Omi-PGTCN:A Whole Process Optimized Model for Pan-cancer Classification and Biomarker Identification

Xin Chen[1], Yun Tie[1], Fenghui Liu[2], Feng Wang[3], and Lin Qi[1]

[1] School of Electrical and Information Engineering, Zhengzhou University, Zhengzhou, China
cxregion@gs.zzu.edu.cn, {ieytie, ielqi}@zzu.edu.cn
[2] Respiratory and Sleep Department, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China
lfh111@163.com
[3] Department of Oncology, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China
zzuwangfeng@zzu.edu.cn

Abstract. Tumor, which typically makes people fearful when they hear it due to its low curable rate and elevated mortality. The reasons for that mainly come from deficient understanding. We can obtain more comprehensive views of the malady and thereby better to heal the patients by integrating analysis of multi-omics data from one cancer patient because of the dissimilar insights in that. However, biological data have super-great dimension and it is a nut that applies what method sufficiently to explore the interrelationship among omics data. These problems will result in difficult training for machine or deep learning models and poor performance of analysis events. Here we proposed a whole process optimized model called Omi-PGTCN to fulfill dimension reduction, data integrating, classification as well biomarker identification. The Omi-PGTCN model combines variational auto-encoder and pattern fusion as well as graph tree convolution network to conquer "over-fitting" problems and enhance the classification accuracy as well robust of model. In our experiments, Omi-PGTCN gained an average classification accuracy of $97.90 \pm 1.01\%$ after 10-fold cross-validation among 33 types of tumors from the cancer genome atlas program. In addition, the model also identified some vital biomarkers that verified by survival analysis of biomarker. The source code at https://github.com/cx-333/Omi-PGTCN.

Keywords: Multi-omics data · Variational auto-encoder · Pattern fusion · Graph tree convolution network · Pan-cancer classification · Biomarker identification.

1 Introduction

With the fast-speed development of high-throughput biological sequencing technology, various kinds of omics data, such as genomics, epigenomics, transcrip-

tomics, proteomics as well as metabolomics, can be easier to obtain than before [5]. The formation of tumors in the human body will be accompanied by abnormal changes in one or more segments of the gene sequence. However, the forming reason of tumors reflected by a certain omics data are local, one-sided, and inaccurate. Consequently, it can provide more comprehensive analysis and insights for tumor prognostication and therapeutic diagnosis by merging different omics types data as DNA methylation, RNA expression and protein expression.

Unfortunately, omics data is known as "high dimension and small samples" [9] whose features usually up to tens of thousands dimension and the samples, but, are very few comparatively. If these high-dimensional small-sample data are directly sent into the model for analysis and research, the phenomenon of "dimension disaster" and “over-fitting” will occur [3]. To crack this nut, the traditional solution first is to reduce the dimension of original data by applying principal component analysis (PCA) or auto-encoder (AE). And then employ processed signals for downstream analysis. Nevertheless, after dimensionality reduction, the data retain only relatively traits that are different from other omics data and does not contain the internal relationship between each omics data. For bridging the internal relationship of each omics data, the common idea is to directly concatenate the processed data of various types to contain the characteristics of various types of data, but this does not reflect the link between the data of each group. Some researchers developed a similarity network fusion (SNF) [20] method for multi-omics data integration again to explore different group's interactive relation. But SNF treats all types of data equally.

In this paper, we proposed an Omi-PGTCN model based on variational auto-encoder (VAE), pattern fusion (PF) and graph tree convolution network (GTCN) for features dimension reduction and integration of multi-omics data, tumor type classification as well as biomarker identification. The work in the article mainly includes three parts: 1) Firstly, preprocessing virgin data according to specific criteria that filters these probes of DNA methylation which do not match that in the human reference genome (hg38) annotation, removes samples of each omics data which have missing values (N/A) more than 10% respectively. Then the few void values in the remaining samples are substituted by the mean value corresponding to probes in the samples. Then normalize the RNA expression and protein expression data to the range of 0 to 1 and decrease the dimensionality of the preprocessed data using VAE. 2) Secondly, concatenating signals of the previous step and integrating it by PF so as to generate a sparse matrix which includes the interactive information of various input features. 3) Finally, making tumor type classification and important biomarker identification based on GTCN utilizing the concatenated features and sparse matrix.

