

Survival stratification for colorectal cancer via multi-omics integration using an autoencoder-based model

Hu Song¹, Chengwei Ruan², Yixin Xu¹, Teng Xu¹, Ruizhi Fan¹, Tao Jiang¹, Meng Cao¹ and Jun Song¹ 

¹Department of Gastrointestinal Surgery, the Affiliated Hospital of Xuzhou Medical University, Xuzhou, Jiangsu 221002, PR China;

²Department of Anorectal Surgery, the Affiliated Hospital of Xuzhou Medical University, Xuzhou, Jiangsu 221002, PR China

Corresponding author: Jun Song. Email: songjunmed@126.com

Impact statement

There is a dire need to better understand the pathophysiology and evaluate the survival prediction for CRC to increase the array of treatment options and improve prognosis. We used deep learning algorithms to integrate multi-omics for predicting the prognosis of CRC. We identified two survival-specific groups of CRC and validated these groups in independent validation cohorts. Overall, our study provides novel insights into the differential mechanisms of the two survival groups.

Abstract

Prognosis stratification in colorectal cancer helps to address cancer heterogeneity and contributes to the improvement of tailored treatments for colorectal cancer patients. In this study, an autoencoder-based model was implemented to predict the prognosis of colorectal cancer via the integration of multi-omics data. DNA methylation, RNA-seq, and miRNA-seq data from The Cancer Genome Atlas (TCGA) database were integrated as input for the autoencoder, and 175 transformed features were produced. The survival-related features were used to cluster the samples using k-means clustering. The autoencoder-based strategy was compared to the principal component analysis (PCA)-, t-distributed random neighbor embedded (t-SNE)-, non-negative matrix factorization (NMF)-, or individual

Cox proportional hazards (Cox-PH)-based strategies. Using the 175 transformed features, tumor samples were clustered into two groups (G1 and G2) with significantly different survival rates. The autoencoder-based strategy performed better at identifying survival-related features than the other transformation strategies. Further, the two survival groups were robustly validated using “hold-out” validation and five validation cohorts. Gene expression profiles, miRNA profiles, DNA methylation, and signaling pathway profiles varied from the poor prognosis group (G2) to the good prognosis group (G1). miRNA–mRNA networks were constructed using six differentially expressed miRNAs (let-7c, mir-34c, mir-133b, let-7e, mir-144, and mir-106a) and 19 predicted target genes. The autoencoder-based computational framework could distinguish good prognosis samples from bad prognosis samples and facilitate a better understanding of the molecular biology of colorectal cancer.

Keywords: Autoencoder, deep learning, K-means clustering, multi-omics, survival

Experimental Biology and Medicine 2022; 247: 898–909. DOI: 10.1177/15353702211065010

Introduction

Colorectal carcinogenesis involves the accumulation of complicated histological, morphological, and genetic changes over time.¹ Although routine screening decreases the incidence and mortality rate² of colorectal cancer (CRC), this malignancy remains one of the most severe and deadly cancers worldwide, with an overall survival (OS) of less than three years for patients with advanced CRC.³ Accordingly, there is a dire need to better understand the pathophysiology and evaluate the survival prediction for CRC to increase treatment options and improve the prognosis of patients.

Deep learning (DL) algorithms have attracted wide attention in processing medical image and computer vision, owing to their stronger computing power and lower hardware cost.⁴ DL techniques have been used for the detection, classification, segmentation, and survival predictions of CRC based on histopathological slides and radiological or colonoscopic images.^{5,6} The Cancer Genome Atlas (TCGA) database collects multi-omics data, including genomic, transcriptomic, proteomic, and epigenomic data of more than 30 cancer types, thereby providing multiple views of the same patients.⁷ DL algorithms offer a potential solution to the integrative analysis of multi-omics data and

have been successfully applied to explore the molecular mechanisms of cancers.⁸ Multi-omics integration using DL algorithms has shown great promise for the survival prediction of breast cancer⁹ and neuroblastoma.¹⁰ However, only few studies have been published on the application of DL procedures for prognosis prediction in CRC.

An autoencoder is an unsupervised DL method that includes input, hidden, and output layers. Using the input data, the autoencoder creates different representative features in the hidden layer and regenerates the output data.¹¹ The autoencoder-based model has identified survival-related multi-omics features of liver cancer and successfully defines two survival-sensitive subtypes.¹² Our study employed an autoencoder-based model to integrate DNA methylation, RNA-seq, and miRNA-seq data from CRC patients and extract representative features. Based on the survival-related features, we identified two survival-specific groups and verified these groups in the validation cohorts. We also assessed the differential molecular mechanisms of the two survival groups.

Materials and methods

Datasets and preprocessing

We downloaded coupled methylation beta value data (Methylation Illumina 450k BeadChip), RNA-seq FPKM data (RNA-seq Illumina HiSeq 2000 RNA Sequencing platform), and miRNA-seq RPM data of 379 CRC samples (miRNA-seq Illumina HiSeq platform), together with the corresponding clinical characteristics from TCGA (<https://gdc-portal.nci.nih.gov/>) database. The retrieved data were defined as the training cohort of our study (TCGA cohort).

To preprocess raw data, we removed the probes or genes missing values across more than half of all samples, samples missing more than 20% three-omics features, and input features whose values were zero in all samples. The methylation value of a gene promoter region that covers 1500 base pairs (bp) from the transcription start sites (TSS) was annotated using IlluminaHumanMethylation450kanno.ilmn12.hg19 package¹³ and calculated using average methylation beta values of all CpG islands within the gene promoter region. Additionally, missing values were filled out using the impute package of R software (<https://www.r-project.org/>, version 3.5.2).¹⁴ We obtained five validation cohorts containing RNA expression profiles of CRC samples from the ArrayExpress (<https://www.ebi.ac.uk/>

[arrayexpress/](https://www.ebi.ac.uk/arrayexpress/)) database. These cohorts included E-GEOD-17538 (N = 232, <https://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-17538/A-AFFY-44> platform, <https://www.ebi.ac.uk/arrayexpress/arrays/A-AFFY-44/?ref=E-GEOD-17538>), E-GEOD-28722 (N = 125, <https://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-28722/A-GEOD-13425> platform: <https://www.ebi.ac.uk/arrayexpress/arrays/A-GEOD-13425/?ref=E-GEOD-28722>), E-GEOD-38832 (N = 122, A-AFFY-44 platform; www.ebi.ac.uk/arrayexpress/arrays/A-AFFY-44/?ref=E-GEOD-38832), E-GEOD-39582 (N = 558, A-AFFY-44 platform: <https://www.ebi.ac.uk/arrayexpress/arrays/A-AFFY-44/?ref=E-GEOD-39582>), and E-GEOD-41258 (N = 141, A-AFFY-33 platform, <https://www.ebi.ac.uk/arrayexpress/arrays/A-AFFY-33/?ref=E-GEOD-41258>). First, the probes were converted into gene symbols. Thereafter, the gene expression data were calculated by averaging the probes corresponding to the same gene. Table 1 shows the demographic and clinical data of TCGA and validation cohorts.

Features transformation

The autoencoder was implemented as described previously¹² using the TCGA cohort. DNA methylation, RNA-seq, and miRNA-seq data were unit-norm scaled by sample as follows:

For a given sample vector $v = (v_1, \dots, v_n)$

$$v_{normed} = v \cdot \frac{1}{\|v\|_2}$$

where $\|v\|_2$ is the l_2 norm of v .

The three matrices were stacked as input features for the autoencoder, and defined as the following functions:

Assuming that the input with n-dimensional features is denoted as

$$x = (x_1, \dots, x_n)$$

The purpose of the autoencoder is to reshape x by output x' through continuous hidden layers. The input x of each layer and the output y in layer i was connected using tanh as the activation function

$$y = f_i(x) = \tanh(W_i \cdot x + b_i)$$

Table 1. Demographic and clinical characteristics of TCGA set and five validation datasets.

Clinical feature	TCGA (N = 379)	E-GEOD-17538 (N = 232)	E-GEOD-28722 (N = 125)	E-GEOD-38832 (N = 122)	E-GEOD-39582 (N = 558)	E-GEOD-41258 (N = 141)
Gender (male/female)	157/131	122/110	–	–	308/250	71/70
Age (mean ± sd)	65.46 ± 13.26	64.73 ± 13.43	65.33 ± 12.95	–	66.81 ± 13.32	64.18 ± 13.48
OS (years, mean ± sd)	2.61 ± 2.42	3.95 ± 2.56	5.39 ± 3.53	3.37 ± 2.68	–	–
OS status (alive/dead)	219/69	139/93	55/70	94/28	–	–
DFS (years, mean ± sd)	2.40 ± 2.31	3.65 ± 2.86	4.98 ± 3.76	3.84 ± 2.77	4.06 ± 3.37	6.12 ± 4.06
DFS status (0/1)	203/62	145/55	92/33	83/9	380/177	105/36
Tumor stage (I/II/III/IV)	44/112/82/40	28/72/76/56	23/64/31/5	18/35/39/30	32/261/201/60	27/48/47/19

OS: overall survival; DFS: disease-free survival.

where x and y are two vectors of size d and p , and W_i is the weight matrix of size $p \times d$.

