Integrating multi-omics data with deep learning for predicting cancer prognosis

Hua Chai[1], Xiang Zhou[1], Zifeng Cui[2], Jiahua Rao[1], Zheng Hu[2], Yutong Lu[1], Huiying Zhao[3*], Yuedong Yang[1,4*]

[1]School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510000, China

[2]The First Affiliated Hospital, Sun Yat-sen University, Guangzhou 510000, China

[3]Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou 510000, China

[4]Key Laboratory of Machine Intelligence and Advanced Computing (Sun Yat-sen University), Ministry of Education, China

*To whom correspondence should be addressed.

Abstract

Motivation: Accurately predicting cancer prognosis is necessary to choose precise strategies of treatment for patients. One of effective approaches in the prediction is the integration of multi-omics data, which reduces the impact of noise within single omics data. However, integrating multi-omics data brings large number of redundant variables and relative small sample sizes. In this study, we employed Autoencoder networks to extract important features that were then input to the proportional hazards model to predict the cancer prognosis.

Results: The method was applied to 12 common cancers from the Cancer Genome Atlas. The results show that the multi-omics averagely improves 4.1% C-index for prognosis prediction over single mRNA data, and our method outperforms previous approaches by at least 7.4%. A comparison of the contribution of single omics data show that mRNA contributes the most, followed by the DNA methylation, miRNA, and the copy number variation. In the case study for differential gene expression analysis, we identified 161 differentially expressed genes in the cervical cancer, among which 77 genes (65.8%) have been proven to be associated with cancer. In addition, we performed the cross-cancer test where the model trained on one cancer was used to predict the prognosis of another cancer, and found 23 pairs of cancers have a C-index larger than 0.5, with the largest value of 0.68. Thus, this study has provided a deep learning framework to effectively integrate multiple omics data to predict cancer prognosis.

1 Introduction

Clinical studies show that significant variations in prognosis happen among patients of the same tumor type. The variations are caused by genetic heterogeneity in subpopulations of cells, which contribute most to hinder the development of effective therapies for cancers (Dagogo-Jack and Shaw, 2018). Therefore, it is necessary to distinguish high-risk patients from low-risk patients for choosing appropriate treatment and surveillance.

Currently, many studies for cancer prognosis risk prediction have been designed based on single omics data (Kourou, et al., 2015). The most frequently used data is gene expression measured by microarray (Beer, et al., 2002; Calon, et al., 2015). With the development of next generation sequencing techniques other types of genomic data are becoming popular to be employed, including DNA methylation (Stirzaker, et al., 2015), miRNA (Volinia and Croce, 2013), and copy number variation (CNV) (Wu, et al., 2018). However, each type of the omics data represents a single view for patients, and is difficult to obtain accurate prediction. In order to achieve a comprehensive view of patients in genomics, many studies have sequenced multiple types of omics data from the same patient. Systematic studies have been performed by using the Cancer Genome Atlas (TCGA) that provides more than ten thousands of samples over 33 cancer types (Tomczak, et al., 2015). The valuable data enables an integrated analysis based on multiple omics data for comprehensive analysis for cancer prognosis.

Though multi-omics data analysis could reduce the impact of noise from single source of omics data (Li, et al., 2018), it is challenging to effectively integrate the high-dimensional data. In the past years, many statistical methods have been developed to utilize multi-omics data for dealing with different biological questions. For example, Rohart *et al* designed a general package based on sparse partial least square-discriminant analysis (Rohart, et al., 2017); Mariette *et al* used an unsupervised multiple kernel framework for predicting breast cancer clinical outcomes (Mariette and Villa-Vialaneix, 2018); Kim *et al* designed a grammatical evolution neural networks to evaluate ovarian cancer prognosis (Kim, et al., 2017); Ahmad proposed a hierarchical Bayesian graphical model that combines a Gaussian mixture model with an accelerated failure time model to find the breast cancer clinically relevant disease subtypes

(Ahmad and Frohlich, 2017); Corett identified differentially expressed genes of different cancer risk subtypes by combining sparse correlation matrix estimator and maximum likelihood estimator algorithm (Coretto, et al., 2018). However, these traditional statistical methods are limited to capture effective features from thousands of variables through dozens or hundreds of samples.

Recently, deep learning techniques have been proven to be powerful in many fields including bioinformatics (Min, et al., 2017). Chaudhary et al integrated RNA-seq, miRNA-seq, and DNA methylation data by unsupervised Autoencoder model to rebuild representative composite features (Chaudhary, et al., 2018). However, the unsupervised classification algorithm used in the study can only separate the patients into two groups, and has failed to directly link the composite features to the survival time. As a result, the model achieved limited accuracy in the prediction of survival prognosis. In addition, the discrete prediction of only two groups could cause problems for patients of intermediate risks, which required to predict the continuous risk scores for the patients.