2 Related Work

Pan-cancer classification recently utilized the multi-omics data for classification. Due to the high dimension of multi-omics data, researchers normally first reduce the length of input data. Then integrate multi-omics data to an entirety and fulfill

classification using it. For the past few years, various integrating and dimension reduction methods [2,4,6,19] have been applied in the pan-cancer analysis, which gains excellent results.

According to the nature of the existing pan-cancer analysis techniques, we classify them into two main categories: machine learning based methods [10, 12] which usually process multi-omics data step by step and have good interpretability, and deep learning based methods [13,16,18,21,22,24] which is an end-to-end model with mystique. Compared to the two methods, the former has better interpreted but non-ideal performance and the latter only the reverse.

However, many researches including the above work only focus on a process in the studies of multi-omics such as integrating, dimensionality reduction or cancer classification among certain types of cancers. And all of these studies only focus on partial optimization in classification without making full use of complementary materials as other types of omics data and the methods are too traditional and simple as well as poor interpretability.

3 Metarials and Methods

Omi-PGTCN, as a whole process optimized model for pan-cancer classification and biomarker identification, which has a set of refined processing for multi-omics data, takes multi-omics data (DNA methylation, RNA expression and protein expression) as the input. And the overall workflow of Omi-PGTCN shown in Figure 1. Omi-PGTCN is a procedural model that provides good interpretability and macroscopic treating processes. In the following, we describe each component of the model in detail, separately.

3.1 Data Preparation

In this paper, we gained virgin pan-cancer datasets including three omics data (DNA methylation, RNA expression and protein expression) which can be downloaded in the website (https://xenabrowser.net/) from the cancer genome atlas (TCGA). Due to the existence of noise, bias signals, missing values, unmatched samples and unit values in the gained datasets, it is essential to preprocessing input data before applying them.

Targeting DNA methylation data, firstly, aligning the probes in that with counterpart in human reference genome (hg38) and removing the probes that do not match to hg38. And then dropping the samples that the missing values (N/A) have more than 10% in total probes. Finally, the surplus NAN values in the samples were substituted by the mean value of corresponding probes in all samples. For RNA and protein expression data, the preliminary processing process in them was same as DNA methylation data excluding alignment. Then normalized their values to the interval 0 and 1. In the end, reserved the samples mutual contained in three omics data and dismissed the remanent samples.

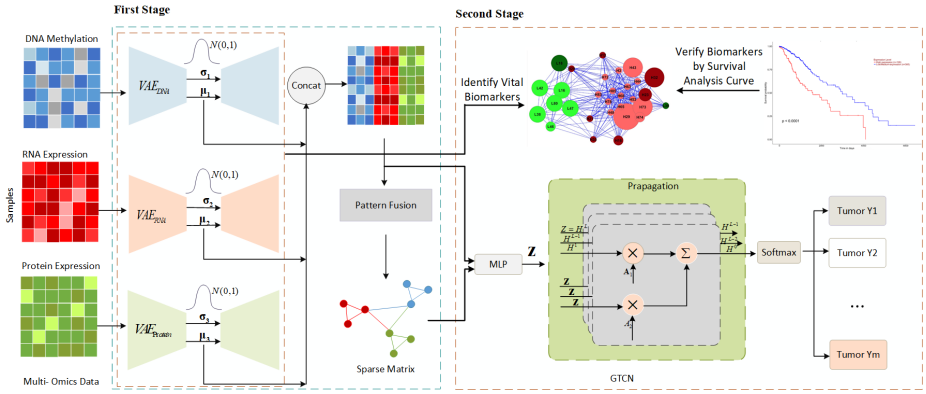

Fig. 1. The workflow schematic of Omi-PGTCN. Omi-PGTCN combines VAE, PF and GTCN for multi-omics data dimensionality reduction, integration as well as classification. The model training process was split into two stages. In the first stage, multi-omics data was encoded to suitable dimension and then concatenated and merged processed signals into new input data and a sparse matrix that include interactive information across each data, respectively, by VAE and PF orderly. In the second stage, GTCN was trained to fulfill tumor classification utilizing fresh input data produced by stage one.

3.2 Variational Auto-Encoder (VAE)

Variational Auto-Encoder (VAE) [14] is a probability inference model based on incomplete data which learning the traits of input data by probability and mapping original characters to latent space. Therefore, VAE has strong representational ability over information, this makes high-dimensional features can be mapped into low-dimensional space by distortion as smaller as possible.

