Gene-MOE: A Sparsely Gated Cancer Diagnosis and Prognosis Framework Exploiting Pan-Cancer Genomic Information

Xiangyu Meng, Xue Li, Qing Yang, Huanhuan Dai, Lian Qiao, Hongzhen Ding, Long Hao, and Xun Wang

***Abstract*—Improved cancer genomic diagnosis and prognosis are vital to accurate medical therapy. Deep learning methods offered an end-to-end solution to enhance the precision of analysis. With the fast pace of pre-trained Transformer models, it remains uncertain whether some novel approaches such as the sparsely gated mixture of expert (MOE) and self-attention mechanisms can further improve the precision of cancer prognosis and classification. In this paper, we introduce a novel sparsely gated cancer diagnosis and prognosis framework called Gene-MOE exploiting the potential of the MOE layers and the proposed mixture of attention expert (MOAE) layers to enhance the analysis accuracy. Additionally, we address overfitting challenges by integrating pan-cancer information from 33 distinct cancer types through pre-training. For survival analysis, Gene-MOE achieves the best Concordance Index compared with state-of-the-art models on 12 of 14 cancer types. For cancer classification, the total accuracy of the classification model for 33 cancer classifications reached 95.8%, representing the best performance compared to state-of-the-art models. For cancer subtyping, Gene-MOE achieves the best result on at least one metric of the log10 P-values and the number of significant clinical on seven of nine cancers. These results indicate that Gene-MOE holds strong potential for these downstream tasks.**

***Index Terms*—Survival analysis, cancer classification, genomic analysis, mixture of expert, self-attention.**

I. Introduction

THERE is a strong need to identify the specific cancer types and carry out accurate survival analysis to help personalize treatment and precision medical therapy [1], [2], [3]. The arrival of the Human Genome Project [4] makes high-throughput gene expression data available. Numerous genomic analysis technologies for a large amount of gene expression data like survival analysis, cancer classification, and co-expression analysis [5], [6], [7], [8], [9], [10], [11] have been developed and proved to be useful for cancer diagnosis and prognosis.

Received 23 December 2023; revised 30 May 2024; accepted 26 December 2024. Date of publication 3 January 2025; date of current version 3 April 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 61972416, in part by the Natural Science Foundation of Shandong Province under Grant ZR2022LZH009, in part by GHfund C under Grant 202407035455, and in part by the National Key R&D Program of China under Grant 2021YFA1000103-3. *(Corresponding author: Xun Wang.)*

The authors are with the Shandong Key Laboratory of Intelligent Oil & Gas Industrial Software, Department of Computer Science and Technology, China University of Petroleum (East China), Qingdao 266580, China (e-mail: x_meng0420@163.com; Xueleecs@gmail.com; s21070069@s.upc.edu.cn; daihuanhuan0901@163.com; s21070055@s.upc.edu.cn; s22070050@s.upc.edu.cn; 1808010425@s.upc.edu.cn; wangsyun@upc.edu.cn).

Digital Object Identifier 10.1109/TCBBIO.2024.3524209

With the capability of learning non-linear functions from high-dimensional genes, deep learning-based algorithms are widely used in genomic diagnosis and prognosis tasks. Representative works includes Cox-nnet model [12], DeepSurv [13], VAECox [14], DeepCC [15], DeepCues [16], and the cancer prognosis and classification method using graph convolutional networks (GCN) [17], [18]. Pioneer to the end-to-end characteristics and powerful fitting capability, deep learning-based genomic diagnosis and prognosis approach is arguably one of the most successful fields in applying to computational biology and bioinformatics. Over the past several years, the pre-trained Transformer models have received widespread attention owing to their powerful feature learning capabilities. Currently, the Transformer-based models are increasingly diverse: The parameters increase drastically due to the improvement of computing power. The model characteristics become various due to the sparsity models and different self-attention methods. Some notable models are dense transformer models [19], [20], and the sparse models based on the mixture of expert (MOE) model [21], [22]. The dense Transformer introduces self-attention mechanisms, which can learn the global semantic correlations of features during the training. Compared to the fully connected layers, introducing self-attention allows the model to prioritize learning of correlated features, thereby improving prediction accuracy. Sparse Transformer replaces FCNs with Mixture of Experts (MOE) modules with multiple expert models. During the training, top K experts are selected to learn the same input feature. Compared to FCN modules, introducing MOE expands the parameter space of the whole model, allowing the model to learn richer feature representations. To date, many pre-trained models have been introduced to solve biological problems and achieved remarkable results. Represent work include BioBERT [23], DNABERT [24], and scBERT [25]. These methods transfer the self-attention mechanism from the Transformer module and the MOE model into gene expression data analysis, which learns significant correlated features of genes and achieves significant results in downstream tasks such as cancer classification and disease prediction.

