Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/MBE

MBE, 18(1): 727–744.
DOI: 10.3934/mbe.2021039
Received: 13 September 2020
Accepted: 06 December 2020
Published: 18 December 2020

Research article

Quantitative integration of radiomic and genomic data improves survival prediction of low-grade glioma patients

Chen Ma, Zhihao Yao, Qinran Zhang and Xiufen Zou*

School of Mathematics and Statistics, Wuhan University, Wuhan 430072, China

* **Correspondence:** Email: xfzou@whu.edu.cn.

Abstract: Glioma is the most common and most serious form of brain tumors that affects adults. Accurate prediction of survival and phenotyping of low-grade glioma (LGG) patients at high or low risk are the key to achieving precision diagnosis and treatment. This study is aimed to integrate both magnetic resonance imaging (MRI) data and gene expression data to develop a new integrated measure that represents a LGG patient's disease-specific survival (DSS) and classify subsets of patients at low and high risk for progression to cancer. We first construct the gene regulatory network by using gene expression data. We obtain twelve network modules and identify eight image biomarkers by using the Cox regression model with MRI data. Furthermore, correlation analysis between gene modules and image features identify four radiomic features. The least absolute shrinkage and selection operator (Lasso) method is applied to predict these image features with gene expression data when lacking MRI data or image segmentation technology. Furthermore, the support vector machine (SVM)-based recursive feature elimination method has been established to predict DSS using gene signatures. Finally, 4 image signatures and 43 gene signatures are recognized to be associated with the patient's prognosis. An integrated measure for combining image and gene signatures is obtained through the PSO algorithm. The concordance index (C-index) and the time-dependent receiver operating characteristic (ROC) analysis are used to evaluate prediction accuracy. The C-index obtained for this integrated measure is 0.8071 and the area under the curve (AUC) of the ROC curve is 0.79, which are higher than any other single features. The 72.1% accuracy of classification of patients is better than the accuracy associated with the published work. These results demonstrate that integration analysis of radiomic and genomic data can improve the accuracy of the prediction of survival for lower grade gliomas.

Keywords: low-grade glioma; MRI; genomic data; lasso model; network module

1. Introduction

Low-grade glioma (LGG) is a uniformly fatal tumor, and the survival from this tumor is approximately 7 years [1]. Because of the heterogeneity in LGG patients, different LGG subtypes increase the difficulty of optimizing management of adult low-grade gliomas [2, 3]. Magnetic Resonance Imaging (MRI) is an imaging technique that can capture tumors of the brain clearly [4]. Clinicians often use MRI images to diagnose the aggressiveness of the tumor. Therefore, the analysis of MRI data and feature extraction are becoming more challenging. To address these issues, many studies have used MRI data to extract prognostic factors for LGG patients. In a study by Pignatti et al. [5], the authors established a score system that can be used to determine the prognostic score. In adult patients with LGG, the age of the patients, the astrocytoma histology, the largest diameter of the tumor, the tumor crossing the midline and the presence of a neurologic deficit before surgery are all important prognostic factors for survival. These factors can be used to identify low-risk and high-risk patients. In a study by Chen et al. [6], the authors developed a computer-assisted algorithm for tumor segmentation and characterization using both kinetic information and morphological features of 3-D DCE-MRI. They differentiated benign and malignant lesions by analyzing 3-D morphological features including shape features and texture features of the segmented tumor. In a study by Agravat et al. [7], the authors implemented the DeepMedic CNN architecture for tumor segmentation and the extracted features are fed to a random forest classifier to obtain 59% overall survival accuracy. In another study by Shboul et al. [8], 40 features were extracted from the predicted brain tumor mask and fed to a random forest regression to predict the overall survival of a glioma patient, with an accuracy of 67% on the training dataset and 57.9% on the testing dataset. In an attempt at prediction of survival [9], the authors extracted 26 image-derived geometrical features and used SVM to predict the risk of death and classify glioma patients into three groups, with an accuracy of 56.8%. In another attempt [10], hundreds of intensity and texture features were extracted from MR images of glioblastoma multiforme, and principal component analysis (PCA) was used to reduce dimensionality. Then, these features were fed to an artificial neural network (ANN). A result with accuracy of 65.1% was obtained based on two classes: short-overall survivor and long-overall survivor. In another study [11], Chato et al. attempted the use of support vector machines (SVMs), k-nearest neighbors (KNNs), linear discriminants, tree, ensembles and logistic regression to classify survivors into two or three classes. The features from segmentations are used to train the linear discriminant for prediction of survival. The texture features resulted in the accuracy of 46%, and histogram features achieved an accuracy of 68.5% for the test dataset.

The above methods predicted survival by using only image information or clinical information. However, the tumor heterogeneity possibly comes from strong phenotypic differences, and it is difficult to predict prognosis accurately by using only medical imaging analysis (see Figure 1), thus motivating the need for integrating another kind of data. Along with the rapid development of deep-sequencing technology, the output of sequencing has made huge progress not only in equality but also in speed [12]. If radiomic data and genomic data can be integrated, this integration will build a bridge between micro and macro and increase the accuracy of the precision diagnosis and treatment of the brain tumor [13]. Grossmann et al. [14] found that prognostic biomarkers performed better in lung cancer when radiomic, genetic, and clinical information was combined. The C-index was 0.73, while the result is only 0.66 when lacking genetic information. Xia et al. [15] created a radiogenomic strategy that can obtain significant associations between imaging features and gene expression patterns in hepatocellular carcinoma. However, similar work is lacking in LGG. Therefore, in this study we integrated two different types of data, i.e., radiomic features of MRI and gene signatures, to develop a new integrated survival prediction measure for LGG.

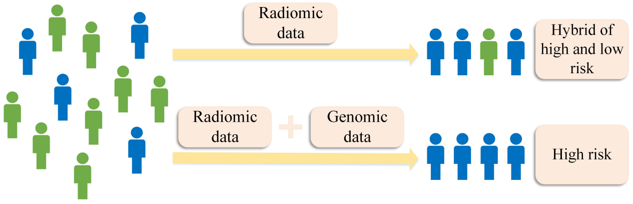


Figure 1. A diagram illustrates why we need to integrate radiomic data and genomic data. Low-risk and high-risk patients are marked in green and blue colors, respectively. Integration will increase the accuracy of recognition of high-risk patients. However, only radiomic data possibly leads to error classification.