In the k -layer autoencoder, x' was defined as

$$x' = F_{1 \rightarrow k}(x) = f_1 \circ \dots \circ f_{k-1} \circ f_k(x)$$

$f_{k-1} \circ f_k(x) = f_{k-1}(f_k(x))$ is a composite function of f_{k-1} and $f_k(x)$. When training the autoencoder, our goal is to minimize the objective function using different weight vectors W_i ; the error between the input x and the output x' was evaluated using *Logloss* as the objective function

$$\text{logloss}(x, x') = \sum_{k=1}^d x_k \log(x'_k) + (1 - x_k) \log(1 - x'_k)$$

To prevent over-fitting, we added the *L1* regularization penalty α_w (0.001) to the weight vector W_i , and *L2* regularization penalty α_a (0.001) to the activation node $F_{1 \rightarrow k}(x)$. The following objective function was used

$$L(x, x') = \text{logloss}(x, x') + \sum_{i=1}^k (\alpha_w \|W\|_{i1} + \alpha_a \|F_{1 \rightarrow k}(x)\|_{22})$$

When training the autoencoder, the TCGA cohort was automatically separated into a training set and a validation set (2:1), and the gradient descent algorithm was used with 50% dropout and 100 epochs (batch size = 10). Supplemental Figure 1 shows that the model reaches convergence with 100 epochs in either the training or validation sets. Using the Python Keras library (<https://github.com/fchollet/keras>), an autoencoder model containing four hidden layers (700, 350, 350, and 700 nodes) and 175 bottleneck layer nodes was implemented. As a result, 175 transformed features were obtained.

K-means clustering

We built a univariate Cox proportional hazards (Cox-PH) model for each transformed feature using the R survival package (<https://cran.r-project.org/web/packages/survival/index.html>) with a log-rank $p < 0.01$ as the significance cutoff. K-means clustering was then applied to the TCGA cohort based on the survival-related features using the R NbClust package.¹⁵ The optimal number of clusters was selected by calculating the Calinski-Harabasz criterion and the silhouette index. A Kaplan-Meier (KM) curve was plotted for each cluster (risk group). The robustness of the predicted risk groups was assessed using the C-index, Log-rank P-value, and Brier score.

The C-index was considered as the proportion of all samples with corrected ordered survival times¹⁶ and was calculated using the R survcomp¹⁷ package. A C-index score of ≥ 0.7 suggests good performance of a model, whereas 0.5 suggests random background.

Log-rank P-value of Cox-PH regression reflects the survival difference of different risk groups with KM survival curves and was calculated using the R survival package.

A smaller log-rank p-value suggested better performance of the survival prediction model.

Brier score was another metric used in the survival analysis to evaluate the inaccuracy between the predicted and actual survival beyond a certain time.¹⁸ Brier score (range: 0-1) was obtained using the function sbrier.score2-proba in the R survcomp package, with larger scores indicating higher inaccuracy.

Alternative transformation approaches to the autoencoder framework

The autoencoder framework was compared to three other dimensionality-reduction techniques, including principal component analysis (PCA), t-distributed random neighbor embedded (t-SNE), and non-negative matrix factorization (NMF). As the autoencoder framework had 175 bottleneck layer nodes, PCA, t-SNE, and NMF transformed the initial features into 175 transformed features using the Python Keras library. Likewise, the survival-related features were obtained from the transformed features using univariate Cox-PH models and then used to perform K-means clustering analysis on TCGA samples, as depicted in Figure 1 (b). These methods are defined as alternative transformation strategies.

The autoencoder framework was compared to the individual Cox-PH-based strategy. Specifically, the uni-variate Cox-PH model was used for the omics data of TCGA cohort. The Top N features according to C-index were selected to cluster all TCGA samples using K-means clustering, as mentioned above. N refers to the number of survival-related features in the autoencoder-based strategy.

Hold-out validation and supervised classification

The robustness of the obtained risk groups was evaluated using hold-out validation with the R caret package as previously described (<http://topepo.github.io/caret/index.html>).¹² Briefly, all TCGA samples were randomly split into a training set and a test set randomly using a ratio of 60%:40%. Analysis of variance (ANOVA) was performed on each omics data. For RNA-seq data, DNA methylation data, miRNA-seq data, or the 3-omics data in the TCGA cohort, according to the ANOVA F-value, the top 85 mRNA features, top 30 methylation features, top 30 miRNA features, or survival-related transformed features were selected to construct separate support vector machine (SVM) models and predict the risk subgroup labels of samples in the test set.

The TCGA data were cross-validated 10 times according to the splitting strategy. Arithmetic means of the C-index and Brier score, and the geometric mean of the log-rank P-value were generated to assess the model performance. With the radial basis function as the kernel function, the svmfs function in R penalizeSVM package¹⁹ was used to construct the SVM model. For the SVM model (S), the svmfs function performs grid search to determine the optimal hyperparameters using five-fold CV and penalties, including L1 norm, elastic net (L1 + L2 norms), smooth clipped absolute deviation (SCAD), and ELastic SCAD (SCAD + L1 norm).

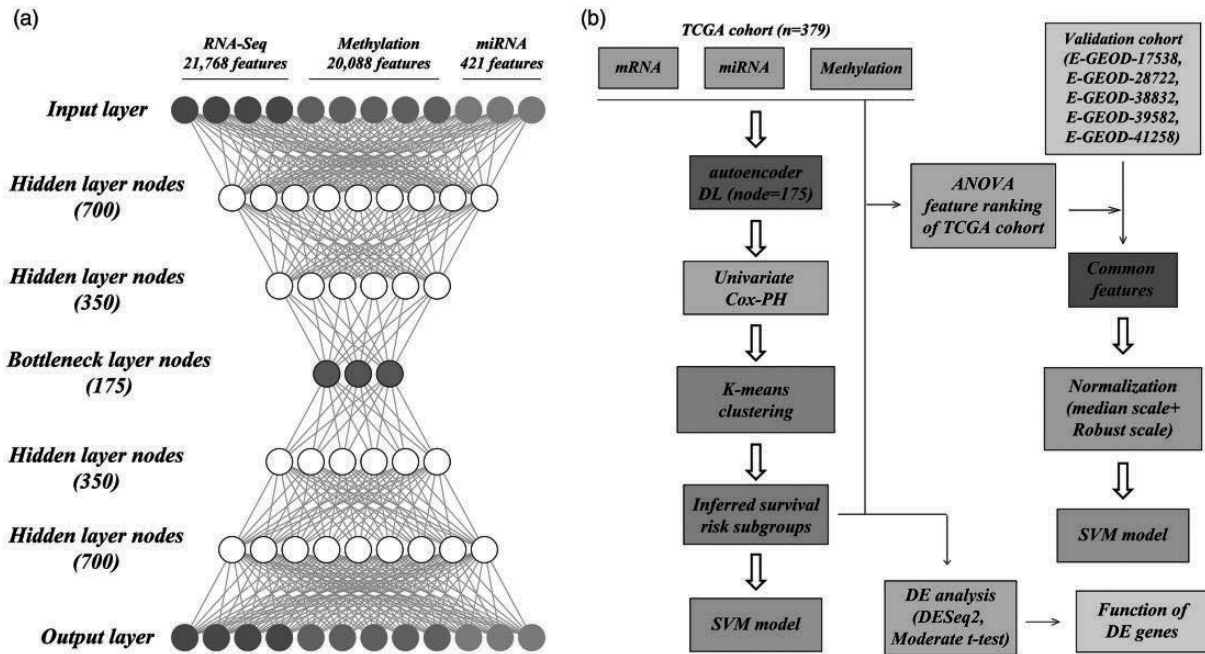


Figure 1. Overall design. (a) autoencoder framework; (b) an analytic pipeline including the autoencoder, uni-variate Cox-PH model, K-means clustering, and construction of the SVM models. (A color version of this figure is available in the online journal.)

