that reproducibility and non-redundancy provide informative and prognostic biomarkers for NSCLC.

Keywords

Radiomics, CV Model, Feature Selection, Sparse Representation Coefficient.

1. Introduction

Lung cancer is the leading cause of cancer death and the second most diagnosed cancer in the United States [1]. In 2016, the disease is expected to cause approximately 158,000 deaths in the United States .And non–small cell lung cancer (NSCLC) accounts for approximately 85% of all lung cancers [2].The need to diagnose NSCLC at an early and potentially curable stage is thus obvious. And the database established by radiomics has great significance for the prediction and diagnosis of cancer especially for NSCLC. Early detection of the disease of lung cancer through adopting lung computed tomography (CT) as the standard modality [3].

In the last decade, extraordinary progressing biomedical science and technology have contributed to the understanding of the lung cancer, especially for NSCLC [4]. In spite of the survival rate of patients with lung cancer has not changed significantly, CT imaging has greatly enhanced the accuracy of imaging,

decreased rotation, and better reconstruction methods. All these improvements have result in better capturing of the anatomic structure for the regions of interest (ROIs).

In recent years, the concept of "radiomics" has been proposed. "Radiomics" is a process of extracting a great quantity of descriptive features from biomedical images that can be used for prognosis [5].In the previous work, it has been proposed that radiomics has an ability of capturing tumor heterogeneity and connecting with genomics .The radiomics enterprise is divided into four processes: (i) Image acquisition and preprocessing; (ii) Image segmentation and acquisition of region of interest(ROIs); (iii)Image feature extraction and quantization; and (iv) databases and data sharing. (Fig.1 shows the research processes of radiomics.) Each of these steps poses discrete challenges that have to be met. In addition to the heterogeneity of tumor, the prediction of tumor can be revealed, which can also be related to the type of gene expression. Therefore, how to convert these mineable data with represented and non-redundancy is increasingly important.

In prior work, it has been found that the reproducibility of features depend on the robust of segmentation algorithms, which should provide an accurate and reproducible results [6]. Furthermore, semi-automated segmentation have a better similarity index (SI)than manual segmentation(the SI of machine-segmented lesions was >0.93, whereas the SI of manual segmentation was 0.73) [7].In this article, a semi-automated segmentation method based on CV model with shape constraint was used twice (obtained 2 data set, test and retest) to get tumor region for CT images of 35 NSCLC patients(15 adenocarcinoma and 20 epidermoid carcinoma). According to this segmentation method , the problem for gray scale inhomogeneous of tumors is solved by adding the shape constraint to the CV model [8].In terms of comparing the intra-class correlation coefficient (ICC) [9] for test/retest ($ICC=0.86\pm0.15$) ,it is obvious that a series features of better reproducibility could be extracted by this reliable robust means.

In this study, a set of quantitative features which contain several of information of tumor heterogeneity was computed, including size, shape, gray gradient histogram, gray level co-occurrence matrix, laws, and Gabor wavelet features [10]. A number of features with high reproducibility were extracted by filtering with concordance correlation coefficient (CCC) and dynamic range (DR) in the repeat scans (test and retest) [11,12].Whereas these features are redundant. In the past few years, a variety of measures have been developed for feature correlation in feature selection. For example, Pearson Correlation Coefficient (PCC) [13] and Symmetrical Uncertainty (SU),and they all can gets used for the evaluation of the correlation between features effectively without taking interaction between the correlation of categories into account, which leads to feature interaction is neglected during the procedural of feature selection [14]. Therefore, a feature section method based on sparse representation coefficient (SRC) was proposed for removing the redundant features effectively, which has a significant difference in reflecting the relevance between some feature and category variable in the dataset. Then support vector machine (SVM) classifier is employed to classify for epactal 43 NSCLC patients (18 adenocarcinoma and 25 epidermoid carcinoma) that can validate the effectiveness of feature selection method, and compare with the method based on SU and PCC individually. These Reproducibility and non-redundancy features played an important role in diagnosis and treatment for lung cancer patients. Thus quantitative image features that reproducibility and non-redundancy provided informative and prognostic biomarkers were obtained for NSCLC. (Fig.2)

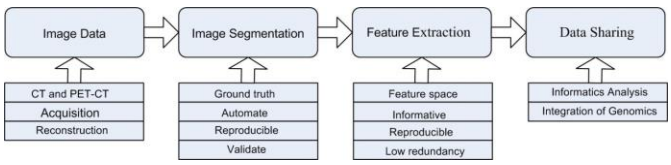


Fig. 1 The research processes of radiomics

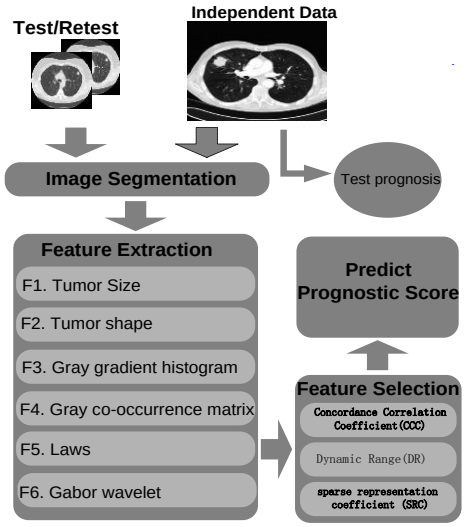


Fig.2 The workflow to obtain representative features.