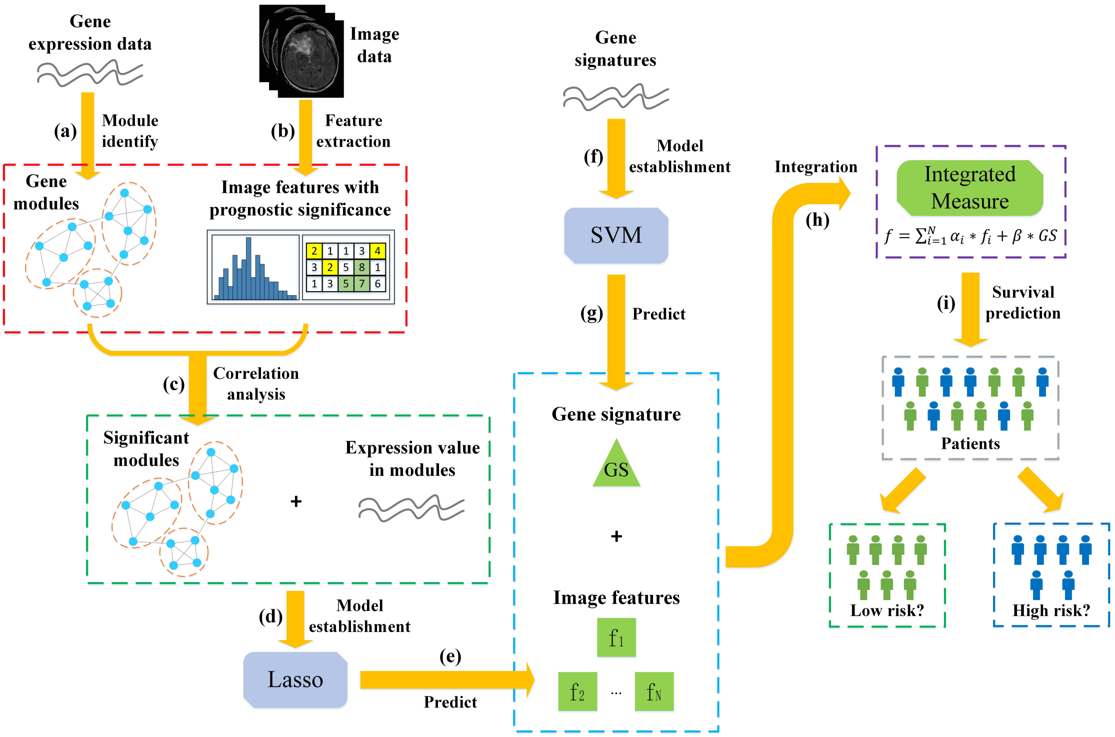


Figure 2. The framework of this study. (a) Construction of a gene regulatory network and identification of modules. (b) Extracting image features associated with patient survival. (c) Module analysis to select gene modules that have a connection with significant image features. (d) Establishing a Lasso model to identify gene signatures. (e) Predicting the significant image features using Lasso. (f) Identifying gene signatures using the SVM-based recursive feature elimination method and training the SVM model. (g) Survival prediction by SVM. The result could be treated as a survival prediction index. (h) A new integrated measure (IM) for combining image features and gene features is obtained through particle swarm optimization (PSO). (i) The IM that is obtained is used to predict survival.

The framework of this study is shown in Figure 2. First, we used gene expression data to construct a gene regulatory network and identify network modules and then used imaging data to extract significant radiomic biomarkers that are associated with the survival of the patient (Parts (a) and (b)), respectively. Then, we calculated the correlation between gene modules and image features to obtain a small number of gene signatures that are connected with these image features (Part (c)). Furthermore, we established a Lasso (least absolute shrinkage and selection operator) model to predict the image features with only gene expression values (Parts (d) and (e)). Based on gene expression data, we used support vector machines (SVMs) to identify the gene signatures (Parts (f) and (g)). We combined the predicted image features and the gene signatures to establish an integrated measure that can predict survival of the LGG patient (Parts (h) and (i)). The results show that the integrated measure performed better on survival prediction than any other single index.

2. Materials and methods

2.1. Collections of datasets

Computer-aided and manually corrected segmentation labels for the preoperative multi-institutional scans of 65 LGG patients and 724 radiomic features along with the corresponding skull-stripped and coregistered multimodal (i.e., T1, T1-Gd, T2, T2-FLAIR) MRI data were collected from the Cancer Imaging Archive (TCIA) [16–18]. The corresponding RNA-seq data and Disease Free Survival (DSS) data for these 65 patients were also obtained from The Cancer Genome Atlas (TCGA) database. These data were used in this study as the training dataset.

The gene expression data and the corresponding DSS data of 455 LGG patients were downloaded from TCGA and used in this study as the validation dataset.

2.2. Network construction and module identification

A gene coexpression network was constructed using gene expression data in the training dataset. We deleted genes that express in less than 20% of the patients or have no expression values. Then, we retained genes that have the highest 25% variance. A pairwise correlation matrix was calculated, and then we adjusted the matrix by raising it to the power of five using the R package WGCNA [19, 20]. The minimum module size was set to 50, and the minimum height for merging modules was set to 0.25.

2.3. A multivariate Cox model for identifying image biomarkers

We identified significant image features that are associated with patient DSS by training a multivariate Cox regression model [21] on the training dataset. Image features were filtered with the standard that p value must be less than 0.01. Then, these image features were treated as image biomarkers and survival prediction indexes. For each image feature, we divided patients on the validation dataset into two groups—high-risk group and low-risk group—by taking the median value of the feature as the threshold and plotted the Kaplan-Meier curves. The concordance index (C-index) [22] and the log-rank test were also used to assess the prognostic prediction performance.

The basic formula of the multivariate Cox regression model is described as follows:

$$h(t, X) = h_0(t) \cdot exp(\beta_1 X_1 + \beta_2 X_2 + ... + \beta_m X_m) \tag{1}$$

$h(t, X)$ represents the hazard function and $h_0(t)$ is the baseline hazard function. The factor X_1, X_2, ... , X_m correspond to the image features here and β_1, β_2, ... , β_m are the corresponding regression coefficients.

2.4. *Correlation between gene modules and image features*

We calculated Pearson correlation coefficients and their statistical significance to obtain the correlations between gene modules and selected image features. Because there are many genes in each module, the principal component analysis (PCA) was used to reduce the dimension of gene expression data of 65 patients in the training dataset. Then, image features were filtered. Features that showed significant correlation (p value less than 0.05) with at least one gene module were retained, and others were removed. Then, gene modules associated with the same image feature were integrated. The enrichment analysis was performed to identify the significantly enriched molecular pathways on these modules.

2.5. *Lasso model for further evaluating association between gene signatures and image features*

We established a radiogenomic map by identifying gene signatures associated with the prognostic imaging features. Lasso (least absolute shrinkage and selection operator) is a regression analysis method that performs both variable selection and regularization [23, 24]. This method can enhance the prediction accuracy and interpretability of the statistical model it produces.

$$Q(\beta) = ||y - X\beta||^2 + \lambda||\beta||_1 \tag{2}$$

Among the above formulas, X is the variable and y is the label. β is the coefficient that we want to optimize. $Q(\beta)$ is the objective function that we want to minimize. Compared with the method of least squares, the objective function in the Lasso model has a regularization term $\lambda||\beta||_1$. With this L_1 norm regularization term, Lasso can control the number of variables used and improve the generalization ability of the model. For each image feature remaining in the gene module analysis, Lasso was trained to select gene signatures from related gene modules and make a prediction on image features with MRI data and gene expression data in the training dataset. We determined the regularization coefficient λ by minimizing the MSE (mean squared error) of the model.

2.6. *SVM model for identifying gene signatures associated with survival*

In this step, we obtained a survival prediction index using only gene signatures, without the information of image features. SVMs (support vector machines) are supervised learning models that can be used for classification and regression problems [25–27]. For a classification problem, the optimal hyperplane is searched to separate data into two classes with the max margin. For new data, the trained hyperplane is used to predict the label or the probability of each class. Sometimes, data may not be separated completely, and a soft margin [25] can be used by adding a penalty parameter C and slack variables ξ_i to obtain the minimum error. The SVM optimization problem is

$$\min_{\omega,C} \frac{1}{2}||\omega||^2 + C\sum_{i=1}^{N}\xi_i \tag{3}$$

subject to

$$y_i f(x_i) \geq 1 - \xi_i, \ and \ \xi_i \geq 0 \tag{4}$$

The vector ω is the vector orthogonal to the hyperplane. x_i, y_i are an observation pair of data points, and $f(x_i)$ is the label of x_i predicted by the SVM. SVM-RFE (support vector machine-recursive feature elimination) [28] is a powerful feature selection algorithm based on SVM that can avoid overfitting when the number of features is high. In each iteration, features are scored and sorted through model training and the least important feature is removed. Remaining features are used for a next training, and the above step is repeated. The score for sorting of the i^{th} feature is defined as

$$c_i = \omega_i^2 \tag{5}$$

ω_i is the i^{th} dimension of the hyperplane orthogonal vector ω in SVM. Finally, the optimal number of features that have the minimum error is determined.

