Predicting Lung Cancer Patients' Survival Time via Logistic Regression-based Models in a Quantitative Radiomic Framework

Shayesteh S. P.[1,2], Shiri I.[3], Karami A. H.[2], Hashemian R.[4], Kooranifar S.[5], Ghaznavi H.[6]*, Shakeri-Zadeh A.[2]

[1]PhD, Department of Physiology, Pharmacology and medical physics, Faculty of Medicine, Alborz University of Medical Sciences, Karaj. Iran
[2]PhD, Department of Medical Physics, School of Medicine, Iran University of Medical Sciences (IUMS), Tehran, Iran
[3]MSc, Department of Medical Physics, School of Medicine, Iran University of Medical Sciences (IUMS), Tehran, Iran
[4]MD, PhD, US oncology Inc, Cincinnati, OH, USA
[5]MD, Department of Pulmonary Sciences, Hazrat Rasoul Akram Hospital, Iran University of Medical Sciences (IUMS), Tehran, Iran
[6]MD, Zahedan University of Medical Sciences (ZaUMS), Zahedan, Iran

ABSTRACT

Background: Selection of the best treatment modalities for lung cancer depends on many factors, like survival time, which are usually determined by imaging.

Objectives: To predict the survival time of lung cancer patients using the advantages of both radiomics and logistic regression-based classification models.

Material and Methods: Fifty-nine patients with primary lung adenocarcinoma were included in this retrospective study and pre-treatment contrast-enhanced CT images were acquired. The patients lived more than 2 years were classified as the 'Alive' class and otherwise as the 'Dead' class. In our proposed quantitative radiomic framework, we first extracted the associated regions of each lung lesion from pre-treatment CT images for each patient via grow cut segmentation algorithm. Then, 40 radiomic features were extracted from the segmented lung lesions. In order to enhance the generalizability of the classification models, the mutual information-based feature selection method was applied to each feature vector. We investigated the performance of six logistic regression-based classification models.

Results: It was observed that the mutual information feature selection method can help the classifier to achieve better predictive results. In our study, the Logistic regression (LR) and Dual Coordinate Descent method for Logistic Regression (DCD-LR) models achieved the best results indicating that these classification models have strong potential for classifying the more important class (i.e., the 'Alive' class).

Conclusion: The proposed quantitative radiomic framework yielded promising results, which can guide physicians to make better and more precise decisions and increase the chance of treatment success.

Citation: Shayesteh SP, Shiri I, Karami AH, Hashemian R, Kooranifar S, Ghaznavi H, Shakeri-Zadeh A. Predicting Lung Cancer Patients' Survival Time via Logistic Regression-based Models in a Quantitative Radiomic Framework. *J Biomed Phys Eng.* 2020;10(4):479-492. doi: 10.31661/jbpe.v0i0.1027.

Keywords

Radiomics; Computed tomography; Lung Cancer; Survival analysis; Computed Tomography

Introduction

Cancer is one of the most common causes of death in the world [1]. Among all types of cancer, lung cancer has the highest mortality rate with 15% survival in 5 years [1]. There are 2 main types of lung cancer, namely non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC), between which ~85% of diagnosed lung cancers are related to NSCLCs [2]. Selection of the best treatment modalities for

*Corresponding author:
H. Ghaznavi
Zahedan University of Medical Sciences (ZaUMS), Zahedan, Iran
E-mail: ghaznavih@yahoo.com

Received: 25 September 2018
Accepted: 19 October 2018

lung cancer depends on many factors, including survival time, type, location and stage of cancer, which are usually determined by computed tomography (CT) and/or positron emission tomography (PET).

CT imaging plays a critical role in early diagnosis, stage prediction, and follow-up of the patients suffering from lung cancer [3]. The standard methods, considered for evaluation of the tumor response to a given treatment modality in bi-dimensional and uni-dimensional, are World Health Organization (WHO) and Response Evaluation Criteria in Solid Tumors (RECIST) guidelines, respectively [4,5]. These guidelines use some general factors, including shape, size and growth of tumor to quantify response evaluation. This quantification methods cannot reflect complexity of tumor behavior and predict therapeutic value of tumor treatment accurately [4,5]. In other hand, RECIST and WHO guidelines only can evaluate the tumor response after treatment process and cannot make any prediction about response to treatment.

Survival of the patients at different stages and response to treatment process strongly depend on the stage at the time of cancer diagnosis [6]. There are different types of biomarkers such as genomic, proteomic and metabolomic providing prognostic and diagnostic information for cancer therapeutic assessment in clinical practice [7-9]. The ability to select the best treatment modality for cancer patients is highly important and of clinical significance. Currently, there is no non-invasive biomarker to predict the survival time of cancer patients at different stages [10]. Radiomics focuses on quantitative analysis of medical images to better analyze the patients' conditions and also tries to help physicians make the best clinical decision [11-12]. With high throughput computing, we can extract quantitative information from tomographic images. Analysis of relationships between quantitative information and tumor response in a quantitative radiomic framework may increase the predictive value of medical images [13].

Although we believe that our study is the first attempt to systematically predict the survival time of patients by analyzing the performance of different logistic regression-based classification models, there exist only few studies in a similar vein. The first study is Hawkins *et al*. which compared the performance of four feature selection and classification methods [14]. Their cohort consisted of 40 patients, and the cutoff survival time was selected in a way that their training dataset was converted into a balanced one [14]. In real world scenarios, nonetheless, this assumption is not correct in the sense that many real world datasets are imbalanced, which makes the classification process more challenging.