To integrate multi-omics data for predicting the prognosis risk of cancer patients, we proposed a deep learning method named as DCAP. In this method, the multi-omics data was input into Autoencoder to obtain representative composite features. These features were then input to the Cox proportional hazards (Cox-PH) model to predict the patients' prognosis. The tests on 12 common cancers demonstrated the DCAP outperforms all previous methods. In order to reduce the number of features in the prognosis predicting model, we further selected important features to re-establish the model by XGboost algorithm, which shows a competitive performance.

2 Methods

2.1 Datasets

In this study, cancer datasets were download from the TCGA portal (https://tcga-data.nci.nih.gov/tcga/) by the *R* package "*TCGA-assembler*"(v1.0.3, (Wei, et al., 2018)). Totally, four types of multi-omics data: mRNA, miRNA, DNA methylation, and copy number variation (CNV) data were employed. Here, "mRNA" was RNA sequencing data generated by UNC Illumina HiSeq_RNASeq V2; Level 3, "miRNA" was miRNA sequencing data obtained by BCGSC Illumina HiSeq miRNASeq, DNA methylation data was generated by USC HumanMethylation450, and CNV data that generated by BROAD-MIT Genome wide SNP_6. All these data are in TCGA data level 3.

The inputs of DNA methylation and CNV were the average values of copy number variations and DNA methylation of CpG sites, respectively. The missing values were imputed using a similar way as the previous study(Chaudhary, et al., 2018). Briefly, one feature would be excluded if it was missed in more than 20% of patients. On the other hand, the patients were excluded from the study if they missed more than 20% features. For the remained data, the missing values were imputed by *R* package "*imputeMissings*" (Bokde, et al., 2018). We selected 12 cancers for facilitating comparison with methods reported in other articles. The numbers of samples for 12 cancer types range from 132 to 613, which include 59774 to 61255 genomic features. The details of the datasets are shown in Table S1.

2.2 The architecture for cancer prognosis prediction

Figure. 1 shows the architecture of the method for predicting cancer prognosis. The high dimensional features from multi-omics data were inputted into a three-hidden layers Autoencoder network to obtain representative features. And then the generated features were input into Cox-PH model for cancer prognosis prediction. To further reduce the number of input features, we utilized XGboost to select the most important features and re-establish the prediction model.

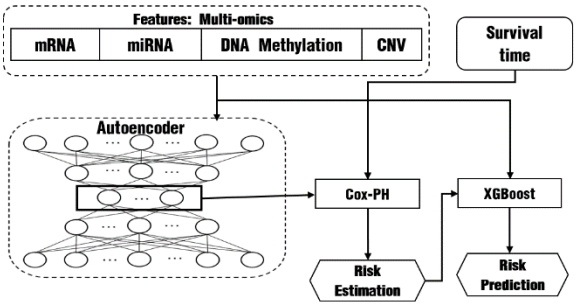

Figure. 1 The workflow of DCAP. Firstly, inputting multi-omics into a three-hidden layers Autoencoder to reconstructing the representative composite features; Next, estimating the cancer prognosis by Cox-PH model using the reconstructed features; Finally, building the XGboost regression model for cancer prognosis prediction based on the estimated risks by Cox-PH model.

2.3 Autoencoder to rebuild representative composite features

Autoencoder is a kind of artificial neural network to learn an efficient representation of the input data in an unsupervised manner (Burbank, 2015). Supposing $x = (x_1, \dots, x_n)$ are high dimensional features, which were reconstructed by replacing *x* with *x'*. *x'* is the output of the Autoencoder in the same dimension as *x*. *x'* is obtained by the *tanh* that is used as the activation function for all layers. The cross-entropy function is used for the loss function:

$$logloss(x, x') = \sum_{i=1}^{n}\left(x_i log(x'_i) + (1 - x_i) log(1 - x'_i)\right) \quad (1)$$

In this study, regularization penalties were added to the Autoencoder to control overfitting:

$$L(x, x') = logloss(x, x') + \sum_{i=1}^{k}(\alpha\|W_i\|_1 + \gamma\|F_{1\to i}(x)\|_2^2) \quad (2)$$

where *k* was set 5 for layers (input, output, and 3 hidden layers), α and γ were the coefficients for L1 and L2-norm regularization penalties, here they were both set as 0.0001. In this study, the number of nodes in the three hidden layers were set as 500, 200, and 500 respectively. The Autoencoder was trained by back-propagation via Adam optimizer with a dropout rate of 0.5.