We use SVM-FRE to select gene signatures and train a classification SVM model with expression data of these selected gene signatures and DSS data in the training dataset. The patient labels are set to 0 or 1 based on their prognostic situation—survival or death. Then, the predicted probability is treated as a survival prediction index. Survival curve and C-index are used to access the prediction performance.

2.7. *An optimization model and algorithm for obtaining an integrated measure for predicting LGG patient survival*

Further, we consider a combination of selected image biomarkers and the index calculated by SVM with gene signatures. To ensure improvement of the new aggregated index, we transform the calculation of optimal combination coefficients of all features into an optimization problem. Specifically, suppose that N image features are considered to be associated with DSS independently—which are recorded as f_1, f_2, ..., f_N and the gene index value from SVM is recorded as g. The integrated measure we want to determine is recorded as f. The optimization problem needing to be solved can be described as follows.

$$\max_{f} C_f \tag{6}$$

subject to

$$f = \sum_{i=1}^{N} \alpha_i \cdot f_i + \beta \cdot g \tag{7}$$

$$\sum_{i=1}^{N} \alpha_i + \beta = 1 \tag{8}$$

where C_f is the C-index of integrated measure f on the training dataset. Our goal is to search optimal parameters α_1, α_2, ..., α_N and β in Eq (7) to maximize the C_f in (6).

The Particle Swarm Optimization (PSO) algorithm [29] is used to solve the optimization problem (6) in this study. PSO is an evolutionary computation algorithm inspired by bird activities that can solve any optimization problem. Initial population with some random particle is created first. For each particle, the position represents a solution, and the corresponding fitness means a value of target function. The object of PSO is to find the optimal particle that has the minimized fitness by updating the velocity and position of particle as the following formula:

$$v_i = \omega \cdot v_i + c_1 \cdot r_1 \cdot (pbest_i - x_i) + c_2 \cdot r_2 \tag{9}$$

$$x_i = x_i + v_i \tag{10}$$

x_i, v_i is the position and velocity of the i^{th} particle. $pbest_i$ is the best position of the i^{th} particle in history and $gbest$ is the best position of all particles currently. r_1, r_2 are random numbers between 0 and 1. ω is the inertia weight, and c_1, c_2 are the acceleration constants.

3. Results

3.1. Prognostic image feature identification

We take a log-rank test on 724 image features using DSS data of 65 patients in TCIA and filter these features with a standard that the p value is less than 0.01. Then, 21 features remain. Features with high similarity to each other are removed: we calculate the Pearson correlation coefficient between features and remove the one that has the bigger log-rank p value if the Pearson correlation coefficient between two image features is greater than 0.8. After this step, 6 features are removed, and 15 features remain. Based on the above univariable analysis, we first implement the proportional hazard test [21]. Each image feature meets the proportional hazard assumption (detailed information is shown in Additional file 1: Table S1). Then, we train a multivariate Cox regression model on these remaining image features with gene expression data and DSS data in the training dataset. The result is shown in Table 1, and eight features marked with * are considered to be independently correlated with DSS ($p < 0.05$).

Table 1. Image features for survival analysis.

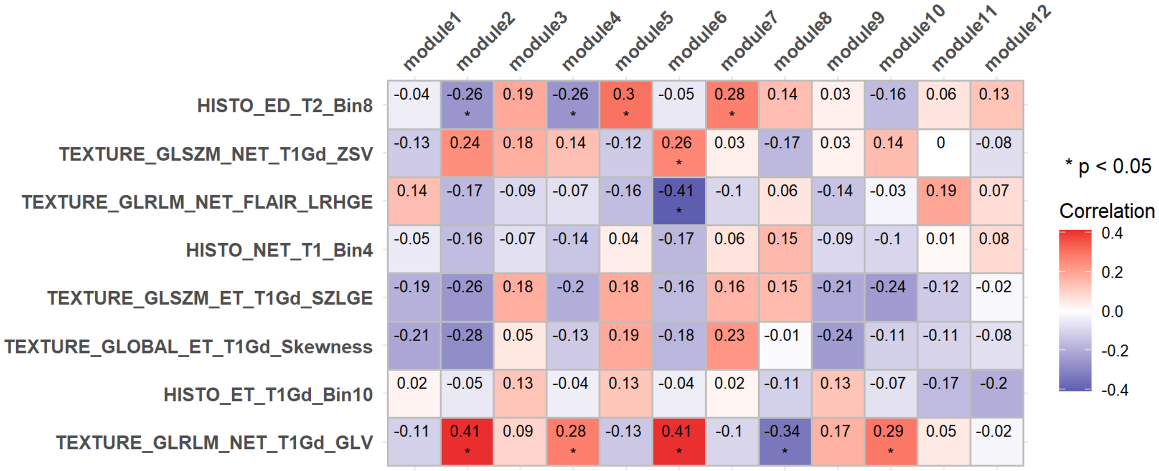
Image features	exp(coef)	exp(coef) lower 95%	exp(coef) upper 95%	Wald test	p value
TEXTURE_GLSZM_ET_T1Gd_SZLGE*	0	0	0	−3.12	0.00178
HISTO_ED_T2_Bin8*	0.7	0.55	0.88	−3.02	0.00254
TEXTURE_GLOBAL_ET_T1Gd_Skewness*	3.03E+05	47.29	1.94E+09	2.82	0.00477
TEXTURE_GLRLM_NET_FLAIR_LRHGE*	1	0.99	1	−2.66	0.0078
HISTO_NET_T1_Bin4*	0.89	0.8	0.98	−2.42	0.01559
HISTO_ET_T1Gd_Bin10*	1.19	1.03	1.38	2.35	0.01877
TEXTURE_GLSZM_NET_T1Gd_ZSV*	0	0	0	−2.34	0.01906
TEXTURE_GLRLM_NET_T1Gd_GLV*	2.00E+42	2230.45	1.79E+81	2.13	0.0333
HISTO_ET_T1_Bin10	0.81	0.63	1.05	−1.59	0.11219
TEXTURE_GLCM_ET_T2_SumAverage	0	0	9.95E+83	−1.57	0.11569
TEXTURE_GLRLM_NET_T1_LGRE	0	0	2.73E+38	−1.49	0.13526
TEXTURE_GLRLM_ED_T1_RLV	inf	0	inf	1.4	0.16185
HISTO_ED_T2_Bin4	0.96	0.88	1.05	−0.91	0.36241
TEXTURE_GLCM_ED_FLAIR_Energy	inf	0	inf	0.78	0.43387
TEXTURE_GLSZM_NET_T1_LZLGE	0.99	0.96	1.03	−0.39	0.69976

3.2. Gene signatures associated with image features

A gene coexpression network is constructed using gene expression data of 65 patients in the training dataset. We delete genes that express in less than 20% of the patients or have no expression values (n = 1875). Then, we retain genes that have the highest 25% variance (n = 4663). A pairwise correlation matrix is calculated, and then we adjust the matrix by raising it to the power of five using the R package WGCNA [19, 20]. The minimum module size is set to 50 and the minimum height for merging modules is set to 0.25. Then, we get 12 gene modules. Detailed information on the modules is shown in Additional file 2: Table S2.