More recently, Parmar *et al*., investigated the performance of different feature selection methods and classifiers [15]. In 2016, Hayano *et al.,* demonstrated that CT texture analysis provides favorable imaging biomarkers to predict the survival of patients with advanced NSCLC [16]. In addition, Dennie *et al*., reported that texture analysis of CT images has considerable sensitivity and specificity in obtaining prognostic information about patients with NSCLC [17]. In comparison to these studies, the methods in our research are different in many components of the quantitative radiomic framework, such as classifiers and segmentation algorithm. Moreover, in this study we investigated the performance of the predictors in a precise manner with respect to the importance of different probable error types which may occur in the classification process.

In the current study, the main purpose was to assess the value of CT image features combined with different logistic regression-based machine learning methods to predict the patients' survival time. To this end, we incorporated different logistic regression-based classification models into a quantitative radiomic framework. These classifiers are found to yield favorable outcomes in many supervised clas-

sification problems [18,19]. As a result, we experimentally evaluated their performance as well. We categorized the survival time of patients into two distinct groups with an acceptable cutoff time of 2 years similar to Parmar et al., and other studies [15,20]. In our study, the patients lived more than 2 years were labeled as 'Alive' and otherwise as 'Dead'. The implemented classification models via the extracted radiomic features were used to automatically predict these class labels for new patients. Consequently, our prediction approach can be used as a useful and non-invasive application in clinical oncology to improve the accuracy of treatment modality selection.

Material and Methods

To elaborate the various sections of the retrospective study, we made a schematic illustrati on for the proposed quantitative radiomic predictor system (Figure 1).

Patient CT images

Fifty-nine patients with primary lung adenocarcinoma, who were treated in the thoracic oncology program at the H. Lee Moffitt cancer center and research institute and the Maastricht radiation oncology clinic (MAASTRO), were included in this retrospective study [21]. For each patient, pretreatment contrast-enhanced CT scans were acquired at MAASTRO between years 2006 to 2009. Clinical data, including demographics, diagnosis, TNM (tumor, nodes and metastases status) stage, and patient survival were obtained for each patient [22]. Based on the clinically provided diagnostic data, 24 patients were labeled as 'Dead' and 35 as 'Alive' in this dataset. More details about the patients can be found in research carried out by Grove *et al*., [22].

Preprocessing of images

Because texture features are sensitive to different image acquisition which is unavoidable such as noise, the full intensities of each CT image were resampled to 64 gray levels. This process helps to efficiently and practically compute the texture features.

Segmentation method

Lung lesions were segmented using competitive region-growing based algorithm [23]. Grow cut algorithm uses a competitive region

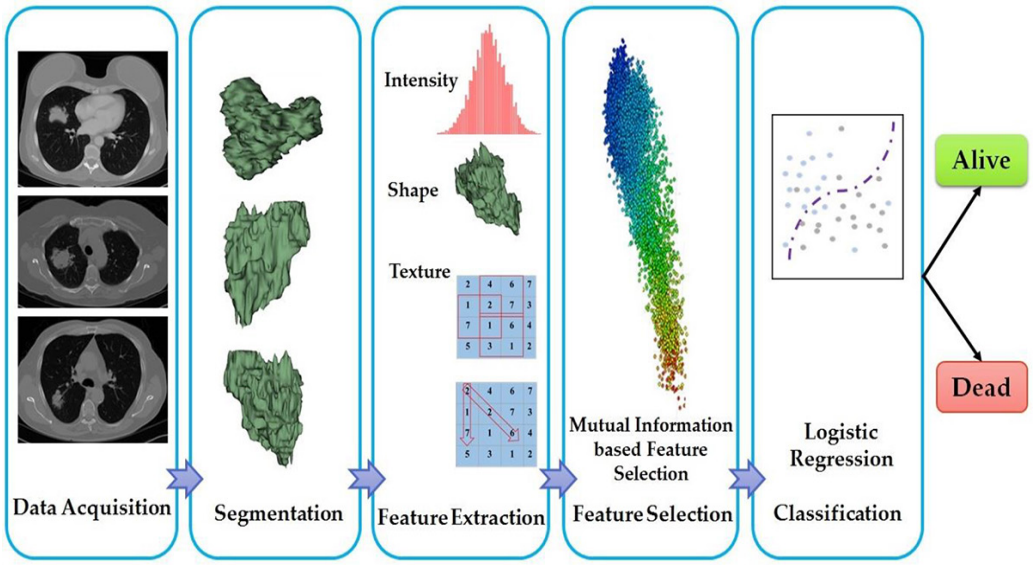

Figure 1: Overall quantitative radiomic framework for predicting survival time of lung cancer patients.

growing approach with given initial points and cellular automation [23]. All lung lesions were segmented by grow cut algorithm used in 3D-Slicer software.

Radiomic features extraction

Following the delineation and segmentation of the lesions, 40 quantitative radiomic features, including histogram-based, shape-based, and texture-based features were extracted from the 3D-tumor volume of each lesion. The histogram-based features roughly calculate the first order statistics of the tumor lesion's voxel intensities [10]. The tumor shape complexity descriptors quantitatively characterize lung adenocarcinomas via convexity and entropy ratio [22]. The texture features such as gray level co-occurrence matrices (GLCM), gray level run length matrices (GLRLM), neighboring gray-level dependence matrix (NGLDM), and gray-level zone length matrix (GLZLM), mathematically quantify the spatial positioning of intensities in the segmented regions of CT images [15,24]. These 40 quantitative radiomic features made up our feature vector, and thereby represented our CT images in the classification task.

Mutual information-based feature selection

Up to this section, we represented each patient's CT image by a set of 40 quantitative radiomic features. In order to obtain more accurate results from the model, it needs to get a few features which have the highest relationship with the class label. In fact, feature selection methods aims to rank the extracted features according to their discrimination power, and select their subset with the highest discrimination power [25].

