**Radiomics Nomogram for the Preoperative Prediction of Lymph Node Metastasis in Pancreatic Ductal Adenocarcinoma**

**Authors**

Yun Bian, MD, PhD \*; Shiwei Guo, MD, PhD †; Hui Jiang, MD, PhD ‡; Suizhi Gao †; Chengwei Shao \*; Kai Cao, MD, PhD \*; Xu Fang \*; Jing Li, MD, PhD \*; Li Wang, MD, PhD \*; Chao Ma \*; Jianming Zheng, MD, PhD ‡; Gang Jin, MD, PhD †; Jianping Lu, MD, PhD \*

\*Department of Radiology, Changhai Hospital, The Navy Military Medical University, Shanghai, China

†Department of Pancreatic Surgery, Changhai Hospital, The Navy Military Medical University, Shanghai, China

‡Department of Pathology, Changhai Hospital, The Navy Military Medical University, Shanghai, China

**Contents**

[Supplementary 1 Image preprocessing 3](#_Toc88861559)

[Supplementary 2 Radiomics Features 4](#_Toc88861560)

[Supplementary 3. the Least Absolute Shrinkage and Selection Operator (LASSO) Algorithm 17](#_Toc88861561)

[Supplementary 4. Decision Curve Analysis (DCA) 18](#_Toc88861562)

[Supplementary 5. Radiomic Features Selected by LASSO Regularization. 19](#_Toc88861563)

## Supplementary 1 Image preprocessing

Step 1- bias field correction: Differences in medical imaging factors can cause inconsistencies in the image intensity information from tissues of the same nature. We used the following formula for intensity normalization (where x represents the original intensity; f (x) indicates the normalized intensity; indicates the average value; refers to variance; and s is an optional scaling ratio, which has been set to 1 by default). While retaining the intensity difference of the diagnostic value, the image intensity inconsistency caused by the difference in imaging parameters is reduced or even eliminated for subsequent imaging radiomics analysis.

Equation 1 (<https://pyradiomics.readthedocs.io/en/latest/radiomics.html#radiomics.imageoperations.normalizeImage>)

Step 2- Resampling: All sets were isotropically resampled in all 3 dimensions via linear interpolation to ensure consistent voxel sizes and resolutions across all the machines and patients. As a result, the voxel dimensions of each case were 0.66 × 0.66 × 0.66 mm3. This step of processing was performed on all annotation masks as well to ensure that the masks remained in correspondence with the volumes.

## Supplementary 2 Radiomics Features

In this study we explored a feature-based approach to extract and quantify meaningful and reliable information from images. A total of 1029 quantitative imaging features were extracted. The methods of feature extraction used in this study included two categories: original feature classes and filter classes. Filter Classes included five categories: Wavelet, Square, Square Root, Logarithm, and Exponential. A total of 1029 2D and 3D features from primary tumors in each scan phase. The imaging traits were described in detail below.

**1.1 Firstorder**

Firstorder included 19 features.