However, challenges persist in the application of pre-trained Transformer models to high-dimensional genetic data. First, the number of genes far exceeds the number of patient samples, exacerbating the risk of over-fitting during training. Second, most genes contain no useful information. Appropriate data

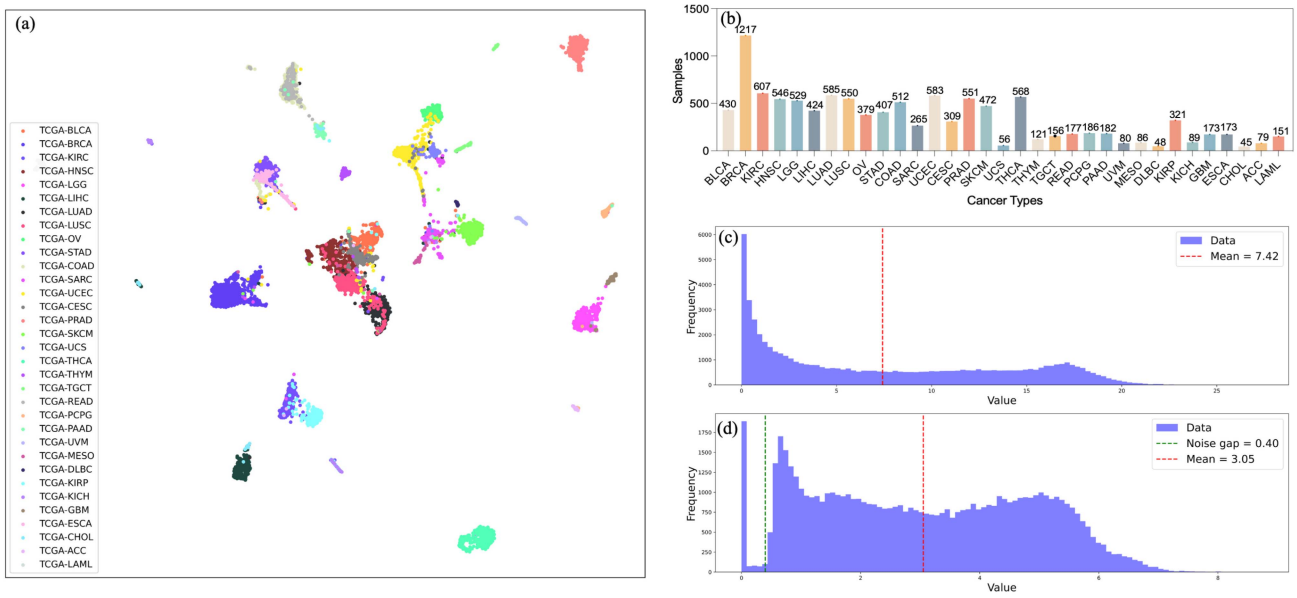


Fig. 1. Statistical analysis of Pan-Cancer dataset with 33 cancer types. (a) UMAP dimensional reduction and visualization for each patient of the Pan-Cancer dataset. (b) Number of patients for each cancer type. (c)Histogram statistics of the mean expression of each gene in the dataset. (d) Histogram statistics of the standard deviation of gene expression in the dataset.

preprocessing and augmentation methods must be designed to avoid meaningless information as much as possible. third, there is a large number of genes with similar expression patterns in the human genome with functional relevance and reproducibility [26], [27]. Extracting such coexpression features from high-dimensional genetics requires special consideration to design appropriate methods tailored to genomic data.

In this work, we combine the principles of the MOE structure to create a pre-trained diagnosis and prognosis model called Gene-MOE for high-dimensional RNA-seq gene expression data. This model combines the characteristics of MOE and the proposed mixture of attention experts (MOAE) to learn the deep correlation features of high-dimensional genes. According to the experiments, the Gene-MOE model achieves good performance for both diagnosis and prognosis. Specifically, the main contributions are as follows:

1) We propose a sparsely gated RNA-seq analysis framework called Gene-MOE. Gene-MOE exploits the MOE layers to extract the features from high-dimensional RNA-seq genes. Furthermore, the self-attention mechanism is added to construct the MOAE model to further learn the deep semantic relationship inside the genetic features.
2) We use a novel self-supervised pre-training strategy to make Gene-MOE learn the common features of 33 cancers and then transfer the pre-trained weight to the specific analysis including survival analysis and cancer classification.
3) According to the survival analysis results on 14 cancer types, the Gene-MOE achieves the best concordance index on 12 cancer types. Moreover, the classifier using pre-trained Gene-MOE achieves accurate classification of 33 cancer types, with a total accuracy of 95.8%.