The techniques of correlation feature selection based on conventional statistics can detect only linear relationships. However, mutual information can detect nonlinear relationships as well as linear ones [25]. Having considered these reasons, we applied the mutual information feature selection method on our dataset. Mutual information measures, from the information theory viewpoint, how much information about a feature contains about the class. The mutual information of two continuous random variables X and Y is defined as

$$I(x,y) \equiv D_{KL}(p(x,y)\|p(x)p(y)) = \int_{Y}\int_{X} p(x,y)\log\frac{(p(x,y)}{p(x)p(y)}dxdy$$

Where D_{KL} is the *Kullback-Leibler* (KL) divergence, p(x,y) is the joint probability density function (PDF) of X and Y, and p(x) and p(y) are the marginal probability density functions of X and Y, respectively [26]. In other words, mutual information tells us how much p(x,y) is different from a hypothetical PDF with independent X and Y via the KL divergence. KL divergence is a measure of the difference between two PDFs [26]. From the properties of KL divergence, we can find out that I(x,y)≥0, and that if a feature's distribution is the same in a class as it is in the other classes, then I(x,y)=0. Mutual information also reaches its highest value if the feature is a perfect indicator of class membership. It occurs if and only if the feature is present in samples belonging to only one class [26].

We need to compute the overall mutual information using the mutual information function of feature x_i of different classes. To this end, we define $I(x_i,y_i)$ as the mutual information between feature x_i and a particular class y_i. In order to compute the overall mutual information, we use the average values of $I(x_i,y_i)$ over all the existing classes (*i.e.*, 'Alive' and 'Dead'). Finally, we select the subset of features with the maximum value of overall mutual information score.

Lastly, before applying a classifier to our selected features, the features will be zero-centered, which is common in machine learning. More precisely, to accelerate the convergence rate of the model, and also prevent some features from dominating the other features, the average of each feature is computed and then every feature's value is subtracted from their respective means.

Logistic regression-based classifiers

In this section, we introduce six different logistic regression-based classifiers in a common framework. The framework has two major components: 1) the linear score function which converts the selected features to class scores, and 2) the objective function that measures the extent to which the predicted labels match the clinical annotated labels (*i.e.*, ground truth labels). Next, in order to find the optimum parameters of our model, we consider the problem as an optimization one which we overcome via adopting a *maximum a posteriori* (MAP) estimation [27].

The linear score function maps the values of selected radiomic features onto our two classes' confidence scores using a matrix multiplication [27], as follows:

$$S = f\left(x^{(i)}; W\right) = Wx^{(i)}$$

Where the matrix W is the weight matrix, and the bias vector is eliminated in order to simplify our notation (*i.e.*, we use the bias trick in equation (3)). It means that we extend $X^{(i)}$ by one extra dimension with the value of 1 (*i.e.*, the default bias dimension). As a result, the size of $X^{(i)}$ equals to $[(d'+1)\times 1]$ and W is a matrix with the size of $[2\times(d'+1)]$. The result of the score function S is a vector with the size of 2×1, which we denote as $S=[S_1 \quad S_2]^T$. It means that the model's confidence score for the i^{th} class is the i^{th} element of the result vector (*i.e.*,S_i).

The second major component is the objective function (or cost function). The objective function will be high if we predict the truth class label for most of the training samples; conversely, it will be low if the predicted confidence scores are different from the ground truth [27]. The logistic regression-based models treat the computed confidence scores as the unnormalized log probabilities of each class. Using the softmax function, which normalizes these probabilities between 0 and 1, the objective function for the i^{th} sample can be defined as [27]

$$J_i = logP\left(y^i x^i; W\right) = \log\left(\frac{\exp\left(S_{y^{(i)}}\right)}{\sum_{k=1}^{2}\exp(S_K)}\right) = S_{y^{(i)}} - \log\left(\sum_{k=1}^{2}\exp(S_k)\right)$$

The total objective function consists of two parts: 1) the mean of J_i for all training samples, and 2) the regularization term. More precisely, the first part computes the mean of the values of objective functions for all training samples indicating the data objective. The second part (*i.e.*, the regularization term/penalty) prevents the overfitting problem by enforcing some constraints on the weight matrix W. Consequently, the total objective function, which measures the quality of the classification, is defined as:

$$J(W) = \frac{1}{n}\sum_{j=1}^{n} J_i - R(W) = J(W) - R(W)$$

Where $J(W)$ is data objective and $R(W)$ indicates regularization term [27]. Now, by subtracting the regularization term from the total objective function, we can estimate the weight matrix W via the MAP estimation as follows:

$$W_{MAP} = arg_w \max J(W) = arg_w \max[J(W) - R(W)]$$

In fact, according to the above-mentioned estimation, we can interpret the regularization term $R(W)$ as some priors on the weight matrix. This MAP estimation can be achieved using different optimization algorithms (*e.g.*, Newton's method, which is known as *iteratively reweighted least squares* (IRLS) in the literature, *stochastic gradient descent* (SGD), etc.) [27]. If one uses the IRLS method to accomplish the MAP estimation, then in the machine learning context, the overall method is called logistic regression, which we denote as LR in our experiments. If the number of class labels is more than two, the generalized logistic regression is called multinomial logistic regression (or softmax regression) [27].

Due to the success of the coordinate descent optimization method in solving the dual form of linear models (*e.g.*, support vector machines (SVM)), Yu *et al.*, presented the dual coordinate descent method for logistic regres-

sion (DCD-LR) [28]. They formulated the dual form of logistic regression model (instead of its primal form), and applied the coordinate descent optimization method to solve it [28]. The DCD-LR is the second model, which we used in our experiments in this paper.

The next four models used in this study are different from one another in terms of their priors on regularization term. These priors include Gaussian, Laplacian, Cauchy, and Uniform. For brevity, we denote these methods as G-LR, L-LR, C-LR, and U-LR, respectively. For some of these priors, the regularization term is non-differentiable, and as a result, the IRLS method cannot easily handle these priors [29]. Carpenter used the (online) SGD method for optimization [29]. In our implementation, the online SGD method uses just one training sample to update weights in each iteration.