|  |  |  |
| --- | --- | --- |
| 1 | Energy |  |
|  |  | Here, *c* is optional value, defined by “voxelArrayShift”, which shifts the intensities to prevent negative values in **X**. This ensures that voxels with the lowest gray values contribute the least to Energy, instead of voxels with gray level intensity closest to 0. |
| 2 | TotalEnergy |  |
|  |  | Here, *c* is optional value, defined by “voxelArrayShift”, which shifts the intensities to prevent negative values in **X**. This ensures that voxels with the lowest gray values contribute the least to Energy, instead of voxels with gray level intensity closest to 0. |
| 3 | Entropy |  |
|  |  | Here, is an arbitrarily small positive number (*≈* 2*:*2 *×* 10*-*16) |
| 4 | Minimum | *minimum* = min(**X**) |
| 5 | 10Percentile |  |
| 6 | 90Percentile |  |
| 7 | Maximum | *maximum* = max(**X**) |
| 8 | Mean |  |
| 9 | Median | The median gray level intensity within the ROI. |
| 10 | InterquartileRange |  |
|  |  | Here **P**25 and **P**75 are the 25*th* and 75*th* percentile of the image array, respectively. |
| 11 | Range | *range* = max(**X**) *-* min(**X**) |
| 12 | MeanAbsoluteDeviation (MAD) |  |
|  |  | Mean Absolute Deviation is the mean distance of all intensity values from the Mean Value of the image array. |
| 13 | RobustMeanAbsoluteDeviation (rMAD) |  |
| 14 | RootMeanSquared (RMS) |  |
|  |  | Here, *c* is optional value, defined by “voxelArrayShift”, which shifts the intensities to prevent negative values in **X**. This ensures that voxels with the lowest gray values contribute the least to RMS, instead of voxels with gray level intensity closest to 0. |
| 15 | StandardDeviation |  |
| 16 | Skewness |  |
|  |  | Where *µ*3 is the 3*rd* central moment. |
| 17 | Kurtosis |  |
|  |  | Where *µ*4 is the 4*th* central moment. |
| 18 | Variance |  |
| 19 | Uniformity |  |
| Notations:  **X** is an image of *N* voxels included in the ROI.  *Pi* is the first order histogram with *Nl* discrete intensity levels, in which *Nl* is the number of non-zero bins.  *pi* is the normalized first order histogram and equal to (This definition is the same for the following sections).  **10Percentile**  The 10*th* percentile of **X**.  **90Percentile**  The 90*th* percentile of **X** | | |

**1.2 Shape features**

Shape features describe the morphological property of the tumor region and were features rated from only the image without filtration. They included 13features.

|  |  |  |
| --- | --- | --- |
| 1 | Volume | The volume of the ROI is approximated by multiplying the number of voxels in the ROI by the volume of a single voxel |
| 2 | SurfaceArea |  |
|  |  | N is the number of triangles forming the surface mesh of the volume (ROI) and are the edges of the ith triangle formed by points , and  Surface Area is an approximation of the surface of the ROI in mm2, calculated using a marching cubes algorithm. |
| 3 | SurfaceVolumeRatio |  |
|  |  | Here, a lower value indicates a more compact (sphere-like) shape. This feature is not dimension less, and is therefore (partly) dependent on the volume of the ROI. |
| 4 | Sphericity |  |
|  |  | Sphericity is a measure of the roundness of the shape of the tumor region relative to a sphere. It is a dimensionless measure, independent of scale and orientation.  The value range is 0 *< sphericity ≤* 1, where a value of 1 indicates a perfect sphere (a sphere has the smallest possible surface area for a given volume, compared to other solids). |
| 5 | Compactness1 |  |
| 6 | Compactness2 |  |
| 7 | SphericalDisproportion |  |
| 8 | Maximum3DDiameter | Maximum 3D diameter is defined as the largest pairwise Euclidean distance between surface voxels in the ROI.  Also known as Feret Diameter. |
| 9 | Maximum2DDiameterColumn | Maximum 2D diameter (Column) is defined as the largest pairwise Euclidean distance between tumor surface voxels in the row-slice (usually the coronal) plane. |
| 10 | Maximum2DDiameterRow | Maximum 2D diameter (Row) is defined as the largest pairwise Euclidean distance between tumor surface voxels in the column-slice (usually the sagittal) plane |
| 11 | MajorAxis | The line through the foci is called the major axis, and the line perpendicular to it through the center is called the minor axis. Least axis is seen as the sum of the axis |
| 12 | MinorAxis |
| 13 | LeastAxis |
| 14 | Elongation |  |
|  |  | Here, *λ*major and *λ*minor are the lengths of the largest and second largest principal component axes. The values range between 1 (where the cross section through the first and second largest principal moments is circle-like (non-elongated)) and 0 (where the object is a single point or 1 dimensional line). |
| 15 | Flatness |  |
|  |  | Here, *λ*major and *λ*least are the lengths of the largest and smallest principal component axes. The values range between 1 (non-flat, sphere-like) and 0 (a flat object). |