The rest of this paper is organized as follows. Section II illustrates our method, including the dataset preparation, Gene-MOE framework illustration, and training strategy. Section III presents our experimental results for Gene-MOE. Finally, in Section IV, we provide our conclusion.

II. Methods

A. Dataset Preparation

In this study, we primarily utilize the pan-cancer RNA-seq database from the TCGA dataset of the Pan-Cancer Atlas project, which consists of 33 cancer types. Furthermore, we download the specific RNA-seq dataset for each TCGA cancer type. UCSC Xena already includes the preliminary processing for these 34 datasets, so we directly use the RNA-seq data after the FPKM-UQ normalization from UCSC Xena. In the initial dataset, each patient is associated with 60,484 genes, a majority of which exhibit null expression and have no relevance to cancer. According to Fig. 1(a) and (b), we find an imbalance in the number of patient samples for each type of cancer. Meanwhile, some cancer types exhibit significant overlap, which indicates clear correlations between some types of cancer in the Pan-Cancer dataset. Then we perform further analysis of the dataset. Fig. 1(c) and (d) shows the histogram of the mean and standard deviation of 60,484 genes. We find that plenty of genes have a low mean expression and standard deviation, which indicates these genes have no contribution to the cancer. Furthermore, there is a valley at 0-0.4 in Fig. 1(d), which indicates that the standard deviation of typical genes is not within this range. According to these obstacles, we introduce a filter scheme to remove these non-contributing genes. We first delete the empty genes and select overlapping genes among 34 datasets. Next, we filter out genes with variances less than 0.4 and mean values less than 0.8. Through preparation, we select 25,182 genes for each dataset. Finally, before feeding the data into the network, we perform min-max normalization.

