IJAMML 8:1 (2018) 23-40 March 2018
ISSN: 2394-2258
Available at http://scientificadvances.co.in
DOI: http://dx.doi.org/10.18642/ijamml_7100121932

RADIOMICS FEATURES APPLICATION TO LUNG CANCER CLASSIFICATION IN CT IMAGE AND ITS CLINICAL EXPLANATION

Yang Hou

Department of Mathematics, Jinan University, Tianhe, Guangzhou, P. R. China

Abstract

For overcoming many challenges for lung cancer classification in CT image, such as (1) the size and shape of lung cancer in different parts of the body are uneven; (2) the locations of the edge of the lung cancer confirmed by different doctors (or the same doctor at different times) are different; (3) the same type of lung cancer in different patients has different growth directions; and (4) it is difficult to judge the classification effect etc. In this paper, we firstly extract a total of 5 categories of 385 radiomics features, and explain the significance of pathological organization of each type of features by the imaging principle. Then we select the logistic regression with L1 regularized model as our predictor. When the segmentation is not too fine for lung squamous cell carcinoma and adenocarcinoma, the classification also achieved a good result with an ACC of 75% and an AUC of 75.3%. To further illustrate the usefulness of these five broad categories of features in overcoming the difficulty of lung cancer imaging classification, we select a more detailed segmented lung cancer for classifying images of benign and malignant images, eventually we can get a better result with ACC of 85% and AUC of 90.7%.

*Corresponding author.
E-mail address: yanghou1991@sina.cn (Yang Hou).

2010 Mathematics Subject Classification: 00A06.
Submitted by Omer Faruk Ertugrul.
Received February 24, 2018; Revised March 8, 2018

Keywords: radiomics features, lung cancer, classification, CT image, logistic regression, clinical explanation, L1 regularization.

1. Introduction

Image classification is already a very mature area, the most primitive method of it is to extract many features of computer vision from the image, and then build a model by machine learning to get an optimal classifier, such as [1, 2]. Later, the image classification problem evolved in feature coding of images, and then use machine learning methods to identify or classify, such as SIFT [3], HOG [4], etc.; currently due to the dramatic increase in the amount of data, the most advanced method is to use deep learning method for image classification, such as ImageNet [5] etc.

In the classification of medical images, the amount of data is very small, and most clinicians are expecting to interpret the features with clinical and pathological significance, so deep learning for classification of medical images is still at the basic research stage and hard to be accepted by clinicians. However, the traditional method is very useful, that is, the method of machine learning used to classify samples by extracting features from medical images, such as [6, 7, 8].

There are many challenges for medical image classification according to [7]. In particularly, the classification of lung cancer CT imaging images is very difficult, mainly manifested in:

(1) The size and shape of the same type of lesions may not uniform;

(2) The edge of the lesion of the lung cancer may be different from time to time by different doctors or even by the same doctor;

(3) The same type of lung cancer has different growth directions;

(4) It is hard to judge the classification effect etc.

For the classification of lung cancer in CT imaging images, we hope to extract some type of features such that they not only can effectively distinguish different types of lung cancer, such as squamous cell carcinoma versus adenocarcinoma and benign and malignant of lung cancer, but also are not too sensitive to the same type of lung cancer growth heterogeneity in different patients. Traditional methods of image classification just extract only one type of features for classification, so they are difficult to overcome these difficulties in medical image classification. For overcoming those difficulties of CT imaging classification of lung cancer, we extract five major categories of image features called radiomics features in [6], which are:

(1) Histogram features that are less sensitive to the size of the lesion (ROI) and shape-independent;

(2) GLCM and GLRLM based features which are less marginal impact and better clinical and pathological interpretation;

(3) Morphological features which capture the shape of the different parts of the lesion;

(4) GLSZM based features which are the best to explain the clinical pathology and can avoid the effects of growth direction in the same type of images at different patients.

Because different dimensionality reduction methods and different modelling methods will produce different results, to evaluate these models effectively, we select the logistic regression with L1 regularity model to classify the lung cancer CT images. The dynamic dimensionality reduction is achieved, and the final classification effect can also be effectively evaluated by the area under the ROC curve (AUC) and the accuracy (ACC).