**1.3 GLCM features**

GLCM features included 28 features

|  |  |  |
| --- | --- | --- |
| 1 | Autocorrelation |  |
| 2 | AverageIntensity |  |
| 3 | ClusterProminence (CP) |  |
| 4 | ClusterShade |  |
| 5 | ClusterTendency (CT) |  |
| 6 | Contrast |  |
| 7 | Correlation |  |
| 8 | DifferenceAverage |  |
| 9 | DifferenceEntropy |  |
| 10 | DifferenceVariance |  |
| 11 | Dissimilarity |  |
| 12 | Energy |  |
| 13 | Entropy |  |
| 14 | Homofeaturesity1 |  |
| 15 | Homofeaturesity2 |  |
| 16 | Imc1 |  |
| 17 | Imc2 |  |
| 18 | Idm |  |
| 19 | Idmn |  |
| 20 | Id |  |
| 21 | Idn |  |
| 22 | InverseVariance (IV) |  |
| 23 | MaximumProbability |  |
| 24 | SumAverage |  |
| 25 | SumEntropy |  |
| 26 | SumVariance |  |
| 27 | SumSquares |  |
| Notations:  **P**(*i; j*) is the co-occurence matrix for *δ* (distance) and *α* (angle)  *p*(*i; j*) is the normalized co-occurence matrix  *Ng* is the number of discrete intensity levels in the image  is the marginal row probabilities  is the marginal column probabilities  is the mean gray level intensity of *px*  is the mean gray level intensity of *py*  *σx* is the standard deviation of *px*  *σy* is the standard deviation of *py*  , where *i* + *j* = *k*  , where  is the entropy of *px*  is the entropy of *py*  is the entropy of *p*(*i; j*) | | |

**1.4 GLSZM features**

GLSZM features included 16 features

|  |  |  |
| --- | --- | --- |
| 1 | SmallAreaEmphasis (SAE) |  |
| 2 | LargeAreaEmphasis (LAE) |  |
| 3 | GrayLevelNonUniformity (GLN) |  |
| 4 | GrayLevelNonUniformityNormalized (GLNN) |  |
| 5 | SizeZoneNonUniformity (SZN) |  |
| 6 | SizeZoneNonUniformityNormalized (SZNN) |  |
| 7 | ZonePercentage (ZP) |  |
| 8 | GrayLevelVariance (GLV) |  |
|  |  |  |
| 9 | ZoneVariance (ZV) |  |
| 10 | ZoneEntropy (ZE) |  |
| 11 | LowGrayLevelZoneEmphasis (LGLZE) |  |
| 12 | HighGrayLevelZoneEmphasis (HGLZE) |  |
| 13 | SmallAreaLowGrayLevelEmphasis (SALGLE) |  |
| 14 | SmallAreaHighGrayLevelEmphasis (SAHGLE) |  |
| 15 | LargeAreaLowGrayLevelEmphasis (LALGLE) |  |
| 16 | LargeAreaHighGrayLevelEmphasis (LAHGLE) |  |
| **Note:**  P (*i; j*) is the size zone matrix  *p* (*i; j*) is the normalized size zone matrix  *Ng* is the number of discrete intensity values in the image  *Ns* is the number of discrete zone sizes in the image  *Np* is the number of voxels in the image | | |