2.4 Cox proportional hazard model for risk estimation

The 200 features from the middle hidden layer were used for building Cox proportional hazard (Cox-PH) model to estimate the cancer prognosis risks. The Cox-PH model was implemented by the "*glmnet*" package in *R* (Simon, et al., 2011). First, the univariate Cox model was used to select significant features that were able to distinguish the high-risk and the low-risk patients with log-rank p-value <0.05. These selected features were input to the multivariate Cox-PH model to estimate the patients' risks. The multivariate Cox proportional hazard model was defined as

$$h(t|X_i) = h_0(t)\theta_i \quad (3)$$

where $h_0(t)$ was the underlying baseline hazard function to describe how the risk changes at time *t*, and $\theta_i = \exp(\beta X_i)$ was used to describe how the hazard varies in response between coefficients vector β and covariates vector X_i by patient *i*.

The probability of the death for the patient *i* at the time t_i was written as:

$$L_i(\beta) = \frac{h_0(t_i)\theta_i}{\sum_{j:t_j>t_i} h_0(t_i)\theta_j} \quad (4)$$

Hence the corresponding log partial likelihood function was given as:

$$l(\beta) = \log\left(L(\beta \prod_i L_i(\beta))\right) = \sum_i \left(X_i\beta - log \sum_{j:t_j>t_i} \theta_j\right) \quad (5)$$

This partial likelihood function was solved by using the Newton-Raphson algorithm. The computed β can be used to estimate the risk scores in the Cox-PH model.

2.5 Application of XGboost in feature selection

In order to reduce the number of features for cancer prognosis prediction, we employed XGboost for feature selection. XGboost is an ensemble of k regression trees $(T_1(X,Y) \dots T_k(X,Y)$, where X is the features and Y is the corresponding patients' risks (Chen and Guestrin, 2016). If we supposed the multi-omics dataset containing n samples and p features, the multi-omics data of all the patients can be described as $\mathcal{D} = \{(x_i, y_i)\}$ $(|\mathcal{D}| = n, x_i \in X, y_i \in Y)$. The XGboost model with K trees was used to select the best features in predicting the patients' risks as described by Equation (6).

$$\hat{y}_i = \emptyset(x_i) = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathcal{F} \quad (6)$$

where $\mathcal{F} = \{f(x) = w_{q(x)}\}$ $(q: \mathbb{R}^m \to T, w \in \mathbb{R}^T)$ represents the space of the regression trees, q was the tree structure and T was the number of leaves in each tree. And each f_k represented a regression tree structure q with weight w. This method was implemented by "*XGboost*" package in *R* (Chen, et al., 2015).

2.6 Cross Validation

The parameters of the above Cox-PH and XGboost models have been optimized by 10-fold cross validation to avoid over-training. Here, the patients were randomly divided into 10 folds, from which nine folds were used to train a model and the rest fold was used for test. This procedure was repeated for 10 times, and results were collected for model evaluation. We also performed multiple tests with different random seeds in dividing folds to test robustness of the method.

2.7 Evaluating cancer prognosis prediction

We evaluated the prediction of cancer prognosis by two measurements, Concordance Index (C-index) and log-rank p-value. The C-index represents the fraction of all pairs of individuals whose predicted survival times are correctly ordered based on the Harrell's C statistics (Van Belle, et al., 2011). A C-index 0.5 means a random prediction, and the higher C-index means the better performance of the prediction. The log-rank p-value is obtained from significance of the risk scores in separating the patients into high-risk and low risk groups.

After the prognosis prediction, patients were divided into high-risk and low-risk groups by Cox-PH model. In these groups of patients, genes expressed significantly different were defined as associated with prognosis. The gene expression was analyzed by the "*DESeq2*" package (Love, et al., 2014) in *R*. The genes with log2 fold change >1 and FDR <0.05, were considered as related to cancer prognosis. The enriched pathways of these genes were obtained by the online tool Metascape (http://metascape.org). The pathway databases used in this study including GO Biological Processes, KEGG Pathways, Reactome Gene Sets, Canonical Pathways and CORUM. The enriched pathways (p-value<0.05, a minimum count=3) were collected and grouped into clusters based on their membership similarities. The most significant pathway in statistics within a cluster was chosen to represent the cluster.

3 Results

3.1 Predicting cancer prognosis with multi-omics data

For cancer prognosis prediction, high dimensional multi-omics features were put into Autoencoder networks to construct 200 representative composite features. Based on the representative features, we utilized Cox-PH model to predict the prognosis risks of 12 types of cancers. As shown in Table 1, DCAP achieved C-index values between 0.661 and 0.871, with an average of 0.711. The highest C-index was 0.871 in prediction of KIRP. The lowest C-index 0.661 was achieved on prediction of PAAD. The P-values of subgroups estimated by DCAP were ranged from 3.0E-11 to 1.2E-5, with a median value of 4.3E-8. By comparison, when we also employed k-means, an unsupervised method, to cluster the patients according to the same representative features generated by Autoencoder. The method, namely DCAP-kmeans, only produced an average C-index value of 0.646, which was 9.1% lower than DCAP. The corresponding P-values by DCAP-kmeans were also less significant with a median of 2.2E-4. Based on the subgroups estimated by DCAP and DCAP-kmeans, the survival curves were shown in Figure. 2. The curves by DCAP for the high and low risk groups were consistently separated better than those by DCAP-kmeans.