2. Methods

2.1 Tumor segmentation

One of the main challenges of Radiomics is tumor segmentation, which should provide accurate and reproducible results. In previous study, it has been shown that semi-automated segmentation approaches are fast and reduce inter-observer variability. Where manual delineation is time consuming and has risk to inter-observer variability [16].In this paper, we use semi-automated segmentation algorithm based on the segmentation results with CV model for ROIs.

Although the algorithm of CV model is a partial solution to the problem that the boundary of the object is unclear, there is still a difficult for gray scale inhomogeneous of tumors of in CT image. In order to enhance the accuracy of segmentation and achieve the tumor region repetitively, it is essential to enlarge the image constraint force for the object edge. The CV energy model segments an input image $m:\Omega \to R$ by minimizing the functional which makes the active contour C approaching the boundary of the object.

$$\begin{aligned} E(\phi,c_1,c_2) &= E_{cv}(\phi,c_1,c_2)+\alpha E_{shape}(\phi) \\ &= \mu\int_{\Omega}\delta(\phi)|\nabla\phi|dxdy+\nu\int_{\Omega}H(\phi)dxdy \\ &+\lambda_1\int_{\Omega}|m(x,y)-c_1|^2\,H(\phi)dxdy \\ &+\lambda_2\int_{\Omega}|m(x,y)-c_2|(1-H(\phi)dxdy \\ &+\alpha\int_{\Omega}(\phi-\phi_0)^2dxdy \end{aligned} \tag{1}$$

Where, c_1 and c_2 are the average intensities inside and outside the contour C respectively, and λ_1, λ_2, ν, μ are some positive parameters, α is a positive parameter which affects curve evolution with the shape-driven energy. The gradient descent equation for ϕ is given by:

$$\frac{\partial\varphi}{\partial t}=\delta(\varphi)[\mu div(\frac{\nabla\varphi}{|\nabla\varphi|})-\nu-\lambda_1(m-c_1)^2+(m-c_2)^2]-2\alpha(\varphi-\varphi_0) \tag{2}$$

The segmentation results are then obtained by minimizing the proposed energy functional. Fig.3 shows the compare with CV for tumor regions.

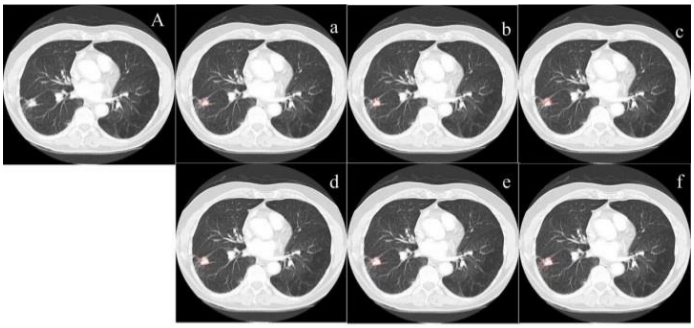


Fig.3 The segmentation results with CV model for ROIs. A: The original images of CT. a, b, and c is the results of manual segmentation. d, e, and f is the results of The segmentation results with CV model for ROIs. The first row is original images.

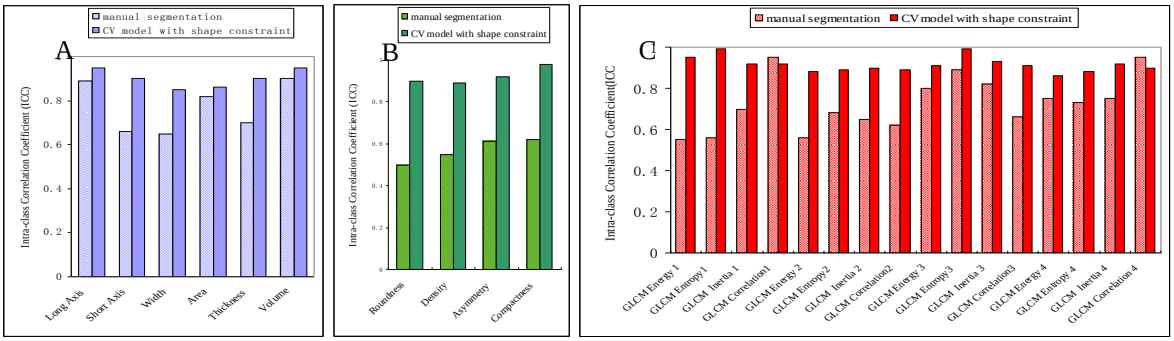
Once the segmentation of all target lesions was deemed accurate sufficiently, statistics for each lesion, such as size, shape, and average density, all readily available .In order to evaluate the reproducibility of the features extracted from the CT image, we assessed 217 radiomic features (125 3D and 92 2D) by calculating intra-class correlation coefficient (ICC) for 35 patients .Intra-class correlation coefficient (ICC) was calculated to quantify the reproducibility of these features .McGraw and Wong defined ICC values for two-way random and Single Absolute [9] agreement to measure the absolute agreement as

$$ICC(A,1) = \frac{MS_R - MS_E}{MS_R + (k-1)MS_E + \frac{k}{n}(MS_C - MS_E)} \tag{3}$$