Validation in five validation cohorts

The robustness of the obtained risk groups was verified in five validation cohorts of RNA expression data: E-GEOD-17538, E-GEOD-28722, E-GEOD-38832, E-GEOD-39582, and E-GEOD-41258. We did not find any dataset of methylation or miRNA data with the corresponding clinical information. Specifically, common features between the TCGA cohort and each validation cohort were separately selected and subjected to median scale normalization and robust scale normalization using R software.¹² The Top 85 mRNA features according to ANOVA F value were used to construct SVM models and predict risk subgroup labels of samples in each validation cohort using the above-mentioned procedure (Figure 1(b)).

Statistical analysis

Uni- and multivariate Cox regression analyses were performed to identify prognostic factors ($p < 0.05$). Using the TCGA data, differentially expressed genes and miRNAs (DEGs and DE miRNAs) were screened between different risk groups using the R *DESeq2* package,²⁰ with $|\log_2FC| > 1$ and $FDR < 0.05$ as the strict cutoff of significance. Differentially methylated genes (DMGs) ($|\beta$ difference > 0.1 and $FDR < 0.05$) were identified using the R *limma* package²¹ and moderate t -test.

The correlations between DNA methylation data and mRNA expression data from the TCGA cohort were analyzed using Pearson correlation analysis (Pearson correlation < -0.5 , p -value < 0.001). Target genes were predicted for the identified DE miRNAs using the miRDB database²² (prediction score > 80 , version 6.0, <http://mirdb.org/index.html>) and TargetScan database

(probability of conserved targeting) > 0.8 , http://www.targetscan.org/vert_72/, version 7.2). The DEGs that were predicted to be target genes of the DE miRNAs by both databases were selected to construct the miRNA-mRNA networks, which were visualized using Cytoscape software (version 3.7.1). Kyoto Encyclopedia of Genes and Genomes (KEGG) signaling pathway enrichment analysis was carried out for upregulated and downregulated DEGs, respectively, using KOBAS software²³ ($FDR < 0.05$).

The source code files of bioinformatics analyses are provided in Supplemental Files.

Results

The two survival groups obtained using the autoencoder-based strategy

The autoencoder model was trained using 421 miRNA features, 21,768 mRNA features, and 20,088 methylation features obtained from the TCGA data. Of the resulting 175 transformed features, 24 survival-related features with Log-rank P -value < 0.05 were found using the univariate Cox-PH model and used to cluster samples of the TCGA cohort using K-means clustering analysis. The optimal cluster number K was determined to be 2, and samples were dichotomized into two clusters (G1 and G2 risk groups, Supplemental Table 1). Figure 2(a) shows that the G2 group had a poorer prognosis with a significantly shorter OS than the G1 group (C-index = 0.781, Brier score = 0.198, Log-rank P -value = 1.53×10^{-7}). Therefore, the G2 group was suggested to be the more aggressive subtype, while the G1 group was regarded as the less aggressive subtype.

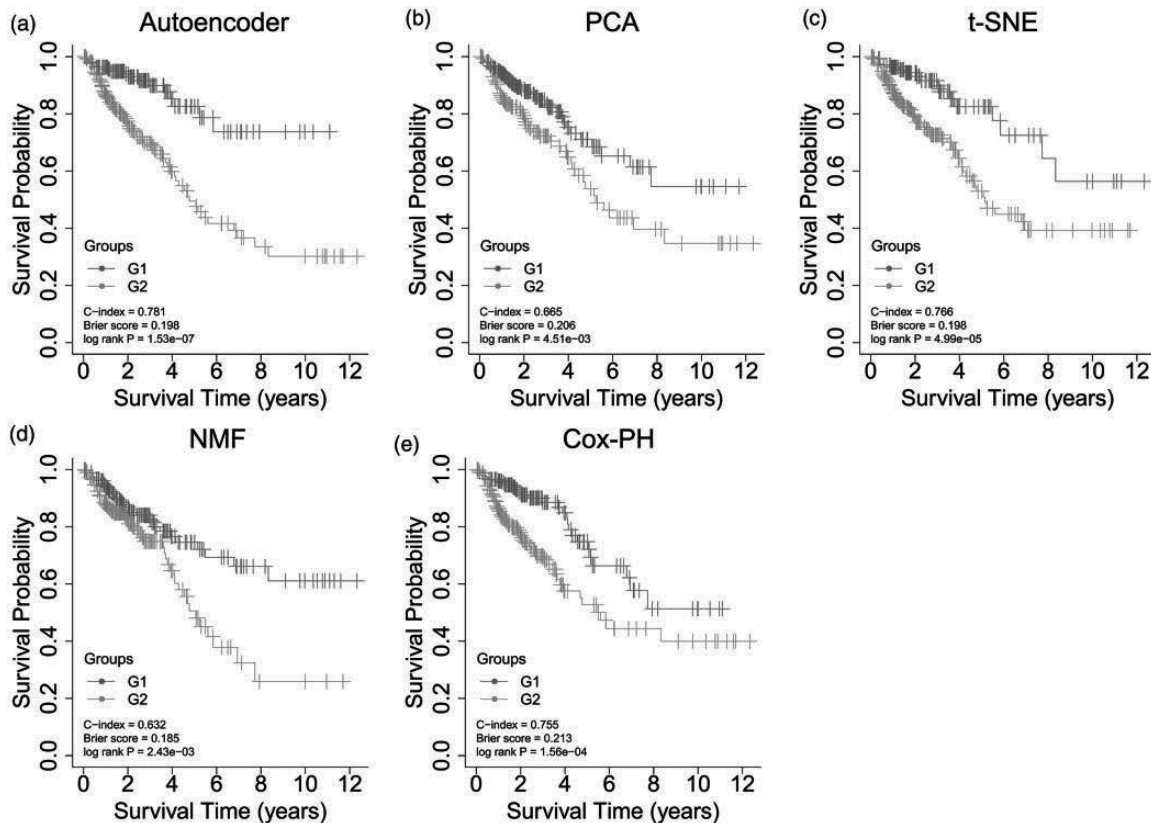


Figure 2. KM curves for TCGA set by the autoencoder- (a), PCA- (b), t-SNE- (c), NMF- (d), or individual Cox-PH-based strategy (e). (A color version of this figure is available in the online journal.)

Autoencoder-based strategy performed better than other strategies

The autoencoder-based strategy was compared with other strategies. Using PCA, t-SNE, or NMF, 175 transformed features were obtained. Among the 175 transformed features, 24 features by autoencoder, 9 features by PCA, 8 features by NMF, and 14 features by t-SNE were significantly related to survival in univariate Cox-PH regression analysis (Supplemental Table 8). Such finding indicates that the autoencoder could better capture survival-related features than other approaches.

The PCA-based strategy dichotomized samples into G1 and G2 groups and yielded a C-index of 0.665, a Brier score of 0.206, and a Log-rank P-value of 4.51e-3 (Figure 2(b)). The t-SNE-based strategy resulted in a C-index of 0.766, a Brier score of 0.198 and a Log-rank P-value of 4.99e-5 (Figure 2(c)). The NMF-based strategy generated a C-index of 0.632, a Log-rank P-value of 2.43e-3, and a Brier score of 0.185 (Figure 2(d)). The cluster labels of samples predicted by each strategy are presented in Supplemental Table 9. Comparative analysis of the clustering results using different strategies was conducted using the chi-square test. Clustering results using the autoencoder-based strategy were consistent with those of the PCA-based strategy ($\chi^2 = 30.03$, $P = 4.25e-08$) or the t-SNE-based strategy ($\chi^2 = 5.44$, $p = 0.020$, Supplementary

Figure 2). However, the NMF-based strategy did not yield a significant p-value ($\chi^2 = 2.09$, $P = 0.149$, Supplementary Figure 2).

With regard to the individual Cox-PH-based strategy, the top 24 mRNA features, miRNA features, and methylation features according to C-index were combined to cluster samples using K-means clustering analysis. The individual Cox-PH-based strategy showed a C-index of 0.755, a Brier score of 0.213, and a Log-rank P-value of 1.56e-4 (Figure 2 (e)). These results suggest that the autoencoder-based strategy is superior to other strategies.

The risk group classification by the autoencoder was an independent prognostic factor for CRC

Risk groups (P -value = 2.43E-06), age (P -value = 2.06E-02), pathologic_T (p -value = 1.49E-04), pathologic_N (P -value = 1.29E-04), pathologic_M (P -value = 2.44E-05), tumor stage (P -value = 1.78E-06), additional pharmaceutical therapy (P -value = 1.56E-02), and additional radiation therapy (P -value = 1.53E-03) were significantly related to survival in univariate Cox regression analysis (Table 2). In multivariate Cox regression analysis, the risk groups were further identified as independent prognostic factors (HR = 3.065, 95% CI = 1.746–5.380, P -value = 9.58E-05, Table 2).