Table 1. The C-index and P-values of prognosis prediction on 12 cancers from the TCGA dataset by two methods

	C-index		P-value	
	DCAP	DCAP-kmeans	DCAP	DCAP-kmeans
BLCA	0.672	0.611	1.4E-10	1.2E-4
BRCA	0.677	0.608	1.2E-5	0.0018
CESC	0.742	0.632	7.2E-10	0.0011
KIRC	0.760	0.669	1.0E-5	5.2E-4
KIRP	0.871	0.725	2.0E-8	3.7E-5
LIHC	0.744	0.708	6.5E-10	2.7E-5
LUAD	0.673	0.617	9.2E-7	0.0058
LUSC	0.670	0.624	4.3E-8	0.023
PAAD	0.661	0.615	4.1E-6	3.3E-4
SKCM	0.678	0.622	3.0E-11	1.4E-7
STAD	0.675	0.616	2.6E-8	1.3E-7
UCEC	0.716	0.706	3.8E-6	7.3E-6
Average	0.711	0.646	-	-
Median	0.678	0.623	4.3E-8	2.2E-4

We further detailed the contribution of each omics type in the DCAP method. As shown in Table 2, when using single type of omics data, mRNA performed the best with an average C-index value of 0.683, and CNV had the lowest performance with C-index of 0.65. The methylation and miRNA ranked the 2nd and 3rd, respectively. Consistently, when excluding one omics type from the DCAP, mRNA caused the largest decrease of C-index from 0.711 to 0.687, while the smallest decrease was from an exclusion of CNV. These results indicated that mRNA plays the most important role to discriminate high risk patients while CNV makes the least contribution, and integrating multi-omics averagely improves 4.1% C-index for prognosis prediction over only using mRNA data.

Table 2. The contribution of each omics data for cancer prognosis evaluation by using only one type of omics data or subtracting one type from the final model.

Single Omics	C-index	Multi-omics	C-index
		All	0.711
mRNA	0.683	-mRNA	0.687
miRNA	0.665	-miRNA	0.695
Methylation	0.673	-Methylation	0.693
CNV	0.650	-CNV	0.701

3.2 Comparing to other methods

Table 3 compared DCAP with other state-of-the-art methods that all the methods only used mRNA data. Table 3 also showed the results of the DCAP using multi-omics data at last column. In comparison, the C-index obtained by three traditional methods (general Cox model, Cox model with lasso regularization, and Cox model with elastic net) achieved C-index values between 0.565 and 0.569, which were much lower than those obtained by some advanced methods in recent studies (Cox_DL, Cox_transfer, Cox_TRACE, Cox_cCMTL) (Cheerla and Gevaert, 2019; Wang, et al., 2017). These advanced methods obtained C-index values between 0.605 and 0.632, with an average of 0.620. The Cox_transfer achieved the highest average C-index value among these

four methods. Nevertheless, the Cox_transfer is a transfer learning framework which requires downloading large numbers of other cancer datasets, and the prediction process would take large amount of computational time. By comparison, the highest average C-index value obtained by Cox-transfer is 0.632, which was 7.4% lower than DCAP-mRNA (C-index=0.683), and 11.1% lower than DCAP (C-index=0.711). As indicated by Table 3, DCAP-mRNA consistently performed better than all Cox methods only using mRNA information in all 12 cancer datasets, and DCAP consistently outperform DCAP-mRNA by using multi-omics data.

3.3 Selecting important features for prognosis prediction using XGboost

DCAP constructed by all multi-omics features may contain redundant variables, which increase the computation time in real-world medical examination. In order to remove the redundancy, we employed XGboost to select features. The features were selected according to the importance computed by XGboost in fitting the patients' risk scores. By using selected features, low redundant models were constructed. The number of features were given in Table 4. The largest number of used features was in BRCA dataset, where the XGboost selected 139 features. And the least one was in UCEC, which only 61 selected features were used to re-establish the model. As shown in Table 4, using XGboost achieved C-index values between 0.602 and 0.829, with an average of 0.657. The differences between the C-index values obtained by DCAP_XGboost and DCAP were ranged from 4.82% to 16.44% with an average of 7.58%. This indicated that although feature selection slightly reduced the accuracy of the prediction, the number of features used is greatly reduced.