ICC values for One-way random and Single Absolute agreement is defined as

$$ICC = \frac{MS_R - MS_W}{MS_R + (k-1)MS_W} \tag{4}$$

Where MS_R =mean square for rows, MS_W =mean square for residual sources of variance, MS_E =mean square error, MS_C =mean square for columns, k =number of observers Involved, n =number of subjects. Fig.4 shows the compare of ICC between manual segmentation method and CV model with shape constraint. It turned out that features extracted by image segmentation algorithm based on CV model with shape constraint have a higher reproducibility significantly, which can be employed for quantitative image feature extraction and image data mining research for prognostic.



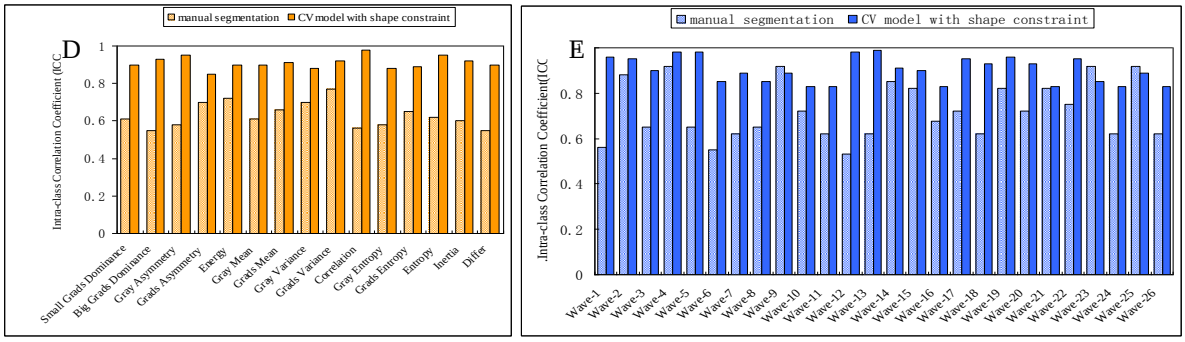


Fig.4 The comparison of Intra-class correlation coefficients (ICC) between CV and CV model with shape constraint segmentations. A: Tumor size. B: Shape C: Gray level co-occurrence matrix features. D: Gray gradient histogram. E: Gabor wavelet.

2.2 Features extraction

In this paper, we extracted a series of tumor features from the region of interest（ROI) of segmented CT image ,which can better reflect the heterogeneity of the tumor areas [17].These features including size, shape, gray gradient, tumor gray histogram features, gray level co-occurrence matrix, Gabor wavelet, laws features [18]. We use MATLAB10.0 and C++ as the image analysis platform.

2.2.1 Tumor size features

Tumor size features contain univariate, bivariate and volume measurements features in pixel units as well as in native resolution. Univariate features include long axis, short axis, width, and other features can represent tumor size. Bivariate features include area, thickness.

2.2.2 Tumor shape features

Tumor shape features category is a kind of measurement of roundness. It contains density, asymmetry, compactness, and largest elliptical fit in the tumor region. The density of the tumor is described by its volume pixel in the three-dimensional space. The density is lower when it is like a filament. The tumor asymmetry was measured in roundness.

2.2.3 Gray gradient histogram features

Gray gradient histogram features analysis can not only use the information of the gray level itself, but also can be used to change the gradient information of the gray level. Gray gradient co-occurrence matrix texture analysis method is used to extract the texture features with the comprehensive information of the gray level and gradient, and the joint statistical distribution of the pixel gray level and the edge gradient is considered represent the varietyof tumor gray level in tumor region with computer the variance, entropy, energy and other statistical features. In this paper, we computed Small Grads Dominance, Big Grads Dominance, Gray Asymmetry, Grads Asymmetry, Energy, Gray Mean, Grads Mean, Gray Variance, Grads Variance, Correlation, Gray Entropy, Grads Entropy, Entropy, Inertia, and Differ Mo.