Experimental setup

In our experiments, we used the 10-fold cross-validation strategy on our dataset. It means that we used 90% of our dataset as a training set and the rest of it (*i.e.*, the remaining 10%) as a testing set. It is important to note that all of the results in this study were obtained by this strategy. The mutual information-based feature selection and all of the six classifiers were implemented in Java and run on a PC with Intel Core2Duo 2.53GHz and 4GB of RAM.

As stated above, predicting the survival time of lung cancer patients via CT images involves a number of inter-related phenomena, each of which affects the overall performance in a particular way. However, since this issue (*i.e.*, predicting the survival time of lung cancer patients) is quite new in the medical literature, there exists only few published papers about it. Most of these papers rarely report results on the same evaluation measures. As a result, making comparison between different models would be stifling.

Using the Machine Learning viewpoint we can analyze our problem as a binary classification task. Therefore, well-known evaluation measures in Machine Learning can be used here. As a result, first of all, in order to objectively evaluate the characteristics of our models, we used the confusion matrices of the models. The *confusion matrix* is a visual table that indicates the number of correct and incorrect predictions made by the classification model in comparison to the actual ground truths in the test data [30]. In fact, using confusion matrix, we can determine the performance of a classifier via quantitative measures.

In the confusion matrix, each column shows the number of test samples which the model predicted their class labels while each row shows the number of test samples according to the ground truth labels [30]. In the confusion matrix, each entry has a specific interpretation in the context of our study, which is defined as follows [30]:

- True positive (TP): the number of test samples, class labels of which are predicted by the classification model as 'Alive' correctly.
- True negative (TN): the number of test samples, class labels of which are predicted by the classification model as 'Dead' correctly.
- False positive (FP): the number of test samples, class labels of which are predicted by the classification model as 'Alive' incorrectly; however, their actual class label is 'Dead'.
- False negative (FN): the number of test samples, class labels of which are predicted by the classification model as 'Dead' incorrectly; however, their actual class label is 'Alive'.

According to these definitions, two fundamental evaluation measures can be defined as:

$$sensitivity = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

The *sensitivity* (also known as *recall* or *true positive rate* (TPR)) is the fraction of 'Alive' samples that were correctly classified [30]. In fact, sensitivity measures how much of the actual 'Alive' class is predicted correctly by the

classification model. Moreover, *precision* is the fraction of the 'Alive' samples predicted correctly. It means that in contrast to sensitivity, precision measures how much of the predicted 'Alive' class is the same as the ground truth. Both of the sensitivity and precision can take values between 0 and 1. If the value is higher, the classification will have the better performance [30].

In some papers [30], the specificity measure is also used for evaluation, which is defined as:

$$Specificity = \frac{TN}{TN + FP}$$

However, this is not considered as an important evaluation measure in this study because we have two error types as follows:

1) Error type 1: in this type of error, the actual class label is 'Alive' but the classification model incorrectly predicts 'Dead'.

2) Error type 2: in this type of error, the actual class label is 'Dead' but the classification model incorrectly predicts 'Alive'.

Based on analyzing these two error types, we can infer that error type 1 is more critical and dangerous than the second one. If error type 1 occurs, it means that the patient will actually be alive (more than 2 years), that we can start the treatment process, and that we expect that the patient will most probably be cured. However, since the classification model incorrectly predicts its class label as 'Dead', we do not start the treatment process (as in this case we mistakenly think that the treatment process would not be effective) and as a result a human's life will be lost. On the other hand, error type 2 will only result in financial loss. Therefore, we can conclude that error type 1 is much more critical than the second one, and that the 'Alive' class is more important. Since the specificity measure focuses on the 'Dead' class, its value is not very important for our experiments. Moreover, in the next section we will show that our best models have less error type 1.

Another useful evaluation measure is accuracy, which specifies the fraction of the total number of classification model's correct predictions [30], and is defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

Although accuracy is a common, widely-used and well-known evaluation measure, it solely is not a reliable and adequate evaluation metric. The reason is that when the dataset is imbalanced (*i.e.*, there is a huge difference between the number of samples in different classes), it may yield unreliable results. On the other hand, the extracted evaluation measure from the confusion matrix allows us to propose better metrics, which can overcome this problem. We know that a perfect classification model requires both high sensitivity and precision values [30]. Thus, the F1 score or harmonic mean of the sensitivity and precision used to determine how well the classification model predicts the ground truth labels is defined as [30]:

$$F1score = \frac{2 \times sentivity \times precision}{sentivity + precision}$$

Finally, in situations where the dataset is balanced, one can use both the accuracy and F1 score to quantify the overall performance of a classifier. If the accuracy and F1 score get higher there will be better match between ground truth and the predicted labels. Since in our dataset the difference between the number of samples in the two different classes is not much (*i.e.*, 11 patients), we used both accuracy and F1 score for a better evaluation. In the next section, a more detailed analysis of the classification models is presented.

Ethics

As stated before, here we used a free dataset published earlier by Clark *et al*., in which the study was conducted in accordance with the Declaration of Helsinki and applicable local regulations [21]. Patients with primary lung adenocarcinoma, who were treated in the thoracic oncology program at the H. Lee Moffitt

cancer center and research institute and the Maastricht radiation oncology clinic (MAASTRO), were included in Clark and co-worker study [21].

Results

In this section, at first, the analysis of classifiers' performance against the number of selected features is discussed. At second, a comprehensive comparison among the six different classification models is presented.

In our framework, the number of selected features via mutual information criterion is one of the key parameters which drastically affects the overall performance of the models. To study the dependency of our classification models on the number of selected features, experiments were conducted on the dataset (using 10-fold cross-validation), varying the number of the selected features in an interval of 2 to 14. In our experiments, we observed that if we increased the number of selected features over 14, the performance of the models would be stable and without any change. The obtained results are shown in Figures 2 and 3.