3.4 Cross-cancer prognosis prediction

To explore the similarities between different cancers in cancer prognosis, we used the DCAP models trained on one cancer to predict the prognosis of another cancer. It should be noted that the tested cancer wasn't included in training model. Figure 3 includes the C-index values of 12 cancer types by cross-cancer prediction. If the prognosis of two cancers were predicted mutually with a C-index value large than 0.5 (a threshold commonly considered to be significant), these two cancers were defined cancer pairs sharing similarities in prognostic risk. In Figure 3, there were 23 cancer pairs have C-index values larger than 0.5. Among these, the model trained on the KIRC achieved the largest C-index (0.68) in prediction the prognosis of BLCA. The value was even slightly higher than the model trained by the data from BLCA. On the other hand, the model trained on KIRC also had an accurate prediction on BLCA with a C-index of 0.60. These results were possibly due to that the bladder cancer and kidney cancer are both urinary system diseases (Gottardo, et al., 2007). Another example is on the prognosis prediction model constructed by the DCAP to predict the prognosis of LUAD, which achieved C-index value 0.6, and the reversed prediction achieved C-index value 0.58. It shows that the prognosis prediction model for the cancers in the same area are similar.

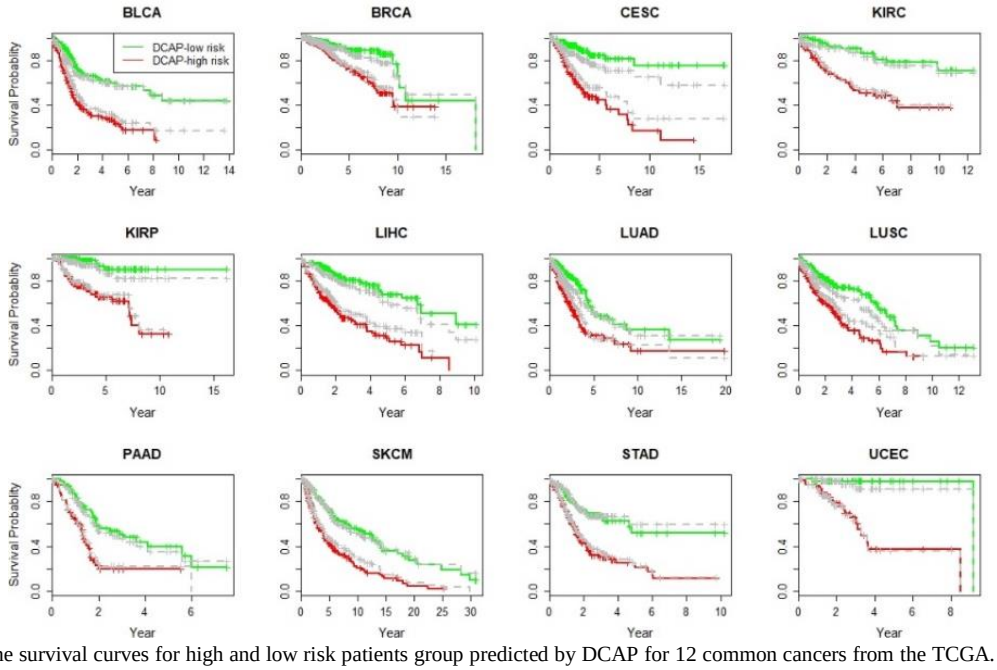

Figure. 2 The survival curves for high and low risk patients group predicted by DCAP for 12 common cancers from the TCGA. The red line represents the high-risk patients and the green line represents the low risk-patients. The x-axis represents the survival time and the y-axis represents the survival probability. The solid lines were drawn by DCAP and the dotted lines (grey) were drawn by DCAP-kmeans.

Table 3. Performance comparison of the DCAP and other existing related methods using C-index values in 12 cancers

	Cox[a]	Cox_lasso[a]	Cox_elastic net[a]	Cox_DL[a]	Cox_transfer[b]	Cox_TRACE[b]	Cox_cCMTL[b]	DCAP_mRNA[c]	DCAP
BLCA	0.561	0.524	0.539	0.600	0.605	0.613	0.621	0.655	0.672
BRCA	0.548	0.588	0.564	0.570	0.627	0.602	0.623	0.648	0.677
CESC	0.572	0.576	0.554	0.670	0.679	0.592	0.630	0.709	0.742
KIRC	0.601	0.629	0.608	0.610	0.670	0.691	0.704	0.711	0.760
KIRP	0.745	0.740	0.749	0.650	0.803	0.794	0.804	0.837	0.871

