SADLN: Self-Attention Based Deep Learning Network of Integrating Multi-omics Data for Cancer Subtype Recognition

Ping Gong (gongping@xzhmu.edu.cn)
Xuzhou Medical University
Qiuwen Sun
Xuzhou Medical University
Lei Cheng
Xuzhou Medical University
Zhiyuan Zhang
Xuzhou Medical University
Shuguang Ge
China University of Mining and Technology
Jie Chen
Affiliated Hospital of Xuzhou Medical University
Longzhen Zhang
Affiliated Hospital of Xuzhou Medical University

Research Article

Keywords: Self-attention, Deep learning, Multi-omics data, Gaussian Mixture Model, Cancer subtype recognition

Posted Date: January 19th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1112753/v1

RESEARCH

SADLN: Self-Attention Based Deep Learning Network of Integrating Multi-omics Data for Cancer Subtype Recognition

Ping Gong[1*], Qiuwen Sun[1], Lei Cheng[1], Zhiyuan Zhang[1], Shuguang Ge[2], Jie Chen[3] and Longzhen Zhang[3]

[*]Correspondence:
gongping@xzhmu.edu.cn
[1]School of Medical Imaging, Xuzhou Medical University, Xuzhou, CN
Full list of author information is available at the end of the article

Abstract

Background: Integrating multi-omics data for cancer subtype recognition is an important task in bioinformatics. Recently, deep learning has applied to recognize the subtype of cancers. However, existing studies almost integrate the multi-omics data simply by concatenation as the single data and then learn a latent low-dimensional representation through deep learning model, which didn't considering the distributes differently of omics data. Moreover, these methods ignore the relationship of samples.

Results: In order to tackle these problems, we proposed SADLN: a self-attention based deep learning network of integrating multi-omics data for cancer subtype recognition. SADLN combined encoder, self-attention, decoder, and discriminator into a unified framework, which can not only integrate multi-omics data but also adaptively model the sample's relationship for learning a accurately latent low-dimensional representation. With the integrated representation learned from the network, SADLN used Gaussian Mixture Model to identify cancer subtypes. Experiments on ten cancer datasets of TCGA demonstrated the advantages of SADLN compared to ten methods.

Conclusions: The Self-Attention Based Deep Learning Network (SADLN) is a effective method of integrate multi-omics data for cancer subtype recognition.

Keywords: Self-attention; Deep learning; Multi-omics data; Gaussian Mixture Model; Cancer subtype recognition

Introduction

Cancer is a heterogeneous disease caused by changes at the transcription, expression, epigenetic and proteomic level of cellular components. That is the same cancer have different subtypes which influence the clinical treatment and prognosis[1][2]. Therefore, cancer subtype recognition is key to an improved and more personalized prognosis and treatment[3]. With the developments of high-throughput biology, there yield large amounts of multi-omics data, such as miRNA expression data, mRNA expression data, DNA methylation data, copy number variation and et al.[4][5]. These multi-omics data can be obtained by some publicly available projects. For example, The Cancer Genome Atlas (TCGA)[6] stories more than 30 cancers over 11,000 patients' data.

How to integrate multi-omics data and construct efficient models has become the major challenges in cancer subtype recognition[7][8][9]. Researchers have proposed

many methods. They can be categorized into three different types: early integration methods, late integration methods and intermediate integration methods[10].

Early integration methods are the most simple approaches. They concatenate different omics's feature matrices to a single matrix, and use the single omics clustering algorithm to subtype the matrix[10]. For example, K-means, LRAcluster[11] and Spectral clustering are all belong to this category. The late integration clusters each omics data individually and integrates the clustering schemes to obtain a single clustering scheme[10]. The intermediate integration methods construct the model to explain all omics. These methods include: (i) Methods that integrate sample similarities. For example, similarity network fusion(SNF)[12] as a widely mentioned algorithm, constructs the samples's similarity network of each omics data and iteratively updates the network until they converge, finally uses spectral clustering to partition them. NEighborhood based multi-omics clustering (NEMO)[13] is much simpler than SNF, it has represented the similarity between all omic samples. (ii) Methods that use joint dimension reduction. For example, a classical dimensionality reduction algorithm Canonical correlation analysis (CCA), projecting the two omics data into a low dimensions space and making the projected data has maximum correlation. Multiple canonical correlation analysis (MCCA)[14] extended CCA to more than two omics. The two methods combine different biological characteristics to obtain a low dimensional space. (iii) Methods that use statistical modeling of the data. For example, iCluster[15] is an effective method, which constructs the Gaussian latent variable model.

As artificial intelligence development, deep learning has been widely used in health care, such as imaging based computer aided diagnosis, digital pathology, drug design, prediction of hospital admission, classification of cancer and so on. Compared with traditional methods, deep learning can automatically extract features from the original data and learn more abstract and useful features. Recently, some deep learning models have been applied in cancer subtype recognition[16][17]. AutoEncoder (AE) and Variational Autoencoders (VAE) as two variants of deep learning model have demonstrated good performance for generating meaningful feature representation of omics data[18]. Some studies have highlighted their utility in integrating different omics data for identifying prognostic cancer traits such as breast[19], liver[20] and neuroblastoma cancer subtypes[21]. However, these methods are almost based on early integration, which combining different omics data into a signal data and learning a common latent low dimensional representation[22]. They ignore the distributions of different omics and increase the dimension of input data, which may lead to do not understanding or over fitting of complex processes.

In order to solve these problem, few researchers have proposed deep learning based middle integration methods[23][24][25]. These methods separately learned each omics data through some sub network, and then integrated the output of every sub network into a unified representation. For example, Tong et al. proposed ConcatAE, a method concatenating features learned from each omics using autoencoder[26]. Yang et al. proposed Subtype-GAN, an approach used multi-input multi-output neural network to separately model multi-omics data[27]. Although these methods have demonstrated good performance in cancer subtype recognition, they ignore the relationship between samples when learning valuable features representation.

More recently, attention mechanism has become a new technology in the field of deep learning. The dominant thought is to measure the similarity between the Key and the Query[28]. Attention mechanism has been applied in speech NLP, image and other fields[29][30][31][32], since it can select most informative features of an input, adaptively consider the importance of a single feature and allow the model to make more accurate judgements. As a special, self-attention[33][34], which calculating the response at a position in the sequence by attending to all positions within the same sequence has achieved notable success in modeling complicated relations[35]. For instance, it displays the superiority in machine translation[36], sentence embedding[37] of modeling arbitrary word dependency, and has been successfully applied to capture node similarities in graph embedding[38]. Research shows that the attention-based encoder is more fit for learning high-level features[39].

Motivated by these works, we proposed SADLN : a self-attention based deep learning network of integrating multi-omics data for cancer subtype recognition. Firstly, it used independent subnetwork to learn a compact representation of each omics data and concatenated them to a concatenation representation. Then a self-attention model was used to learn the similarities of samples on concatenation representation, the final integrated feature representation obtained by summing the similarity of samples and the concatenation representation. Finally, the learned representation was used as the input of the Gaussian Mixture Model (GMM) for cancer subtyping recognition. In addition, we added decoder to reconstruct the original multi-omics data from the integrating representation. In order to fit the learned representation distributions to the Gaussian distribution, we added discriminator to the network.

The main contribution summarized as follows:

(1) We proposed novel deep learning method, SADLN, which combine encoder, self-attention, decoder, and discriminator into a unified framework. It can simultaneously integrate multi omics representation and samples' relations.

(2) We firstly introduced the self-attention into the deep learning based method for cancer subtyping recognition task which allow the model to autonomously learn the similarity of samples for better representation.