Table 2. Results of uni-and multi-variate Cox regression analysis.

Clinical variable	Uni-variate				Multi-variate			
	HR	95% CI	z	p	HR	95% CI	z	p
Group				1.53E-07				
G1	1.000	–	–	–	1.000	–	–	–
G2	3.588	2.109–6.103	4.714	2.43E-06	3.065	1.746–5.380	3.901	9.58E-05
Age				2.06E-02				
≥60	1.759	1.066–2.902	2.211	2.70E-02	2.162	1.243–3.760	2.731	6.31E-03
<60	1.000	–	–	–	1.000	–	–	–
Pathologic_T				1.49E-04				
T1	0.161	0.010–2.582	–1.290	1.97E-01	0.066	0.001–7.678	–1.120	2.63E-01
T2	0.250	0.030–2.085	–1.281	2.00E-01	0.141	0.002–10.881	–0.883	3.77E-01
T3	0.390	0.054–2.831	–0.932	3.52E-01	0.075	0.001–5.721	–1.171	2.42E-01
T4	1.410	0.188–10.606	0.334	7.38E-01	0.172	0.002–13.652	–0.789	4.30E-01
Pathologic_N				1.29E-04				
N0	0.420	0.057–3.083	–0.853	3.94E-01	12.603	0.156–1017.827	1.131	2.58E-01
N1	0.769	0.104–5.693	–0.257	7.97E-01	7.654	0.092–633.555	0.903	3.66E-01
N2	1.391	0.189–10.256	0.323	7.46E-01	11.759	0.142–970.721	1.095	2.74E-01
Pathologic_M				2.44E-05				
M0	0.578	0.330–1.012	–1.917	5.53E-02	0.625	0.328–1.190	–1.431	1.52E-01
M1	2.119	1.124–3.995	2.321	2.03E-02	1.363	0.280–6.630	0.383	7.01E-01
Stage				1.78E-06				
I	0.131	0.038–0.450	–3.230	1.24E-03	0.127	0.024–0.673	–2.423	1.54E-02
II	0.242	0.102–0.571	–3.237	1.21E-03	0.232	0.063–0.855	–2.196	2.81E-02
III	0.422	0.183–0.975	–2.020	4.34E-02	0.570	0.161–2.014	–0.873	3.83E-01
IV	0.999	0.425–2.350	–0.002	9.98E-01	0.597	0.124–2.873	–0.644	5.20E-01
Additional pharmaceutical therapy				1.56E-02				
No	2.249	1.252–4.040	2.712	6.68E-03	0.560	0.203–1.542	–1.122	2.62E-01
Yes	1.738	0.981–3.078	1.894	5.82E-02	0.396	0.143–1.094	–1.787	7.39E-02
Additional radiation therapy				1.53E-03				
NO	2.382	1.508–3.763	3.720	1.99E-04	3.296	1.349–8.052	2.618	8.86E-03
YES	1.707	0.730–3.988	1.234	2.17E-01	2.586	0.856–7.807	1.685	9.20E-02

Table 3. CV-based performance metrics on training and test set of TCGA cohort.

Dataset	10-fold cv	C-index	Brier score	Log-rank-p (geo.mean)
Training	3-omics training (60%)	0.74 ± 0.08	0.20 ± 0.01	1.53e-4
	RNA only	0.72 ± 0.06	0.20 ± 0.02	5.10e-4
	miRNA only	0.72 ± 0.09	0.20 ± 0.01	1.16e-3
	Methylation only	0.69 ± 0.10	0.18 ± 0.02	6.90e-4
Test	3-omics test (40%)	0.65 ± 0.11	0.20 ± 0.02	4.40e-2
	RNA only	0.69 ± 0.10	0.20 ± 0.02	1.31e-2
	miRNA only	0.71 ± 0.08	0.20 ± 0.01	2.74e-2
	Methylation only	0.63 ± 0.11	0.18 ± 0.03	2.77e-2

Robustness of the risk groups based on hold-out validation

We used hold-out validation to verify the robustness of the obtained risk groups using 3-omics data and each single omic data of the TCGA cohort. In the training set, an SVM model was built and then in the test set the risk group labels were predicted. Both the 3-omics training set and the 3-omics test set generated a high C-index (0.74 ± 0.08; 0.65 ± 0.11), low Brier score (0.20 ± 0.01; 0.20 ± 0.02), and significant log-rank P-value (1.53e-4; 4.40e-2, Table 3).

When single omic data were employed, the top 85 RNA features, top 30 methylation features, and top 30 miRNA features according to ANOVA F-value were selected to construct the SVM model (Supplemental Table 2). When the methylation data were tested, the training set yielded a

C-index of 0.69 ± 0.10, a log-rank P-value of 6.90e-4, and a Brier score of 0.18 ± 0.02, while the test set yielded a C-index of 0.63 ± 0.11, a log-rank P-value of 2.77e-2, and a Brier score of 0.18 ± 0.03 (Table 3). The SVM model also displayed good performances for the training set and test set of RNA-seq and miRNA-seq (Table 3). It demonstrates the robustness of the obtained risk groups and suggests that multi-omics data outperform single omics data at predicting survival.

Successful verification of the obtained risk groups in five validation cohorts

With regard to the five validation cohorts, E-GEOD-17538 (N = 232), E-GEOD-28722 (N = 125), E-GEOD-38832 (N = 122), E-GEOD-39582 (N = 558), and E-GEOD-41258

(N=141), shared 12974, 12486, 11759, 12974, and 11628 common mRNA features with TCGA cohort, respectively. The Top 85 common mRNA features according to ANOVA F-values were selected to build the SVM model to predict the risk group label of samples in the corresponding validation cohort. E-GEOD-17538 had a C-index of 0.651, a Log-rank P-value of $1.33\text{e-}2$, and a Brier score of 0.201 for OS (Figure 3(a)) and a C-index of 0.662, a Log-rank P-value of $3.33\text{e-}3$, and a Brier score of 0.149 for disease-free survival (DFS) (Figure 3(b)). In E-GEOD-28722, significantly different DFS was found for the two risk groups (C-index=0.717, Brier score=0.187, Log-rank P-value = $2.37\text{e-}2$, Figure 3(c)). The log-rank P-value of E-GEOD-38832, the smallest cohort, was marginally significant ($8.70\text{e-}2$, Figure 3(d)), which might be attributed to the small sample size. Either E-GEOD-39582 (C-index = 0.609, Brier score = 0.217, Log-rank P-value = $4.76\text{e-}3$, Figure 3(e)) or E-GEOD-41258 (C-index = 0.678, Brier score = 0.188, Log-rank P-value = $3.03\text{e-}2$, Figure 3(f)) was partitioned into two risk groups with significantly different DFS. These results reveal that the autoencoder-based strategy for risk stratification can be generalized to other cohorts of patients with CRC.

Function analysis of the survival groups of the TCGA cohort

The TCGA cohort was divided into G1 and G2 groups using the autoencoder-based strategy. We identified 708

significant DEGs ($|\log_2\text{FC}| > 1$ and $\text{FDR} < 0.05$), including 368 upregulated genes and 340 downregulated genes in the group G2 relative to the group G1 (Supplemental Table 3). A total of 31 DE miRNAs met the cutoff of $|\log_2\text{FC}| > 0.585$ and $\text{FDR} < 0.05$, including 22 upregulated and 9 downregulated DE miRNAs in the group G2 compared to the group G1 (Supplemental Table 4). Moreover, SLAMF6 and CASP5 with $\text{FDR} < 0.05$ and $|\text{delta methylation}| > 0.1$ were significantly hypermethylated in the group G2 compared to the group G1 (Supplemental Table 5). The top 10 DEGs, top 10 DE miRNAs, and top 10 DMGs according to FDR value are displayed in Figure 4. Distinctive expression patterns of these features were observed between the G1 and G2 groups.

These results indicate that the two survival groups have distinct genomic properties. Gene expression was not affected by promoter methylation in any gene using Pearson correlation analysis (correlation coefficient < -0.5 , $P < 0.001$).