LIHC	0.534	0.552	0.545	0.640	0.650	0.606	0.651	0.715	0.744
LUAD	0.497	0.490	0.498	0.630	0.569	0.555	0.597	0.653	0.673
LUSC	0.578	0.550	0.562	0.500	0.571	0.601	0.600	0.632	0.670
PAAD	0.528	0.563	0.567	0.570	0.580	0.545	0.557	0.623	0.661
SKCM	0.536	0.578	0.527	0.560	0.654	0.616	0.596	0.666	0.678
STAD	0.463	0.499	0.486	0.630	0.554	0.485	0.527	0.661	0.675
UCEC	0.637	0.542	0.576	0.630	0.626	0.644	0.655	0.681	0.716
Average	0.567	0.569	0.565	0.605	0.632	0.612	0.630	0.683	0.711

[a]: The results reported in reference (Wang, et al., 2017)
[b]: The results reported in reference (Cheerla and Gevaert, 2019)
[c]: DCAP using only mRNA

Table 4 The prognosis prediction performances obtained by XGboost compared with the DCAP in different TCAG datasets

	BLCA	BRCA	CESC	KIRC	KIRP	LIHC	LUAD	LUSC	PAAD	SKCM	STAD	UCEC	AVE
Biomarkers	112	139	92	68	128	112	116	114	67	109	118	61	103
DCAP_XGboost	0.602	0.623	0.700	0.635	0.829	0.705	0.625	0.610	0.637	0.631	0.606	0.674	0.657
DCAP	0.672	0.677	0.742	0.76	0.871	0.744	0.673	0.670	0.661	0.678	0.675	0.716	0.711
Difference	10.41%	7.97%	5.66%	16.44%	4.82%	5.24%	7.13%	8.95%	9.64%	6.93%	10.22%	5.86%	7.59%

	BLCA	BRCA	CESC	KIRC	KIRP	LIHC	LUAD	LUSC	PAAD	SKCM	STAD	UCEC
DCAP_BLCA	0.67	0.50	0.54	0.60	0.46	0.49	0.52	0.51	0.48	0.54	0.52	0.51
DCAP_BRCA	0.33	0.68	0.50	0.49	0.45	0.53	0.47	0.49	0.39	0.57	0.37	0.65
DCAP_CESC	0.52	0.52	0.74	0.56	0.65	0.28	0.52	0.57	0.62	0.55	0.47	0.50
DCAP _KIRC	0.68	0.57	0.35	0.76	0.54	0.51	0.45	0.52	0.45	0.34	0.40	0.59
DCAP _KIRP	0.52	0.53	0.60	0.51	0.87	0.43	0.46	0.58	0.24	0.29	0.50	0.42
DCAP _LIHC	0.55	0.46	0.42	0.52	0.57	0.74	0.46	0.55	0.46	0.58	0.44	0.58
DCAP_LUAD	0.41	0.50	0.54	0.41	0.48	0.50	0.67	0.60	0.55	0.40	0.58	0.49
DCAP_LUSC	0.52	0.52	0.53	0.42	0.46	0.45	0.58	0.67	0.53	0.52	0.59	0.59
DCAP_PAAD	0.59	0.52	0.50	0.52	0.42	0.55	0.52	0.56	0.66	0.45	0.51	0.41
DCAP_SKCM	0.57	0.47	0.46	0.49	0.50	0.56	0.56	0.55	0.46	0.68	0.45	0.53
DCAP_STAD	0.53	0.54	0.44	0.51	0.44	0.45	0.49	0.58	0.49	0.49	0.68	0.55
DCAP_UCEC	0.51	0.54	0.46	0.51	0.54	0.73	0.50	0.56	0.54	0.38	0.43	0.72

Figure. 3 The C-index values obtained by DCAP models in cross-cancer prognosis prediction. The DCAP models trained on one cancer are used to predict prognosis on another cancer. Figure.3 indicated that there were 23 pairs of cancers with similarity in predicting prognosis. All the cancer pairs are given in Table S2.

3.5 Case study: Differential gene expression analysis in CESC

DCAP classified the CESC patients into high-risk and low-risk groups. For these two groups of patients, we performed differential gene expression analysis. Totally, 161 genes were identified with FDR<0.05 and log2 fold change >1, among which 116 genes are down-regulated and 45 up-regulated.

The heat map of 161 genes in these patients was shown in Figure 4(a). Among them, 22 genes were reported associated with cervical cancer in previous study (Tables S3). For example, the overexpression of CYP26A1 (p=7.6E-6) was found contributing to the development and progression of cervical cancer (Osanai and Lee, 2014); The MIA (FDR= 2.5E-3) was proved to be associated with the tumor progression and metastasis in the cervical cancer, and was presented as a new biomarker in cancer treatment (Sasahira, et al., 2016). Other 20 genes: CCL18, CD200R1, ZNF683, APOC1, IGF1, WT1, CXCL11, TRPA1, TNF, PPP1R3C, MMP3, TREM1, C1QTNF1, IL1B, STC1, IL1A, FGF13, TFPI2, IL12A and CXCL5 were also reported to be associated with the cervical cancer. Meanwhile, 84 genes (52.2%) were reported associated with other cancers. For example, the promoter region of GFRA3 shows significant hypermethylation in many different types of tumors, and it was found associated with survival and other clinicopathological parameters in cancer patients (Eftang, et al., 2016); PLXNA4 was reported as promoting tumor progression and

angiogenesis by enhancing VEGF and bFGF signaling (Kigel, et al., 2011). Thus, 106 out of the 161 (65.8%) genes have been proven to be associated with cancer.