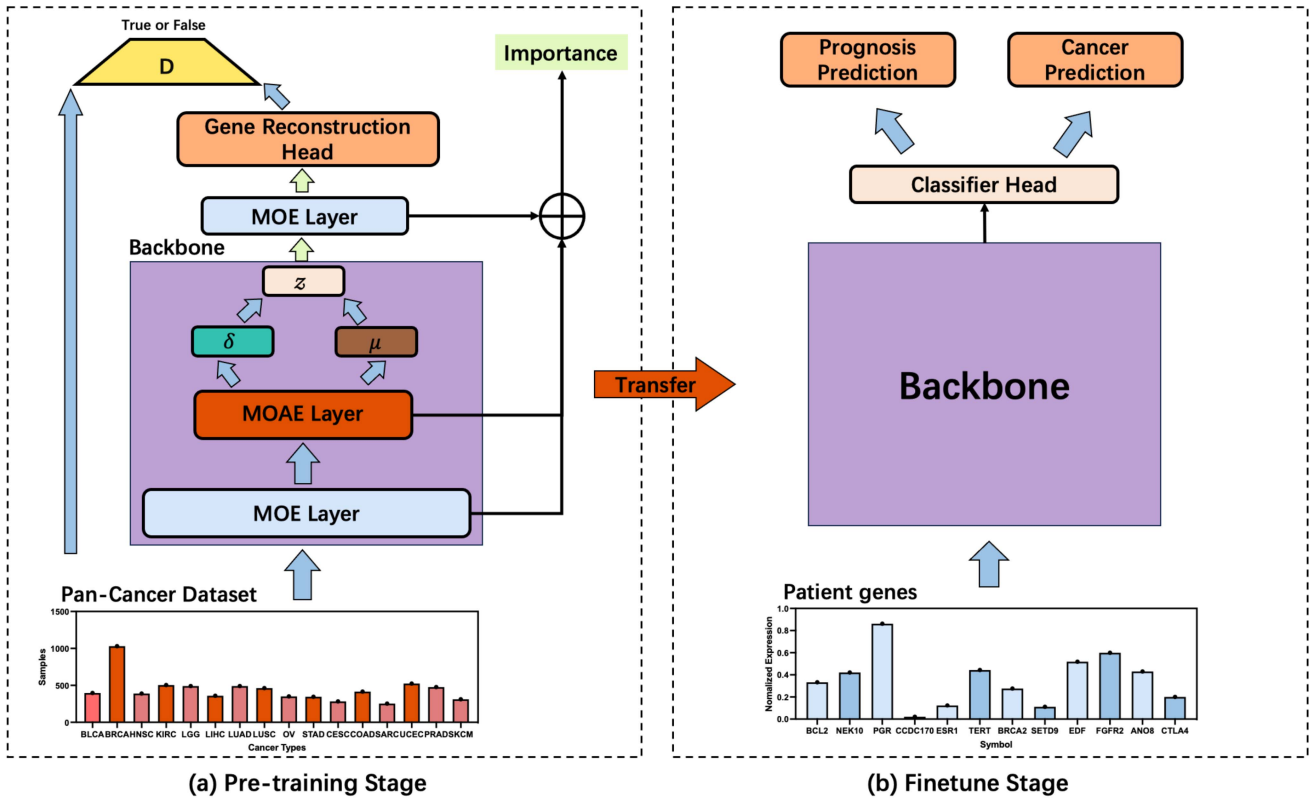


Fig. 2. Framework illustration of Gene-MOE. This model primarily consists of an encoder backbone network and a decoder network. (a) Pre-training stage. During this phase, Gene-MOE takes preprocessed pan-cancer genes as input to train the encoder–decoder network. (b) Fine-tuning stage. During this stage, Gene-MOE transfers the pre-trained backbone and connects to a classification head to achieve accurate downstream prediction tasks such as prognosis prediction, cancer classification, and specific pathways of interest.

B. *Framework Illustration*

Fig. 2 illustrates the framework of Gene-MOE, which comprises two stages: pre-training and fine-tuning. During the first stage, we construct the Gene-MOE model to learn low-dimensional feature encoding of high-dimensional pan-cancer genes. In this stage, we employ a self-supervised learning approach where the training labels are the input genes, aiming to acquire low-dimensional feature encoding of high-dimensional pan-cancer genes. The Gene-MOE model primarily consists of an encoder backbone network and a decoder network. Within the encoder network, we introduce an innovative MOE model based on sparse gating. This method involves multiple experts, enabling the encoder to learn rich feature representations of the genomic information. Moreover, we design a MOAE model, which employs multiple attention mechanisms as distinct experts and uses a learnable sparse gating mechanism to merge attention features adaptively. Note that the Gene-MOE model differs from the current Transformer models (e.g., Switch Transformers). The Gene-MOE uses MOE and the proposed MOAE alone to apply the effects of self-attention and MOE to the entire genome, aiming to facilitate the discovery of potentially correlated genes. In the second stage, the backbone of the Gene-MOE is transferred, and a new classification head is constructed to accomplish a fast, accurate downstream task after fine-tuning. This methodology enables us to fully leverage the Gene-MOE model in two stages, facilitating the learning of rich pan-cancer gene feature encoding and the achievement of excellent performance in multiple downstream tasks.

1) Sparsely Gated MOE Layer: Each patient has over 20,000 genes, among which intricate correlations exist. In previous work, it is common to build one or multiple fully connected networks (FCNs) to learn and extract key features related to genes. However, directly employing FCNs cannot effectively help the model learn these intricate correlations of high-dimensional gene input. The learned features would result in a substantial loss. In contrast to dense layers, the MOE module trains N experts, each of which independently learns and extracts features based on the characteristics of the input data. Compared with FCN, the MOE model can integrate diverse features from multiple expert models, enhancing the feature extraction capabilities of high-dimension genes and the overall model performance. Moreover, the training of each expert is independent, allowing for adjustments based on the characteristics of gene data, thus providing greater flexibility compared with fully connected layers. Given the input feature $x \in \mathbb{R}^{1\times N}$, the MOE layer with E experts is denoted as

$$y = \sum_{i=1}^{E} G(x)_i \cdot D_i(x), \tag{1}$$

where $D_i(x)$ is the output after feeding x to the ith expert network, which is an independent dense layer, and $G(x)_i$ is the

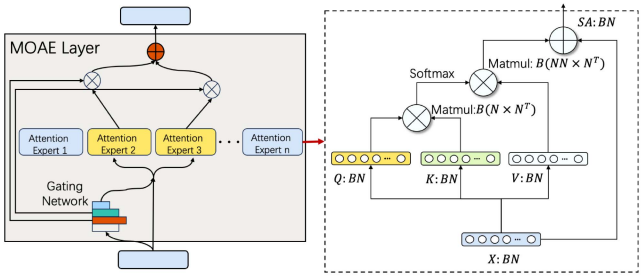


Fig. 3. Framework of mixture of attention expert layer (MOAE) layer.

selection for ith expert predicted by the sparsity gating network. When the $G(x)_i = 1$, we compute the ith expert. When the $G(x)_i = 0$, we do not need to compute the expert. $G(x)$ is expressed as

$$G(x) = Softmax(TopK(H(x), k)), \tag{2}$$

where $TopK(H(x), k)$ denotes a discrete function that maps the input feature $H(x)$ to a mask $m \in \mathbb{R}^{1\times N}$. It is denoted as

$$TopK(H(x), k)_i = \begin{cases} H(x)_i & \text{if } H(x)_i \text{ in the top } K \text{of } H(x). \\ -\infty & \text{otherwise.} \end{cases} \tag{3}$$