Given multi-omics data sets $D = \{\mathbf{X}^i\}_{i=1}^N$, which is composed of DNA methylation $X_{DNA} \in \mathbf{R}^{N\times P1}$, RNA expression $X_{RNA} \in \mathbf{R}^{N\times P2}$ and protein expression $X_{Protein} \in \mathbf{R}^{N\times P2}$, where N represents the number of samples and $P_i(i = 1, 2, 3)$ stand for the dimension of features.Assuming that the probability distribution of $X_.$ is $p(x_.)$, in which $x_.$ denote different input data. And $p(z_.)$ represent the distribution of latent variable $z_.$ corresponding to $x_.$. Therefore, according to Bayesian principle: $p(x_.) = \frac{p(x_.|z_.)p(z_.)}{p(z_.|x_.)} = \frac{p(x_.,z_.)}{p(z_.|x_.)}$.

On account of the intractability of posterior probability distribution $p(z_. \mid x_.)$, we used $q(z_.)$ to approximate $p(z_. \mid x_.)$ and took the logarithm of the above equation while computed their expectation:

$$\ln p(x_.) = \int q(z_.) \ln \frac{p(x_., z_.)}{q(z_.)} dz_. + \int q(z_.) \ln \frac{q(z_.)}{p(z_. \mid x_.)} dz_. \tag{1}$$

Where $L(q) = \int q(z_.) \ln \frac{p(x_.,z_.)}{q(z_.)} dz_.$ denotes the variational lower bound and $KL(q(z_.)\|p(z_. \mid x_.)) = \int q(z_.) \ln \frac{q(z_.)}{p(z_.|x_.)} dz_.$ is called Kullback-Leibler Divergence (KL) which measures the discrepancy of two functions. The smaller KL value, the

more imperceptible difference in objective functions.By minimizing $KL(q(z_.) \mid\mid p(z_. \mid x_.))$ can we obtain the proximate distribution $q(z_. \mid x_.)$ of $p(z_. \mid x_.)$.

Finally, the loss function of the model is calculated as:

$$\text{Loss}_{VAE} = \underset{q(z_.|x_.)}{\operatorname{argmin}} \left[KL(q(z_. \mid x_.) \| p(z_. \mid x_.)) + L_{p(x_.)} \right] \tag{2}$$

Where $L_{p(x_.)} = E_{q(z_.)} \ln p(x_.) - \ln p(x_.)$ denote reconstruct error which was designed as a proper function, such as mean square error (MSE).

3.3 Pattern Fusion (PF)

Pattern Fusion (PF) [17] algorithm extracted the "native pattern" of input data. PF algorithm can also discover the identity in various types of data even though existing noise or interference. In particularity, the PF algorithm can identify important "sample pattern" by optimizing the adjustment of different types of data. The principle of PF is: 1) Obtain the local "sample pattern" of each type of data; 2) Process the local "sample pattern" to gain global "sample pattern" utilizing adaptive optimum strategy. By this time, the "sample pattern" included the information of correlation (similarity).

For k types of input data $X^i (i = 1, 2, \cdots, k, k \geq 2)$, the target function of PF is defined as:

$$\begin{aligned} & min\ tr(Y \Phi Y^T) + \lambda \|W\|_2^2 \\ & s.t. \begin{cases} YY^T = I \\ W^T \mathbf{1} = \mathbf{1} \\ W \geq 0 \end{cases} \end{aligned} \tag{3}$$

where, $W^T = (w_1^1, w_2^1, \cdots, w_{n1}^1, w_1^2, w_2^2, \cdots, w_{n2}^2, \cdots, w_1^k, w_2^k, \cdots, w_{nk}^k), M = n1+n2+\cdots+nk$, and λ are hyper parameters which balance the solving accuracy of matrix Y and the sparsity of weight W. $\mathbf{1}= (1, 1, \cdots, 1)^T$ and Y represent the unit column vector and global "sample pattern" equals X at begin, respectively. Φ can be calculated as follows:

$$\begin{aligned} \Phi = & \sum_{i=1}^{k} \frac{1}{V^i} (I - \frac{\mathbf{11^T}}{n})(I - (Y^i W^i)^\dagger (Y^i W^i)) * \\ & ((I - \frac{\mathbf{11^T}}{n})(I - (Y^i W^i)^\dagger (Y^i W^i)))^T \end{aligned} \tag{4}$$

Where the symbol † denotes the pseudo-inverse operation, I equals the identity matrix. The formula for obtaining correction matrix W^i equals $W^i = diag(sqrt(w_1^i), sqrt(w_2^i), \cdots, sqrt(w_n^i)), w_j^i (i = 1, 2, \cdots, k; j = 1, 2, \cdots, n)$ is a scalar and denotes the weight of the j-th sample in the i-th type of data. V^i is calculated as: $V^i = \|(I - \frac{\mathbf{11^T}}{n})(I - (Y^i W^i)^\dagger (Y^i W^i))\|_F$.