2. Feature Extraction & Pathological Significance

We hope to extract some features that not only reflect the distribution of internal tissue within the ROI but also obtain some internal information, such as texture and heterogeneity of the ROI, and

can overcome many difficulties in the classification of the lung cancer CT images as described above. Therefore, we extract 385 features of 5 categories from lung cancer images and use them to describe the lesion area of lung cancer. They are Histogram features (42), Gray level co-occurrence matrix based features (144), Gray level run-length matrix based features (180), Morphological features (9), Gray level connectivity area size matrix based features (10).

2.1. Histogram features

Histogram is a kind of bar-like graph (see Figure 1) introduced by Karl Pearson [9] to describe the actual distribution of data. It counts the frequency of pixel values in each interval. Ed Sutton [10] first applied histogram theory to digital image processing.

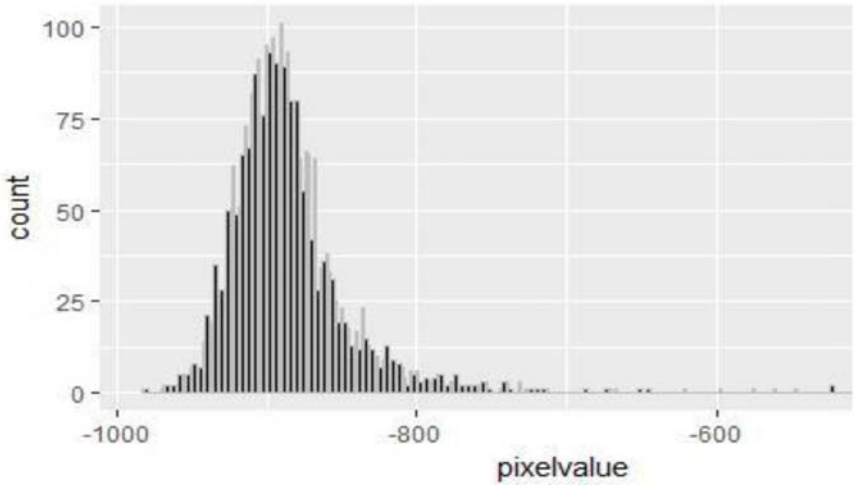


Figure 1. Histogram: A kind of bar-like graph.

Histogram features can reflect the overall distribution of the lesion area. In the two different types of lungs, the overall Gray distribution of lesions may be different, and the features of histograms are mostly based on features of probability, such as mean, variance, skewness, kurtosis, energy and entropy, etc. These features constructed by probability, so

when the number of pixels inside the ROI is larger than a certain value, it is not directly related to the absolute size of each ROI. Therefore, the histogram features can exclude the effect of inconsistency in the ROI size, and they have nothing to do with the shape of the ROI.

Histogram features also can directly reflect some basic information of the image, such as the brightness (mean), the degree of volatility or concentration of brightness (variance, kurtosis), offset of brightness distribution (skewness), complexity of the ROI (energy, entropy) and so on. Meanwhile, different intensities represent different tissues according to the imaging principle, so histogram features can directly reflect the overall distribution of human tissues inside the ROI.

2.2. Gray level co-occurrence matrix based features

Gray-level co-occurrence matrix (see Figure 2) based features is a type of feature proposed by Haralick [1] for image classification, and hence also called Haralick features, which compute the probability of a pixel pair occurring when the image is in a certain direction and step.

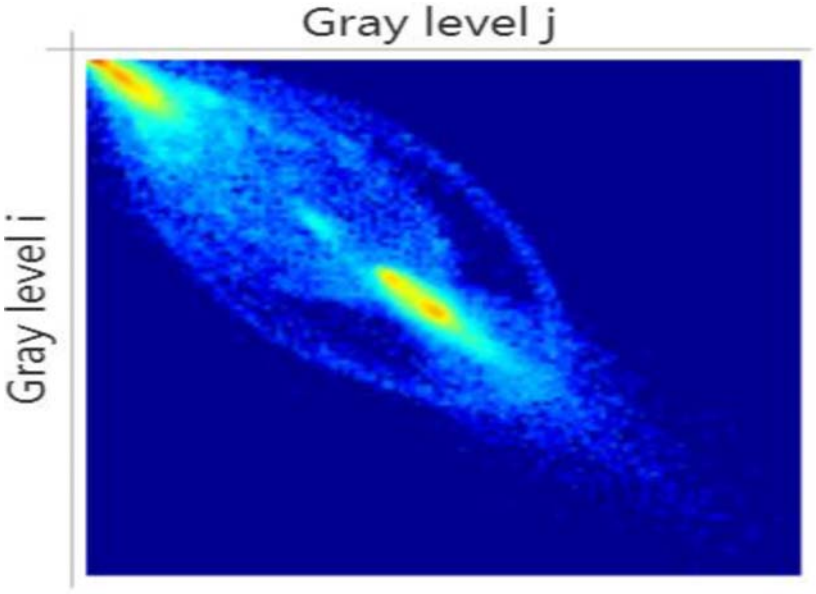


Figure 2. Gray level co-occurrence matrix map.