2.2.4 Gray level co-occurrence matrix features

Gray level co-occurrence matrix features represent the spatial dependence of the gray level, which suggested the spatial relationship of the pixels in a texture pattern. The co-occurrence matrix is a matrix that contains the frequency of one gray level intensity appearing in a specified spatial linear relationship with another gray level intensity within a certain range. Furthermore, co-occurrence matrix defined by the joint probability density of the pixels of the two gray position. It is not only reflects the brightness distribution characteristics, but also reflect luminosity of the same or close to the brightness of the pixels between the distribution characteristics, which is the second-order statistical features of the image brightness changes. We calculate the energy, entropy, moment of inertia and correlation of texture features in 0 °, 45 °, 90 °and 135 °four directions

2.2.5 Laws features

Laws features are structured from a set of five one-dimensional filters which designed to detect a different type of structure in the image, and 125 features are computed on different group kernel and orientation approximately. These 1D filters are represented as E5 (edges), S5 (spots), R5 (ripples), W5 (waves), and L5 (low pass, or average gray value).Through using these 1D convolution filters ,the 2D filters are produced by convolving pairs of these filters, such as L5L5, E5L5, S5L5, W5L5, R5L5, etc. Accordingly, 25 different 2D filters can be generated in the aggregate. 3D Laws filters are structured sim-ilarly by convolving 3 types of 1D filter. Such as L5L5L5, L5L5E5, L5L5S5, L5L5R5, L5L5W5, etc. The quantity of 3D filters is 125.

Table 1. Feature category and description.

Category	Description	2D(No. of Features)	3D(No. of Features)
F1:Tumor size	Size, Volume descriptors	6	
F2:Tumor shape	Roundness	4	
F3: Gray gradient histogram	Statistics on the intensity	15	
F4:Gray level co-occurrence matrix	Energy, Entropy	16	
F5:Gabor wavelet	Wavelet kernel	26	
F6:Laws	Laws kernel	25	125
	Total	92	125

2.2.6 Gabor wavelet features

Gabor wavelet features can decomposes an image into four components, which based on the frequency, content and orientation. The Four components are emerged: a high-pass/high-pass component composing of mostly diagonal structure, a high-pass/low-pass component composing mostly of vertical structures, a low-pass/high-pass component composing mostly of horizontal structure, and a low-pass/low-pass component that represents a blurred version of the original image. Subsequent iterations then repeat the decomposition on the low-pass/low-pass component from the previous iteration, which highlight broader diagonal, vertical, and horizontal textures. We computed the energy feature for each component. By applying the 1D wavelet decomposition along the rows and columns of the image separately, the wavelet decomposition of a 2D image can be achieved. Whereas, 1D wavelet transform is applied along all the three directions (x, y, z) for 3D.The features listed in table.1.

2.3 Feature selection method

Although we began with a large feature sets compared to prior conventional radiologic analyses, it is expected that there may be redundancy in these features due to various limitations on the sample size, texture, and consistency in the population. Thus, to obtained the reproducibility of this informative feature sets, we filtered features on the basis of their reproducibility firstly, i.e., concordance correlation coefficients (CCC) between the repeat scans (test and retest). The CCC is defined as

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)} \tag{5}$$

Where μ_x, μ_y is the mean value of x and y, σ_x, σ_y is the variance of x and y ,and ρ is the correlation coefficient values close to1 are preferred. Then a large number of and diversity features are obtained by using Dynamic Range(DR).The Dynamic Range for a feature is defined as the inverse of the average difference between measurements divided by the entire range of observed values in the sample set as in

$$DR = (1 - \frac{1}{n}\sum_{i=1}^{n}\frac{|Test(i) - \mathrm{Re}test(i)|}{Max - Min}) \quad (6)$$

Where n is the number of data sets. Max and Min are the maximum value and minimum value for a test/retest population of n patient cases, $Test(i) - \mathrm{Re}test(i)$ is the difference of entire sample set. Values close to1 are preferred, and purport that the feature has a large biological range relative to reproducibility. Table 2 shows the feature obtained after CCC and DR produce.

Table 2. Feature obtained after CCC and DR produce.

Category	CCC and DR≥0.85	CCC and DR≥0.90	CCC and DR≥0.95
F1:Tumor size	6	6	5
F2:Tumor shape	4	4	4
F3: Gray gradient histogram	15	13	9
F4:Gray level co-occurrence matrix	15	7	5
F5:Gabor wavelet	34	12	2
F6:Laws	18	15	7
Total	92	57	32

Finding features that are consistent in repeated experiments is a prerequisite step, which is followed by a redundancy reduction step to obtain an informative set. In previous work, we has extracted features that include a large number of quantitative image information, whereas they are almost redundant In recent years, the measure of the correlation of various features is proposed. Symmetrical Uncertainty (SU) and Pearson Correlation Coefficient (PCC) are always used as a correlation measure to evaluate the correlation of features. This kind of measure usually reflect only the correlation between the two features, whereas did not reflect the influence of other features of them. Therefore, we used Sparse Representation Coefficient (SRC) to assess the correlation of these features. SSC is a reflection of the relevance of a feature under the influence of other features, which is a priority compared with other measures.

2.3.1 Pearson Correlation Coefficient (PCC)

Pearson Correlation Coefficient (PCC) is a linear correlation coefficient, which reflects the degree of correlation between the two variables. Suppose X and Y are random variables, the PCC is defined as follows:

$$PCC(X,Y) = \frac{\sum_i (x_i - \bar{x}_i)(y_i - \bar{y}_i)}{\sqrt{\sum_i (x_i - \bar{x}_i)^2}\sqrt{\sum_i (y_i - \bar{y}_i)^2}} \quad (7)$$

Where x_i and y_i is the mean of X , Y and $PCC(X,Y) \in [-1,1]$ .In the case of $PCC = -1$ or $PCC = 1$, the two variables are fully correlated. When $PCC = 0$ indicates that the two variables are linearly independent.