Fixed W, global "sample pattern" Ycan be acquired by computing matrix Φ. The updating rule is:

$$\mathrm{Y} = \Psi_d \tag{5}$$

Where the eigenvalues of matrix Φ are sorted by ascending, matrix Ψ_d consist of d eigenvector corresponding to d eigenvalue from 2 to d+1. $d = min(d^1, d^2, \cdots, d^k)$ and $d^i (i = 1, 2, \cdots, k)$ denotes the dimension of i-th type of data.

Fixed Y, update W by the following criterion:

$$\begin{aligned} &\min \Delta^T W + \lambda \|W\|_2^2 \\ &s.t. \begin{cases} W^T \mathbf{1} = 1 \\ W \geq 0 \end{cases} \end{aligned} \tag{6}$$

Where, $\Delta = (\phi_1, \phi_2, \cdots, \phi_{i*j}, \cdots, \phi_M), M = n1 + n2 + \cdots + nk, \phi_{i*j} = \frac{\|x_j - b - L^i x_j^i\|^2}{V^i}$, and the values in Δ are sorted by ascending, namely $\phi_1 \leq \phi_2 \leq \cdots \leq \phi_M$. x_j^i and x_j represent the j-th sample of the i-th types of data and the global information of the j-th sample, b is d dimension vector and is designed to centralized over matrix Y.

The optimization problem above can be summarized as:

$$\begin{cases} w_m = \frac{\theta - \varphi_m}{2\lambda}, m = 1, \ldots, N \\ w_m = 0, m = N+1, \ldots, M \\ \theta = \frac{2\lambda + \sum_{m=1}^{N} \varphi_m}{N} \\ N = \arg\max_m (\theta - \varphi_m > 0) \end{cases} \tag{7}$$

By fixing W can we update Y at meeting the conditions above and vice versa. Finally, we obtained optimal global "sample pattern" Y when Y reaches convergence.

3.4 Graph Tree Convolution Network (GTCN)

Graph tree convolution network (GTCN) [23] makes message passing by the tree representation of graphs. Different from the schedule of massage passing in traditional graph network, the GTCN represents graphs utilizing the tree structure which delivers information from leaf node to root node, and each node will retain initial information itself before receiving updating from neighboring nodes. The updating of one node in GTCN is fulfilled by aggregating its original value and the renewed values of its neighboring nodes. Figure 2 shows the overall structure of graph tree convolution network.

The graph topology structure $g(v, \epsilon)$ consists of nodes v and edges ϵ which can be described by adjacency matrix A and degree matrix D completely. In addition, $\|v\| = N$ and $\|\epsilon\| = M$ are the number of nodes and edges in graph structure, respectively. $\mathcal{N}_u$ denotes the set of the direct (1-hop) neighbor of node u, and $X \in \mathbf{R}^{N \times D}$ represents features mapping over all nodes. $x_u \in \mathbf{R}^{1 \times D}$ is the feature vector of node u with D dimension. $Y \in \mathbf{R}^{N \times C}$ is the class matrix of all nodes and $y_u \in \mathbf{R}^{1\ times C}$ represents the class vector of node u with C dimension.

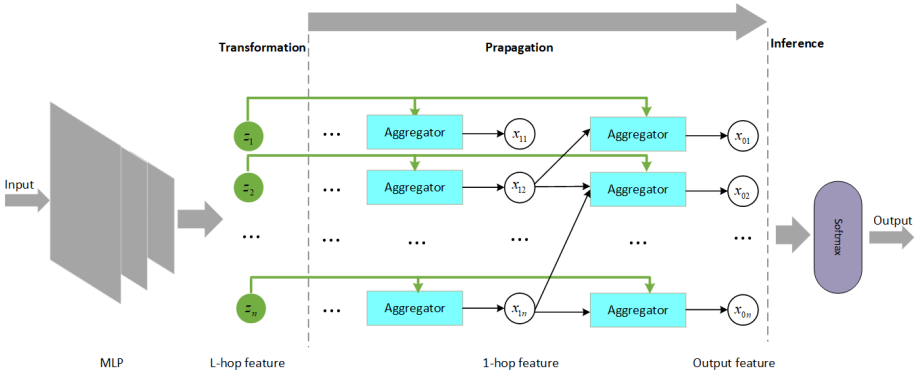

Fig. 2. The overall structure of graph tree convolution network with an input example.