For these 161 genes, we performed pathway analysis to identify their enriched pathways. Totally, 31 pathways were found to be significantly enriched by these genes. Figure 4(b) shows the percentage of enriched pathways in different function categories. The top eight most significant pathways were in the functional category related with immunology that is well known to be associated with cancer prognosis. All these immunology pathways were cancer microenvironment related. Additionally, there were three pathways about the cell growth, two pathways about the metabolism, one pathway associated with drug resistance, and one pathway about the malignant reproductive phenotype. Table S3 detailed all 161 identified genes and the corresponding enriched pathways. Table S4 listed all the differential expressed genes in 12 cancers.

4 Discussion

For a long time, people used single omics data to predict the cancer prognosis. However, these methods ignored the complementary effects and interactions between different omics data, and the results were easy to be affected by noise. The increase of multi-omics data in cancer study promoted the integration of multi-omics to predict cancer prognosis. The multi-omics data analysis could effectively alleviate problems encountered in single data analysis for cancer prognosis. Although several recent studies have been reported to integrate multi-omics data in predicting cancer prognosis, the complex relationship between the omics data and the exponential increase of the computational complexity pose challenges to design suitable methods.

In this study, we proposed DCAP to predict cancer prognosis by integrating the multi-omics data using Autoencoder. DCAP is different from the previous methods because it used Cox-PH model for classifying the patients, and the XGboost for feature selection. Application of DCAP in 12 common cancers achieved significant better performance than the previous approach for cancer prognosis prediction.

One of the important applications of DCAP is on predicting the prognosis of patients in real world. In this study, the most informative features were selected to construct the prediction model by XGboost. The reduced number of features makes the clinical data collection easier and computing faster. Another application of DCAP is in identification of cancer driver genes. In this study, we show 161 cervical cancer associated genes identified by DCAP. The reliability of these genes was tested by the literature searching. The results indicated that about 65.8% of the genes reported as cancer related previously.

We tested the performances of DCAP in prediction of prognosis cross multiple cancers. Among 12 cancers, 24 cancers pairs were identified predicted effectively by the models developed based on other cancers. These results may implicate the similar mechanisms of cancer prognosis.

In the future, we will further improve the model by combining multi-omics data with other biomedical data, such as medical images and clinical data. The comprehensive integration of multimodal data is helpful to reveal the relation of genotypic and phenotypic information, and thus helps to discover new biomarkers determining cancer prognosis.

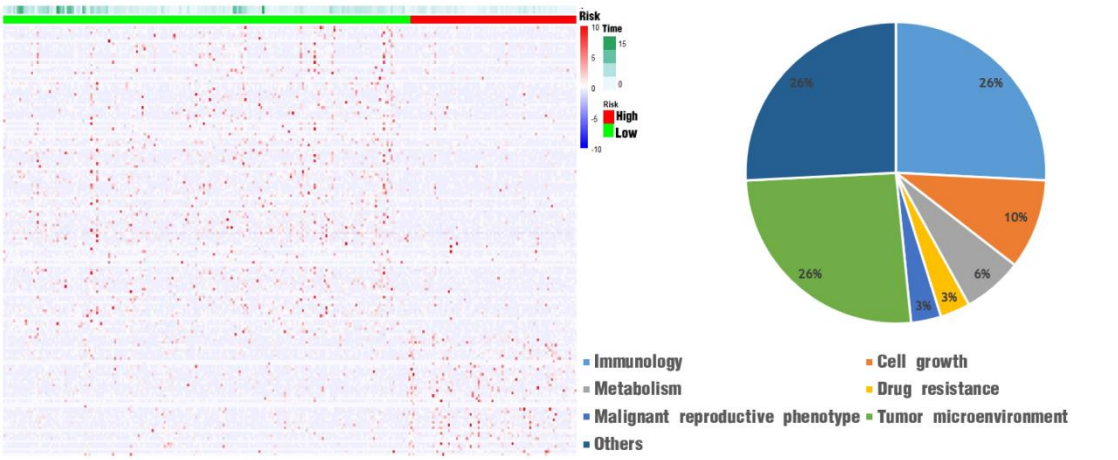

Figure. 4 The biological analysis performance in CESC dataset. (a). The heat map using differential expressed genes in CESC (p-value <0.05 and log2Fold >1); (b) The percentage of enriched pathways in different function categories in CESC.