This equation generates a mask, where the values in $H(x)$ corresponding to elements outside the top K range are set to $-\infty$. When feeding the mask to the softmax function, the $-\infty$ elements are set to 0. The input $H(x)$ denotes the dense layer in the noise gating model. This layer learns to choose the expert during the training. It is denoted as

$$H(x)_i = (x \cdot W)_i + z \cdot Softplus((x \cdot W_{noise})_i), \tag{4}$$

where $z \sim N(0, 1)$ is a random Gaussian noise, W is a trainable weight matrix that learns the sparsity gating feature, and W_{noise} is the weight that controls the noise increment. Different from the normal dense layer, $H(x)$ includes a trainable Gaussian noise to make model load balancing during the distributed training.

2) Mixture of Attention Expert Layer: The human genome exhibits significant correlations. Typically, a mutation in one gene would be associated with multiple types of cancer, and the occurrence of a particular cancer can also be related to modifications in multiple genes. Identifying such strong correlations among 60,484 genes presents numerous challenges for model design. Recently, the attention mechanisms exhibit a significant effect in many fields Compared to traditional FCNs, this approach can effectively learn deep correlations between features. Incorporating attention mechanisms into the model can effectively discover related genes and further improve model performance. Additionally, combining attention mechanisms with MOE can help the model learn rich correlation features from massive datasets. Therefore, we propose a mixture of attention expert (MOAE) models based on the MOE mechanism by replacing FCN experts with residual self-attention networks. Fig. 3 illustrates the framework of MOAE. Note that the purpose of using residual is to reduce the model degradation problem and increase the training effect. Similar to the MOE, we use the same discrete function in (2) to select the top K self-attention models and merge the features.

C. Pre-Training Strategy

We construct a self-supervised pre-training strategy to learn the common features of pan-cancer genes and improve the feature extraction performance of the backbone network.

1) Data Augmentation by Gaussian Noise: The Pan-cancer dataset has 10,000 patients after preprocessing. If we use 0.5 billion parameters, Gene-MOE might cause over-fitting issues. Therefore, we build a data augmentation strategy by introducing Gaussian noise. For an input gene x, the augmented input $\hat{x}$ is expressed as

$$\hat{x} = x + z, \tag{5}$$

where $z \sim N(0, \sigma^2)$. The variance σ^2 is searched by hyperparameter optimization. Moreover, the dropout strategy is introduced during the training.

2) Joint Training Using a Generative Adversarial Network: We use the training strategy of a generative adversarial network (GAN) to pre-train Gene-MOE. We construct a discriminator called D, which is a dense layer network, and the Gene-MOE model is the generator. Then, Wasserstein loss is introduced to train G and D jointly. For G and D, the loss is denoted as

$$\begin{aligned} \mathcal{L}_{gan} = {} & \mathbb{E}_{\hat{x}\sim P_{data}(\hat{x})} D(\hat{x}) - \mathbb{E}_{\hat{x}\sim P_{data}(\hat{x})} D(G(\hat{x})) \\ & - \lambda_{gp} \mathbb{E}_{\hat{x}\sim\mathcal{X}} ||||\nabla_{\hat{x}} D(\hat{x})||_2 - 1||_2, \end{aligned} \tag{6}$$

where λ_{gp} is the hyper-parameter of the gradient penalty, and $\mathcal{X}$ is the sample space of x and $\hat{x}$. Training using (6) can help Gene-MOE learn how to perform dimensional reduction and reconstruct the generated gene.

3) Measuring the Distribution: We further introduce the KL divergence to measure the similarity of the latent code z generated by $\hat{x}$ and the standard Gaussian distribution. It can be denoted as

$$\mathcal{L}_{KL} = \sum_{i=0}^{n} (\mu(\hat{x})^2 + \sigma(\hat{x})^2 - \log(\sigma(\hat{x})^2) - 1), \tag{7}$$

where n represents the dim of z. Using this loss allows z to maintain a standard normal distribution, thus simplifying the training difficulty of Wasserstein loss.