The Pearson correlation coefficient and their statistical significance were calculated between the 12 gene modules and the 8 image features. The result is shown in Figure 3. Four image features that show significant correlation ($p < 0.05$) with at least one gene module were obtained. HISTO_ED_T2_Bin8 is the 8-bin histogram feature of the peritumoral edema in T2-weighted precontrast, TEXTURE_GLSZM_NET_T1Gd_ZSV is the zone size variance of gray level size zone matrix (GLSZM) of the nonenhancing part of the tumor core in T1-weighted postcontrast, TEXTURE_GLRLM_NET_FLAIR_LRHGE is the long run high gray level emphasis of gray level run length matrix (GLRLM) of the nonenhancing part of the tumor core in T2 Fluid-Attenuated Inversion Recovery, and TEXTURE_GLRLM_NET_T1Gd_GLV is the gray level variance of GLRLM of the nonenhancing part of the tumor core in T1-weighted postcontrast. Then, their corresponding gene modules were integrated. The statistical results are shown in Table 2 and the detailed list of genes is shown in Additional file 3: Table S3.

A further KEGG enrichment analysis was performed on integrated gene modules using the Metascape website [30], which is shown in Figure 4. The complete list of biological annotations is shown in Additional file 4: Table S4. Among these, the neuroactive ligand-receptor interaction pathway is mostly enriched in all integrated gene modules with the minimum p value of 1.259×10^{-41}, which is reported to be associated with glioma [31, 32].



	module1	module2	module3	module4	module5	module6	module7	module8	module9	module10	module11	module12
HISTO_ED_T2_Bin8	-0.04	-0.26 *	0.19	-0.26 *	0.3 *	-0.05	0.28 *	0.14	0.03	-0.16	0.06	0.13
TEXTURE_GLSZM_NET_T1Gd_ZSV	-0.13	0.24	0.18	0.14	-0.12	0.26 *	0.03	-0.17	0.03	0.14	0	-0.08
TEXTURE_GLRLM_NET_FLAIR_LRHGE	0.14	-0.17	-0.09	-0.07	-0.16	-0.41 *	-0.1	0.06	-0.14	-0.03	0.19	0.07
HISTO_NET_T1_Bin4	-0.05	-0.16	-0.07	-0.14	0.04	-0.17	0.06	0.15	-0.09	-0.1	0.01	0.08
TEXTURE_GLSZM_ET_T1Gd_SZLGE	-0.19	-0.26	0.18	-0.2	0.18	-0.16	0.16	0.15	-0.21	-0.24	-0.12	-0.02
TEXTURE_GLOBAL_ET_T1Gd_Skewness	-0.21	-0.28	0.05	-0.13	0.19	-0.18	0.23	-0.01	-0.24	-0.11	-0.11	-0.08
HISTO_ET_T1Gd_Bin10	0.02	-0.05	0.13	-0.04	0.13	-0.04	0.02	-0.11	0.13	-0.07	-0.17	-0.2
TEXTURE_GLRLM_NET_T1Gd_GLV	-0.11	0.41 *	0.09	0.28 *	-0.13	0.41 *	-0.1	-0.34 *	0.17	0.29 *	0.05	-0.02

Figure 3. The heatmap of correlation between the image features and the gene modules. Colored checks marked with * means significant Pearson correlation.

Table 2. The statistical results of image features and their corresponding gene modules with significant association.

Image features	Associated gene modules	Number of associated genes
HISTO_ED_T2_Bin8	module2, module4, module5, module7	2794
TEXTURE_GLSZM_NET_T1Gd_ZSV	module6	506
TEXTURE_GLRLM_NET_FLAIR_LRHGE	module6	506
TEXTURE_GLRLM_NET_T1Gd_GLV	module2, module4, module6, module8, module10	1421

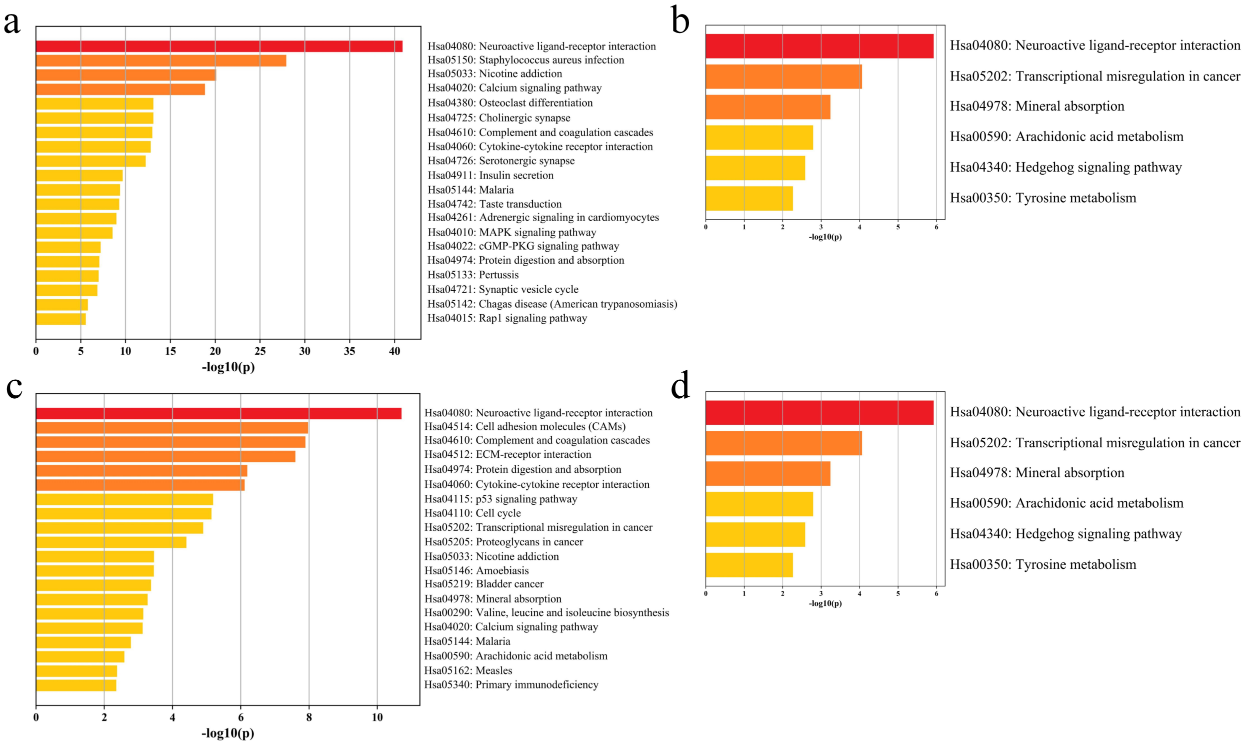


Figure 4. Results of KEGG enrichment analysis: a. Enrichment of modules associated with HISTO_ED_T2_Bin8. b. Enrichment of modules associated with TEXTURE_GLSZM_NET_T1Gd_ZSV. c. Enrichment of modules associated with TEXTURE_GLRLM_NET_T1Gd_GLV. d. Enrichment of modules associated with TEXTURE_GLRLM_NET_FLAIR_LRHGE.