(3) We conducted experiments on ten cancer datasets of TCGA, SADLN achieved outstanding performance compared with ten integration methods. It provided theoretical basis and new method for clinical diagnosis and precise treatment of cancer, which has great theoretical significance and clinical application value.

Materials and methodology

Materials

SADLN was applied to ten cancer datasets which preprocessed by Yang et al[27] from TCGA. The datasets include BRCA, LUAD, BLCA, PAAD, KIRC, STAD, UVM, GBM, SKCM and UCEC. Each cancer dataset contains four types omics data, i.e., copy number,DNA methylation, mRNA and miRNA. The dimensions of the four types omics data are 3105, 3217, 383 and 3139, respectively.

Methodology

The overview architecture of SADLN is depicted in Figure1, it is consist of four key modules: a self-attention based encoder, decoder, discriminator and GMM cluster-

ing. In SADLN, the input is cancer example's four omics data, the output is the number of subtypes.

The self-attention based encoder

The self-attention based encoder, in our SADLN model transforms the multi-omics data into a latent low-dimensional feature representation with distribution $N(\mu, \sigma^2)$ using multiple independent network layers, a self-attention model and two fully-connected layers. For each sub-independent layer, let $\mathbf{x}^m = \{x_1^m, ..., x_N^m\} \in R^{N \times D_m}$ denotes the input of the network for the m-th omics data, $\mathbf{y}^m = \{y_1^m, ..., y_N^m\} \in R^{N \times d_m}$ denotes the output of the m-th omics through the sub-independent layer, N is the number of data samples, D_m and d_m are the feature dimension of the input data and the output data respectively. We used Dense network to extract each omics feature, that's $\mathbf{y}^m$ can be express as:

$$\mathbf{y}^m = \mathbf{w}_m \mathbf{x}^m + \mathbf{b}^m \tag{1}$$

where $\mathbf{w}_m$ is the weight matrix, $\mathbf{b}_m$ is the bias.

In order to integrate different omics features, firstly, we concatenate four omics features matrices into a feature representation matrix. The representation feature matrix $\mathbf{Y}$ can be expressed as:

$$\mathbf{Y} = Concat(\mathbf{y}^1, ..., \mathbf{y}^4) \tag{2}$$

After concatenation, the output of the matrix is became $N \times 4d$. In order to avoid over fitting, batch normalization layers were added and the Gaussian Error Linear Unit(GELU) was used as the activation function. That is:

$$\mathbf{Y}^{'} = GELU(\mathbf{Y}) \tag{3}$$

Although the concatenation operation can integrate multi omics data, the relationship between samples is not considered. In order to further integrating the relationship between samples, we introduced self-attention to construct the similarity between samples. We treated each sample's multi-omics features as a word in a sentence.

Let $d_k = 4d$, $\mathbf{K} = [k_1, k_2, ..., k_N] \in R^{N \times d_k}$ is a set of keys, $\mathbf{Q} = [q_1, q_2, ..., q_N] \in R^{N \times d_k}$ is a set of queries, $\mathbf{V} = [v_1, v_2, ..., v_N] \in R^{N \times d_k}$ is a set of values, $\mathbf{K} = \mathbf{Q} = \mathbf{V} = \mathbf{Y}'$, $\mathbf{K} = \mathbf{Y}'\mathbf{W}^K$, $\mathbf{Q} = \mathbf{Y}'\mathbf{W}^Q$, $\mathbf{V} = \mathbf{Y}'\mathbf{W}^V$. $\mathbf{W}^K$,$\mathbf{W}^Q$,$\mathbf{W}^V$ are the parameters of linear projection layers. $\mathbf{z} = \{z_1, z_2, ..., z_N\} \in R^{N \times d_k}$ denotes the finally integrating representation, the jth feature vector z_j is computed as the following steps[40]. Firstly, we use the dot-product between q_i and k_j to compute the similarity of the sample i and j. In order to ensure the result does not get excessively large, we scale it by $\sqrt{d_k}$. That is:

$$r_{i,j} = \frac{q_i \times k_j^T}{\sqrt{d_k}} \tag{4}$$

Secondly, softmax function was used to obtain the similarity weight. That is:

$$\omega_i = softmax\{\frac{q_i \times k_1^T}{\sqrt{d_k}}, \frac{q_i \times k_2^T}{\sqrt{d_k}}, ..., \frac{q_i \times k_N^T}{\sqrt{d_k}}\} \tag{5}$$

Thirdly, the final integrated feature vector z_i of sample i can be obtained by a weighted sum of the values. That is:

$$\mathbf{z}_i = Attention(q_i, \mathbf{K}, \mathbf{V}) = \sum_{j=1}^{N} \omega_i v_j \tag{6}$$

Finally, the integrated feature representation can be express as:

$$\mathbf{z} = Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = [\mathbf{z}_1, \mathbf{z}_2, ...\mathbf{z}_N] \in R^{N \times d_k} \tag{7}$$

In order to keep the data distribution unchanged, we added batch normalization layers after self-attention model.

Suppose $\mathbf{z}$ obeys Gaussian distribution $\mathbf{z} \sim N(\mu, \sigma^2)$, where μ is the mean and σ^2 is the variance. In this paper, we obtained μ and σ^2 through two fully-connected layers.

Decoder

Decoder, in our SADLN model attempts to reconstruct the original multi-omics data from the integrating representation $\mathbf{z}$. As shown in the upper right halves of Figure1 , it contains fully connected layers and output layer. Let $\mathbf{X}_I = \{\mathbf{x}_I^1, \mathbf{x}_I^2, \mathbf{x}_I^3, \mathbf{x}_I^4\}$denotes the input of encoder, $\mathbf{X}_O = \{\mathbf{x}_O^1, \mathbf{x}_O^2, \mathbf{x}_O^3, \mathbf{x}_O^4\}$ denotes the output of decoder. In order to minimize the error between the input $\mathbf{X}_I$ and the output $\mathbf{X}_O$[41], the square euclidean distance was applied to calculate the loss L_1[27], it can be expressed as:

$$L_1 = \|\mathbf{X}_I - \mathbf{X}_O\|_2^2 = \frac{1}{4}\sum_{k=1}^{4} \| \mathbf{x}_I^k - \mathbf{x}_O^k \|_2^2 \tag{8}$$

Discriminator

In order to force the posterior distribution $S(z)$ of final integrated representation matches the prior Gaussian distribution $P(z)$, we added a discriminator D to the model, which is a part of GAN network. A typical GAN network composed of a generator G and a discriminator D. In this work, we regards the self-attention base encoder part as the generator, the input of the discriminator D is the output of the encoder part and the randomly sampled data with standard normal distribution. The discriminator D is used to distinguish the samples from $P(z)$ or the $S(z)$[27]. Through adversarial learning, $S(z)$ is as close to $P(z)$ as possible.

The objective function optimization of discriminator D adopts the method of maximization and minimization. It can be express as:

$$\min_S \max_D E_{z' \sim P(z)}[log_2 D(z')] + E_{z \sim S(z)}[log_2(1 - D(S(z)))] \tag{9}$$

Where E represents the expected value of the distribution function. We use binary_crossentropy function to train the discriminator learning process. The cost of the discriminator is:

$$L_2 = -E_{z' \sim P(z)}[logD(z')] - E_{z \sim S(z)}[log(1 - D(S(z)))] \tag{10}$$

Our model parameters of the whole network are jointly trained by minimizing the following total loss:

$$L = \lambda_1 L_1 + \lambda_2 L_2 \tag{11}$$

where L_1 and L_2 are defined in Eq.(8) and Eq.(10) respectively. λ_1 and $\lambda_2 \in [0, 1]$ are trade-off parameters.