In Figures 2 and 3, the performance of the models in terms of accuracy and F1 score are respectively plotted against the number of selected features to find out which logistic regression-based classification model achieves the highest F1 score or accuracy for a specific number of selected features. It is worth mentioning that due to the random nature of the classification models, we reported the average results of 30 runs per model. With respect to the plots in Figures 2 and 3, we observed that increasing the number of selected features would result in the improvement of the overall performance of the G-LR, C-LR, L-LR, and U-LR models, which enforce a prior function on the regularization term. On the contrary, the performance of the LR and DCD-LR models will decrease about 5% as the number of selected features increases. By and large, the best results are obtained via the LR and DCD-LR

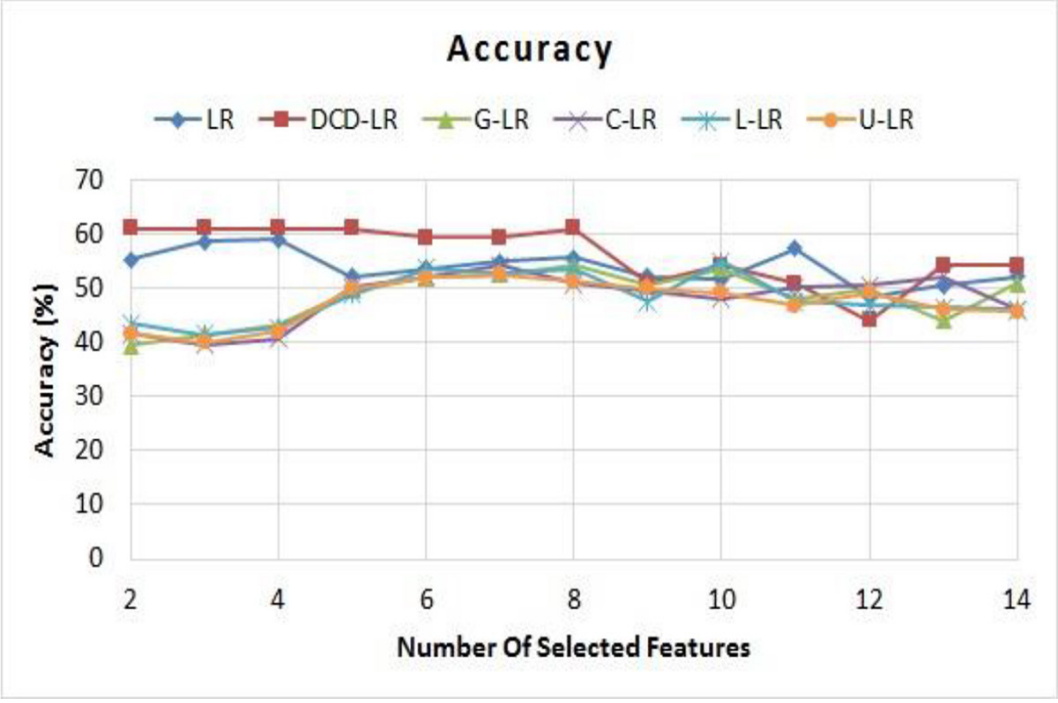

Figure 2: Plot of accuracy measure against the number of selected features.

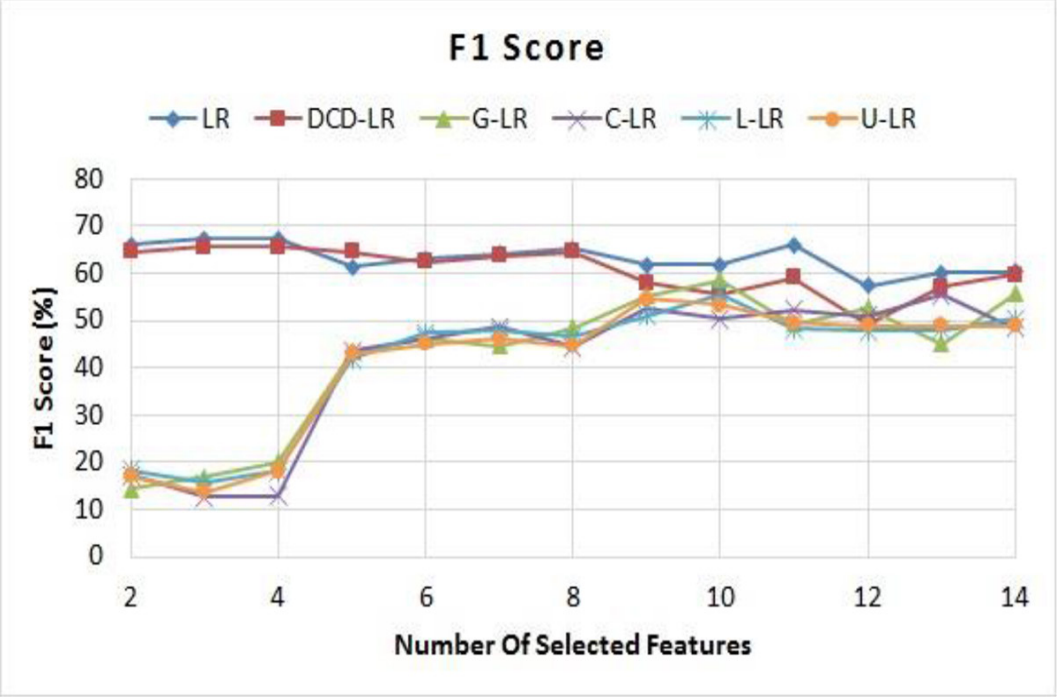

Figure 3: Plot of F1 score against the number of selected features.

models with 4 selected features. This indicates that the mutual information feature selection method can help the classifier to achieve better predictive results. A comprehensive comparison among the best results of the classification models in terms of accuracy and F1 score is illustrated in Figure 4.