Of the 708 identified DEGs, 19 were predicted to be targets of six DE miRNAs using the miRDB database (prediction score > 80) and the TargetScan database ($\text{Pct} > 0.8$) (Supplemental Table 6). These miRNA-mRNA pairs were used to construct miRNA-mRNA networks, which included eight target genes (*B3GNT7*, *SCUBE3*, *GDF6*, *BNC2*, *PCDH19*, *PEG10*, *PLA2G3*, and *CPEB1*) of upregulated let-7c, five target genes (*SCN2B*, *NAV3*, *KCNK3*, *INA*, and *CPLX2*) of upregulated mir-34c, two target genes

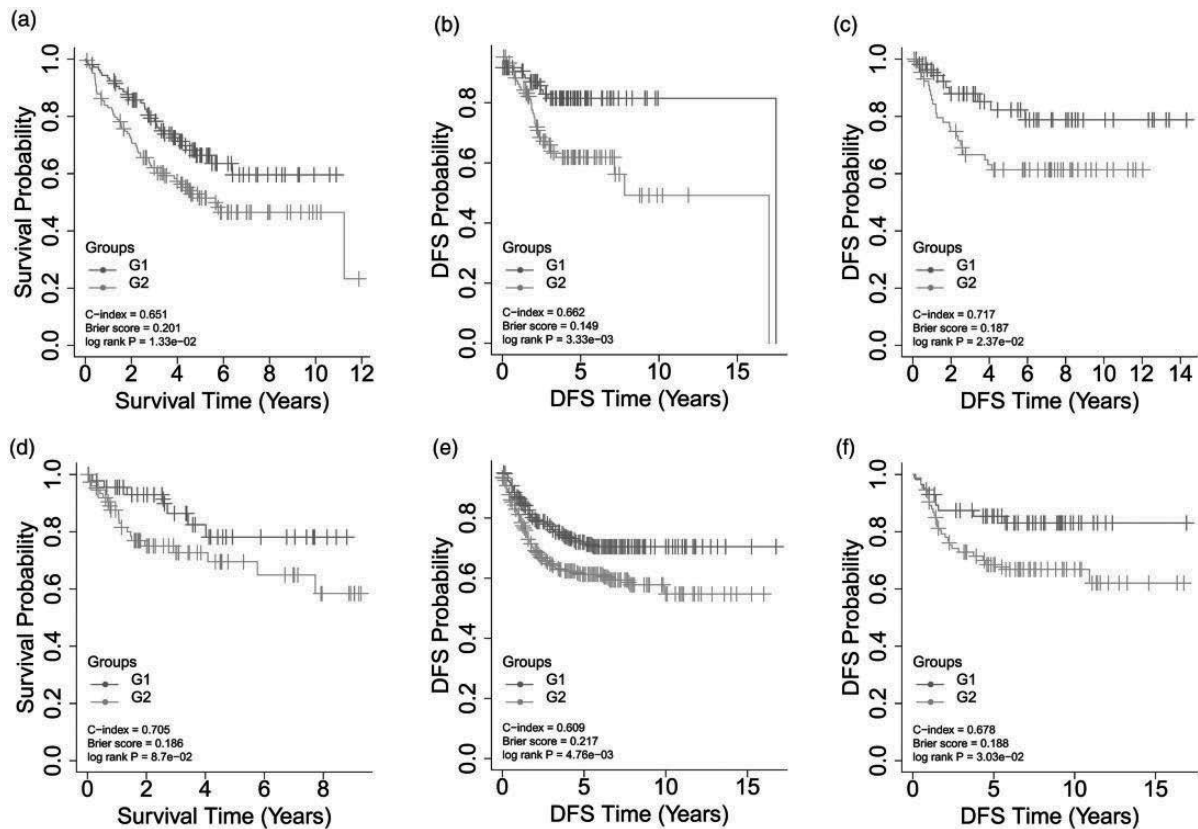


Figure 3. KM curves for the five validation datasets. (a) KM curves for OS in E-GEOD-17538; (b) KM curves for DFS in E-GEOD-17538; (c) KM curves for DFS in E-GEOD-28722; (d) KM curves for OS in E-GEOD-38832; (e) KM curves for DFS in E-GEOD-39582; (f) KM curves for DFS in E-GEOD-41258. (A color version of this figure is available in the online journal.)

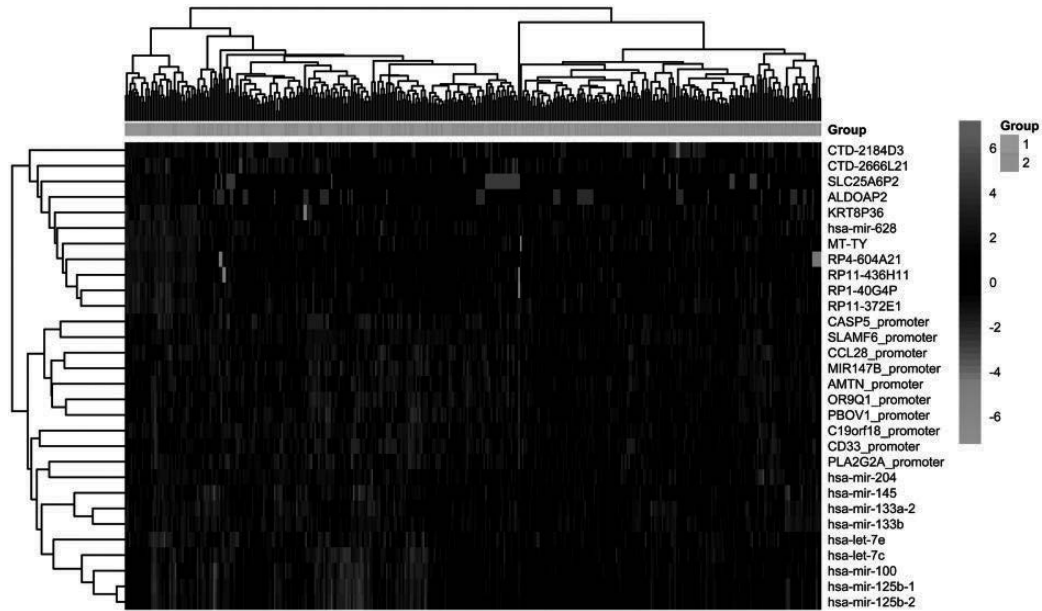


Figure 4. Heatmap of the clustering analysis results of the top 10 DEGs, top 10 DEmiRNAs, and top 10 DMGs. Blue and orange bars represent the G1 and G2 groups, respectively. (A color version of this figure is available in the online journal.)

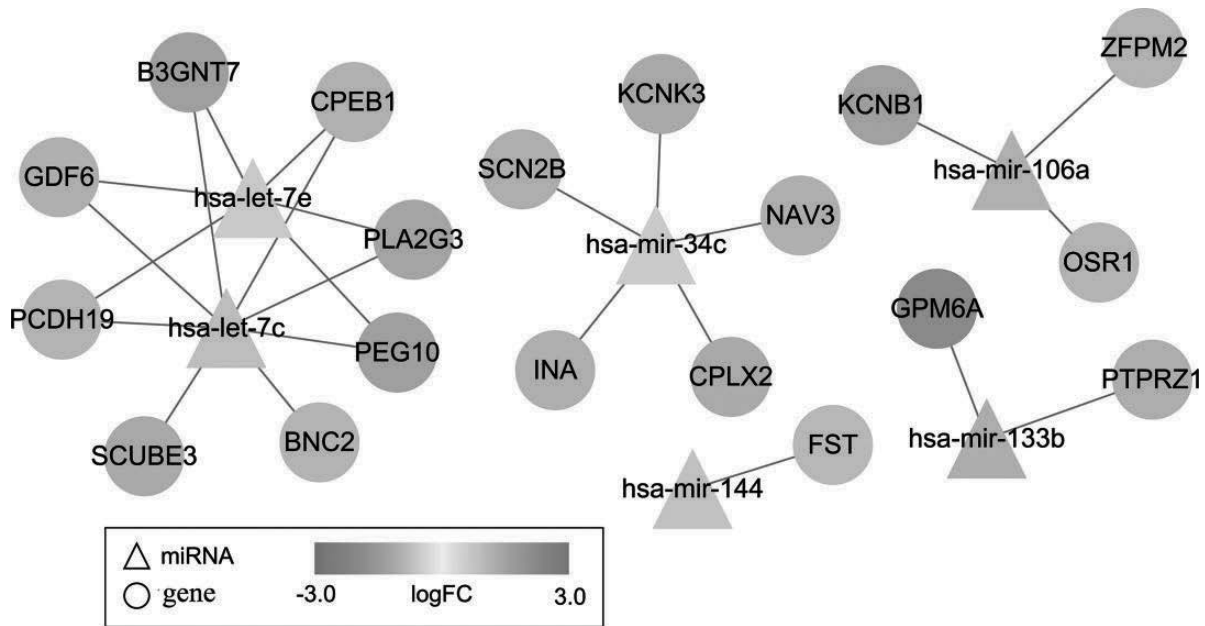


Figure 5. miRNA-gene networks. Six DEmiRNAs and 19 predicted target DEGs are included in the networks. (A color version of this figure is available in the online journal.)