**2.5 GLRLM features**

GLRLM features included 16 features

|  |  |  |
| --- | --- | --- |
| 1 | ShortRunEmphasis (SRE) |  |
| 2 | LongRunEmphasis (LRE) |  |
| 3 | GrayLevelNonUniformity (GLNNN) |  |
| 4 | GrayLevelNonUniformityNormalized (RLNN) |  |
| 5 | RunLengthNonUniformity (RLN) |  |
| 6 | RunLengthNonUniformityNormalized (RLNN) |  |
|  |  | RLNN measures the similarity of run lengths throughout the image, with a lower value indicating more homofeaturesity among run lengths in the image. This is the normalized version of the RLN formula |
| 7 | RunPercentage (RP) |  |
| 8 | GrayLevelVariance (GLV) |  |
|  |  | Here,  GLV measures the variance in gray level intensity for the runs. |
| 9 | RunVariance (RV) |  |
|  |  | Here,  RV is a measure of the variance in runs for the run lengths. |
| 10 | RunEntropy (RE) |  |
|  |  | Here, is an arbitrarily small positive number (*≈* 2*:*2 *×* 10*-*16).  RE measures the uncertainty/randomness in the distribution of run lengths and gray levels. A higher value indicates more heterofeaturesity in the texture patterns |
| 11 | LowGrayLevelRunEmphasis (LGLRE) |  |
| 12 | HighGrayLevelRunEmphasis (HGLRE) |  |
| 13 | ShortRunLowGrayLevelEmphasis (SRLGLE) |  |
| 14 | ShortRunHighGrayLevelEmphasis (SRHGLE) |  |
| 15 | LongRunLowGrayLevelEmphasis (LRLGLE) |  |
| 16 | LongRunHighGrayLevelEmphasis (LRHGLE) |  |
| Notations:  is the run length matrix of direction *θ*  is the normalized run length matrix  *Ng* is the number of discrete intensity values in the image  *Nr* is the number of discrete run lengths in the image  *Np* is the number of voxels in the image | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Table 1 1029 Radiomics Feathers | | | | |
|  | **Firstorder** | **Shape** | **GLCM** | **GLSZM** | **GLRLM** |
| 1 | Energy | Volume | Autocorrelation | SmallAreaEmphasis | ShortRunEmphasis |
| 2 | TotalEnergy | SurfaceArea | AverageIntensity | LargeAreaEmphasis | LongRunEmphasis |
| 3 | Entropy | SurfaceVolumeRatio | ClusterProminence | GrayLevelNonUniformity | GrayLevelNonUniformity |
| 4 | Minimum | Sphericity | ClusterShade | GrayLevelNonUniformityNormalized | GrayLevelNonUniformityNormalized |
| 5 | 10Percentile | Compactness1 | ClusterTendency | SizeZoneNonUniformity | RunLengthNonUniformity |
| 6 | 90Percentile | Compactness2 | Contrast | SizeZoneNonUniformityNormalized | RunLengthNonUniformityNormalized |
| 7 | Maximum | SphericalDisproportion | Correlation | ZonePercentage | RunPercentage |
| 8 | Mean | Maximum3DDiameter | DifferenceAverage | GrayLevelVariance | GrayLevelVariance |
| 9 | Median | Maximum2DDiameterColumn | DifferenceEntropy | ZoneVariance | RunVariance |
| 10 | InterquartileRange | Maximum2DDiameterRow | DifferenceVariance | ZoneEntropy | RunEntropy |
| 11 | Range | MajorAxis | Dissimilarity | LowGrayLevelZoneEmphasis | LowGrayLevelRunEmphasis |
| 12 | MeanAbsoluteDeviation | MinorAxis | Energy | HighGrayLevelZoneEmphasis | HighGrayLevelRunEmphasis |
| 13 | RobustMeanAbsoluteDeviation | LeastAxis | Entropy | SmallAreaLowGrayLevelEmphasis | ShortRunLowGrayLevelEmphasis |
| 14 | RootMeanSquared | Elongation | Homofeaturesity1 | SmallAreaHighGrayLevelEmphasis | ShortRunHighGrayLevelEmphasis |
| 15 | StandardDeviation | Flatness | Homofeaturesity2 | LargeAreaLowGrayLevelEmphasis | LongRunLowGrayLevelEmphasis |
| 16 | Skewness |  | Imc1 | LargeAreaHighGrayLevelEmphasis | LongRunHighGrayLevelEmphasis |
| 17 | Kurtosis |  | Imc2 |  |  |
| 18 | Variance |  | Idm |  |  |
| 19 | Uniformity |  | Idmn |  |  |
| 20 |  |  | Id |  |  |
| 21 |  |  | Idn |  |  |
| 22 |  |  | InverseVariance |  |  |
| 23 |  |  | MaximumProbability |  |  |
| 24 |  |  | SumAverage |  |  |
| 25 |  |  | SumEntropy |  |  |
| 26 |  |  | SumVariance |  |  |
| 27 |  |  | SumSquares |  |  |