2.3.2 Symmetrical Uncertainty (SU)

Symmetrical Uncertainty (SU) is a kind of nonlinear correlation measure based on entropy definition, which can reveal the nonlinear correlation between the two variables. The entropy of random variable X is defined as

$$H(X) = -\sum_i P(x_i)\log_2 P(x_i) \quad (8)$$

After observing the random variable Y, the entropy of random variable X is defined as

$$H(X|Y) = -\sum_j P(y_j) \sum_i P(x_i|y_j) \log_2 P(x_i|y_j) \quad (9)$$

Where, $P(x_i)$ is the probability of $X = x_i$, $P(x_i|y_i)$ is the probability of $X = x_i$ under $Y = y_i$. After observing the random variable Y, the reduction of the entropy of X becomes the information gain, that is

$$IG(X|Y) = H(X) - H(X|Y), s_i \in R^N \quad (10)$$

And X is a kind of normalized information gain.

$$SU(X,Y) = 2\left[\frac{IG(X|Y)}{H(X)+H(Y)}\right] \quad (11)$$

$SU(X,Y) \in [0,1]$.In the case of $SU(X,Y) = 1$, the two variables is fully correlated; in the case of $SU(X,Y) = 0$ indicates that the two variables are independent of each other.

2.3.3 Sparse Representation Coefficient (SRC)

Given M elementary signals (atoms) : $s_i (i = 1,\ldots,M)$, where $M \geq N$, a target Signal $\phi \in R^N$,and an over complete dictionary $S = [s_1, s_2, \ldots s_i]$ .The problem of sparse representation is to find an $M \times 1$coefficient vector β ,such that $S\beta = \phi$, and β should as little non-0 elements as possible in the meantime. Sparse Representation Coefficients (SRC) is the elements in β. In this study, we apply for basis pursuit, which commands the sparse representation problem as the following constrained l-norm minimization problem:

$$\min \| \beta \|_1 \quad \text{subject to } S\beta = \phi \quad (12)$$

Problem (12) can be solved using linear programming, and it ensures the global optimal solution of β can be achieved. This means interaction between atoms is considered and the optimal combination of atoms for approaching ϕ is found out in the process of solving problem (12). Thus the absolute value of i-th SRC in β reflects the relevance between i-th atoms and ϕ under the influence of all other atoms.

Given a data set $P = \{(x_i, y_i) | i = 1, 2, \cdots, N\}$ Where $x_i = [f_1, f_2, \cdots, f_M]^T \in R^M$, $y_i \in \{1, 2, \cdots, k\}$, k is class number of sample set . Let $X = [x_1, x_2, \cdots, x_N]^T$, $Y = [y_1, y_2, \cdots, y_N]^T$ .If the class label as the target signal will lead to the characteristics of the correlation will be the subject of change and change. Therefore, we extracted the first component of Margin Maximizing Discriminant Analysis (MMDA) as the target signal on the X . Since the first component extracted by MMDA is in the direction of the projection data of the optimal interface method, it is not only does not change with the value of the class label, but also the classification of the two types of samples is also the best of them.

For two-class problem, let D is the first component extracted from MMDA on X .To obtain the vector of SRC, θ, which can reveal the importance of sparse representation class variable. The formula is calculated as follows:

$$\min \|\theta\|_1, \quad \text{s.t.} \quad X\theta = D \quad (13)$$

After θ is calculated, the relevance of feature-category for feature $f_j, j = 1, 2, \cdots, M$, is defined as $SRC(f_j, y) = |\theta_j|$.

For k-class problems, let D^i be the 1-th component of extracting from X in MMDA, where $i = 1, 2, \cdots, k$ is the k-th two-class problem. SRC corresponding vector for i-th two-class problem θ^i can calculated from the formula as following:

$$\min \|\theta^i\|_1, \quad \text{s.t.} \quad X\theta^i = D^i \quad (14)$$

After the θ is calculated, the relevance of feature-category for feature f_j , $j = 1, 2, \cdots, M$ is defined as

$$SRC(f_j, y) = \sum_{i=1}^{k} |\theta_j^i| \qquad j = 1, 2, \cdots, M \tag{15}$$

Let $f_i, i \in \{1, 2, \cdots, M\}$ be the target feature. $F_i = [X_{1,i}, X_{2,i}, \cdots, X_{N,i}]^T$ is the vector of the i-th feature from all of the samples. Features from other samples except f_i is denoted as $x_i^j = [f_1, f_2, \cdots, f_{i-1}, f_{i+1}, \cdots, f_M]^T$, $j \in \{1, 2, \cdots, \mathrm{N}\}$ .Corresponding samples is denoted as $X' = [x_i^1, x_i^2, \cdots, x_i^N]^T$ .The other sparse representation f_i coefficient obtained by calculating the formula as follows

$$\min \| \theta_{f_i} \|_1, \qquad s.t. \qquad X'_i \theta_{f_i} = F_i \tag{16}$$