The GMM Clustering

GMM is a probabilistic clustering method, which also belongs to the generative model. It assumes that all the data points are generated from a mixture of a finite number of Gaussian distributions[42]. GMM model has excellent clustering performance. In this paper, we use GMM as the clustering module. Let K denotes the number of clusters, $\pi = (\pi_1, \pi_2, ..., \pi_k)$ represent the weight of each cluster, $\mu = (\mu_1, \mu_2, ..., \mu_k)$ is the mean vector, $\sum = (\sum_1, \sum_2, ..., \sum_k)$ is the covariance vector, $Z = \{z_n\}_{n=1}^{N}$ is the final integrated feature representation, $p(z_n)$ is the probability distribution function as a mixture of K Gaussian distributions. That is:

$$p(z_n) = \sum_{k=1}^{K} \pi_k p_k(z_n) = \sum_{k=1}^{K} \pi_k N(z_n \mid \mu_k, \sum\nolimits_k) \tag{12}$$

GMM used the EM algorithm to update the parameters π, μ and $\sum$. According to the maximum probability density of sample in different clusters, the most suitable subtype labels is obtained.

Experiments

Comparison algorithms and evaluation metrics

In order to verify the performance of SADLN, we compared it with ten methods. Three deep learning based methods include AE, VAE and Subtype-GAN. Seven non-deep learning based methods include K-means, LRAcluster, iCluster, Spectral, NEMO, MCCA and SNF. The ten chosen methods can represent different types approaches of multi-omics integrating. K-means, LRAcluster, Spectral clustering, AE and VAE are belong to early integration methods. SNF is used as similarity based algorithms. MCCA is used as dimension reduction algorithms.

We used the P values based on Cox log-rank model[43] to measure differential survival between the obtained subtype. With the decrease of P value, the subtype's survival rate is more significant, the clustering effect is more obvious. Meanwhile, the enrichment of clinical labels[44] were used to test the clustering results. Six clinical labels including pathologic T, age at diagnosis, pathologic M, gender, pathologic N and pathologic stage were used for testing. The ten cancers' clinical parameters were not all available, such as GBM and UCEC only had two clinical parameters.

Table 1 The network structure of SADLN

Architectures	SADLN	
Self-Attention Based Encoder	3105+3217+383+3139(Input)	
	25+25+25+25(concatenate)	
	100(Batch Normalization)	
	100(Activation)	
	100(Attention)	
	100(Batch Normalization)	
	100(Fully-Connected)	
	100(Fully-Connected,Mean)	100(Fully-Connected,VAR)
	100(Output)	
Decoder	100(Input)	
	100(Fully-Connected)	
	100(Batch Normalization)	
	100(Activation)	
	3105+3217+383+3139(Output)	
Discriminator	100(Input)	
	1(Fully-Connected)	
	1(Sigmoid)	
	1(Output)	

Network structure and hyperparameter setting

The SADLN network has 19 layers, including 10 layers encoder, 5 layers decoder, 4 layers discriminator. The specific network structure of SADLN is shown in Table1. The SADLN is built based on python 3.6.12, Keras 2.2.4, TensorFlow 1.14.0 (the CPU version). The operating system is Windows 10. In terms of hardware, the CPU is Intel(R) Core (TM) i7-105 10U.

Optimizing hyperparameters are the key to training neural network models. Choosing appropriate hyperparameters can significantly improve the performance of the model. In this paper, the hyperparameters of the SADLN model mainly include the feature dimension of the independent network (d), the initial epoch, batch size, random seed, optimizer, activation function, learning rate and loss. Table2 shows the hyperparameter settings of the SADLN model.

Table 2 Hyperparameter settings of SADLN model

Hyperparameter	Setting
d	25
Epoch	600
Batch size	64
Random seed	2
Optimizer	Adam Optimizer
Activation function	Gaussian Error Linear Unit
Learning rate (lr)	[1e-4,2e-4,3e-4,4e-4,5e-4, 1e-5,2e-5,3e-5,4e-5,5e-5]
Loss λ_1	0.0001
Loss λ_2	0.2499

Experiments Results

Comparison experiments on TCGA datasets

In order to reduce the influence of different clustering numbers on the results of subtyping. The clustering number of BRCA, LUAD, BLCA, PAAD, KIRC, STAD, UVM, GBM, SKCM and UCEC were 5, 3, 5, 2, 3, 3, 4, 3, 4, 4, respectively.

Table3 gives the -log10P-values of survival analysis for eleven methods of ten cancers datasets on TCGA. The clustering results of other ten compared methods comes from Yang's literature[27]. Bold indicates that this method performs best on the corresponding cancer dataset.

As shown in Table3, SADLN achieved the most significant results on PAAD, STAD, UCEC and UVM cancer datasets. Compared with Subtype-GAN, SADLN obtained better value on eight cancer datasets(BLCA, GBM, KIRC, PAAD, SKCM, STAD, UCEC and UVM). Compared with VAE and AE, SADLN obtained the best -log10P value in ten cancer datasets. Compared with non-deep learning based methods, although same methods had best results in specific cancer datasets, but the -log10P value were highest on most cancer datasets.

Table4 give the clinical parameters enrichment analysis result of SADLN and other compared methods of ten cancer datasets. From Table4, we can see that SADLN obtained the best results on four datasets (KIRC, LUAD, STAD, UCEC). Therefore, we believe that SADLN is competitive with other methods in cancer subtype recognition.

Table 3 The -log10P values of survival analysis based on Cox log-rank model of ten cancers datasets on TCGA (bold indicates that this method performs best on the corresponding cancer dataset)

Cancer	SADLN	Subtype -GAN	AE	VAE	K-means	Spectral	LRAcluster	SNF	NEMO	MCCA	iCluster
BLCA	3.8	3.0	0.68	3.6	1.7	3.3	0.6	2.7	**4.2**	2.5	2.4
BRCA	2.7	3.1	0.7	1.2	0.8	0.7	0.9	2.2	2.3	**4.3**	1.9
GBM	2.7	0.9	2.1	2.1	3.7	4.5	1.8	3.5	**3.9**	3.3	3.4
KIRC	9.9	7.7	3.7	8.0	6.1	6.9	8.3	9.8	7.1	**11.5**	5.7
LUAD	2.7	**3.7**	1.2	1.9	1.1	1.1	0.6	2.7	2.7	1.2	1.6
PAAD	**3.7**	2.2	0.1	2.3	2.4	2.4	2.1	3.2	1.8	2.7	1.0
SKCM	6.6	1.4	0.2	3.8	3.9	3.4	3.3	**7.3**	7.0	2.5	2.7
STAD	**2**	0.9	0.2	0.2	0.2	0.6	0.4	1.2	1.5	1.8	0.7
UCEC	**9.7**	9.3	1.1	7.9	8.6	1.6	6.0	6.5	7.4	6.8	2.3
UVM	**5.7**	4.3	4.1	4.2	2.7	3.0	3.7	4.0	3.3	3.9	2.0

Friedman analysis was also used to evaluate the performance (Figure2). From Figure2, we can see that the performance of SADLN is obviously better than the three methods K-means, LRAcluster and VAE ($P < 0.05$), but not better than other methods. We found that the performance of the methods is not exactly consistent under the two evaluation strategies.