**Reference:** van Griethuysen JJ, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, Beets-Tan RG, Fillion-Robin J-C, Pieper S, Aerts HJJCr: Computational radiomics system to decode the radiographic phenotype. 2017;77:e104-e107

## Supplementary 3. the Least Absolute Shrinkage and Selection Operator (LASSO) Algorithm

LASSO is a powerful algorithm for regression analysis with high dimensional predictors. In our study, the LASSO algorithm was combined with the logistic regression model for model development. We used the LASSO logistic regression model to select the most important predictive features and construct a radiomics signature in the training set. This algorithm minimizes a log partial likelihood subject to the sum of the absolute values of the parameters bounded by a constant:

, subject to

where is the obtained parameters, is the log partial likelihood of the logistic regression model, and is a constant.

The LASSO algorithm shrinks some coefficients and reduces others to exactly 0 via the absolute constraint. Thus, LASSO is an outstanding method for feature selection by retaining the good features of both subset selection and ridge regression.

**Reference:** Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological) 1996:267-88

## Supplementary 4. Decision Curve Analysis (DCA)

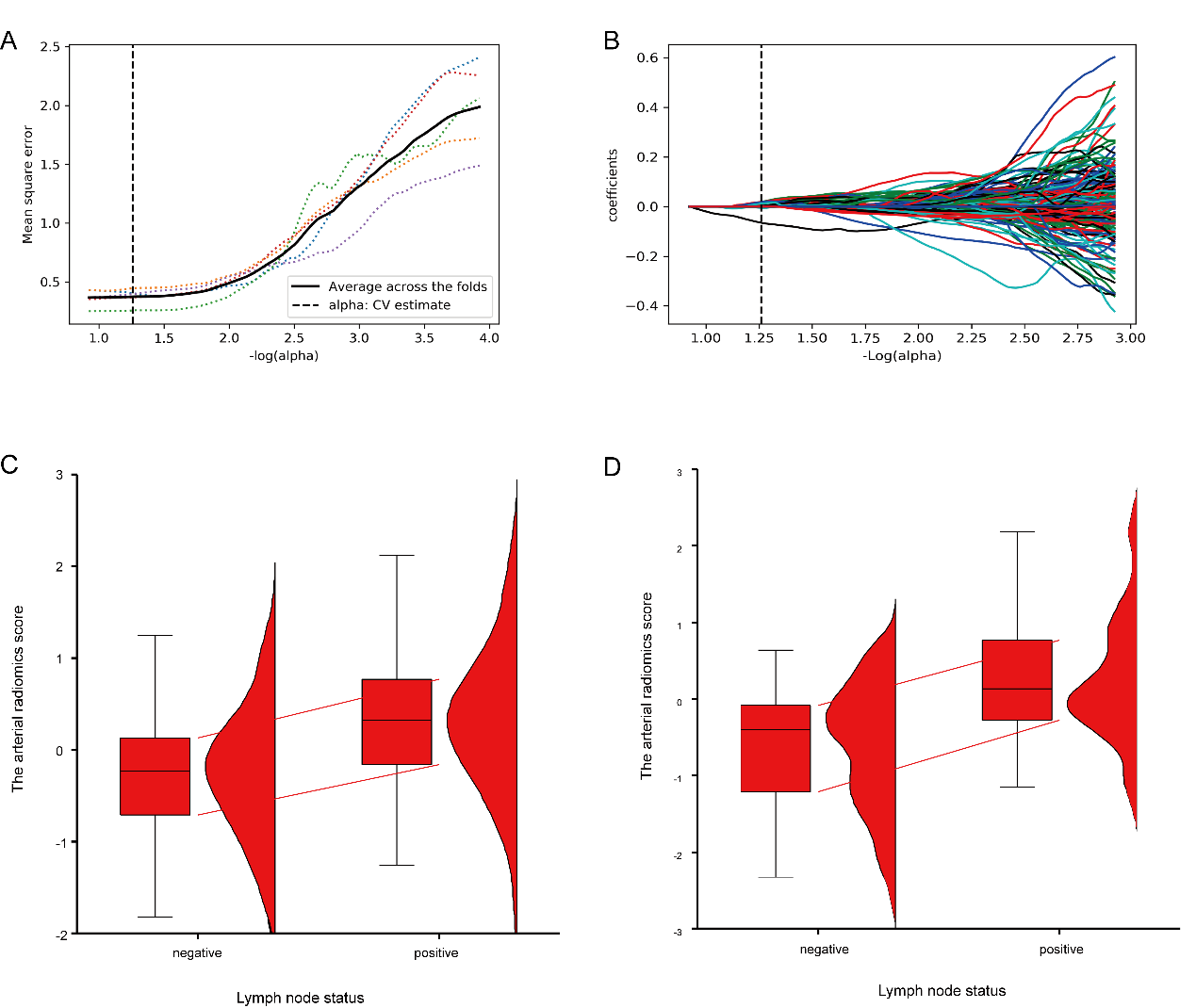
In our study, the DCA method was used to evaluate the clinical utility of the radiomics nomogram. The DCA algorithm assesses prediction models by calculating the range of threshold probabilities in which a prediction or prognostic model was clinically useful. DCA is a compositive method for evaluating and comparing different diagnostic and prognostic models. The theory of DCA can be illustrated by the equation below:

where *d* – *b* represents the influence of unnecessary treatment. If treatment is directed by a prediction model, *d* – *b* is the harm related to a false-positive result compared with a true-negative result. Inversely, *a* – *c* represents the consequence of rejecting beneficial treatment, in other words, the harm from a false-negative result compared with a true-positive result. *Pt* represents where the expected benefit of treatment is equal to the expected benefit of refraining from treatment.

**Reference:** Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006 Nov-Dec;26(6):565-74.

## Supplementary 5. Radiomic Features Selected by LASSO Regularization.

A total of 1029 radiomics features from the arterial phase of CT were extracted and grouped on the basis of the LN metastasis. Consequently, the radiomics characteristics were reduced to 13 features. There was a significant difference in arterial rad-scores between the LN-positive and the LN-negative patients (p =0.002).



**Figure A, B** Radiomic feature selection by using a parametric method, the least absolute shrinkage and selection operator (LASSO). **(A)** Selection of the tuning parameter (λ) in the LASSO model via 5-fold cross validation based on minimum criteria. Mean square error from the LASSO regression cross-validation procedure were plotted as a function of -log(λ). The y-axis indicates mean square error. The lower x-axis indicates the -log(λ). Black line indicate average square error values for each model with a given λ, and colorful lines show mean square error of λ value. The vertical black lines define the optimal values of λ, where the model provides its best fit to the data. The optimal λ value of 0.055 with log(λ) = -1.26 was selected. **(B)** LASSO coefficient profiles of the 1024 texture features. The dotted vertical line was plotted at the value selected using 5-fold cross-validation in A. The thirteen resulting features with nonzero coefficients are indicated in the plot. **C, D** Combo chart of the arterial radiomics scores by the LASSO regression formula. Combo chart included box plot and density plot. 25th and 75th percentiles connecting lines between groups. (**C)** training cohort, (**D)** validation cohort.