4) Measuring the Similarity of Genes: We introduce L1 loss to further measure the similarity of each gene between samples reconstructed by Gene-MOE and the ground-truth samples. It is computed as

$$\mathcal{L}_{L1} = ||G(\hat{x}) - \hat{x}||_1. \tag{8}$$

5) Balancing Expert Utilization: To allow each MOE layer to select each expert in a balanced manner, we introduce importance loss to each sparse gating layer of the MOE. The importance loss can be denoted as

$$\mathcal{L}_{importance} = ||Importance(\hat{x}^k)||_2, \tag{9}$$

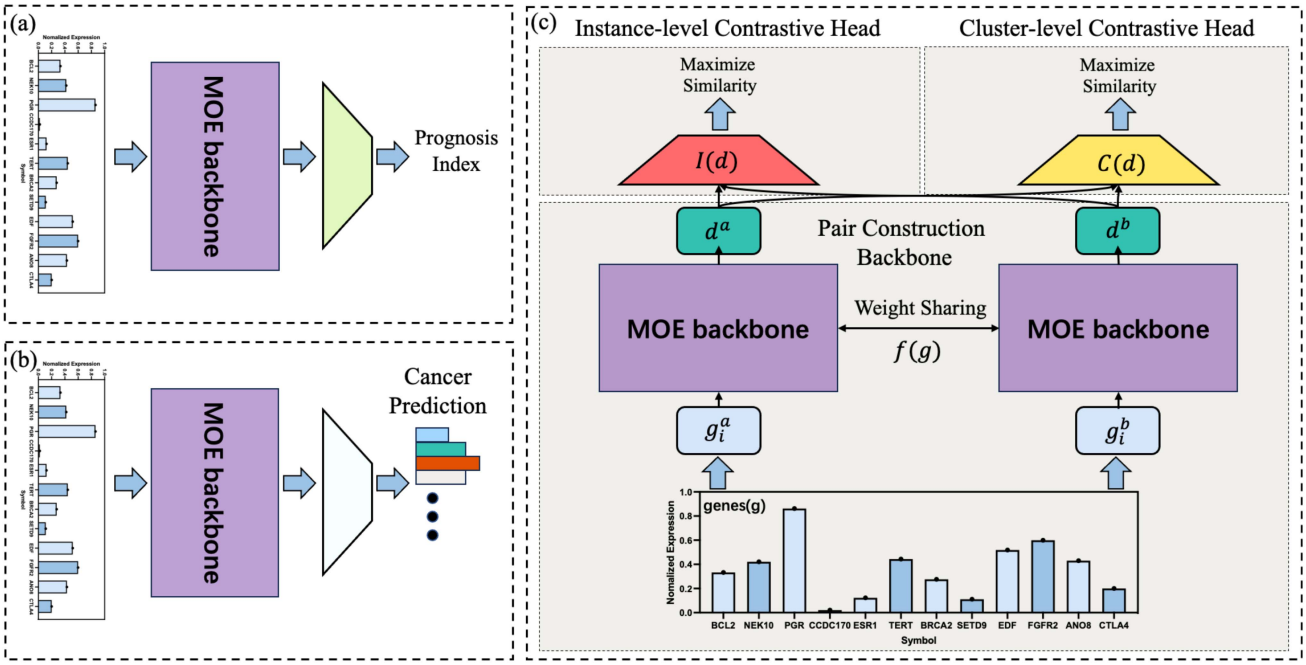


Fig. 4. Overview of the downstream analysis. (a) Survival analysis. This method predicts the hazard ratio to perform cancer prognosis. (b) Cancer classification analysis. This approach performs cancer-type classification based on the gene expression of patients. (c) Cancer subtyping. This approach uses the self-supervised learning method to identify the cancer subtypes.

where $\hat{x}^k$ denotes the input features of kth MOE layer, and $Importance(f(\hat{x}))$ denotes

$$Importance(\hat{x}^k) = \sum_{\hat{x} \in X} G(\hat{x}^k), \tag{10}$$

where $X \in \mathbb{R}^{B \times N}$ denotes the total sample in a batch size. Eq. (10) ensures the same importance of each expert, which makes the model allocate experts for the training data in balance. However, experts may receive imbalanced samples, which would cause memory and performance problems. To further improve the balanced loading, we also introduce the load balance loss $\mathcal{L}_{load}$ in [21], which is denoted as

$$\mathcal{L}_{load} = \left\| \sum_{\hat{x}^k \in X} \Phi \left(\frac{(\hat{x}^k \cdot W)_i - TopK'(H(\hat{x}^k), k, i)}{Softplus(\hat{x}^k \cdot W_{noise})_i} \right) \right\|_2, \tag{11}$$

where the Φ is the CDF of the standard normal distribution, $TopK'$ means selecting the kth highest value of $H(\hat{x}^k)$ excluding itself.