(*GPM6A* and *PTPRZ1*) of upregulated mir-133b, three target genes (*KCNB1*, *OSR1*, and *ZFPM2*) of downregulated mir-106a, six target genes (*B3GNT7*, *PCDH19*, *PEG10*, *GDF6*, *PLA2G3*, and *CPEB1*) of upregulated let-7e, and one target gene (*FST*) of downregulated mir-144 (Figure 5). Only *B3GNT7* and *PLA2G3* were downregulated DEGs, while the other 17 genes were upregulated DEGs. Additionally, let-7c, mir-133b, and let-7e were among the top 10 DEmiRNAs mentioned above.

Using the upregulated and downregulated DEGs, we performed KEGG pathway enrichment analysis,

respectively. Specifically, 27 signaling pathways were significantly enriched with the upregulated DEGs, including Wnt signaling pathway, calcium signaling pathway, and PI₃K-Akt signaling pathway, whereas 25 signaling pathways significantly involved the downregulated DEGs, such as pancreatic secretion, nitrogen metabolism, and intestinal immune network for IgA production pathways (Figure 6, Supplemental Table 7). These results suggest that different signaling pathways are implicated in the carcinogenic mechanisms of the G1 and G2 groups.

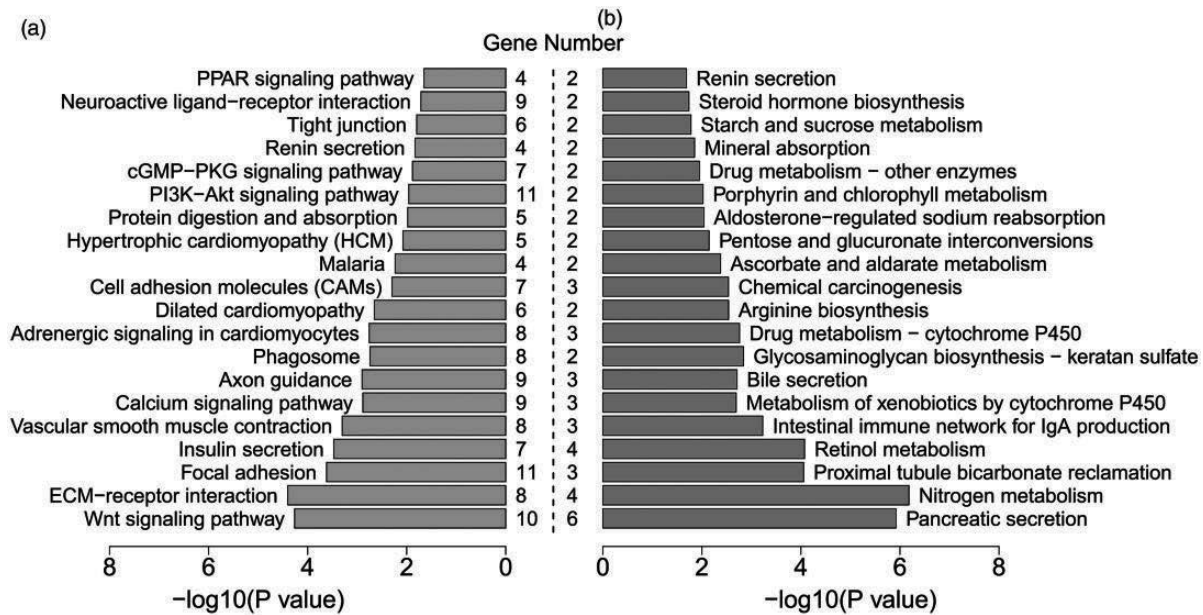


Figure 6. Top 20 significant signaling pathways for upregulated (a) and downregulated (b) DEGs. Gene number suggests the count of genes significantly enriched in a pathway. (A color version of this figure is available in the online journal.)

Discussion

CRC is characterized by high tumor heterogeneity, varied outcomes, and differential drug responses; thus, it is a difficult task to predict the survival of patients with CRC.²⁴ Mortality risk stratification in CRC patients would facilitate a deeper understanding of the molecular biology of CRC and a more precise use of individualized therapies, thereby leading to an improvement in the outcome of patients.²⁵ Our study successfully dichotomized CRC patients into two survival subpopulations using an autoencoder-based model. DNA methylation, RNA-seq, and miRNA-seq data of CRC samples from TCGA were integrated as a single input vector for the autoencoder framework to yield transformed features. Two optimal risk groups were obtained using k-means clustering and were used to build the SVM model. The prognostic value of the autoencoder-based model was robustly verified using hold-out validation and five validation cohorts. The risk groups were not only significantly related to survival, but also had significant prognostic value independent of clinical factors. To the best of our knowledge, this is the first study to employ an autoencoder-based model to capture the survival-related multi-omics features of CRC.

Incorporating multi-omics data enables the extraction of coherent biological features and is of remarkable significance for tailored medicine and management.²⁶ Multi-omics analysis data possess strong predictive power and can overcome the shortcomings of one-omics analysis, such as non-universality, uniqueness, and noisy data.²⁷ In this study, an autoencoder was applied for multi-omics integration to regenerate new representative features of CRC. Compared to other DL algorithms, the autoencoder does not require training data to be labeled, regenerates a smaller size of encodings, and has low computational complexity, with similar input and output data.²⁸ Owing to

these prominent advantages, autoencoders have been frequently used in the field of medical data processing.⁵ Furthermore, the current study showed that the autoencoder framework outperformed PCA, t-SNE, NMF, and the univariate Cox-PH-based strategy in identifying survival-related features of CRC. Our results suggest that the autoencoder is a feasible and reliable tool for multi-omics integration.

Our findings implied that the good prognosis group (G1) had more aggressive behaviors, whereas the poor prognosis group (G2) had less aggressive behaviors. The two groups have distinct carcinogenic mechanisms at the gene, miRNA, methylation, and pathway levels. Our results also indicate that the six DEMiRNAs (let-7c, mir-34c, mir-133b, let-7e, mir-144, and mir-106a) and their predicted 19 target genes in the constructed miRNA-mRNA networks are critical participants in the differential molecular mechanisms of the two groups and are important causes of their distinctive outcomes. Studies have reported that let-7c-5p is associated with the prognosis of CRC and may play a critical role in the progression of CRC.^{29,30} Previous reports have also revealed that another DEMiRNA, let-7e, is downregulated in CRC and exerts suppressive effects on the proliferation and migration of CRC cells.^{31,32} In the miRNA-mRNA networks constructed in our study, let-7c and let-7e shared six common target DEGs, including downregulated *B3GNT7* and *PLA2G3*, and upregulated *PCDH19*, *PEG10*, *GDF6*, and *CPEB1*. These findings imply that let-7c and let-7e promote the progression and migration of CRC by downregulating *B3GNT7* and *PLA2G3*, and upregulating *PCDH19*, *PEG10*, *GDF6*, and *CPEB1*. Beta-1, 3-N-acetylglucosaminyltransferase 7 (*B3GNT7*) is suppressed in colon oncogenic processes to promote the cancer metastasis by increasing the expression of cell surface sialyl Lewis x and x antigens.³³ Phospholipase A2 group III (*PLA2G3*) plays an oncogenic

and pro-inflammatory role in the biology of CRC and is a risk factor for lymph node metastasis as well as a prognostic biomarker.³⁴ Downregulation of protocadherin 19 (*PCDH19*), a tumor suppressor gene, promotes the metastasis and proliferation of hepatocellular carcinoma and is a predictor of poor prognosis.^{35,36} Paternally expressed gene 10 (*PEG10*) is involved in rectal adenocarcinoma metastasis³⁷ and contributes to tumor cell proliferation and invasion.³⁸ Growth differentiation factor 6 (*GDF6*), a member of the transforming growth factor (TGF)- β family, was significantly enriched in the TGF- β signaling pathway in our study, which is involved in modulating cell growth, differentiation, apoptosis, and homeostasis.³⁹ Cytoplasmic polyadenylation element-binding protein 1 (*CPEB1*), encoding a key component of tight junctions, mediates epithelial-to-mesenchymal transition (EMT) and the metastasis of breast cancer by regulating matrix metalloproteinase 9.⁴⁰

miR-34c-5p promotes the proliferation of cancer cells and CRC metastasis.⁴¹ In the miRNA-mRNA networks, miR-34c was predicted to target *SCN2B*, *NAV3*, *KCNK3*, *INA*, and *CPLX2*, indicating that the five target genes mediate the promoting effect of miR-34c on CRC. Sodium voltage-gated channel beta subunit 2 (*SCN2B*) encodes cell adhesion molecules and is associated with the adhesion and migration of breast cancer cells.⁴² Navigator 3 (*NAV3*) acts as a target gene of p73 to inhibit colon cancer metastasis⁴³ and is associated with colon cancer development-related inflammation.⁴⁴ TWIK-related acid-sensitive potassium channel 1 (*TASK-1*) encoded by potassium channel subfamily K member 3 (*KCNK3*) participates in the regulation of apoptosis and proliferation of lung cancer cells.⁴⁵ α -internexin (*INA*) inhibits microtubule polymerization in early stage CRC.⁴⁶ Complexin-2 (*CPLX2*) is upregulated in high-grade lung neuroendocrine tumors and serves as a potential prognostic biomarker.⁴⁷