Then, the Lasso method described in section 2.5 was used to select gene signatures from the related gene modules and establish a map from genes to image features. We determined the regularization coefficient λ by minimizing the MSE (mean squared error) of the model. The process is shown in Figure 6. The optimal coefficient λ and the corresponding RMSE (root mean squared error) of 65 patients are shown in Table 3. The number of selected gene signatures is also shown. The detailed list of gene signatures is shown in Additional file 5: Table S5.

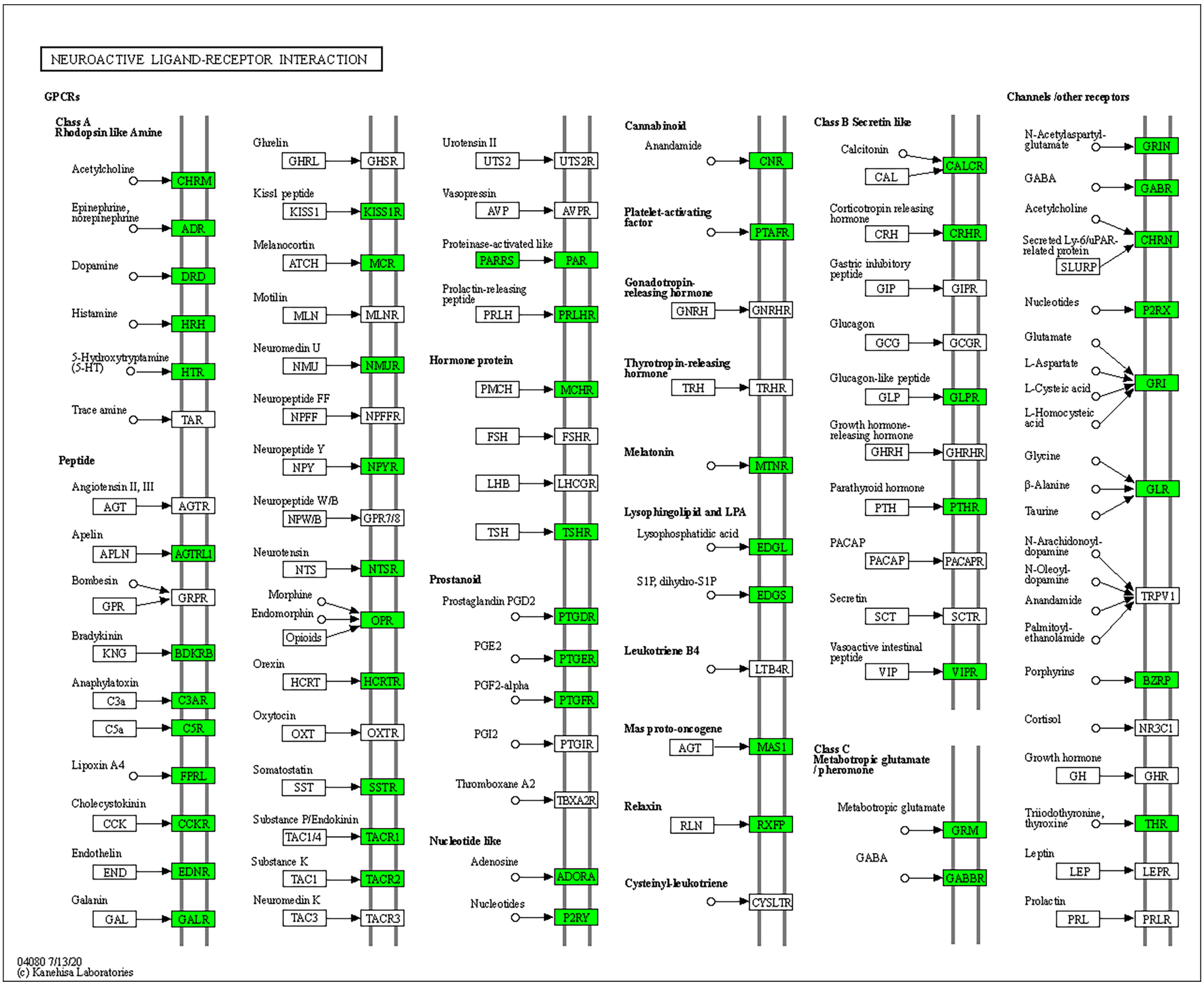


Figure 5. The chart of the neuroactive ligand-receptor interaction pathway. Genes appearing in associated modules are marked in green.

Table 3. The optimal parameters of Lasso and number of selected gene signatures for four image features.

Image feature	Number of genes in associated modules	Optimal λ	RMSE	Number of genes selected by Lasso
HISTO_ED_T2_Bin8	2794	1.6627	6.0847	12
TEXTURE_GLSZM_NET_T1Gd_ZSV	506	7.81E-06	2.0195E-5	3
TEXTURE_GLRLM_NET_FLAIR_LRHGE	506	163.9677	528.16	6
TEXTURE_GLRLM_NET_T1Gd_GLV	1421	3.01E-03	0.0120	18

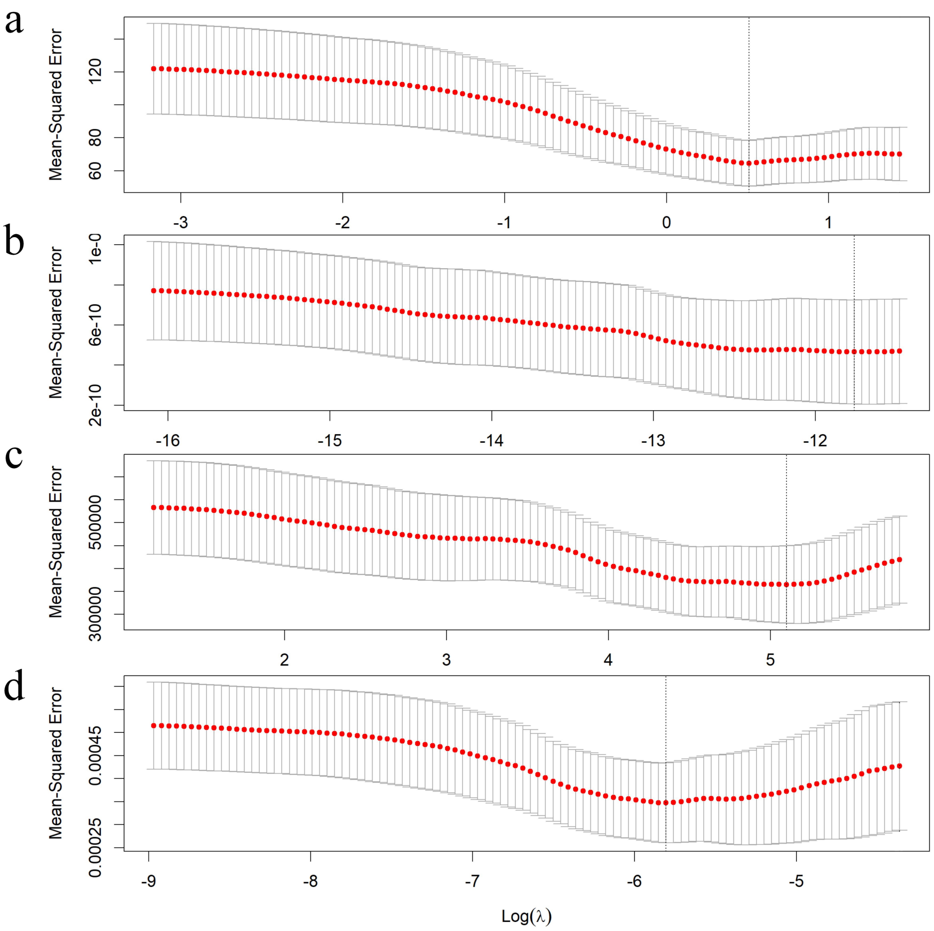


Figure 6. The value and 95% confidence interval of MSE for each regularization coefficient λ. The dotted line marks λ with the minimal MSE. All Lasso models were trained on 65 patients in the training dataset. a. λ and corresponding MSE of the Lasso model, mapping from gene signatures to HISTO_ED_T2_Bin8. b. λ and corresponding MSE of Lasso model, mapping from gene signatures to TEXTURE_GLSZM_NET_T1Gd_ZSV. c. λ and corresponding MSE of Lasso model, mapping from gene signatures to TEXTURE_GLRLM_NET_FLAIR_LRHGE. d. λ and corresponding MSE of Lasso model, mapping from gene signatures to TEXTURE_GLRLM_NET_T1Gd_GLV.