Table 4 The clinical parameters enrichment analysis of SADLN and other methods of ten cancer datasets on TCGA(bold indicates that this method performs best on the corresponding cancer dataset)

Methods	BRCA	LUAD	BLCA	PAAD	KIRC	STAD	UVM	GBM	SKCM	UCEC
SADLN	5	**5**	5	1	**6**	**3**	1	**1**	3	**1**
Subtype-GAN	**6**	5	5	2	6	2	**2**	1	**4**	1
AE	0	1	0	1	5	1	0	1	0	0
VAE	5	2	**6**	1	6	2	1	0	1	1
K-means	5	1	3	0	6	2	0	1	1	1
Spectral	3	1	4	0	6	2	0	1	2	1
LRAcluster	5	1	3	1	6	1	0	0	0	1
SNF	5	3	6	2	4	1	0	0	4	1
NEMO	5	4	6	2	5	1	1	1	3	1
MCCA	5	4	3	**4**	3	2	1	1	0	1
iCluster	4	1	1	0	4	2	0	1	1	1

Survival curves can also be used to express heterogeneity of different subtypes. Figure3 shows the ten cancers' Kaplan Meier survival alanalysis curves. From Figure3, we can see that different clusters have significantly differences in survival curves(P-value<0.05). Take BRCA cancer for example(Figure3a), C1 has the longest average survival time. C2 and C5 have the poor survival time, C3 and C4 are at the intermediate level, this is consistent with the classification standards established by the St Gallen International Breast Conference in 2013, C1 represents basal-like, C2 represents normal-like, C3 represents luminal-B, C4 represents luminal-A, C5 represents HER2-enriched.

Visualization of clustering results

In order to visualize the clustering results, we used t-SNE embedding method to display the final integrated feature representation of the SADLN (Figure4). From Figure4 we can see that samples of the same cluster are almost grouped together, samples of different clusters are almost departed. Different subtypes have apparent boundaries.

Comparison of multiple omics data and single omics data on subtyping results

SADLN integrated mRNA, miRNA, copy number and DNA methylation four omics data. In order to demonstrate the necessity of integrating multiple omics data for subtyping, we compared multiple omics data and single omics data of SADLN(denoted as SADLN-single) on subtyping results.

Firstly, random forest(RF) method[27] was used to analyze the contribution of different omics data on the subtyping results of SADLN. Figure5 gives the four omics data's contribution results of SADLN on ten cancer datasets. From Figure5 we can see that the greatest contribution of BRCA, BLCA, LUAD, SKCM, STAD and UVM datasets were mRNA data, the greatest contribution of GBM, KIRC and UCEC were CNV data and the greatest contribution of PAAD were DNA methylation data. For different cancers, we choose the greatest contribution omics data as the input of SADLN-single. The settings of parameters remain the same as SADLN. We also use the metric of P -value of survival analysis in Cox log-rank model to compare the performance of SADLN and SADLN-single(Table5).

From Table5, we can see that the P values of SADLN are all smaller than the values of SADLN-single on ten cancer datasets. These results demonstrated that integration of multiple omics data can help improve the performance of subtyping.

Table 5 The P values of survival analysis in Cox log-rank model of SADLN based multiple omics data and single omics data

Cancer	SADLN	SADLN-single
BRCA	**2.00e-03**	7.00e-02
BLCA	**1.46e-04**	2.23e-04
LUAD	**2.00e-03**	3.00e-02
SKCM	**2.80e-07**	2.30e-02
STAD	**9.00e-03**	5.63e-01
UVM	**1.82e-06**	6.74e-01
GBM	**2.00e-03**	6.97e-01
KIRC	**1.37e-10**	2.00e-03
UCEC	**1.99e-10**	5.70e-02
PAAD	**2.16e-04**	3.64e-01

Identify the key biomarkers in each cancer

In order to identify the key biomarkers that determine the subtyping results in each cancer, we ranked the importance of mRNA features of each cancer datasets using the clustering labels of SADLN and random forest method to achieve the five most essential biomarkers. For each cancer, Table6 gives the five biomarkers most relevant to the classification.

For BRCA as example, the five key biomarkers are: ALDH8A1, SRPK3, FUT6, BEX2 and KIRREL2. By literature review, we found that the BEX2 gene[45] affects the prognosis of ER- breast cancer patients. SRPK3 gene[46] and ALDH8A1 gene[47] have been found to influence the prognosis of triple-negative breast cancer

Table 6 The five biomarkers most relevant to ten cancers

Cancers	Biomarkers
BRCA	ALDH8A1, SRPK3, FUT6, BEX2, KIRREL2
BLCA	GDA, IGSF21, F2RL2, KCNJ12, IL12RB2
GBM	CCL22, IP6K3, CPZ, FCGR2B, CTCFL
KIRC	TNFRSF10D, IL12RB2, PCDHGA5, DMKN, PTPRZ1
LUAD	PON1, SIX1, S100A1, APOD, DOC2A
PAAD	RIMS1, HAL, PAX8, THEM5, EDN2
SKCM	BMPR18, ARC, ROBO2, HOXA4, CD38
STAD	LOC100302650, MUC6, KCNJ16, MFSD6L, TNFSF11
UCEC	SLC13A5, KCNH2, KCNJ12, SELE, RXRG
UVM	TNFRSF11B, CPZ, PCP4, S100A1, CLDN4

patients. Some genes have not been found in BRCA, but have shown that they act on the pathogenesis and development of other cancers. For example, FUT6 gene[48] has been found to be occupied an important position in the metastasis of colorectal cancer. In addition, study has shown that the expression of KIRREL2 gene[49] may cause congenital nephrotic syndrome. All of these literature review demonstrated the results of SADLN on BRCA dataset are reliable.

Discussion

Recently, integrating multi-omics data for cancer subtyping is an important task in bioinformatics. In this paper, we proposed SADLN, a novel deep learning based integrated method for cancer subtyping. The method firstly introduced self-attention into the encoder-decoder based network architecture. It attempted to describe complex and diverse multi-omics data accurately and adaptively build the samples' relationship when learning a shared low-dimensional representation during molecular subtyping. Compared with three deep learning and seven non-deep learning based integration algorithms, SADLN has two characteristics:(1) Unlike the early methods such as AE and VAE, SADLN characterizes multi-omics data respectively which enable the model to effectively describe different omics data with distinct distributions, meanwhile, the output integrating representation fits the prior distribution. (2) The self-attention module in SADLN taking full use of sample's multi-omics information, can automatically learn the weight matrix between samples and make the results of feature integrating more convincing.

We demonstrated the power of SADLN using ten datasets of TCGA. The experiments of survival analysis and Friedman analysis show that SADLN has a good clustering consequence. Meanwhile, the experiments of SADLN and SADLN-single show that integrating multiple omics data is necessity and useful. The BRCA results indicated that SADLN can efficiently distinguish cancer subtypes.

SADLN found 50 biomarkers for each cancer. Some biomarkers have been verified in previous studies. In clinical research, researchers can conduct more subtype analysis studies on related cancers based on the biomarkers obtained by SADLN. For example, SADLN believes that MUC6 is an important biomarker of stomach adenocarcinoma. The study[50] has shown that MUC6 is a new prognostic biomarker of stomach adenocarcinoma clinical outcome and immune infiltration, and may be a promising therapeutic target.

Although SADLN has enhanced the performance of cancer subtyping recognition, it also has limitations. Firstly, it is unsuited to integrate binary data. Secondly, it couldn't find the genes modules which affect each subtype. Thirdly, the relationship

between omics data was not considered. For next research, we will continue our efforts to develop an attention based method to simultaneously learn the relationship between multi-omic and samples to explore cancer heterogeneity.

Conclusion

In this paper, we proposed Self-Attention Based Deep Learning Network(SADLN) of integrating multi-omics data for cancer subtype recognition. The novel method based on recent advances on deep learning and self-attention. It can jointly learning different multi-omic data representations and relations between samples. Experiments on ten datasets of TCGA have demonstrated the effectiveness of SADLN to the state-of-the-art methods.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Availability of data and materials
The input omics data and python code of SADLN were available through https://github.com/gpxzmu/SADLN

Competing interests
The authors declare that they have no competing interests.