The Haralick features reflect the differences in texture and Gray level inside the ROI. When analysing two different types of lung cancer, the extracted Haralick features usually make a good distinction between the hierarchical differences of lung cancer tissues and the complexities of the texture, the direction of the texture, and the similarity. Because different clinicians do not have the same criteria for the same lesion (Region of Interest) or the same physician at different times, the margins of the ROI may be unclear, resulting in more or less unrelated areas being the lesion or some related area be excluded as a lesion, but the confirmation of the edge of the gray value of the difference is not too large, and the number of pixels will not be too much, then the ratio of the number of those increased or reduced pixels to the total will be small. Therefore, the Haralick features associated with the probability of pixel pairs are able to overcome the effects of edge ambiguity.

Because when the different tissues are scanned, the pixel values are different, while the probability of different pairs of pixels is exactly recorded by the co-occurrence matrix. Therefore, the Haralick feature can reflect the relationship between different human tissues inside the ROI. For example, when the heterogeneity of the ROI is very large, different organizations are separated by a certain distance, the relationship between different organizations is more complicated, the distribution of pixel pairs is more complicated, Haralick features can effectively measure the complexity between organizations.

2.3. Gray level run length matrix based features

Gray level run length matrix (GLRLM) (see Figure 2) based features is started with a study of coding methods called Run-length encoding [11], and Galloway [2] applied it to the texture analysis of images, which calculated the probability of the number of pixels that the pixel values were identical and connected together in one direction.

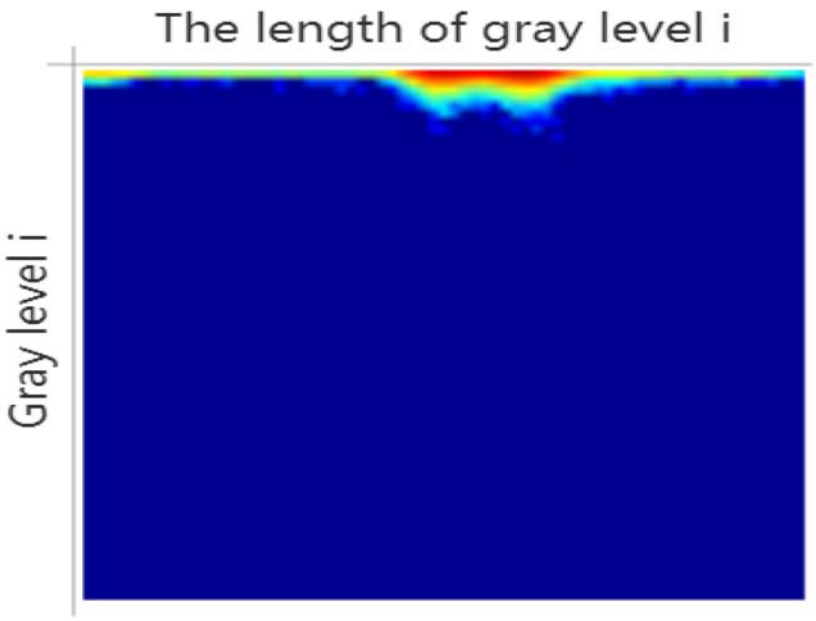


Figure 3. Gray level co-occurrence matrix map.

GLRLM features can express the different pathological texture characteristics. This type of feature usually reflects the difference between the two types of pathological images in the stripes, the depth of the texture gully and so on. Similar to the Haralick features, they are also a type of features related to probability, so its can also reduce the impact brought by unclear edges.

The body of the same organization after scanning, in the pixel value of image is the same, GLRLM happens to count the length of the tissue in a certain direction of the information, so this type of feature can measure the complexity of internal tissue within ROI. If there is a large heterogeneity within the ROI and a large difference between adjacent tissues, then many adjacent pixels in this direction will not be the same value. On the contrary, if the ROI is a benign lesion or normal tissues, then they may be the same organization between adjacent tissues in a certain direction; therefore, the GLRLM features can effectively quantify this heterogeneity.