Accumulating evidence shows that miR-133b is a tumor suppressor involved in the regulation of CRC cell proliferation and apoptosis.^{48,49} We found that miR-133b was upregulated in the poor prognosis group relative to the good prognosis group and was predicted to target *GPM6A* and *PTPRZ1*. Glycoprotein M6A (*GPM6A*), an oncogene, is related to EMT and cell migration in gonadotroph pituitary adenomas.⁵⁰ *GPM6A* expression at the mRNA and protein levels is negatively regulated by miR-133b during prenatal stress,⁵¹ which aligns with our results. Protein tyrosine phosphatase receptor-like type Z polypeptide 1 (*PTPRZ1*) shows increased expression in CRC tissues and contributes to carcinogenesis.⁵²

miR-106a, a tumor suppressor, suppresses cell proliferation and strengthens cell apoptosis in CRC,⁵³ which aligns with our results that Hsa-mir-106a is downregulated in the poor prognosis group, targeting upregulated *KCNB1*, *OSR1*, and *ZFPM2*. Potassium channel subfamily B member 1 (*KCNB1*) expression is decreased in gastric and colorectal cancers and may be a promising prognostic biomarker.⁵⁴ Oxidative stress-responsive kinase 1 (*OSR1*) is related to the angiogenesis and proliferation of hepatoma cells.⁵⁵ The expression of zinc finger protein, FOG family member 2 (*ZFPM2*), a glioma susceptibility gene, is associated with the incidence and severity of glioma.⁵⁶

Additionally, the present study showed that downregulated miRNA-144 targeted FST, which was significantly enriched in the TGF- β signaling pathway. Consistently, there is a rich body of evidence that miRNA-144 functions as a tumor suppressor in CRC.^{57,58} Follistatin (FST) acts as an antagonist of the TGF- β family and participates in the tumorigenesis of solid tumors.⁵⁹ FST and TGF- β are inferred to mediate the inhibitory effect of miRNA-144 on CRC. Our study supports these miRNAs and target genes as potential prognostic biomarkers and therapeutic targets of CRC, providing more insights into the complicated regulatory mechanisms of CRC.

Our study illustrates an applicable analytical model for the survival prediction of CRC based on the integration of multi-omics data. However, some limitations of this study should not be ignored. Further improvements, such as the inclusion of clinical information as an additional modality, are beneficial for improving the performance of our autoencoder-based model before its application to other cancers. Further, the prognostic miRNAs and target genes should be validated using *in vivo* and *in vitro* experiments in future studies. Herein, the model was trained using coupled DNA methylation, RNA-seq, and miRNA-seq data from 379 tumor samples. Larger datasets with more tumor samples are expected to yield better results.

AUTHORS' CONTRIBUTIONS

All authors participated in the design, interpretation of the studies, analysis of the data, and review of the manuscript; HS and JS wrote the article.


DECLARATION OF CONFLICTING INTERESTS

The author(s) declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

FUNDING

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Project of Invigorating Health Care through Science, Technology and Education for Jiangsu Provincial Medical Youth Talent (grant number QNRC2016790).

ORCID ID

Jun Song  <https://orcid.org/0000-0003-4588-6522>

SUPPLEMENTAL MATERIAL

Supplemental material for this article is available online.

REFERENCES

1. Harada S, Morlote D. Molecular pathology of colorectal cancer. *Adv Anat Pathol* 2020;27:20-6
2. Ladabaum U, Dominitz JA, Kahi C, Schoen RE. Strategies for colorectal cancer screening. *Gastroenterology* 2020;158:418-32