After getting θ_{f_i} , it changes each element of θ_{f_i} into absolute value to get $\theta_{f_i}, i = 1, 2, \cdots, M$, Let $W = [\theta_{f_1}, \theta_{f_2}, \cdots, \theta_{f_M}]$. $W(i, j)$ reflects the importance of f_i to f_j ,and $W(j, i)$ reflect the importance of f_j to f_i .We defined the sparse representation feature-feature correlation between f_i and f_j as

$$SRC(f_i, f_j) = \frac{W(i, j) + W(j, i)}{2} \tag{17}$$

2.3.4 Feature select ion methods based on sparse representation coefficient (SRC)

In the feature selection processing, a feature is selected by using SRC to assess the coefficient firstly, and then features that redundant is removed by using approximate Markov Blanket that defined by SRC.For approximate Markov Blanket: Since f_i is the approximate Markov Blanket of f_j , it follows that $SRC(f_i, y) \geq SRC(f_j, y)$ and $SRC(f_i, f_j) \geq SRC(f_j, y)$.

The steps of feature selection method based on sparse representation coefficient are as follows (the input is sample set D and threshold value K ; the output is feature sequences P :

(1) Correlation Analysis.

1) We use $SRC(f_i, y)$ to evaluate the correlation between each feature f_i , $i = 1, 2, \cdots, k$ and category.

2) According to the correlation of category, the features are ranked in descending order.

3) Then the top K features are extracted, adding them to P one by one.

(2) Redundancy Analysis.

4） Let $i = 1$,N is the number of features.

5） For each f_j , $j = i+1, i+2, \cdots, \mathrm{N}$,in case $SRC(f_i, y) \geq SRC(f_j, y)$, f_j is removed from P and let $N = N - 1$.

6） If $i < N$, $i = i+1$,and go to step 5）

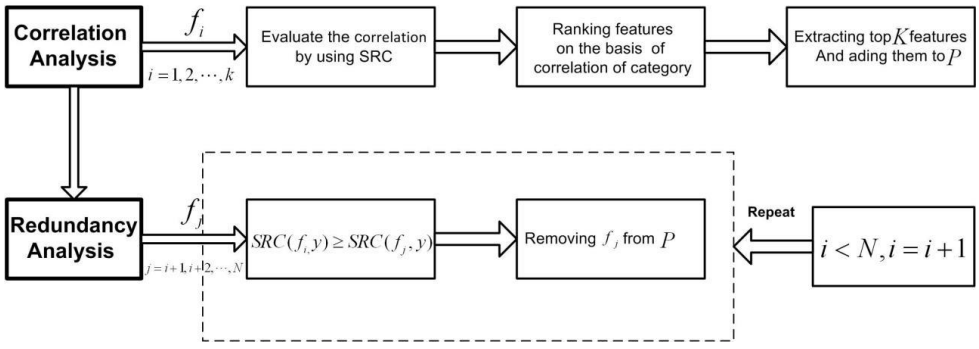


Fig.5 The process of feature selection.

3. Results

In this study, CV model with shape constraint algorithm was used for 35 non–small cell lung cancer (NSCLC) patients to get the region of interesting, and 125 3D and 92 2D features from the ROI were extracted. Intra-class correlation coefficient (ICC) was calculated to quantify the reproducibility of these features. Fig.4 shows the compare of ICC between manual segmentation method and CV model with shape constraint. It is indicated that the method of CV model with shape constraint algorithm has a good effect on feature extracted, which is reproducibility. In order to conduct the concordance correlation coefficient (CCC) and dynamic range (DR), we scan the CT images for the CT images twice. A set of representative reproducible image features were obtained with CCC and DR≥ 0.90 after concordance correlation coefficient (test and retest) and dynamic range procedure. Tables.2 lists the feature counts.