2.4. Morphological features

Morphological features (MF), derived from a study called Form Factor Design [12], is a shape-based (see Figure 4) feature and hence also called a shape-base or morphological features that measure the spherical nature and compactness of an ROI with a ratio of the ROI's surface area to some combination of the volume.

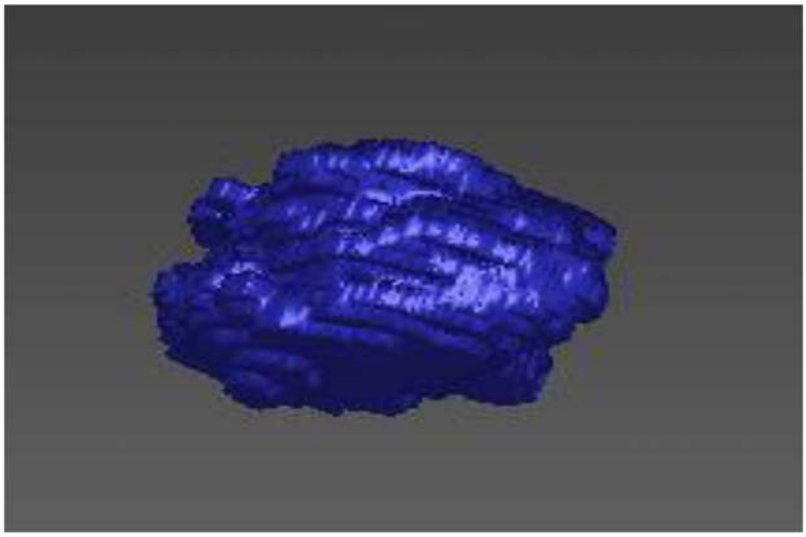


Figure 4. Morphology of a 3D ROI.

The MF can extract trait information such as the size, shape, distribution uniformity of the lesion area. If the two types of pathological images have different morphologies, such as compactness, spherical shape, etc., then this type of feature will play a very large role.

When considering image classification for lung cancer, we also want a class of features that specifically quantified the different shapes of the ROI, because different types of lung cancer may have different shape characteristics. For example, early cancers are usually spheroidal in size, and then they will grow wild in all directions after malignant transformation. The more malignant the tumors are, the weaker the shape and the worse the spherical shape will be, and so on.

2.5. Gray level size zone matrix based features

Gray level size zone matrix (see Figure 5) based features come from the study of the connected components of graphs in graph theory [13], and was introduced into image processing by Guillaume Thibault et al. for the classification of nuclei [14, 15] and DNA [16]. It counts the size of each organization of the human body inside the ROI.

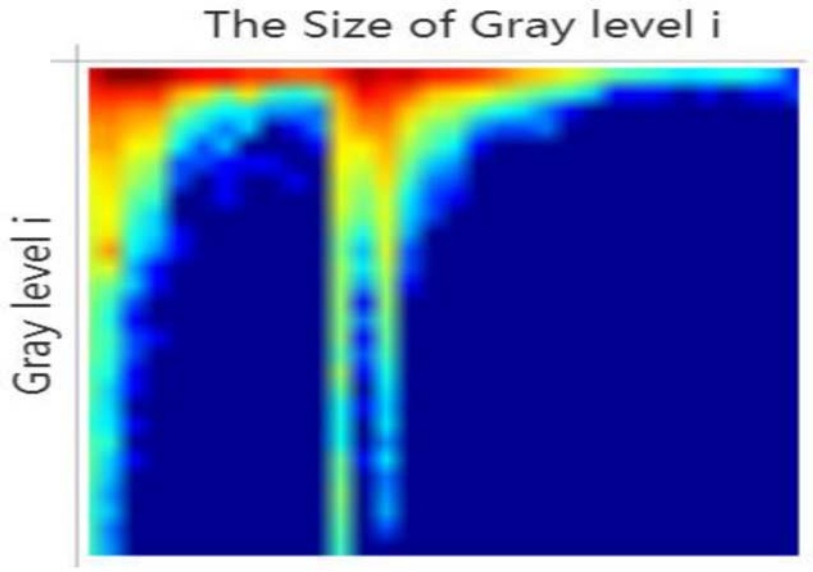


Figure 5. Gray level size zone matrix map.