3. Dekker E, Tanis PJ, Vleugels JLA, Kasi PM, Wallace MB. Colorectal cancer. *Lancet (London, England)* 2019;**394**:1467–80
4. Chan HP, Samala RK, Hadjiiski LM, Zhou C. Deep learning in medical image analysis. *Adv Exp Med Biol* 2020;**1213**:3–21
5. Pacal I, Karaboga D, Basturk A, Akay B, Nalbantoglu U. A comprehensive review of deep learning in colon cancer. *Comput Biol Med* 2020;**126**:1–33
6. Sánchez-Peralta LF, Bote-Curiel L, Picón A, Sánchez-Margallo FM, Pagador JB. Deep learning to find colorectal polyps in colonoscopy: a systematic literature review. *Artif Intell Med* 2020;**108**:1–23
7. Hutter C, Zenklusen JC. The cancer genome atlas: creating lasting value beyond its data. *Cell* 2018;**173**:283–5
8. Ma T, Zhang A. Integrate multi-omics data with biological interaction networks using multi-view factorization AutoEncoder (MAE). *BMC Genom* 2019;**20**:1–11
9. Huang Z, Zhan X, Xiang S, Johnson TS, Helm B, Yu CY, Zhang J, Salama P, Rizkalla M, Han Z, Huang K. SALMON: survival analysis learning with multi-omics neural networks on breast cancer. *Front Genet* 2019;**10**:1–13
10. Zhang L, Lv C, Jin Y, Cheng G, Fu Y, Yuan D, Tao Y, Guo Y, Ni X, Shi T. Deep learning-based multi-omics data integration reveals two prognostic subtypes in high-risk neuroblastoma. *Front Genet* 2018;**9**:1–9
11. Xu X, Gu H, Wang Y, Wang J, Qin P. Autoencoder based feature selection method for classification of anticancer drug response. *Front Genet* 2019;**10**:1–10
12. Chaudhary K, Poirion OB, Lu L, Garmire LX. Deep learning-based multi-omics integration robustly predicts survival in liver cancer. *Clin Cancer Res* 2018;**24**:1248–59
13. IlluminaHumanMethylation450kanno.HKilmn12. hg19: annotation for illumina's 450k methylation arrays. 2015. *R package, version 02*.
14. Moorthy K, Jaber AN, Ismail MA, Ernawan F, Mohamad MS, Deris S. Missing-values imputation algorithms for microarray gene expression data. *Meth Mol Biol (Clifton, NJ)* 2019;**1986**:255–66
15. Charrad M, Ghazzali N, Boiteau V, Niknafs A. NbClust: an R package for determining the relevant number of clusters in a data set. *Journal of Statal Software* 2014;**061**:1–36
16. Cheung LC, Pan Q, Hyun N, Katki HA. Prioritized concordance index for hierarchical survival outcomes. 2019;**38**:2868–82
17. Schröder MS, Culhane AC, Quackenbush J, Haihe-Kains B. Survcomp: an R/bioconductor package for performance assessment and comparison of survival models. *Bioinformatics (Oxford, England)* 2011;**27**:3206–8
18. Gerds TA, Schumacher M. Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biom J* 2010;**48**:1029–40
19. Becker N, Werft W, Toedt G, Lichter P, Benner A. penalizedSVM: a R-package for feature selection SVM classification. *Bioinformatics* 2009;**25**:1711–2
20. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:1–21
21. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* 2015;**43**:1–13
22. Chen Y, Wang X. miRDB: an online database for prediction of functional microRNA targets. *Nucleic Acids Res* 2020;**48**:D127–d31
23. Xie C, Mao X, Huang J, Ding Y, Wu J, Dong S, Kong L, Gao G, Li CY, Wei L. KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Res* 2011;**39**:316–22
24. Lim LC, Lim YM. Proteome heterogeneity in colorectal cancer. *Proteomics* 2018;**18**:1–39
25. Punt CJ, Koopman M, Vermeulen L. From tumour heterogeneity to advances in precision treatment of colorectal cancer. *Nat Rev Clin Oncol* 2017;**14**:235–46
26. Huang S, Chaudhary K, Garmire LX. More is better: recent progress in multi-omics data integration methods. *Front Genet* 2017;**8**:1–12
27. Wang F, Han J. Multimodal biometric authentication based on score level fusion using support vector machine. *Opto-Electronics Rev* 2009;**17**:59–64
28. Yu S, Príncipe JC. Understanding autoencoders with information theoretic concepts. *Neural Netw* 2019;**117**:104–23
29. Zhou XG, Huang XL, Liang SY, Tang SM, Wu SK, Huang TT, Mo ZN, Wang QY. Identifying miRNA and gene modules of colon cancer associated with pathological stage by weighted gene co-expression network analysis. *Oncotargets Ther* 2018;**11**:2815–30
30. Yan M, Song M, Bai R, Cheng S, Yan W. Identification of potential therapeutic targets for colorectal cancer by bioinformatics analysis. *Oncol Lett* 2016;**12**:5092–8
31. Li Z, Pan W, Shen Y, Chen Z, Zhang L, Zhang Y, Luo Q, Ying X. IGF1/IGF1R and microRNA let-7e down-regulate each other and modulate proliferation and migration of colorectal cancer cells. *Cell Cycle (Georgetown, Tex)* 2018;**17**:1212–9
32. Shan Y, Liu Y, Zhao L, Liu B, Li Y, Jia L. MicroRNA-33a and let-7e inhibit human colorectal cancer progression by targeting ST8SIA1. *Int J Biochem Cell Biol* 2017;**90**:48–58
33. Lu CH, Wu WY, Lai YJ, Yang CM, Yu LC. Suppression of B3GNT7 gene expression in colon adenocarcinoma and its potential effect in the metastasis of colon cancer cells. *Glycobiology* 2014;**24**:359–67
34. Murase R, Taketomi Y, Miki Y, Nishito Y, Saito M, Fukami K, Yamamoto K, Murakami M. Group III phospholipase A(2) promotes colitis and colorectal cancer. *Sci Rep* 2017;**7**:1–13
35. Zhang T, Guan G, Chen T, Ji, Zhang L, Yao M, Qi X, Zou J, Chen J, Lu F, Chen X. Methylation of PCDH19 predicts poor prognosis of hepatocellular carcinoma. 2018;**14**:e352–e58
36. Yao X, Zhang H, Liu Y, Liu X, Wang X, Sun X, Cheng Y. miR-99b-3p promotes hepatocellular carcinoma metastasis and proliferation by targeting protocadherin 19. *Gene* 2019;**698**:141–9
37. Hua Y, Ma X, Liu X, Yuan X, Qin H, Zhang X. Identification of the potential biomarkers for the metastasis of rectal adenocarcinoma. *Apmis* 2017;**125**:93–100
38. Watson KM, Gardner IH, Byrne RM, Ruhl RR, Lanciault CP, Dewey EN, Anand S, Tsikitis VL. Differential expression of PEG10 contributes to aggressive disease in early versus late-onset colorectal cancer. *Dis Colon Rectum* 2020;**63**:1610–20
39. Itatani Y, Kawada K. Transforming growth factor- β signaling pathway in colorectal cancer and its tumor microenvironment. *Int J Mol Sci* 2019;**20**:1–16
40. Nagaoka K, Fujii K, Zhang H, Usuda K, Watanabe G, Ivshina M, Richter JD. CPEB1 mediates epithelial-to-mesenchyme transition and breast cancer metastasis. *Oncogene* 2016;**35**:2893–901
41. Gu J, Wang G. SATB2 targeted by methylated miR-34c-5p suppresses proliferation and metastasis attenuating the epithelial-mesenchymal transition in colorectal cancer. *Cell Prolif* 2018;**51**:1–12
42. Chioni AM, Brackenbury WJ, Calhoun JD, Isom LL, Djamgoz MB. A novel adhesion molecule in human breast cancer cells: voltage-gated Na⁺ channel beta1 subunit. *Int J Biochem Cell Biol* 2009;**41**:1216–27
43. Uboveja A, Satija YK, Siraj F, Sharma I, Saluja D. p73-NAV3 axis plays a critical role in suppression of colon cancer metastasis. *Oncogenesis* 2020;**9**:1–15
44. Carlsson E, Ranki A, Sipilä L, Karenko L, Abdel-Rahman WM, Ovaska K, Siggberg L, Aapola U, Ässämäki R, Häyry V, Niiranen K, Helle M, Knuutila S, Hautaniemi S, Peltomäki P, Krohn K. Potential role of a navigator gene NAV3 in colorectal cancer. *Br J Cancer* 2012;**106**:517–24
45. Leithner K, Hirschmugl B, Li Y, Tang B, Papp R, Nagaraj C, Stacher E, Stiegler P, Lindenmann J, Olschewski A, Olschewski H, Hrzjenjak A. TASK-1 regulates apoptosis and proliferation in a subset of non-small cell lung cancers. *PLoS One* 2016;**11**:1–18
46. Li Y, Bai L, Yu H, Cai D, Wang X, Huang B, Peng S, Huang M, Cao G, Kaz AM, Grady WM, Wang J, Luo Y. Epigenetic inactivation of α -inter-nexin accelerates microtubule polymerization in colorectal cancer. *Cancer Res* 2020;**80**:5203–15
47. Komatsu H, Kakehashi A, Nishiyama N, Izumi N, Mizuguchi S, Yamano S, Inoue H, Hanada S, Chung K, Wei M, Suehiro S, Wanibuchi H. Complexin-2 (CPLX2) as a potential prognostic biomarker in human lung high grade neuroendocrine tumors. *Cbm* 2013;**13**:171–80
48. Lv LV, Zhou J, Lin C, Hu G, Yi LU, Du J, Gao K, Li X. DNA methylation is involved in the aberrant expression of miR-133b in colorectal cancer cells. *Oncol Lett* 2015;**10**:907–12

49. Lv L, Li Q, Chen S, Zhang X, Tao X, Tang X, Wang S, Che G, Yu Y, He L. miR-133b suppresses colorectal cancer cell stemness and chemoresistance by targeting methyltransferase DOT1L. *Exp Cell Res* 2019;**385**:1–25
50. Falch CM, Sundaram AYM, Øystese KA, Normann KR, Lekva T, Silamikelis I, Eieland AK, Andersen M, Bollerslev J, Olarescu NC. Gene expression profiling of fast- and slow-growing non-functioning gonadotroph pituitary adenomas. *Eur J Endocrinol* 2018;**178**:295–307
51. Monteleone MC, Adrover E, Pallarés ME, Antonelli MC, Frasc AC, Brocco MA. Prenatal stress changes the glycoprotein GPM6A gene expression and induces epigenetic changes in rat offspring brain. *Epigenetics* 2014;**9**:152–60
52. Laczmanska I, Karpinski P, Gil J, Laczmanski L, Bebenek M, Sasiadek MM. High PTPRQ expression and its relationship to expression of PTPRZ1 and the presence of KRAS mutations in colorectal cancer tissues. *Anticancer Res* 2016;**36**:677–81
53. Huang Q, Ma Q. MicroRNA-106a inhibits cell proliferation and induces apoptosis in colorectal cancer cells. *Oncol Lett* 2018;**15**:8941–4
54. Farah A, Kabbage M, Atafi S, Gabteni AJ, Barbirou M, Madhioub M, Hamzaoui L, Mohamed MA, Touinsi H, Kchaou AO. Selective expression of KCNA5 and KCNB1 genes in gastric and colorectal carcinoma. *BMC Cancer* 2020;**20**:1–9
55. Sie ZL, Li RY, Sampurna BP, Hsu PJ, Liu SC, Wang HD. WNK1 kinase stimulates angiogenesis to promote tumor growth and metastasis. *Cancers* 2020;**12**:1–20
56. Tsang SY, Mei L, Wan W, Li J, Li Y, Zhao C, Ding X, Pun FW, Hu X, Wang J, Zhang J, Luo R, Cheung ST, Leung GK, Poon WS, Ng HK, Zhang L, Xue H. Glioma association and balancing selection of ZFPM2. *PLoS One* 2015;**10**:e0133003
57. Xiao R, Li C, Chai B. miRNA-144 suppresses proliferation and migration of colorectal cancer cells through GSPT1. *Biomed Pharmacother* 2015;**74**:138–44
58. Sheng S, Xie L, Wu Y, Ding M, Zhang T, Wang X. MiR-144 inhibits growth and metastasis in colon cancer by down-regulating SMAD4. *Biosci Rep* 2019;**39**:1–8
59. Shi L, Resaul J, Owen S, Ye L, Jiang WG. Clinical and therapeutic implications of follistatin in solid tumours. *Cancer Genomics Proteomics* 2016;**13**:425–35

(Received July 27, 2021, Accepted November 18, 2021)