Funding
This work was supported in part by the Natural Science Foundation of China (No.82001987), and in part by the Postgraduate Research & Practice Innovation Program of Jiangsu Province (No.KYCX21-2646), and in part by the Xuzhou Key Research and Development Program Project (No.KC20148).

Consent for publication
Not applicable.

Authors' contributions
PG participated in study design,conceived the study and organized documents. QWS carried out data analysis and organized documents. LC and ZYZ carried out data analysis. SGG, JC and LZZ participated in study analysis. All authors read and approved the final manuscript.

Acknowledgements
Thanks to Yang et al. for the preprocessed multi-omics data of TCGA.

Author details
[1]School of Medical Imaging, Xuzhou Medical University, Xuzhou, CN. [2]School of Information and Control Engineering, China University of Mining and Technology, Xuzhou, CN. [3]Department of Radiation Oncology, Affiliated Hospital of Xuzhou Medical University, Xuzhou, CN.

References

1. Siegel, R., Miller, K., Jemal, A.: Cancer statistics, 2020. CA: A Cancer Journal for Clinicians **70**, 7–30 (2020)
2. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R., Torre, L.A., Jemal, A.: Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians **68**, 394–424 (2018)
3. Zhao, W.-h., Luo, J., Jiao, S.: Comprehensive characterization of cancer subtype associated long non-coding rnas and their clinical implications. Scientific Reports **4**, 6591 (2014)
4. Huang, D.-s., Premaratne, P., Goebel, R., Tanaka, Y.: Intelligent Computing Methodologies. Science and Business Media LLC 2020, ??? (2020)
5. Song, M., Greenbaum, J., Luttrell, J., Zhou, W., Wu, C., Shen, H., Gong, P., Zhang, C., Deng, H.-W.: A review of integrative imputation for multi-omics datasets. Frontiers in Genetics **11** (2020)
6. Sayáns, M.P., Petronacci, C.C.C., Pouso, A.L.L., Iruegas, E.P., Carrión, A.B., Peñaranda, J.M.S., García, A.G.: Comprehensive genomic review of tcga head and neck squamous cell carcinomas (hnscc). Journal of Clinical Medicine **8**, 1896 (2019)
7. Picard, M., Scott-Boyer, M.-P., Bodein, A., Périn, O., Droit, A.: Integration strategies of multi-omics data for machine learning analysis. Computational and Structural Biotechnology Journal **19**, 3735–3746 (2021)

8. Simidjievski, N., Bodnar, C., Tariq, I., Scherer, P., Andrés-Terré, H., Shams, Z., Jamnik, M., Lio', P.: Variational autoencoders for cancer data integration: Design principles and computational practice. bioRxiv (2019). doi:10.3389/fgene.2019.01205
9. Xu, J., Wu, P., Chen, Y., Meng, Q., Dawood, H., Dawood, H.: A hierarchical integration deep flexible neural forest framework for cancer subtype classification by integrating multi-omics data. BMC Bioinformatics **20** (2019)
10. Rappoport, N., Shamir, R.: Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic Acids Research **46**, 10546–10562 (2018)
11. Wu, D., Wang, D., Zhang, M.Q., Gu, J.: Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification. BMC Genomics **16**, 1022 (2015)
12. Cavalli, F.M.G., Remke, M., Peacock, J., Shih, D.J.H., Luu, B., Garzia, L., Torchia, J., Nor, C.: Intertumoral heterogeneity within medulloblastoma subgroups. Cancer cell **316**, 737–7546 (2017)
13. Rappoport, N., Shamir, R.: Nemo: cancer subtyping by integration of partial multi-omic data. Bioinformatics **35**, 3348–3356 (2019)
14. Witten, D.M., Tibshirani, R.: Extensions of sparse canonical correlation analysis with applications to genomic data. Statistical Applications in Genetics and Molecular Biology **8**, 1–27 (2009)
15. Shen, R., Olshen, A.B., Ladanyi, M.: Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics **25 22**, 2906–12 (2009)
16. Poirion, O.B., Chaudhary, K., Garmire, L.: Deep learning data integration for better risk stratification models of bladder cancer. AMIA Summits on Translational Science Proceedings **2018**, 197–206 (2018)
17. Guo, Y., Shang, X., Li, Z.: Identification of cancer subtypes by integrating multiple types of transcriptomics data with deep learning in breast cancer. Neurocomputing **324**, 20–30 (2019)
18. Adem, K., Kiliçarslan, S., Cömert, O.: Classification and diagnosis of cervical cancer with stacked autoencoder and softmax classification. Expert Systems With Applications **115**, 557–564 (2019)
19. Xu, J., Xiang, L., Liu, Q., Gilmore, H., Wu, J., Tang, J., Madabhushi, A.: Stacked sparse autoencoder (ssae) for nuclei detection on breast cancer histopathology images. IEEE Transactions on Medical Imaging **35**, 119–130 (2016)
20. Chaudhary, K., Poirion, O.B., Lu, L., Garmire, L.: Deep learning–based multi-omics integration robustly predicts survival in liver cancer. Clinical Cancer Research **24**, 1248–1259 (2017)
21. 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. Frontiers in Genetics **9**, 477 (2018)
22. Chai, H., Zhou, X., Zhang, Z., Rao, J., Zhao, H., Yang, Y.: Integrating multi-omics data through deep learning for accurate cancer prognosis prediction. bioRxiv (2019). doi:10.1101/807214
23. Sharifi-Noghabi, H., Zolotareva, O., Collins, C., Ester, M.: Moli: Multi-omics late integration with deep neural networks for drug response prediction. bioRxiv, 501–509 (2019)
24. Picard, M., Scott-Boyer, M.-P., Bodein, A., Périn, O., Droit, A.: Integration strategies of multi-omics data for machine learning analysis. Computational and Structural Biotechnology Journal **19**, 3735–3746 (2021)
25. Adossa, N., Khan, S., Rytkönen, K., Elo, L.: Computational strategies for single-cell multi-omics integration. Computational and Structural Biotechnology Journal **19**, 2588–2596 (2021)
26. Tong, L., Mitchel, J., Chatlin, K., Wang, M.D.: Deep learning based feature-level integration of multi-omics data for breast cancer patients survival analysis. BMC Medical Informatics and Decision Making **20**, 225 (2020)
27. Yang, H., Chen, R., Li, D., Wang, Z.: Subtype-gan: a deep learning approach for integrative cancer subtyping of multi-omics data. Bioinformatics (2021). doi:10.1093/bioinformatics/btab109
28. Mercer, E. Robert Neufeld: Advances in Artificial Intelligence and Security. Springer, ??? (2021)
29. Yuan, S., Zhang, Y., Tang, J., Shen, H., Wei, X.: Modeling and predicting popularity dynamics via deep learning attention mechanism. ArXiv **abs/1811.02117** (2018)
30. Li, M., Wang, Y., Wang, Z., Zheng, H.: A deep learning method based on an attention mechanism for wireless network traffic prediction. Ad Hoc Networks **107**, 102258 (2020)
31. Liu, C., Zhang, L., Niu, J., Yao, R., Wu, C.: Intelligent prognostics of machining tools based on adaptive variational mode decomposition and deep learning method with attention mechanism. Neurocomputing **417**, 239–254 (2020)
32. Luo, L., Yang, Z., Yang, P., Zhang, Y., Wang, L., Lin, H., Wang, J.: An attention-based bilstm-crf approach to document-level chemical named entity recognition. Bioinformatics **34**, 1381–1388 (2018)
33. Shaw, P., Uszkoreit, J., Vaswani, A.: Self-attention with relative position representations, 464–468 (2018)
34. Hou, Y., Ma, Z., Liu, C., Loy, C.C.: Learning lightweight lane detection cnns by self attention distillation. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 1013–1021 (2019)
35. Gao, H., Li, Y., Wang, X., Han, J., Li, R.: Ensemble attention for text recognition in natural images. 2019 International Joint Conference on Neural Networks (IJCNN), 1–8 (2019)
36. Zhang, Z., Wu, S., Chen, G., Jiang, D.: Self-attention and dynamic convolution hybrid model for neural machine translation. 2020 IEEE International Conference on Knowledge Graph (ICKG), 352–359 (2020)
37. Li, Y., Long, G., Shen, T., Zhou, T., Yao, L., Huo, H., Jiang, J.: Self-attention enhanced selective gate with entity-aware embedding for distantly supervised relation extraction. ArXiv **abs/1911.11899** (2020)
38. Mustafa Abualsaud, M.D.S.: Proceedings of the 28th acm international conference on information and knowledge management. Proceedings of the 28th ACM International Conference on Information and Knowledge Management (2019)
39. Chen, C., Zha, Y., Zhu, D., Ning, K., Cui, X.: Hydrogen bonds meet self-attention: all you need for general-purpose protein structure embedding. bioRxiv (2021)
40. Yang, H., Wang, M., Liu, X., Zhao, X., Li, A.: Phosidn: an integrated deep neural network for improving protein phosphorylation site prediction by combining sequence and protein–protein interaction information.