Massage passing in GTCN is calculated as:

$$h_u^k = \sum_{v \in N_u} \hat{A}_{uv} h_u^{k+1} + \hat{A}_{uu} Z_u \tag{8}$$

Here, $\hat{A} = \hat{D}^{-1/2}\hat{A}\hat{D}^{-1/2}$ denotes symmetric normalized adjacency matrix, and $\hat{A} = (A + I)$ is the addition of adjacency matrix with self-loop. $\hat{D}$ represents the corresponding degree matrix, and $\hat{D}_{uu} = \sum_v \hat{A}_{uv}, Z_u = MLP(x_u)$.

As to the real label Y of input data, loss function constructed as:

$$L_{GTCN} = -\log \frac{exp(\hat{y}_{i,y_i})}{\sum_{c=0}^{C-1} exp(\hat{y}_{i,c})} \tag{9}$$

Where, y_i denotes true label of i-th input data, and $\hat{y}_i$ is i-th prediction vector via softmax function by GTCN output.

4 Experiments

We compared the classification performance of Omi-PGTCN with latest multi-omics integration and classification models. And we also demonstrated the advancement of components in Omi-PGTCN which can be substituted by other methods by ablation studies. The network structure was implemented in the same way as figure 1. The model optimized parameters by Adam [7], and the learning rate was set to and batch size was set to 32. For the robust performance comparison of different models, a stratified 10-fold cross-validation evaluation idea was applied to all models in this paper as to avoid bias from single testing set.

4.1 First Stage

As shown in Figure 1, this stage includes VAE and PF, which is utilized to reduce dimension and fusion of data, respectively. We combined a variety of

dimension reduction models with a classifier to compare their performance. The result obtained after stratified 10-fold cross-validation is demonstrated in Table 1. At this moment, the output dimension of the model is 128.

We selected six kinds of method, principal component analysis (PCA) [1], kernel principal component analysis (kPCA) [8], factor analysis (FA) [11], auto-encoder (AE) [15] and our proposed model, to reduce dimension of multi-omics data. They are applied to decrease dimension of multi-omics features across 33 types of tumors with a total of 6133 samples. As we can see in the table, our proposed dimension reduction model outperformed others in all the four-classification metrics.

Table 1. Classification performance of representations with different dimension reduction models.

Method	Key Performance Indicator			
	Accuracy	F1 score	Precision	Recall
PCA+classifier	$86.23 \pm 0.63\%$	$85.98 \pm 0.61\%$	$83.86 \pm 0.86\%$	$81.43 \pm 0.45\%$
kPCA+classifier	$86.54 \pm 0.55\%$	$86.23 \pm 0.59\%$	$84.52 \pm 0.65\%$	$81.76 \pm 0.41\%$
FA+classifier	$88.05 \pm 0.34\%$	$86.26 \pm 0.53\%$	$85.55 \pm 0.73\%$	$80.43 \pm 0.55\%$
AE+classifier	$90.04 \pm 0.50\%$	$89.04 \pm 0.64\%$	$85.96 \pm 0.80\%$	$83.01 \pm 0.23\%$
VAE+classifier	$\mathbf{96.19 \pm 0.05\%}$	$\mathbf{91.2 \pm 0.40\%}$	$\mathbf{88 \pm 0.30\%}$	$\mathbf{89 \pm 0.10\%}$

4.2 Second Stage

In the first stage, we do not need to train PF because of its whole mathematical theory just like PCA. In this stage, we compared the performance of the GTCN model which needs the input of both data and corresponding adjacency matrix with the latest classification model in 33 types of tumors. Table 2 displayed the experimental outcome with eight excellent models. As shown in the table, the classification performance of graph structural network is better than that of classifier that only requires inputting data, which reflects the superiority of graph neural network. In addition, the GTCN model outperformed those that have homogeneous structure.

To demonstrate the necessity and advancement of PF component, we replaced PF with similarity network fusion (SNF) and removed it in common classification methods. Table 3 shows the experimental results after substituting the PF component.

Besides the dual-stage training method, the Omi-PGTCN model has another end-to-end learning way which is more convenient but accounts for extra storage than the first mothed. Figure 3shows the classification performance of various models in which Omi-PGTCN trained by end-to-end way and the accuracy of Omi-PGTCN in different hyperparameters, respectively.