3.3. Survival analysis with image signatures

We made a prediction on the 4 image features using Lasso with gene expression data of 455 patients in TCGA as the validation dataset. We then took the value of each image feature as a survival prediction index. We calculated the C-index and plotted the Kaplan-Meier curves on the validation dataset. The result is shown in Figure 7. The C-index of these four survival prediction indexes are 0.6945, 0.7321, 0.7926, and 0.7985. These results indicate that these four image features perform well in survival prediction.

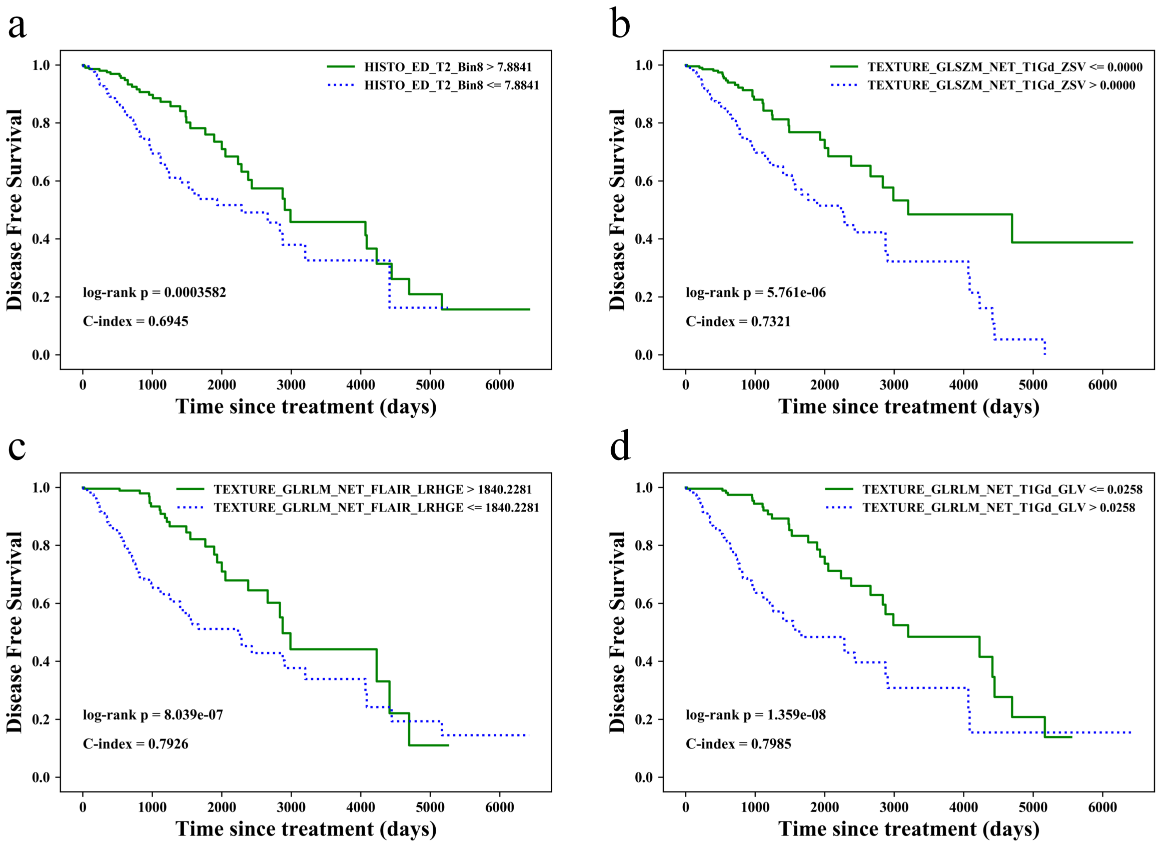


Figure 7. Kaplan-Meier curves of DSS and C-index. a. HISTO_ED_T2_Bin8. b. TEXTURE_GLSZM_NET_T1Gd_ZSV. c. TEXTURE_GLRLM_NET_FLAIR_LRHGE. d. TEXTURE_GLRLM_NET_T1Gd_GLV.

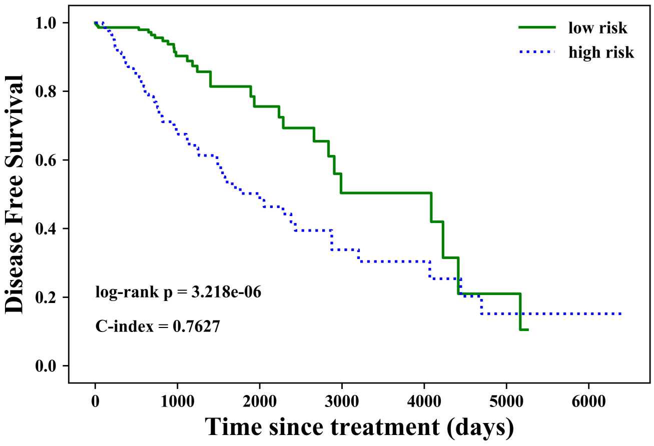


Figure 8. Kaplan-Meier curve of DSS and C-index of the index from SVM.

3.4. Survival analysis with gene signatures

From the selected 4663 genes with high variance, we fed gene expression data and DSS data of 65 patients in TCIA to SVM-FRE and obtained 43 gene signatures (shown in Additional file 6: Table S6). Then, we trained a classification SVM model with these selected genes. The variables were gene expression data of 65 patients, and the labels were set to 0 or 1 based on the patient prognostic situation—survival or death. Penalty parameter C was set to 2 and 5-fold cross-validation was used to evaluate the error in the recursive feature elimination process. We trained the SVM model and took the predicted probability of survival as a survival prediction index. C-index and survival curve are shown in Figure 8. The C-index is 0.7627.

3.5. Integration of different features

We took a linear combination of four significant image features and the index calculated by SVM with gene signatures. A better integrated measure was obtained that represents patient survival situation. Set $N = 4$ in formula (7). Four normalized image feature values were recorded as f_1, f_2, f_3, and f_4, and the index value from SVM was recorded as g. The integrated measure is recorded as f. Then, we get

$$f = \sum_{i=1}^{4} \alpha_i \cdot f_i + \beta \cdot g \tag{11}$$

We used PSO algorithm to calculate the optimal coefficient to maximize the C-index of 65 patients in the training dataset, with parameters ω, C_1 and C_2 of 0.8, 0.5 and 0.5. The initial population size was set to 20, 25, 30, 35 and 40, and the corresponding iteration number was set to 30 to ensure the convergence of PSO. We repeated numerical experiments 10 times and recorded the average result for different parameters. Detailed results of each experiment are shown in Additional file 7: Table S7. For each population size, we then brought the coefficients into formula (11) and obtained integrated measure f with different forms. C-index was calculated using gene expression data on the validation dataset. The validation result is shown in Table 4.