Bioinformatics **37**, 4668–4676 (2021)
41. Artificial neural networks and machine learning – icann 2019: Workshop and special sessions: 28th international conference on artificial neural networks, munich, germany, september 17–19, 2019, proceedings. Artificial Neural Networks and Machine Learning – ICANN 2019: Workshop and Special Sessions (2019)
42. Gu, L., Zhang, X., Li, K., Jia, G.: Using molecular fingerprints and unsupervised learning algorithms to find simulants of chemical warfare agents. (2020)
43. Ge, S., Wang, X., Cheng, Y., Liu, J.: Cancer subtype recognition based on laplacian rank constrained multiview clustering. Genes **12**, 526 (2021)
44. Ryall, S., Zapotocky, M., Fukuoka, K., Nobre, L.F., Stucklin, A.S.G.: Integrated molecular and clinical analysis of 1,000 pediatric low-grade gliomas. Cancer cell **374**, 569–5835 (2020)
45. Naderi, A.: Molecular functions of the androgen receptor and bex2 in breast cancer (2018)
46. Wu, S., Wang, J., Zhu, X., Chyr, J., Zhou, X., Wu, X., Huang, L.: The functional impact of alternative splicing on the survival prognosis of triple-negative breast cancer. Frontiers in Genetics **11**, 604262 (2020)
47. Qi, F., Qin, W.-x., Zang, Y.: Molecular mechanism of triple-negative breast cancer-associated brca1 and the identification of signaling pathways. Oncology letters **173**, 2905–2914 (2019)
48. Deschepper, F.M., Zoppi, R., Pirro, M., Hensbergen, P., Dall'Olio, F., Kotsias, M., Gardner, R.A., Spencer, D., Videira, P.: L1cam as an e-selectin ligand in colon cancer. International Journal of Molecular Sciences **21**, 8286 (2020)
49. Jia, X., Yamamura, T., Gbadegesin, R., McNulty, M.T.: Common risk variants in nphs1 and tnfsf15 are associated with childhood steroid-sensitive nephrotic syndrome. Kidney international **985**, 1308–1322 (2020)
50. Yamada, S., Okamura, T., Kobayashi, S., Tanaka, E., Nakayama, J.: Reduced gland mucin-specific o-glycan in gastric atrophy: A possible risk factor for differentiated-type adenocarcinoma of the stomach. Journal of Gastroenterology and Hepatology **30**, 1478–1484 (2015)

Figure Legends

Figure 1 The overview architecture of SADLN

Figure 2 The P-values of Friedman test on ten cancer datasets

Figure 3 The Kaplan–Meier survival curves of ten cancer datasets.(a)BRCA,(b)BLCA, (c)GBM, (d)KIRC, (e)LUAD, (f)PAAD, (g)SKCM, (h)STAD, (i)UCEC, (j)UVM

Figure 4 t-SNE visualization of the final integrated features by SADLN on ten cancer datasets.(a)BRCA,(b)BLCA, (c)GBM, (d)KIRC, (e)LUAD, (f)PAAD, (g)SKCM, (h)STAD, (i)UCEC, (j)UVM

Figure 5 Contribution of mRNA, miRNA, CNV and DNA methylation to the subtyping results of SADLN on ten cancer datasets