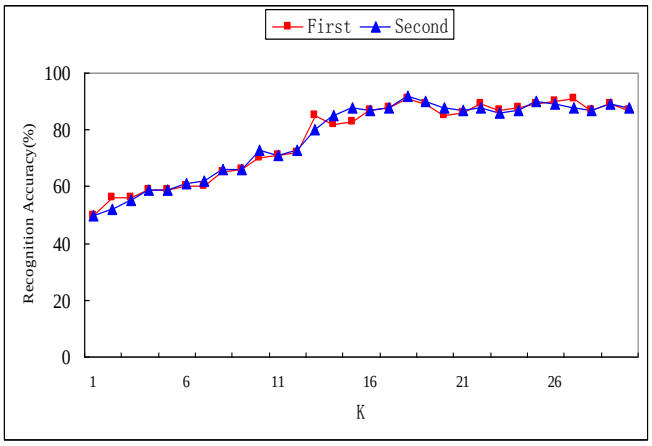


Fig.6 The influence of K on the performance of SVM

In order to assess the ability of feature section method for non-redundant features, We tested the ability of classify on an independent 43 NSCLC, 18 adenocarcinoma and 25 epidermoid carcinoma samples. The classifier is support vector machine (SVM).In this study, the value of K determine the number of feature selection in correlation analysis and result in the scale of redundancy of analysis. In this experiment, let $0 < K \leq 30$. Fig.6 showed the influence of K on the performance of SVM. It is suggested that recognition rate curve is stabilized gradually with K in 8 to 26 based on two experiments. Table 3 shows the recognition rate results of the classifiers compared the classification ability of feature selection method based on SRC, SU and PCC respectively. The experiment proves that the result of SRC is effective for feature selection. According to the experiment, we obtained a set of representative features that reproducibility and non-redundancy. We list it in Table.4.

In this experiment, all programs were realized by Matlab and C++.Running software environment is MAtlab7.1in windows 7 on a i3-2310M CPU and 8 GB RAM.

Table 2. The recognition rate results of the classifiers compared the classification ability of feature selection method based on SRC, SU and PCC respectively.

Feature selection Method	Parameter g	Penalty Factor C	Recognition Accuracy (%)
Symmetrical Uncertainty (SU)	0.142	300	80.3215
Pearson Correlation Coefficient (PCC)	0.128	100	85.7658
Sparse Representation Coefficient (SRC)	0.091	150	89.6528

Table3. The representative features obtained by SRC feature selection method.

Representative Features (CCC and DR ≥ 0.9)
F1:Tumor size() C1:Area;C2:Width;C3:Tickness;C4:Volume
F2:Tumor shape(4) C5:Roundness;C6:Density;C7:Asymmetry;C8:Compactness
F3:Gray gradient histogram(4) C9:Energy;C10:Correlation;C11:Entropy;C12:Differ
F4:Gray level co-occurrence matrix(5) C13:GLCM Enpropy1;C14: GLCM Energy 1;C15:GLCM Inertia 1;C16:GLCM Correlation1
F5:Laws(3) C17:Laws-E5 E5 E5;C18:Laws- E5 S5 L5 Layer 1;C19:Laws- R5 S5 W5 Layer 1
F6:Gabor wavelet(2) C20:Wave-6;C21:Wave-22

4. Discussion

Medical imaging is a component in modern medicine that as an important instrument for cancer staging and treatment response monitoring. In previous studies, it is generally utilized sub-fraction of tumor feature category to analyzed preliminarily for lung cancer patient, especially for NSCLC [1,2]. With the rapid development of modern medical technology, using data mining to quantify tumor heterogeneity for establishing potential imaging biomarkers was investigated. Referring to radiomics in medical imaging, a large number quantitative features was extracted from tumor region of interesting (ROIs)[18], which provided information in tumor heterogeneity, and expand the scope of clinical oncology. For example, it has been suggested 78% of the gene expression variability in hepatocellular carcinoma can be predicted by extracting features from CT image [3].

The reproducibility of radio graphical feature extracted from CT image of NSCLC depended on the segmentation of accuracy for ROI [22]. Despite high segmentation accuracy and sensibility of human eyes to texture features, manual segmentation method is time consuming and high operator input. Prior work has shown numerous algorithm for image segmentation, for example, region grow, quaternary tree and level set. In addition, it has been shown that the segmentation method based on CV module for the region of tumor is widen significantly. However, it is a little effect for gray scale inhomogeneous of tumors, which has an important role for lung cancer. Therefore, In order to enhance the accuracy of segmentation and achieve the tumor region repetitively, it is essential to enlarge the image constraint force for the object edge, which is the CV model with shape constraint. And in order to access the robustness of segmentation, intra-class correlation coefficient (ICC) [19], which can be used to evaluate the coherence and reliability for measuring. The results in divided that the method of CV model with shape constraint compared with manual segmentation method has a better consequence in most features (See Fig.4).

Prior work has augmented RECIST and WHO measures to infer concordance consistency for automatically segmented lung lesions, which seem to be limited in describing the complex nature of the tumor [20]. In this study, we extracted a large number of features from ROI using the different categories as follows: size, shape, gray gradient histogram, gray co-occurrence matrix, laws, Gabor wavelet.(table.1) For the set of features, consistency in the repeat scans (test, retest) was tested and filtered for representative features by using concordance correlation coefficients (CCC), it has used a correlation coefficient with 0.9 to distinguish highly correlated features, which is a stringent measure of reproducibility [21].The dynamic range for a feature is defined as the inverse of the average difference between measurements divided by the entire range of observed values. In a parallel research, screening

for a large DR means greater variability in the repeat scans. In this article, after filtering the features according to CCC and DR>0.9 (See Table 2) 65 features were obtained.