The average classification accuracy on different types of tumors (33 classes) achieved by Omi-PGTCN model is 97.90% after 10-fold cross-validation and the maximal standard deviation is 1.01%.

Table 2. Classification performance of different models in 33 types of tumors from TCGA.

Method	Key Performance Indicator			
	Accuracy	F1 score	Precision	Recall
VAE+LR	86.04 ± 0.65%	87.53 ± 0.52%	85.52 ± 1.05%	80.76 ± 0.91%
VAE+SVM	89.65 ± 0.54%	87.66 ± 0.50%	85.95 ± 0.93%	80.53 ± 0.85%
VAE+MLP	90.44 ± 0.5%	89.86 ± 0.74%	88.06 ± 0.88%	83.51 ± 0.13%
VAE+PF+GNN	92.01 ± 0.7%	92.11 ± 0.65%	88.52 ± 1.04%	81.99 ± 1.00%
VAE+PF+GTN	91.12 ± 0.6%	90.01 ± 0.88%	83.45 ± 1.09%	83.75 ± 1.01%
VAE+PF+GCN	89.99 ± 1.1%	89.15 ± 0.32%	89.51 ± 0.55%	85.99 ± 0.81%
VAE+PF+GTAN	91.80 ± 0.7%	91.04 ± 0.65%	86.36 ± 1.1%	83.03 ± 1.0%
VAE+PF+GTCN (Omi-PGTCN)	**97.90 ± 0.6%**	**92.03 ± 0.3%**	**90 ± 1.01%**	**91 ± 1.0%**

Table 3. Classification performance of different models in which PF substituted by SNF.

Method	Key Performance Indicator			
	Accuracy	F1 score	Precision	Recall
VAE+SNF+GNN	90.10 ± 0.8%	90.01 ± 0.55%	86.12 ± 1.0%	81.59 ± 1.01%
VAE+SNF+GTN	90.02 ± 0.5%	89.21 ± 0.80%	81.50 ± 0.99%	83.15 ± 1.02%
VAE+SNF+GCN	89.90 ± 1.0%	88.10 ± 0.32%	88.01 ± 0.85%	85.49 ± 0.78%
VAE+SNF+GTAN	90.90 ± 0.8%	90.34 ± 0.65%	85.06 ± 1.1%	82.13 ± 1.0%
VAE+SNF+GTCN	**96.90 ± 0.7%**	**91.03 ± 0.5%**	**89.23 ± 1.0%**	**89.05 ± 1.0%**

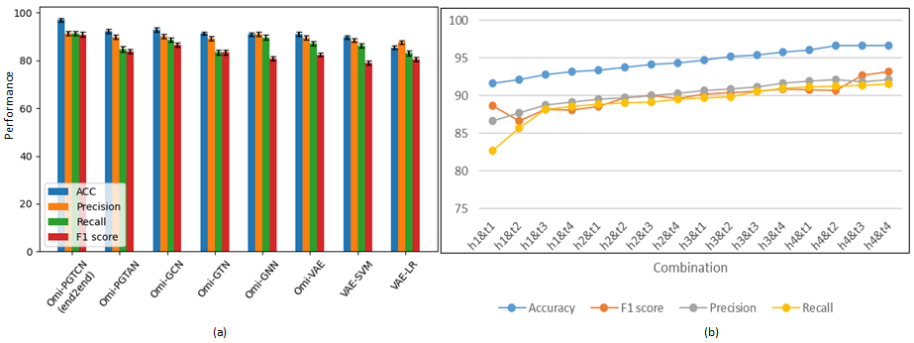

Fig. 3. The classification performance including comparison of different models and the Omi-PGTCN in different hyperparameters. (a) demonstrated the classification of performance of Omi-PGTCN model trained by end-to-end mode with alternatives including Omi-PGTAN, Omi-GCN, Omi-GTN, Omi-GNN Omi-VAE, VEA-SVM and VAE-LR. (b) shown the evaluation of Omi-PGTCN in different hops and thresholds, in which h1 to h4 represent 5, 10, 15, 20, respectively, and t1 to t4 indicate 0.7, 0.8, 0.9, 0.95, separately.

4.3 Biomarker Identification

According to the first-layer weights of improved VAE trained, we can get the rankings of vital biomarkers identified by Omi-PGTCN. In addition, the number of top influential genes can be selected on the basis of first-layer weights. In the paper, we set top-n equal to 10 and verify these genes of different types of omics data by survived analysis. The top-n genes discerned by Omi-PGTCN are shown in table 4.