The GLSZM feature allows the size of the internal tissue of the lesion area to be described. This type of feature can better reflect the structural differences within different organizations. Because there must be some structural differences between different lung cancer tissues, those features related to GLSZM may be useful.

The same human tissue shows the same pixel value on the image according to the imaging principle. Therefore, GLSZM is able to calculate the size of each tissue, more effectively describe the complexity of the tissues inside the ROI, and overcome the impact that the same pathological organization in different people's bodies grows in the different directions. If the internal heterogeneity of the ROI is very large

and there is a big difference between the adjacent tissues, then there will not be a lot of adjacent pixels in each small local area that will be the same value, so the value of the small area of GLSZM is large.

3. Data

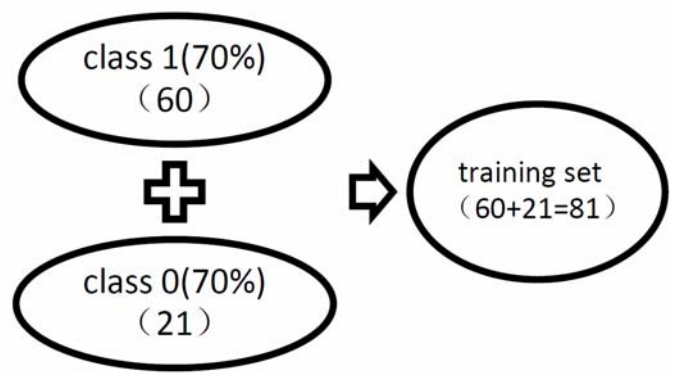
3.1. Datasets

3.1.1. Dataset (1): TCGA dataset [17]. We download 422 data on lung cancer from the TCGA public database, of which 117 are labelled (ROI and pathology), 87 are squamous cell carcinoma (labelled 1) and 30 are adenocarcinoma (labelled 0). We use it to make the classification of medical images of squamous cell carcinoma and adenocarcinoma. And the data is used to build the model.

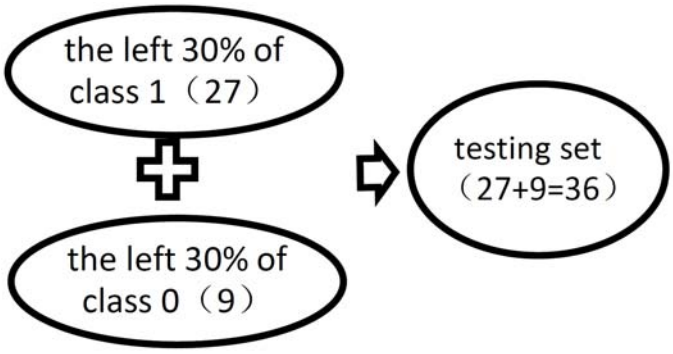
3.1.2. Dataset (2): Data from the First Affiliated Hospital of Guangzhou Medical University. A total of 222 cases of data, of which 141 cases of malignant lung cancer, 81 cases of benign lung tumors. We use it to make the benign and malignant classification of the image of the lungs. The data is used for the discussion section.

3.2. Data division

Stratified sampling: Because of the small sample size: only 117 cases in dataset 1, 87 cases in class 1 and 30 cases in class 0, and that the two types of samples are seriously unbalanced, which cannot meet the 1: 1 requirements between training set and testing set, we use stratified sampling, taking 70% (total 81 cases) from class 1 and class 0 respectively as training and the remaining 30% (total 36 cases in total) as external validation, as shown in Figure 6.



(a)



(b)

Figure 6. Data division (a) training set; (b) testing set.

4. Model Building

4.1. Logistic regression [18] with L1 regularity [19] model

Logistic regression is very useful for identifying classification problems [20, 21] and [22], and for the unbalance problem, the model can be evaluated using the AUC value [23, 24] under the ROC curve [25]. So we choose logistic regression as our model.

However, we have a total of 81 training data, and the number of features of each sample is 385. If we use logistic regression model directly will result in over-fitting (that is, performed well on the training set, but in the external validation set effect is poor), so we need to select the features (dimensionality reduction).

We adopt the L1 regularity method to generate sparse solutions [26] (that is, control the complexity of the logistic model so that the weights of many features are zero), which can not only achieve the purpose of dimensionality reduction, but also take the features possible relationships into account.