6) Overall Pre-Training Loss: Combining these losses, we can express the overall loss of MOE as

$$\begin{aligned} \mathcal{L}_{total} = \mathcal{L}_{gan} + \lambda_{KL} \cdot \mathcal{L}_{KL} + \lambda_{l1} \cdot \mathcal{L}_{L1} \\ + \lambda_{balance} \cdot (\mathcal{L}_{importance} + \mathcal{L}_{load}), \end{aligned} \tag{12}$$

where λ_{KL}, $\mathcal{L}_{KL}$, λ_{l1}, and $\lambda_{balance}$ are the hyper-parameters. Therefore, the pre-training stage aims to fit the optimal parameters θ_G^* of MOE by solving

$$\theta_G^* = \arg\min_G \max_D \mathcal{L}_{total}. \tag{13}$$

D. Survival Analysis

Survival analysis approach helps researchers to apply further treatment and improve the cancer prognosis. In this section, we use Gene-MOE to perform COX survival analysis on specific cancers. Fig. 4(a) shows the framework of the survival model. We use the pre-trained Gene-MOE backbone as the encoder to extract the useful features of the high-dimensional genes. Subsequently, the gene features are fed into the FCN classifier to predict the prognosis index of the patients. The training strategy of survival analysis adopts the Cox-ph model, which is denoted as

$$h(t|x_i) = h_0(t)\exp(\theta^T \cdot x_i), \tag{14}$$

where $h_0(t)$ is the baseline hazard function, θ^T refers to the trainable parameters of the Cox model, and x_i represents the hazard ratio of patients, which is the low-dimensional feature generated by the backbone of Gene-MOE. Training the Cox model is aimed at solving

$$\theta^* = \arg\min_\theta \sum_{C(i)=1} \left(\theta^T \cdot x_i - \log \sum_{t_j \geq t_i} \theta^T \cdot x_j \right), \tag{15}$$

where t is the survival time of the patient sample, and $C(i)$ indicates whether the patient sample i is censored. For each cancer type, we train the individual survival model and evaluate it to prove its accuracy.

E. Cancer Classification

Accurate cancer prediction based on gene expression can help to understand the disease mechanisms and leverage the customized treatment for specific cancer types. We then chose

cancer classification analysis as the downstream task to further evaluate the performance of Gene-MOE. Fig. 4(b) shows the framework of the classification model. Similar to the survival model, we transfer the weight of the Gene-MOE backbone to extract the representative features of genes. Then the features are fed into the FCN classification head. The classifier head of the classification model then predicts the probability of 33 distinct cancer types. The classification head is designed as a multi-class classification model, directly predicting the probability of 33 cancer types, and The Focal loss [28] is employed during the training process to mitigate the imbalance issue and enhance the model performance.

F. Cancer Subtyping

Cancer subtyping methods are crucial to cancer diagnosis, prognosis, and customized treatment. We choose this typical field to carry out the evaluations. We then transfer the encoder backbone and add the Subtype-DCC [29] head, which is a decoupled contrastive clustering method that applies the deep clustering algorithm and contrastive learning. Fig. 4(b) shows the framework of the cancer subtyping model using Gene-MOE. The patient gene g is first applied to three data augmentations including noise, mask, and dropout in [30], resulting in two correlated samples denoted as g_i^a and g_i^b. Then the correlated samples are fed into a shared MOE backbone to extract the feature embeddings d_i^a and d_i^b. Then the d_i^a and d_i^b are fed into the Instance-level contrastive head (ICH) and Cluster-level contrastive head (CCH) to predict the subtype and the clustering.

ICH is an FCN model aiming to maximize the similarities of positive pairs, which maps d_i^a and d_i^b to subspace z_i^a and z_i^b. Since no prior labels are available in subtyping, positive and negative sample pairs are constructed from pseudo-labels generated by data augmentations [31]. The generated samples augment from the same patient form positive pairs, while others denote the negative pairs. We optimize the instance-level head by adding the Decoupled Contrastive Learning (DCL) objective [32] to address the coupling phenomenon, which is denoted as

$$\mathcal{L}_i^a = -S_{i,i}^{a,b}/\tau_I + \log \sum_{j=1,j\neq i}^{N} (\exp(S_{i,j}^{a,a}/\tau_I) + \exp(S_{i,j}^{a,b}/\tau_I)), \tag{16}$$

where τ_I denotes the instance-level temperature parameter to control the softness. $s(z_i^a, z_j^b)$ is the similarity of paired samples, which is denoted as

$$S_{i,j}^{x,y} = \frac{z_i^x \cdot (z_j^y)^T}{||z_i^x|| \cdot ||z_j^y||}, \tag{17}$$

where $x, y \in a, b$ and $i, j \in [1, N]$. Then, the decoupled instance-level contrastive loss on each augmented sample is denoted as

$$\mathcal{L}_{ins} = \frac{1}{2N} \sum_{i=1}^{N} (\mathcal{L}_i^a + \mathcal{L}_i^b). \tag{18}$$