Funding

The work was supported in part by the National Key R&D Program of China (2018YFC0910500), National Natural Science Foundation of China (U1611261, 61772566, and 81801132), Guangdong Frontier & Key Tech Innovation Program (2018B010109006, 2019B020228001) and Introducing Innovative and Entrepreneurial Teams (2016ZT06D211).

Authors' contributions

CH and YY conceived the study. CH, ZX, CZ, and RJ performed the data analysis. CH, HZ, LY, ZH, and YY interpreted the results. CH, ZH, and YY wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

References

Ahmad, A. and Frohlich, H. Towards clinically more relevant dissection of patient heterogeneity via survival-based Bayesian clustering. *Bioinformatics* 2017;33(22):3558-3566.

Beer, D.G., *et al.* Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med* 2002;8(8):816-824.

Bokde, N., *et al.* A novel imputation methodology for time series based on pattern sequence forecasting. *Pattern Recognit Lett* 2018;116:88-96.

Burbank, K.S. Mirrored STDP Implements Autoencoder Learning in a Network of Spiking Neurons. *PLoS Comput Biol* 2015;11(12):e1004566.

Calon, A., *et al.* Stromal gene expression defines poor-prognosis subtypes in colorectal cancer. *Nat Genet* 2015;47(4):320-329.

Chaudhary, K., *et al.* Deep Learning-Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. *Clin Cancer Res* 2018;24(6):1248-1259.

Cheerla, A. and Gevaert, O.J.b. Deep Learning with Multimodal Representation for Pancancer Prognosis Prediction. 2019:577197.

Chen, T. and Guestrin, C. Xgboost: A scalable tree boosting system. In, *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM; 2016. p. 785-794.

Chen, T., *et al.* Xgboost: extreme gradient boosting. 2015:1-4.

Coretto, P., Serra, A. and Tagliaferri, R. Robust clustering of noisy high-dimensional gene expression data for patients subtyping. *Bioinformatics* 2018;34(23):4064-4072.

Dagogo-Jack, I. and Shaw, A.T. Tumour heterogeneity and resistance to cancer therapies. *Nat Rev Clin Oncol* 2018;15(2):81-94.

Gottardo, F., *et al.* Micro-RNA profiling in kidney and bladder cancers. *Urol Oncol* 2007;25(5):387-392.

Kim, D., *et al.* Using knowledge-driven genomic interactions for multi-omics data analysis: metadimensional models for predicting clinical outcomes in ovarian carcinoma. *J Am Med Inform Assoc* 2017;24(3):577-587.

Kourou, K., *et al.* Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J* 2015;13:8-17.

Li, Y., Wu, F.X. and Ngom, A. A review on machine learning principles for multi-view biological data integration. *Brief Bioinform* 2018;19(2):325-340.

Love, M.I., Huber, W. and Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;15(12):550.

Mariette, J. and Villa-Vialaneix, N. Unsupervised multiple kernel learning for heterogeneous data integration. *Bioinformatics* 2018;34(6):1009-1015.

Min, S., Lee, B. and Yoon, S. Deep learning in bioinformatics. *Brief Bioinform* 2017;18(5):851-869.

Rohart, F., *et al.* mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Comput Biol* 2017;13(11):e1005752.

Simon, N., *et al.* Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *J Stat Softw* 2011;39(5):1-13.

Stirzaker, C., *et al.* Methylome sequencing in triple-negative breast cancer reveals distinct methylation clusters with prognostic value. *Nat Commun* 2015;6:5899.

Tomczak, K., Czerwinska, P. and Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol (Pozn)* 2015;19(1A):A68-77.

Van Belle, V., *et al.* Support vector methods for survival analysis: a comparison between ranking and regression approaches. *Artif Intell Med* 2011;53(2):107-118.

Volinia, S. and Croce, C.M. Prognostic microRNA/mRNA signature from the integrated analysis of patients with invasive breast cancer. *Proc Natl Acad Sci U S A* 2013;110(18):7413-7417.

Wang, L., *et al.* Multi-task survival analysis. In, *2017 IEEE International Conference on Data Mining (ICDM)*. IEEE; 2017. p. 485-494.

Wei, L., *et al.* TCGA-assembler 2: software pipeline for retrieval and processing of TCGA/CPTAC data. *Bioinformatics* 2018;34(9):1615-1617.

Wu, Y., *et al.* Genome-wide Association Study (GWAS) of Germline Copy Number Variations (CNVs) Reveal Genetic Risks of Prostate Cancer in Chinese population. *J Cancer* 2018;9(5):923-928.