Therefore, we finally choose the logistic regression model with L1 regularity.

The realization of the model, we use linear model in scikit-learn, a python-based machine learning library [27].

4.2. Results

According to the training set, we plot the change of the coefficient of the logistic regression plus L1 regularity model with the control coefficient C of the regular item (Figure 7).

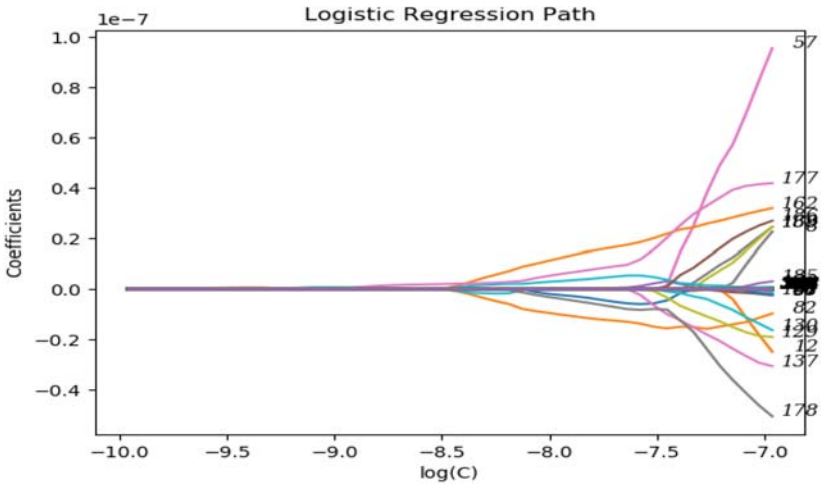


Figure 7. Logistic regression path of training set 1.

Then we test the set of external validation sets and draw the ROC curve with an accuracy of 0.75 and an AUC of 0.75308 (see Figure 8), which is a well performed model here according to the literature [23, 24] and [25].

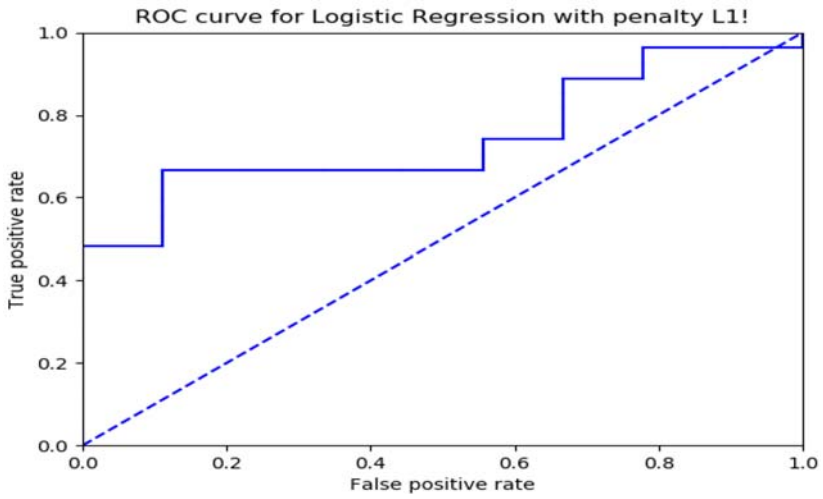


Figure 8. ROC curve of testing set in dataset 1.

5. Discussion

5.1. Different categories purpose

We perform the above steps using dataset 2 to analyse the effect of our extracted features on the classification of benign and malignant images of lung tumours. The accuracy of the model obtained on the external test set is ACC = 0.85 and AUC = 0.907, as shown in Figure 9, which shows that the features we extract are very effective for the classification of lung images.