As seen in Figure 4, the DCD-LR and LR models achieved the best results compared to the other models in terms of accuracy and F1 score, respectively. In the experiments, the DCD-LR model obtained an average accuracy of 61.02% and 65.67% F1 score using its coordinate descent optimization method to optimize the dual form of the objective function. It indicates that similar to other classification tasks [18,28], this robust model (*i.e.*, DCD-LR) can lead to better results. In addition, the LR model obtained the best average F1 score compared to the other models. From Figure 4, it is evident that the G-LR, L-LR, C-LR, and U-LR models, which enforce a prior function on the regularization term, yielded worse results than LR and DCD-LR did. In spite of the fact that enforcing a prior function on the regularization term generally prevents the overfitting problem and enhances the generalization ability in many cases, in some cases (*e.g.*, when the size of the test dataset is small or the number of the selected features is scant) it cannot significantly improve the classification results. This justifies the results of the four models obtained in this study. By and large, in our study the LR and DCD-LR models achieved the best results, indicating that these classification models have a strong potential for predicting the survival time of lung cancer patients in the quantitative radiomic framework.

In order to better explain the behavior of the LR and DCD-LR models, two examples of the results obtained by these models are presented in Figure 5. In these examples, the LR model totally achieved better results in terms of F1 score, accuracy, and sensitivity. However, due to the fact that the DCD-LR model has less FP (*i.e.*, it has the stronger

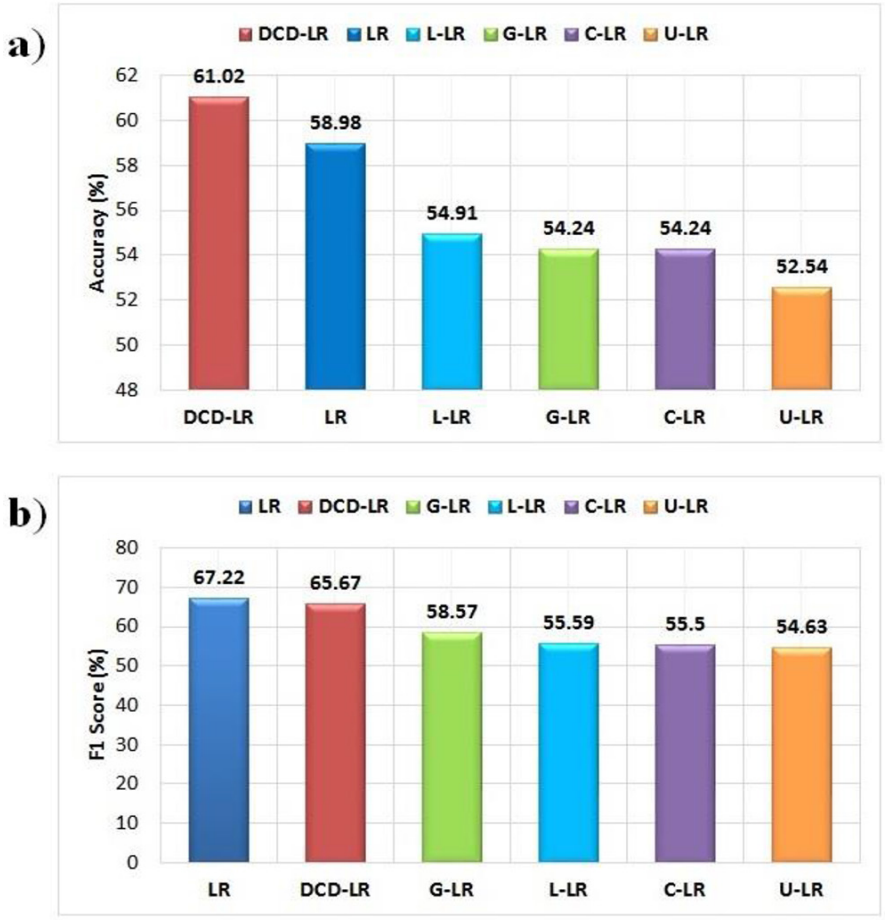

Figure 4: Comparison of performance of six classification models in terms of: a) accuracy and b) F1 score.

Confusion Matrix		Predicted	
		Alive	Dead
Actual	Alive	TP = 28	FN = 7
	Dead	FP = 14	TN = 10

Results of LR Model (with 4 selected features):

Sensitivity = 80%
Precision = 66.67%
Specificity = 41.67%
Accuracy = 64.41%
F1 Score = 72.73%

(a)

Confusion Matrix		Predicted	
		Alive	Dead
Actual	Alive	TP = 23	FN = 12
	Dead	FP = 10	TN = 14

Results of DCD-LR Model (with 4 selected features):

Sensitivity = 65.71%
Precision = 69.70%
Specificity = 58.33%
Accuracy = 62.71%
F1 Score = 67.65%

(b)

Figure 5: Two examples of the results: (a) LR classification model and (b) DCD-LR classification model.

ability to classify the 'Dead' class than the LR model), it achieved a better performance in terms of specificity and precision (see Figure 5). As mentioned in experimental setup section, in our framework error type 1 is much more important than the second error type, thus the value of the specificity is not a very important evaluation measure to quantify the overall performance of the classifier. The low value of the specificity of the LR model (*i.e.*, 41.67%) does not significantly undermine its classification performance. Moreover, Figure 5 indicates that both of the LR and DCD-LR models have an acceptable accuracy in classifying the more important class label (*i.e.*, 'Alive' class). As a result, we can conclude that these models, which have obtained better results than the other logistic regression-based models, can better handle error type 1 and achieve high sensitivity values.