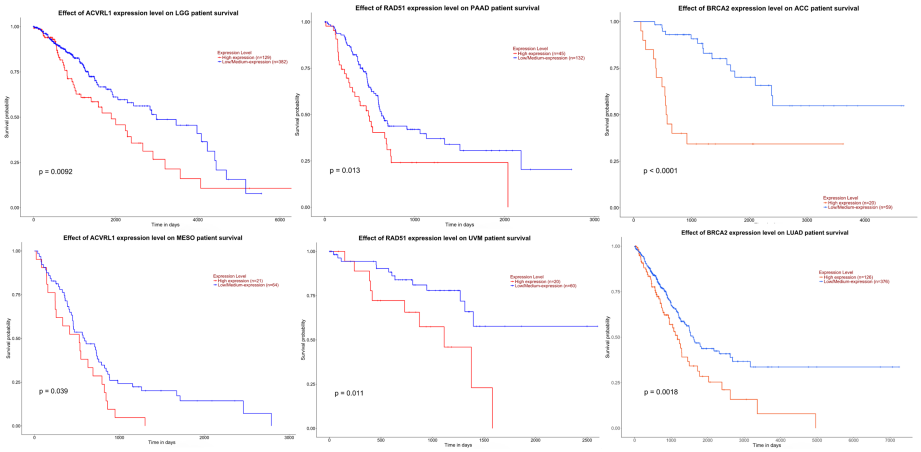

Fig. 4. The survival analysis curve of important biomarkers in different cancers (partial).

Then, the survived analysis curse of various biomarkers graphed in figure 4, separately. As we can see from these pictures, the vital genes in table 4 have great influence on one or more types of tumors and perhaps are driver genes of a caner. Hence, it is obvious that the significant biomarkers recognized by Omi-PGTCN are instructive for cancer clinical doctors or researchers.

Table 4. Classification performance of representations with different dimension reduction models.

Omics data type	Identified Biomarkers
DNA methylation	TMC4, HYAL2, TTC15, GPR37L1, OR1J4, ATP10B, TMEM207, CDH26, MT1DP, AGA
RNA expression	NPNT, CDK18, APLN, SYTL1, ARRDC2, DSG1, UGT8, SOX10, PI3, SERPINB5
Protein expression	ACVRL1, RAD51, BRCA2, MSH6, CDK1, BID, ATM, SMAD4, IGFBP2, BCL2

5 Conclusion

It makes integrating analysis of various types of genes complicated due to the extreme dimension of the features of multi-omics data. Even some simple tasks like tumor classification are also difficult to accomplish. Omi-PGTCN, a whole process optimized model, which achieves convincing tumor classification and biomarker identification outcome by optimizing each process of processing multi-omics data. In the first stage, Omi-PGTCN reduces the dimension of multi-omics data, DNA methylation, RNA expression and protein expression, by improving the VAE model to avoid the phenomenon of 'dimensionality curse' and address the problem of 'over-fitting' caused by insufficient samples. At the same time, the PF method in the Omi-PGTCN model is used to integrate these omics data for maximal extracting the interactive information of multi-omics data and utilizing the signals sufficiently. In the second stage, the GTCN component of Omi-PGTCN classifies given samples to 33 types of cancer by signals that dimension decreased by improved VAE and interactive message of multi-omics data integrated by PF. Finally, Omi-PGTCN achieved an average classification accuracy of 97.06% on 33 kinds of tumors and excellent ability to identify critical biomarker after 10-fold cross-validation. Moreover, the Omi-PGTCN model can also visualize processing of each component because of its decomposability so as to understand the mechanism of happening of different tumors and the impact of various omics features on the formation and treatment of a tumor.