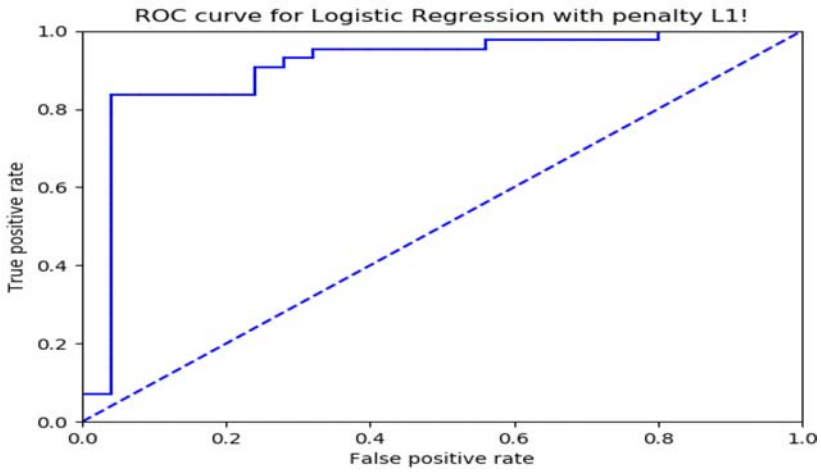


Figure 9. ROC curve of testing set in dataset 2.

5.2. The impact of segmentation

In the data we selected, there are many cases in which the ROI is not clearly defined, that may lead to bad value of features under the segmentation, because many parts of the ROI are non-diseased, as shown in Figure 10. However, judging from our results (ACC = 75%, AUC = 0.753), we obtain a very good taxonomic classification, indicating that the features we extracted are not too sensitive to segmentation.

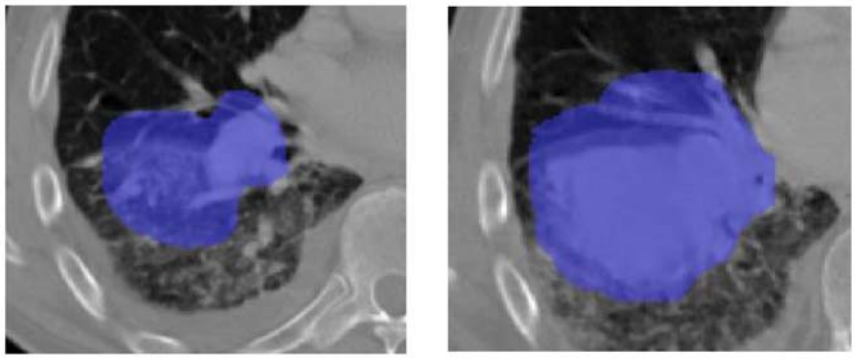


Figure 10. Many parts of the ROI are non-diseased.

5.3. Evaluation

For overcoming the difficulties in the classification of lung cancer CT images data, we extract five major type of features of the lung cancer CT images and analyse their differences in benign vs malignant lung cancer and squamous cell carcinoma vs adenocarcinoma, and try to contact the difference of tissues with the principle of imaging to explain their pathological significance, which is very useful to clinicians.

6. Conclusion

The experiments performed on the 5 major types of radiomics features, which is applied to the classification of lung cancer, shows us these radiomics features can basically figure out the differences between benign and malignant of lung tumors and it also have a good result on classifying the squamous cell carcinoma versus adenocarcinoma of lung cancer, although there are many difficulties in the classification of lung cancer CT images data, such as (1) the size and shape of lung cancer in different parts of the body are uneven; (2) the locations of the edge of the lung cancer confirmed by different doctors (or the same doctor at different times) are different; (3) the same type of lung cancer in different patients has different growth directions; and (4) it is difficult to judge the classification effect etc.

Acknowledgements

Special thanks to radiomics software AK provided by GE Healthcare and its feature explanation document, because all my figures in this paper and texture features table are got from AK; thanks to Dr. Hugo for providing case and pathology data.

References

[1] Robert M. Haralick, K. Shanmugam and Its'Hak Dinstein, Textural features for image classification, IEEE Transactions on Systems, Man, and Cybernetics, SMC-3(6) (1973), 610-621.

DOI: https://doi.org/10.1109/TSMC.1973.4309314

[2] M. M. Galloway, Texture analysis using gray level run lengths, Computer Graphics and Image Processing 4(2) (1975), 172-179.

DOI: https://doi.org/10.1016/S0146-664X(75)80008-6

[3] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60(2) (2004), 91-110.

DOI: https://doi.org/10.1023/B:VISI.0000029664.99615.94

[4] N. Dalal and B. Triggs, Histograms of oriented gradients for human detection, Computer Vision and Pattern Recognition 1 (2005), 886-893.

DOI: https://doi.org/10.1109/CVPR.2005.177

[5] Alex Krizhevsky, Ilya Sutskever and Geoffrey E. Hintonr, ImageNet classification with deep convolutional neural networks, Neural Information Processing Systems, Conference (2012), 1097-1105.