Discussion

Predicting the survival time of cancer patients is a critical factor in opting for an appropriate medical treatment. Machine learning methods and medical imaging provide a powerful radiomic framework which yields a quantitative measure of patients' biomarkers in an attempt to predict the survival time of those patients. In this study, we investigated the performance of six different logistic regression-based classification models for predicting the survival time of lung cancer patients. The features used for classification were extracted from the CT images of patients. To achieve better results and enhance the generalization ability of the classifiers, the mutual information feature selection method was also employed.

Since radiomics is a newly emerging filed in medical imaging, very few studies have been carried out in this regard. To the best of our knowledge, there exist only two studies which similarly aimed at predicting the survival time of lung cancer patients via machine learning methods [14,15]. Both of these studies, authored by Hawkins *et al*., and Parmar *et al*., evaluated the performance of different feature selection and classification methods [14,15]. Nevertheless, our methods are totally different from the previous ones. More specifically, the current study focuses on the investigation of the performance of six different logistic regression-based classification models in a quantitative radiomic framework. Moreover, a systematic analysis of the evaluation measures is presented, which considers the varying importance of probable error types according to the proposed radiomic framework. Our analysis demonstrated that one can summarize the results regarding the survival time prediction using F1 score and accuracy only, between which the former is even more precise. Finally, based on different error types, we concluded that sensitivity is a critical evaluation measure in this regard (see experimental setup section).

Experimental results showed that the use of the mutual information feature selection method can increase the accuracy of the models. We also discovered that there was a negative correlation between the performance of the LR and DCD-LR models and the number of selected features. That is, the performance of the LR and DCD-LR models decreased as the number of selected features increased. In contrast, the other four models (*i.e.*, G-LR, L-LR, C-LR, and U-LR) showed a different behavior in the sense that by increasing the number of selected features, their performance also increased. The DCD-LR and LR models, which achieved the best results with 4 selected features and over the 10-fold cross-validation strategy on the dataset, obtained an overall accuracies of 61.02% and 58.98% with F1 scores of 65.67% and 67.22%, respectively. Furthermore, we have shown that these models have an acceptable performance of classifying the more important class (*i.e.*, the 'Alive' class), which results in less error type 1. Finally, we can conclude that the presented methods have a great potential to predict the survival time of lung cancer patients in a quantitative radiomic framework. This yields a type of computer-

aided diagnosis (CAD) system which can assist physicians by offering reliable and accurate treatment decisions.

Conclusion

The implemented methods in this paper have mainly focused on predicting the survival time of lung cancer patients via extracted features form their CT images. Future research can extend these methods to other types of cancer (*e.g.*, prostate cancer, breast cancer, brain cancer, etc.). Moreover, from the viewpoint of prediction, one can employ the radiomic framework to analyze the type as well as stage of tumors. Another interesting avenue for future researchers is the combination of radiomic and genomic features to improve the performance of predictors. By and large, we can conclude that as predicting the survival time of lung cancer patients plays a significant role in the selection of appropriate treatment strategies, we believe that the proposed quantitative radiomic framework is an effective and promising approach in this regard.

Acknowledgment

The authors sincerely thank the Zahedan University Medical Sciences and Imam Khomeni hospital for their collaboration and facilities.

Conflict of Interest

None

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2016. *CA Cancer J Clin.* 2016;**66**:7-30. doi: 10.3322/caac.21332. PubMed PMID: 26742998.
2. Torre LA, Siegel RL, Jemal A. Lung Cancer Statistics. *Adv Exp Med Biol.* 2016;**893**:1-19. doi: 10.1007/978-3-319-24223-1_1. PubMed PMID: 26667336.
3. Bach PB, Mirkin JN, Oliver TK, Azzoli CG, Berry DA, Brawley OW, et al. Benefits and harms of CT screening for lung cancer: a systematic review. *JAMA.* 2012;**307**:2418-29. doi: 10.1001/jama.2012.5521. PubMed PMID: 22610500. PubMed PMCID: PMC3709596.
4. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer.* 2009;**45**:228-47. doi: 10.1016/j.ejca.2008.10.026. PubMed PMID: 19097774.
5. Nishino M, Jackman DM, Hatabu H, Yeap BY, Cioffredi LA, Yap JT, et al. New Response Evaluation Criteria in Solid Tumors (RECIST) guidelines for advanced non-small cell lung cancer: comparison with original RECIST and impact on assessment of tumor response to targeted therapy. *AJR Am J Roentgenol.* 2010;**195**:W221-8. doi: 10.2214/AJR.09.3928. PubMed PMID: 20729419. PubMed PMCID: PMC3130298.
6. DeSantis CE, Lin CC, Mariotto AB, Siegel RL, Stein KD, Kramer JL, et al. Cancer treatment and survivorship statistics, 2014. *CA Cancer J Clin.* 2014;64:252-71. doi: 10.3322/caac.21235. PubMed PMID: 24890451.
7. Ginsburg GS, Willard HF. Genomic and personalized medicine: foundations and applications. *Transl Res.* 2009;**154**:277-87. doi: 10.1016/j.trsl.2009.09.005. PubMed PMID: 19931193.
8. Yuan Y, Van Allen EM, Omberg L, Wagle N, Amin-Mansour A, Sokolov A, et al. Assessing the clinical utility of cancer genomic and proteomic data across tumor types. *Nat Biotechnol.* 2014;**32**:644-52. doi: 10.1038/nbt.2940. PubMed PMID: 24952901. PubMed PMCID: PMC4102885.
9. Spratlin JL, Serkova NJ, Eckhardt SG. Clinical applications of metabolomics in oncology: a review. *Clin Cancer Res.* 2009;**15**:431-40. doi: 10.1158/1078-0432.CCR-08-1059. PubMed PMID: 19147747. PubMed PMCID: PMC2676437.
10. Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun.* 2014;**5**:4006. doi: 10.1038/ncomms5006. PubMed PMID: 24892406. PubMed PMCID: PMC4059926.
11. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, Van Stiphout RG, Granton P, et

al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer*. 2012;**48**:441-6. doi: 10.1016/j.ejca.2011.11.036. PubMed PMID: 22257792. PubMed PMCID: PMC4533986.

12. Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA, Schabath MB, et al. Radiomics: the process and the challenges. *Magn Reson Imaging*. 2012;**30**:1234-48. doi: 10.1016/j.mri.2012.06.010. PubMed PMID: 22898692. PubMed PMCID: PMC3563280.
13. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology.* 2016;**278**:563-77. doi: 10.1148/radiol.2015151169. PubMed PMID: 26579733.PubMed PMCID: PMC4734157.
14. Hawkins SH, Korecki JN, Balagurunathan Y, Gu Y, Kumar V, Basu S, et al. Predicting outcomes of nonsmall cell lung cancer using CT image features. *IEEE access.* 2014;**2**:1418-26.doi: 10.1109/access.2014.2373335.
15. Parmar C, Grossmann P, Bussink J, Lambin P, Aerts HJ. Machine Learning methods for Quantitative Radiomic Biomarkers. *Sci Rep*. 2015;**5**:13087. doi: 10.1038/srep13087. PubMed PMID: 26278466. PubMed PMCID: PMC4538374.
16. Hayano K, Kulkarni NM, Duda DG, Heist RS, Sahani DV. Exploration of Imaging Biomarkers for Predicting Survival of Patients With Advanced Non-Small Cell Lung Cancer Treated With Antiangiogenic Chemotherapy. *AJR Am J Roentgenol.* 2016;**206**:987-93. doi: 10.2214/AJR.15.15528. PubMed PMID: 26934729.
17. Dennie C, Thornhill R, Sethi-Virmani V, Souza CA, Bayanati H, Gupta A, et al. Role of quantitative computed tomography texture analysis in the differentiation of primary lung cancer and granulomatous nodules. *Quant Imaging Med Surg.* 2016;**6**:6-15. doi: 10.3978/j.issn.2223-4292.2016.02.01. PubMed PMID: 26981450. PubMed PMCID: PMC4775240.
18. Hsieh CJ, Yu HF, Dhillon IS. PASSCoDe: Parallel ASynchronous Stochastic dual Co-ordinate Descent. *ICML.* 2015;**15**:2370-9.
19. Dikaios N, Alkalbani J, Abd-Alazeez M, Sidhu HS, Kirkham A, Ahmed HU, et al. Zone-specific logistic regression models improve classification of prostate cancer on multi-parametric MRI. *Eur Radiol.* 2015;**25**:2727-37. doi: 10.1007/s00330-015-3636-0. PubMed PMID: 25680730.
20. Oberije C, Nalbantov G, Dekker A, Boersma L, Borger J, Reymen B, et al. A prospective study comparing the predictions of doctors versus models for treatment outcome of lung cancer patients: a step toward individualized care and shared decision making. *Radiother Oncol.* 2014;**112**:37-43. doi: 10.1016/j.radonc.2014.04.012. PubMed PMID: 24846083. PubMed PMCID: PMC4886657.
21. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J Digit Imaging*. 2013;**26**:1045-57. doi: 10.1007/s10278-013-9622-7. PubMed PMID: 23884657. PubMed PMCID: PMC3824915.
22. Grove O, Berglund AE, Schabath MB, Aerts HJ, Dekker A, Wang H, et al. Quantitative computed tomographic descriptors associate tumor shape complexity and intratumor heterogeneity with prognosis in lung adenocarcinoma. *PLoS One.* 2015;**10**:e0118261. doi: 10.1371/journal.pone.0118261. PubMed PMID: 25739030. PubMed PMCID: PMC4349806.
23. Velazquez ER, Parmar C, Jermoumi M, Mak RH, Van Baardwijk A, Fennessy FM, et al. Volumetric CT-based segmentation of NSCLC using 3D-Slicer. *Sci Rep*. 2013;**3**:3529. doi: 10.1038/srep03529. PubMed PMID: 24346241. PubMed PMCID: PMC3866632.
24. Panth KM, Leijenaar RT, Carvalho S, Lieuwes NG, Yaromina A, Dubois L, et al. Is there a causal relationship between genetic changes and radiomics-based image features? An in vivo preclinical experiment with doxycycline inducible GADD34 tumor cells. *Radiother Oncol.* 2015;**116**:462-6. doi: 10.1016/j.radonc.2015.06.013. PubMed PMID: 26163091.
25. Bennasar M, Hicks Y, Setchi R. Feature selection using joint mutual information maximisation. *Expert Systems with Applications*. 2015;**42**:8520-32. doi: 10.1016/j.eswa.2015.07.007.
26. Torkkola K. Feature extraction by non-parametric mutual information maximiza-

tion. *Journal of machine learning research.* 2003;**3**:1415-38.

27. Li J, Bioucas-Dias JM, Plaza A. Semisupervised hyperspectral image segmentation using multinomial logistic regression with active learning. *IEEE Transactions on Geoscience and Remote Sensing*. 2010;**48**:4085-98. doi: 10.1109/tgrs.2010.2060550.

28. Yu HF, Huang FL, Lin CJ. Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning.* 2011;**85**:41-75. doi: 10.1007/s10994-010-5221-8.

29. Carpenter B. Lazy sparse stochastic gradient descent for regularized multinomial logistic regression. Alias-i, Inc, Tech Rep. 2008:1-20.

30. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. *Inf Process Manag*. 2009;**45**:427-37. doi: 10.1016/j.ipm.2009.03.002.