References

1. Abdi, H., Williams, L.J.: Principal component analysis. Wiley interdisciplinary reviews: computational statistics **2**(4), 433–459 (2010)
2. Acharjee, A., Kloosterman, B., Visser, R.G., Maliepaard, C.: Integration of multi-omics data for prediction of phenotypic traits using random forest. BMC bioinformatics **17**(5), 363–373 (2016)
3. Altman, N., Krzywinski, M.: The curse (s) of dimensionality. Nat Methods **15**(6), 399–400 (2018)
4. Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J.C., Buettner, F., Huber, W., Stegle, O.: Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. Molecular systems biology **14**(6), e8124 (2018)
5. Berger, B., Peng, J., Singh, M.: Computational solutions for omics data. Nature reviews genetics **14**(5), 333–346 (2013)
6. Chaudhary, K., Poirion, O.B., Lu, L., Garmire, L.X.: Deep learning–based multi-omics integration robustly predicts survival in liver cancerusing deep learning to predict liver cancer prognosis. Clinical Cancer Research **24**(6), 1248–1259 (2018)
7. Chen, C., Carlson, D., Gan, Z., Li, C., Carin, L.: Bridging the gap between stochastic gradient mcmc and stochastic optimization. In: Artificial Intelligence and Statistics. pp. 1051–1060. PMLR (2016)
8. González, D., Aguado, J.V., Cueto, E., Abisset-Chavanne, E., Chinesta, F.: kpca-based parametric solutions within the pgd framework. Archives of Computational Methods in Engineering **25**(1), 69–86 (2018)

9. Huang, Z., Zhan, X., Xiang, S., Johnson, T.S., Helm, B., Yu, C.Y., Zhang, J., Salama, P., Rizkalla, M., Han, Z., et al.: Salmon: survival analysis learning with multi-omics neural networks on breast cancer. Frontiers in genetics **10**, 166 (2019)
10. Huda, S., Yearwood, J., Jelinek, H.F., Hassan, M.M., Fortino, G., Buckland, M.: A hybrid feature selection with ensemble classification for imbalanced healthcare data: A case study for brain tumor diagnosis. IEEE access **4**, 9145–9154 (2016)
11. Joliffe, I.T., Morgan, B.: Principal component analysis and exploratory factor analysis. Statistical methods in medical research **1**(1), 69–95 (1992)
12. Li, Y., Kang, K., Krahn, J.M., Croutwater, N., Lee, K., Umbach, D.M., Li, L.: A comprehensive genomic pan-cancer classification using the cancer genome atlas gene expression data. BMC genomics **18**(1), 1–13 (2017)
13. Lyu, B., Haque, A.: Deep learning based tumor type classification using gene expression data. In: Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics. pp. 89–96 (2018)
14. Mahmud, M.S., Huang, J.Z., Fu, X.: Variational autoencoder-based dimensionality reduction for high-dimensional small-sample data classification. International Journal of Computational Intelligence and Applications **19**(01), 2050002 (2020)
15. Ramamurthy, M., Robinson, Y.H., Vimal, S., Suresh, A.: Auto encoder based dimensionality reduction and classification using convolutional neural networks for hyperspectral images. Microprocessors and Microsystems **79**, 103280 (2020)
16. Rhee, S., Seo, S., Kim, S.: Hybrid approach of relation network and localized graph convolutional filtering for breast cancer subtype classification. arXiv preprint arXiv:1711.05859 (2017)
17. Shi, Q., Zhang, C., Peng, M., Yu, X., Zeng, T., Liu, J., Chen, L.: Pattern fusion analysis by adaptive alignment of multiple heterogeneous omics data. Bioinformatics **33**(17), 2706–2714 (2017)
18. Titus, A.J., Bobak, C.A., Christensen, B.C.: A new dimension of breast cancer epigenetics. In: 9th International Conference on Bioinformatics Models, Methods and Algorithms (2018)
19. Van De Wiel, M.A., Lien, T.G., Verlaat, W., van Wieringen, W.N., Wilting, S.M.: Better prediction by use of co-data: adaptive group-regularized ridge regression. Statistics in medicine **35**(3), 368–381 (2016)
20. Wang, B., Mezlini, A.M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., Goldenberg, A.: Similarity network fusion for aggregating data types on a genomic scale. Nature methods **11**(3), 333–337 (2014)
21. Wang, Z., Wang, Y.: Exploring dna methylation data of lung cancer samples with variational autoencoders. In: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 1286–1289. IEEE (2018)
22. Way, G.P., Greene, C.S.: Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. In: PACIFIC SYMPOSIUM ON BIOCOMPUTING 2018: Proceedings of the Pacific Symposium. pp. 80–91. World Scientific (2018)
23. Wu, N., Wang, C.: Gtnet: A tree-based deep graph learning architecture. arXiv preprint arXiv:2204.12802 (2022)
24. Zhang, X., Zhang, J., Sun, K., Yang, X., Dai, C., Guo, Y.: Integrated multi-omics analysis using variational autoencoders: application to pan-cancer classification. In: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 765–769. IEEE (2019)