[6] Philippe Lambin, Emmanuel Rios-Velazquez, Ralph Leijenaar et al., Radiomics: Extracting more information from medical images using advanced feature analysis, European Journal of Cancer 48(4) (2012), 441-446.

DOI: https://doi.org/10.1016/j.ejca.2011.11.036

[7] V. Kumar, Y. Gu, S. Basu et al., Radiomics: The process and the challenges, Magnetic Resonance Imaging 30(9) (2012), 1234-1248.

DOI: https://doi.org/10.1016/j.mri.2012.06.010

[8] R. J. Gillies, P. E. Kinahan and H. Hricak, Radiomics: Images are more than pictures, They are data, Radiology 278(2) (2015), 563-577.

DOI: https://doi.org/10.1148/radiol.2015151169

[9] K. Pearson, Contributions to the mathematical theory of evolution, II: Skew variation in homogeneous material, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 186 (1895), 343-414.

DOI: https://doi.org/10.1098/rsta.1895.0010

[10] Ed Sutton, Histograms and the Zone System, Illustrated Photography.

[11] A. H. Robinson and C. Cherry, Results of a prototype television bandwidth compression scheme, Proceedings of the IEEE 55(3) (1967), 356-364.

DOI: https://doi.org/10.1109/PROC.1967.5493

[12] 'Wikipedia: Form factor'

https://en.wikipedia.org/wiki/Form_factor_(design).

[13] J. Hopcroft and R. Tarjan, Algorithm 447: Efficient algorithms for graph manipulation, Communications of the ACM 16(6) (1973), 372-378.

DOI: https://doi.org/1145/362248.362272

[14] Guillaume Thibault, Bernard Fertil, Claire Navarro et al., Texture indexes and gray level size zone matrix, Application to cell nuclei classification, Pattern Recognition and Information Processing (2009), 140-145.

[15] Guillaume Thibault, Bernard Fertil, Claire Navarro et al., Shape and texture indexes application to cell nuclei classification, International Journal of Pattern Recognition and Artificial Intelligence 27(1) (2013), 23 pages.

DOI: https://doi.org/10.1142/S0218001413570024

[16] Guillaume Thibault, Jesus Angulo and Fernand Meyer, Advanced statistical matrices for texture characterization: Application to DNA chromatin and microtubule network classification, IEEE International Conference on Image Processing (2011), 53-56.

[17] 'Wikipedia: NSCLC-Radiomics',

https://wiki.cancerimagingarchive.net/display/Public/NSCLC-Radiomics.

[18] David A. Freedman, Statistical Models: Theory and Practice, Cambridge University Press (2009), 128.

[19] Christopher M. Bishop, Pattern Recognition and Machine Learning (Corr. printing. ed.). New York: Springer, 2007.

[20] D. R. Cox, The regression analysis of binary sequences (with discussion), Journal of the Royal Statistical Society, Series B 20(2) (1958), 215-242.

[21] S. Biondo, E. Ramos, M. Deiros et al., Prognostic factors for mortality in left colonic peritonitis: A new scoring system, Journal of the American College of Surgeons 191(6) (2000), 635-642.

DOI: https://doi.org/10.1016/S1072-7515(00)00758-4

[22] D. W. Hosmer, T. Hosmer, S. Le Cessie and S. Lemeshow, A comparison of goodness-of-fit tests for the logistic regression model, Statistics in Medicine 16(9) (1997), 965-980.

[23] James A. Hanley and Barbara J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143(1) (1982), 29-36.

DOI: https://doi.org/10.1148/radiology.143.1.7063747

[24] Jorge M. Lobo, Alberto Jiménez-Valverde and Raimundo Real, AUC: A misleading measure of the performance of predictive distribution models, Global Ecology and Biogeography 17(2) (2008), 145-151.

DOI: https://doi.org/10.1111/j.1466-8238.2007.00358.x

[25] Tom Fawcett, An introduction to ROC analysis, Pattern Recognition Letters 27(8) (2006), 861-874.

DOI: https://doi.org/10.1016/j.patrec.2005.10.010

[26] B. K. Natarajan, Sparse approximate solutions to linear systems, SIAM Journal on Computing 24(2) (1995), 227-234.

DOI: https://doi.org/10.1137/S0097539792240406

[27] Scikit-Learn: Logistic Regression CV.

http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegressionCV.html

■