This representative features are also redundant, and SU and PCC is using to dimensionality reduction in former years. But they have not taken attention to r the relevance between feature and category. The method of feature selection based on sparse representation coefficient (SRC) provided a good method of feature selection. Then we used SVM to classify 43 NSCLC, 18 adenocarcinoma and 25 epidermoid carcinoma samples and the recognition accuracy is better than the method based on SU and PCC.(See Table.2) It is considered the correlation between features and category compared with other feature selection methods. This means those features that contain not only category information singly, but also important for sparsely representing the category variable [22]. Therefore, quantitative image features that reproducibility and non-redundancy provide informative and prognostic biomarkers for NSCLC can be obtained.

5. Conclusion

In this paper, we proved that the effective method of feature selection to obtained reproducibility and non-redundancy features from lung CT images. According to using the segmentation method based on CV model with shape constraint, we get a tumor region of interesting for reproducibility features, and it has a better consequence in most features comparing the ICC with manual method. Then we reduced the features from 217 to 57 by combining segmentation differences and biologic range at CCC and $DR \geq 0.9$.A feature selection method base on SRC was proposed to reduce the dimension and 21 features are representative features. This features has a high ability to distinguish the tumor and informative and prognostic biomarkers for lung cancer patient.

References

[1] SEER-NCI (2013). SEER Cancer Statistics Review (CSR) 1975-2010.

[2] American Cancer Society. Cancer Facts & Figures 2016. American Cancer Society. Available athttp://www.cancer.org/acs/groups/content/@research/documents/document/acspc-047079.pdf. Accessed: March 30, 2016.

[3] Jemal, Ahmedin, et al. "Global cancer statistics. " Ca A Cancer Journal for Clinicians 61.2(2011):69-90.

[4] Na F, Wang J, Li C, et al. "Primary tumor standardized uptake value measured on F18-Fluorodeoxyglucose positron emission tomography is of prediction value for survival and local control in non-small-cell lung cancer receiving radiotherapy: meta-analysis. " Journal of Thoracic Oncology Official Publication of the International Association for the Study of Lung Cancer25. 108(2014):1080-7.

[5] Kumar, Virendra, et al. "Radiomics: the process and the challenges."Magnetic Resonance Imaging 30.9 (2012):1234-1248.

[6] Parmar C, Rios V E, Leijenaar R, et al. "Robust Radiomics feature quantification using semiautomatic volumetric segmentation. " Plos One 9.7(2014):e102107-e102107.

[7] Gu, Y., Kumar, V., Hall, L. O., Goldgof, D. B., Li, C. Y., & Korn, R., et al. "Automated delineation of lung tumors from CT images using a single click ensemble segmentation approach." Pattern Recognition 46.3(2013):692-702.

[8] Li, Qianqian, et al. "An Improved Method Based on CV and Snake Model for Ultrasound Image Segmentation." International Conference on Image & Graphics IEEE Computer Society, 2013:160-163.

[9] Reinhold, Müller, and B. Petra. "A critical discussion of intraclass correlation coefficients. " Statistics in Medicine 13.23-24(1994):2465–2476.

[10] Balaji, Ganeshan, et al. "Non-small cell lung cancer: histopathologic correlates for texture parameters at CT. " Radiology 266.1(2013):326-336.

[11] Feng, Changyong, et al. "A note on the concordance correlation coefficient." Advances & Applications in Statistics 15.2(2010):195-205.
[12] Lin, L., and L. D. Torbeck. "Coefficient of accuracy and concordance correlation coefficient: new statistics for methods comparison. " Pda Journal of Pharmaceutical Science & Technology 52.2(1998):55-9.
[13] Wang, Jiguang. Pearson Correlation Coefficient. Springer New York, 2013.
[14] Kannan, S. Senthamarai, and N. Ramaraj. "A novel hybrid feature selection via Symmetrical Uncertainty ranking based local memetic search algorithm." Knowledge-Based Systems 23.6 (2010): 580-585.
[15] Wright J, Ganesh A, Zhou Z, et al.Wright, J., et al. "Demo: Robust face recognition via sparse representation." Automatic Face & Gesture Recognition, 2008. FG '08. 8th IEEE International Conference on 2008:1-2.
[16] Lin, Ting Qiang. "A New Algorithm for Image Segmentation Base On CV Model." Signal Processing (2010).
[17] Balaji, Ganeshan, et al. "Non-small cell lung cancer: histopathologic correlates for texture parameters at CT. " Radiology 266.1(2013):326-336.
[18] Lambin, P, et al. "Radiomics: Extracting more information from medical images using advanced feature analysis." European Journal of Cancer48.4(2012):441-6.
[19] Weir, J. P. "Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. " Journal of Strength & Conditioning Research 19.1(2005):p ágs. 231-240.[20]Iii, Sg Armato, et al. "The Reference Image Database to Evaluate Response to Therapy in Lung Cancer (RIDER) Project: A Resource for the Development of Change-Analysis Software." Clinical Pharmacology & Therapeutics 84.4(2008):448-56.
[21] Lin, L. I. "A concordance correlation coefficient to evaluate reproducibility. " Biometrics 45.1 (1989):255-68.
[22] Rustin, G. J., et al. "Re: New guidelines to evaluate the response to treatment in solid tumors (ovarian cancer). " Journal of the National Cancer Institute 96.18(2004):487-8.