Figures

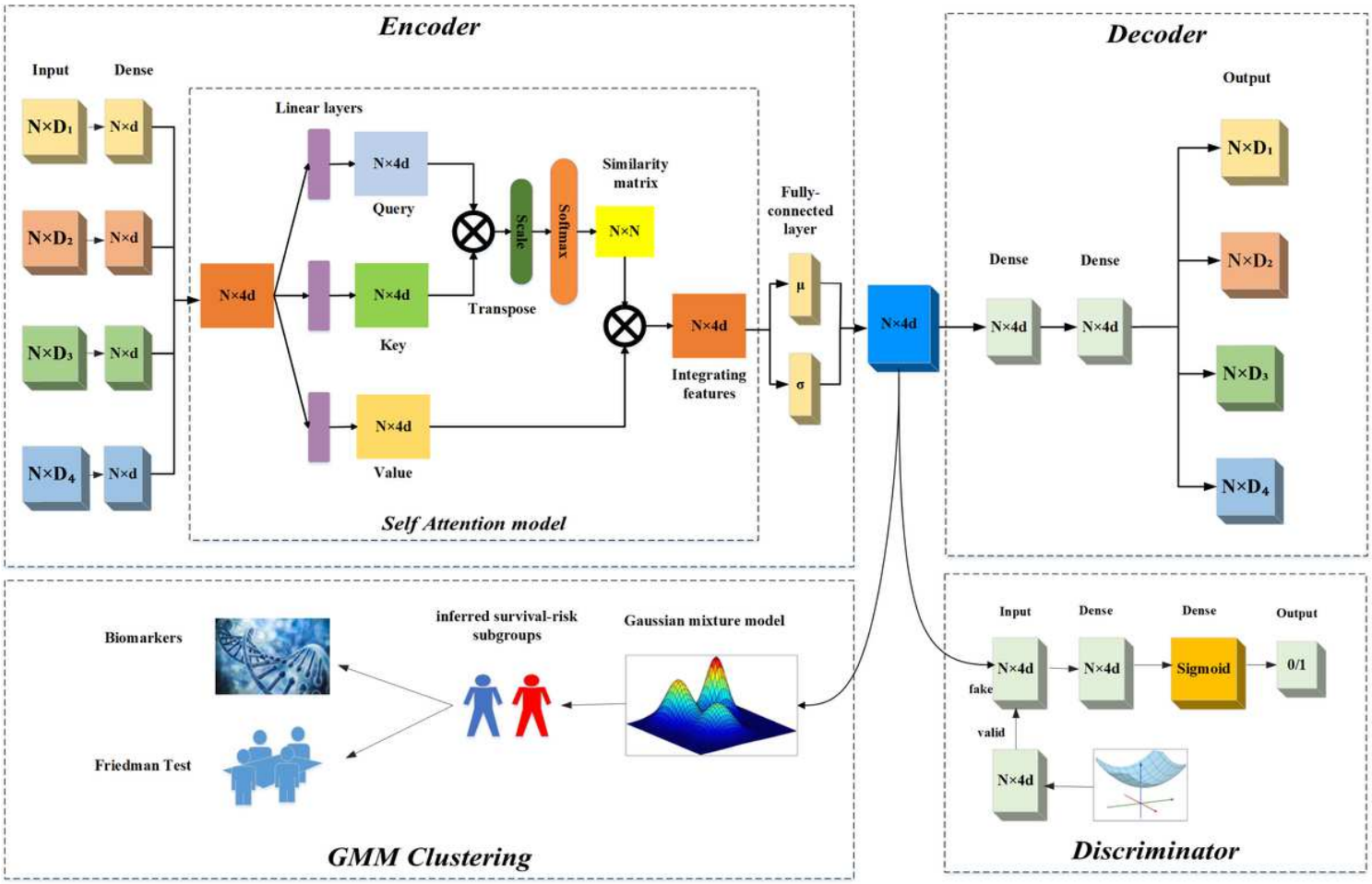

Figure 1

The overview architecture of SADLN

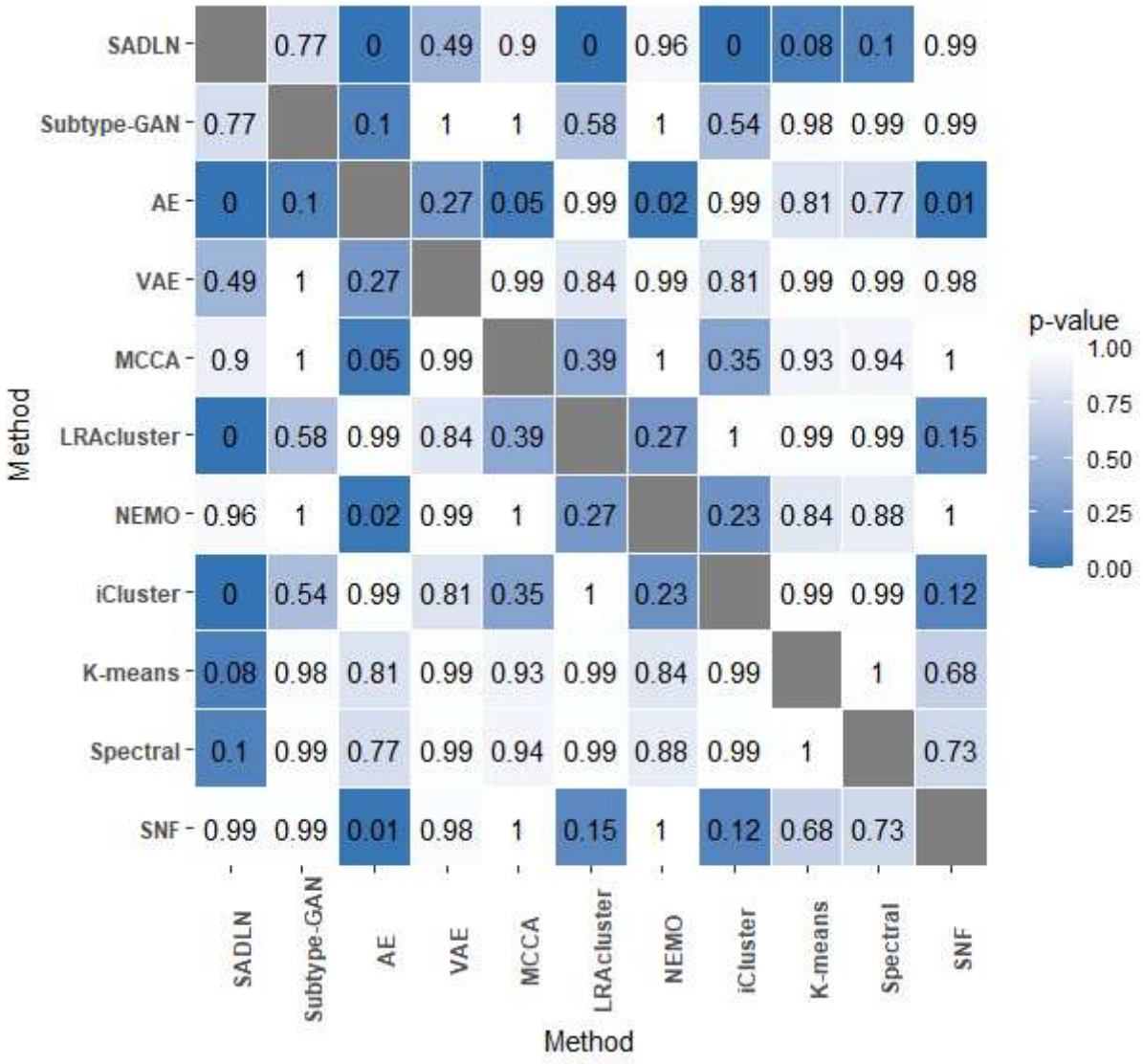

Figure 2

The P-values of Friedman test on ten cancer datasets

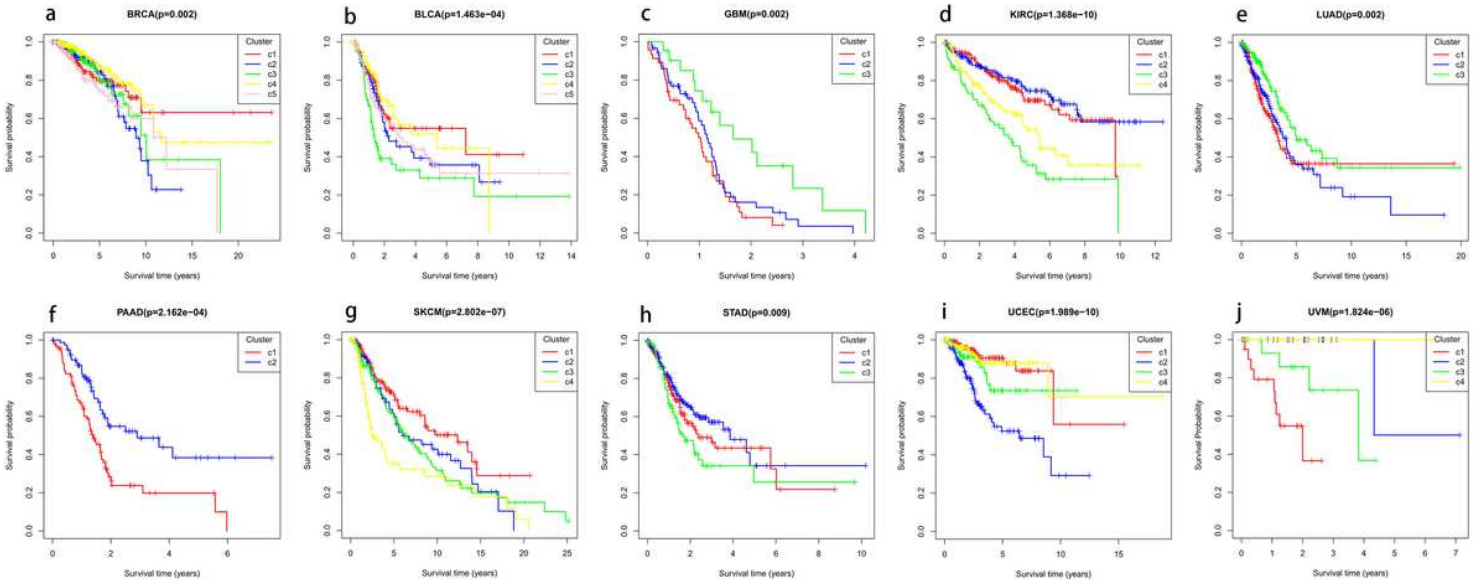

Figure 3

The Kaplan–Meier survival curves of ten cancer datasets.(a)BRCA,(b)BLCA, (c)GBM, (d)KIRC, (e)LUAD, (f)PAAD, (g)SKCM, (h)STAD, (i)UCEC, (j)UVM