Table 4. The mean result of combination coefficients calculated by PSO and C-index with different parameters.

Populations sizes	20	25	30	35	40
α_1	0.2926	0.3187	0.2792	0.3303	0.276
α_2	0.0663	0.0394	0.0739	0.0505	0.068
α_3	0.2171	0.2329	0.2076	0.2214	0.2102
α_4	0.0091	0.019	0.0107	0.0298	0.0159
β	0.4149	0.39	0.4288	0.368	0.4298
C-index	0.8065	0.807	0.8061	0.807	0.8057

From Table 4, we observe that β is more or less than 0.4 with different parameters. Therefore, the proportion of gene signatures in integration is approximately 40%. α_1 is approximately 0.3, α_2 is approximately 0.06 and α_3 is approximately 0.24. α_4 is nearly 0, indicating that the gray level variance of GLRLM of the nonenhancing part of the tumor core in T1-weighted postcontrast can be removed in the integration. We then set parameters α_1, α_2, α_3, α_4 and β to 0.3, 0.06, 0.24, 0 and 0.4. We brought these coefficients into formula (11) and calculated the integrated measure f on the validation dataset. The Kaplan-Meier curve is shown in Figure 9. The C-index of the four independent image features, gene signatures and integrated measures are shown in Table 5.

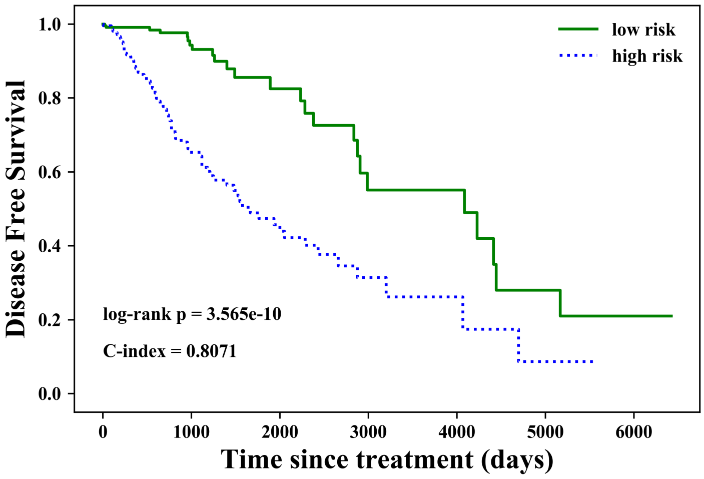


Figure 9. Kaplan-Meier curve of DSS and C-index of integrated measure f.

Table 5. C-index of image features and gene signatures.

Image features	f_1	f_2	f_3	f_4	g	f
C-index	0.6945	0.7321	0.7926	0.7985	0.7627	0.8071

The C-index of the integrated measure f is 0.8071 and is higher than any other measure based on image signatures or gene signature. This result indicates that the integrated measure can improve the prediction accuracy. The integrated measure is recorded as follows.

$$f = 0.3f_1 + 0.06f_2 + 0.24f_3 + 0.4g \tag{12}$$

Furthermore, we use the time dependent Receiver Operating Characteristic (ROC) [33] to further assess the predictive power and compare different prediction models. Time-dependent ROC analysis showed that the integrated measure improved our ability to predict prognosis [AUC, 0.79; and 95% confidence intervals (CI), 0.71 to 0.87] (see Figure 10), when compared with other measures based on image signatures or gene signatures.

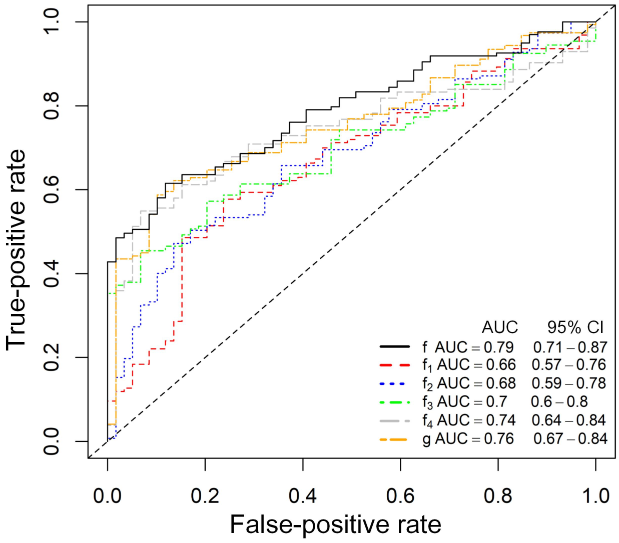
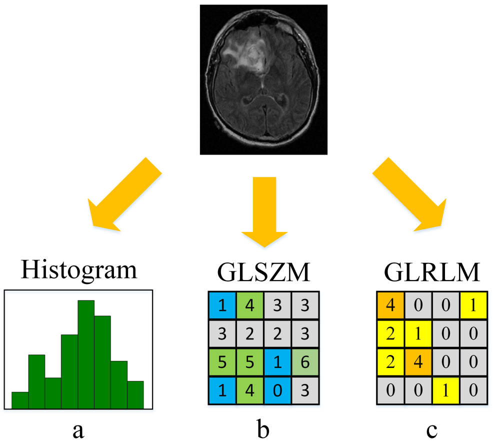


Figure 10. ROC and corresponding AUCs for 5-year survival predicted by f_1, f_2, f_3, f_4, g and f on the 455 patients in validation dataset.



1	4	3	3
3	2	2	3
5	5	1	6
1	4	0	3

4	0	0	1
2	1	0	0
2	4	0	0
0	0	1	0

Figure 11. Image features used for final survival prediction. a The 8-bin histogram feature of the peritumoral edema. b The zone size variance of gray level size zone matrix (GLSZM) of the nonenhancing part of the tumor core. c The long run high gray level emphasis of gray level run length matrix (GLRLM) of the nonenhancing part of the tumor core.

Patients are defined into two groups—high-risk group and low-risk group, based on their prognosis—DSS value in this study, by taking the median value of DSS of 65 patients in the training dataset as a threshold. Then, classification is conducted on 455 patients in the validation dataset by taking a threshold of the median value of the integrated measure in the training dataset. The accuracy is 72.1%, which is higher than the accuracy of the published studies [7–11].

4. Conclusions

The primary goal of phenotyping and classifying a human tumor is to capture tumor heterogeneity and realize personalized precision diagnosis and therapy. In clinical practice, the massive and multiple types of big medical data are available with the rapid development of biomedical engineering and computer application technology. However, one of the biggest challenges in clinical applications is how to integrate these different types of data to extract accuracy information.