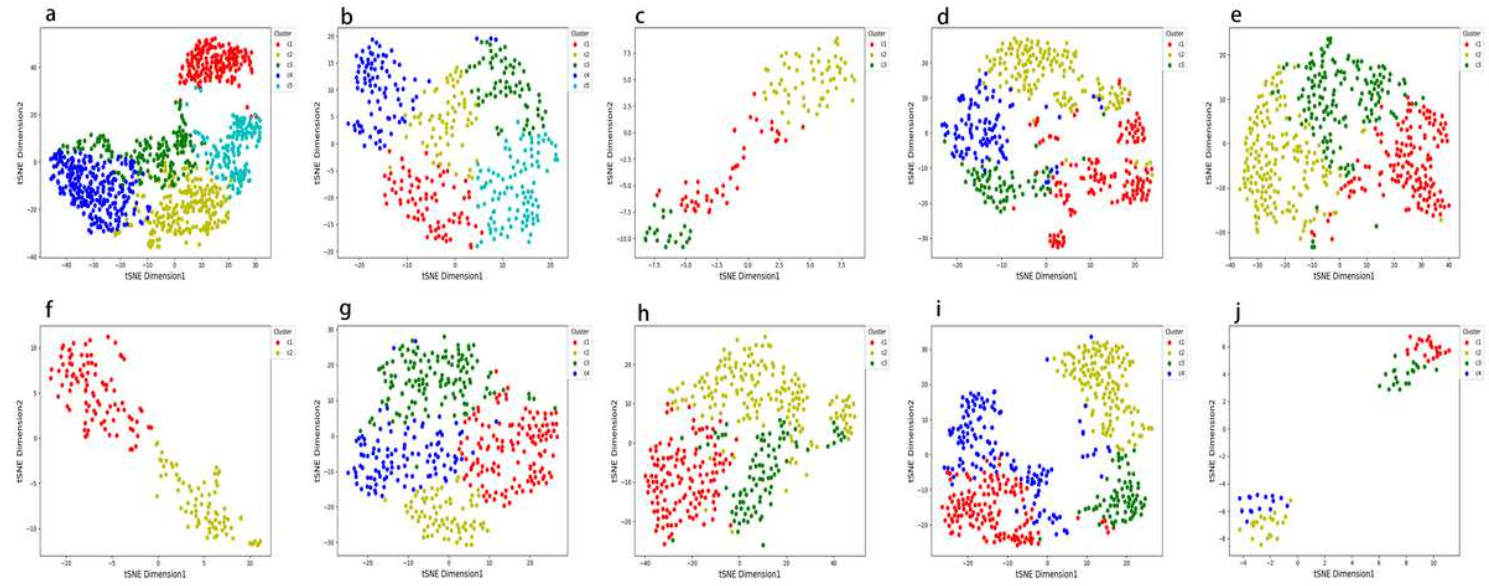

Figure 4

t-SNE visualization of the final integrated features by SADLN on ten cancer datasets.(a)BRCA,(b)BLCA, (c)GBM, (d)KIRC, (e)LUAD, (f)PAAD, (g)SKCM, (h)STAD, (i)UCEC, (j)UVM

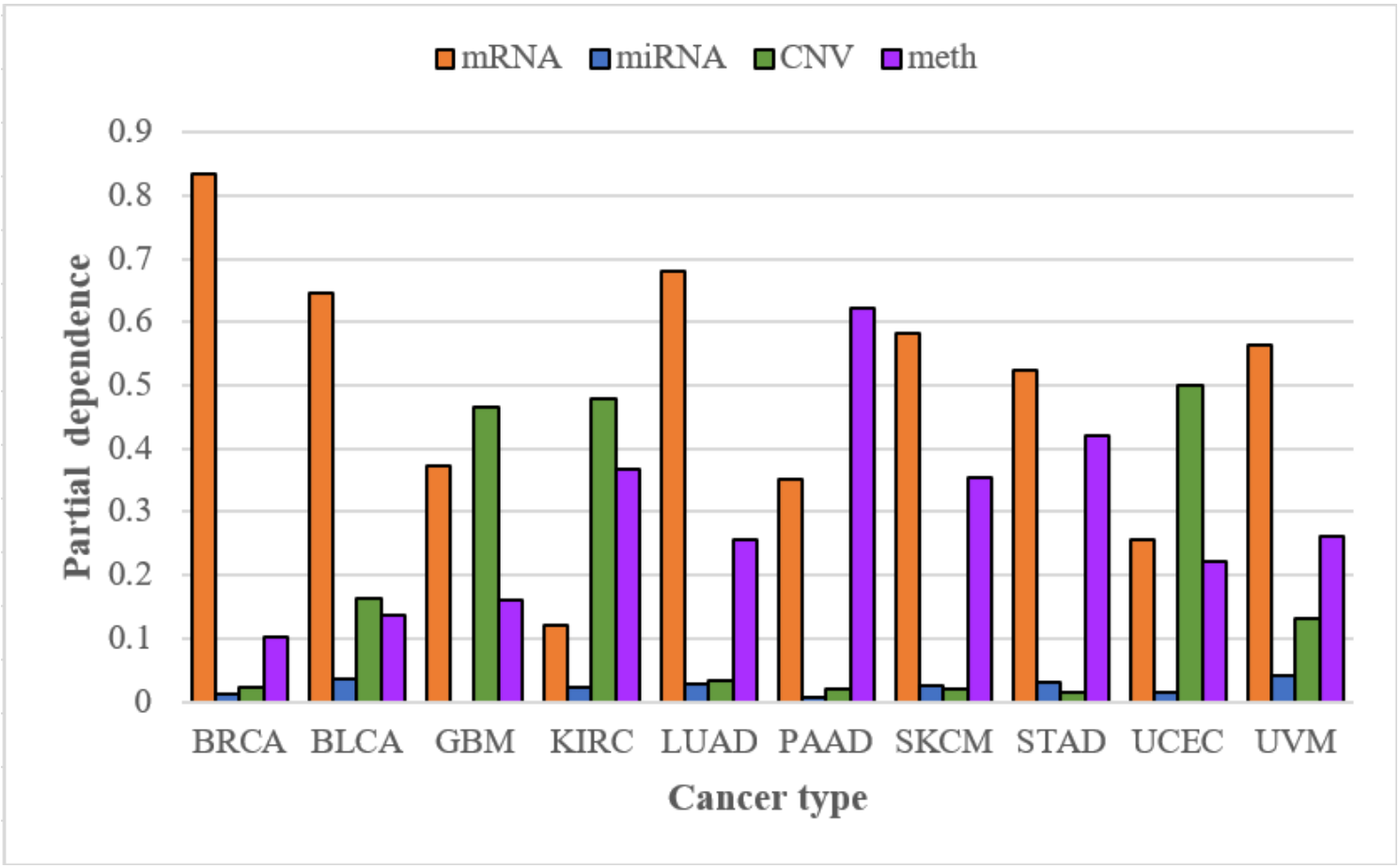

Figure 5

Contribution of mRNA, miRNA, CNV and DNA methylation to the subtyping results of SADLN on ten cancer datasets

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- bmctemplate.rar