In this study, we attempted to integrate both MRI data and gene expression data to propose a new feature measure that could be used to identify subsets of LGG patients at low and high risk for progression to DSS. Based on gene expression data, we first used the WGCNA method to construct the network and identify twelve network modules. With MRI data, eight image biomarkers were obtained by using the Cox regression model. Furthermore, through correlation analysis between gene modules and image features, four radiomic biomarkers were identified. Because MRI data are not available in our test dataset, the Lasso method was applied to build a map from gene expression data to these image features. In addition, we also independently used gene expression data to predict image biomarkers through the SVM method. Finally, an integrated measure (IM) for combining image and gene signatures was obtained through the PSO algorithm. We validated IM with gene expression data and DSS data on 455 patients in the validation dataset. The C-index of IM is 0.8071 and its Area Under Curve (AUC) of the ROC curve is 0.79, higher than any other single measure. The accuracy of classification of patients is 72.1%, which is higher than the accuracy of the published work using only radiomic data [7–11]. The results demonstrate that the proposed IM enhances the prediction accuracy for lower grade gliomas.

In summary, the accuracy of DSS prediction of LGG patients is successfully improved by integrating radiomic features in Macro with the gene expression data in Micro. The proposed method in this study can also be extended to analyze different data sources of other tumors.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (No. 2018YFC1314600), the Key Program of the National Natural Science Foundation of China (No. 11831015) and the Chinese National Natural Science Foundation (No. 61672388).

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

1. E. B. Claus, K. M. Walsh, J. K. Wiencke, A. M. Molinaro, J. L. Wiemels, J. M. Schildkraut, et al., Survival and low-grade glioma: the emergence of genetic information, *Neurosurg. Focus*, **38** (2015), E6.
2. K. Lote, T. Egeland, B. Hager, B. Stenwig, K. Skullerud, J. Berg-Johnsen, et al., Survival, prognostic factors, and therapeutic efficacy in low-grade glioma: a retrospective study in 379 patients, *J. Clin. Oncol.*, **15** (1997), 3129–3140.

3. D. Schiff, P. D. Brown, C. Giannini, Outcome in adult low-grade glioma: the impact of prognostic factors and treatment, *Neurology*, **69** (2007), 1366–1373.

4. Z. P. Liang, P. C. Lauterbur, *Principles of Magnetic Resonance Imaging: A Signal Processing Perspective*, SPIE Optical Engineering Press, 2000.

5. F. Pignatti, M. V. Den Bent, D. Curran, C. Debruyne, R. Sylvester, P. Therasse, et al., Prognostic factors for survival in adult patients with cerebral low-grade glioma, *J. Clin. Oncol.*, **20** (2002), 2076–2084.

6. T. C. Wang, Y. H. Huang, C. S. Huang, J. H. Chen, G. Y. Huang, Y. C. Chang et al., Computer-aided diagnosis of breast dce-mri using pharmacokinetic model and 3-d morphology analysis, *Magn. Reson. Imaging*, **32** (2014), 197–205.

7. R. R. Agravat, M. S. Raval, Prediction of overall survival of brain tumor patients, *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*, 2019.

8. Z. A. Shboul, L. Vidyaratne, M. Alam, K. M. Iftekharuddin, Glioblastoma and survival prediction, *International MICCAI Brainlesion Workshop*, 2017.

9. A. Jungo, R. Mckinley, R. Meier, U. Knecht, L. Vera, J. Pérez-Beteta, et al., Towards uncertainty-assisted brain tumor segmentation and survival prediction, *International MICCAI Brainlesion Workshop*, 2017.

10. J. Sachdeva, V. Kumar, I. Gupta, N. Khandelwal, C. K. Ahuja, Segmentation, feature extraction, and multiclass brain tumor classification, *J. Digital Imaging*, **26** (2013), 1141–1150.

11. L. Chato, S. Latifi, Machine learning and deep learning techniques to predict overall survival of brain tumor patients using mri images, in *2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE)*, 2017.

12. S. D. Kahn, On the future of genomic data, *Science*, **331** (2011), 728–729.

13. H. J. Aerts, E. R. Velazquez, R. T. Leijenaar, C. Parmar, P. Lambin, Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach, *Nat. Commun.*, **5** (2014), 1–9.

14. P. Grossmann, O. Stringfield, N. El-Hachem, M. M. Bui, E. R. Velazquez, C. Parmar, et al., Defining the biological basis of radiomic phenotypes in lung cancer, *Elife*, **6** (2017), e23421.

15. W. Xia, Y. Chen, R. Zhang, Z. Yan, X. Zhou, B. Zhang, et al., Radiogenomics of hepatocellular carcinoma: multiregion analysis-based identification of prognostic imaging biomarkers by integrating gene data—a preliminary study, *Phys. Med. Biol.*, **63** (2018), 035044.

16. S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. Kirby, et al., Segmentation labels and radiomic features for the pre-operative scans of the tcga-lgg collection, *Cancer Imaging Arch.*, **286** (2017).

17. S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. Kirby, et al., Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features, *Sci. Data*, **4** (2017), 170117.

18. K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, et al., The cancer imaging archive (tcia): Maintaining and operating a public information repository, *J. Digital Imaging*, **26** (2013), 1045–1057.

19. P. Langfelder, S. Horvath, Wgcna: an r package for weighted correlation network analysis, *BMC Bioinf.*, **9** (2008), 559.

20. M. Fan, P. Xia, B. Liu, L. Zhang, Y. Wang, X. Gao, et al., Tumour heterogeneity revealed by unsupervised decomposition of dynamic contrast-enhanced magnetic resonance imaging is associated with underlying gene expression patterns and poor survival in breast cancer patients, *Breast Cancer Res.*, **21** (2019), 112.

21. D. R. Cox, Regression models and life tables, *J. R. Stat. Soc.*, **34** (1972), 187–202.

22. F. E. Harrell Jr, K. L. Lee, D. B. Mark, Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors, *Stat. Med.*, **15** (1996), 361–387.

23. F. Santosa, W. W. Symes, Linear inversion of band-limited reflection seismograms, *SIAM J. Sci. Stat. Comput.*, **7** (1986), 1307–1330.

24. R. Tibshirani, Regression shrinkage and selection via the lasso, *J. R. Stat. Soc. Ser. B Methodol.*, **58** (1996), 267–288.

25. C. Cortes, V. Vapnik, Support-vector networks, *Mach. Learn.*, **20** (1995), 273–297.

26. N. Cristianini, J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge university press, 2000.

27. R. E. Fan, P. H. Chen, C. J. Lin, Working set selection using second order information for training support vector machines, *J. Mach. Learn. Res.*, **6** (2005), 1889–1918.

28. I. Guyon, A. J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using svm, *Mach. Learn. J.*, **46** (2002), 389–422.

29. J. Kennedy, R. Eberhart, Particle swarm optimization, *Proceedings of ICNN'95-International Conference on Neural Networks*, 1995.

30. Y. Zhou, B. Zhou, L. Pache, M. Chang, A. H. Khodabakhshi, O. Tanaseichuk, et al., Metascape provides a biologist-oriented resource for the analysis of systems-level datasets, *Nat. Commun.*, **10** (2019), 1–10.

31. J. Pal, V. Patil, A. Kumar, K. Kaur, C. Sarkar, K. Somasundaram, Genetic landscape of glioma reveals defective neuroactive ligand receptor interaction pathway as a poor prognosticator in glioblastoma patients, *AACR*, **77** (2017), 2454–2454.

32. R. Wang, J. Wei, Z. Li, Y. Tian, C. Du, Bioinformatical analysis of gene expression signatures of different glioma subtypes, *Oncol. Lett.*, **15** (2018), 2807–2814.

33. P. J. Heagerty, T. Lumley, M. S. Pepe, Time-dependent roc curves for censored survival data and a diagnostic marker, *Biometrics*, **56** (2000), 337–344.