CCH uses another FCN model to project the embedding matrix d^a into M-dimensional space $c^a \in \mathbb{R}^{N\times M}$. The column $\hat{c}_i^a$ and $\hat{c}_i^b$ in the ith column of c^a and c^b are combined as the positive pair, while the remaining 2M-2 pairs are considered as negative pair. The similarity between subtype cluster pairs is denoted as

$$\hat{S}_{i,j}^{x,y} = \frac{(\hat{c}_i^x)^T \cdot \hat{c}_j^y}{||\hat{c}_i^x|| \cdot ||\hat{c}_j^y||}, \tag{19}$$

where $x, y \in a, b$ and $i, j \in [1, N]$. Based on similarity, The loss function is denoted as

$$\hat{\mathcal{L}}_i^a = \log \frac{\exp(\hat{S}_{i,i}^{a,b}/\tau_C)}{\sum_{j=1}^{M}(\exp(\hat{S}_{i,j}^{a,a}/\tau_C) + \exp(\hat{S}_{i,j}^{a,b}/\tau_C))}, \tag{20}$$

where τ_C is the temperature parameter controlling the softness. Finally, the cluster-level contrastive loss is defined by traversing all clusters, i.e.,

$$L_{clu} = \frac{1}{2M} \sum_{i=1}^{M} (\hat{\mathcal{L}}_i^a + \hat{\mathcal{L}}_i^b) - H(Y), \tag{21}$$

where $H(Y)$ is denoted as

$$H(Y) = -\sum_{i=1}^{M} (P(\hat{c}_i^a)\log P(\hat{c}_i^a) + P(\hat{c}_i^b)\log P(\hat{c}_i^b)), \tag{22}$$

which is the entropy of the subtype cluster to avoid the issue that most instances are assigned to the same subtype cluster.

Finally, the total loss is

$$\mathcal{L}_{total} = \mathcal{L}_{ins} + \lambda \mathcal{L}_{clu}, \tag{23}$$

where λ is the hyperparameter.

G. Experimental Settings

The Gene-MOE is implemented using the PyTorch framework. We train and evaluate Gene-MOE using an NVIDIA Tesla V100 (32GB) GPU. During the pre-training stage, we initially train the Gene-MOE model on the normalized pan-cancer dataset including 33 cancer types. We then randomly divide the pan-cancer data into the train dataset and test dataset as a ratio of 4:1, where the training dataset is used for training and the test dataset is used for hyperparameter search and feature analysis. The Adam optimizer is used in the Gene-MOE training. For better convergence, we introduce a learning rate decay method that sets the learning rate that remains constant for the first half epochs and decays linearly to 0 for the last half epochs. To find the best hyperparameter settings, we perform a random parameter search scheme. We set the noise variance range to [0.2, 0.4, 0.6, 0.8], the epochs range to [100, 200, 300], the learning rate range to [0.002, 0.0002, 0.00002], the λ_{kl} range to [1, 10, 20], the λ_{l1} range to [1, 10, 20], the $\lambda_{balance}$ range to [1, 10, 20], and the λ_{gp} range to [1, 10, 20]. Next, we perform a hyperparameter search and obtain a noise variance of 0.2, epoch of 200, a learning rate of 0.0002, λ_{kl} of 10, λ_{l1} of 20, $\lambda_{balance}$ of 10, and λ_{gp} of 10.

For the downstream tasks, we use the Bayesian Optimization strategy to select the best hyperparameter settings. During the survival analysis phase, we evaluate the survival model in the same way as Cox-nnet [12] and the VAE-Cox [14]. We train

specific Gene-MOE on 14 TCGA datasets of common cancer types. For each dataset, we use the same method to randomly divide the train dataset and test dataset in a ratio of 4:1. The optimal hyperparameters of each model are selected using the Bayesian Optimization strategy. Moreover, we repeat the entire process 5 times and calculate the average result to avoid the bias of the random splitting. In the cancer classification phase, due to the imbalance of samples for each cancer, directly dividing the pan-cancer data into the training set and the test set at a ratio of 4:1 will result in the missing cancer types with a small number of samples in the test set. To solve the above issues, we introduce a new partition method in the sample that each cancer is divided into the train subset and test subset according to the ratio of 4:1, and then the training subsets and test subsets of 33 cancers are combined to obtain the training set and test set of the classification task. During the cancer subtyping stage, we use 6 cancer types for performance evaluation, i.e., BLCA, BRCA, KIRC, LUAD, STAD, and UCEC. Then, we train the corresponding model five times to take the average result as the final result.

H. Evaluation Metric

In the survival analysis phase, the evaluation method we mainly used is the Concordance Index [33], which is widely used in survival analysis models and ranges from 0 to 1. When the Concordance Index $\leq$ 0.5, the model has completed an ineffective survival analysis prediction. When the Concordance index $\geq$ 0.5 and higher, the prediction effect of the model has been better.

In the cancer classification phase, we mainly use the accuracy, precision, recall, and F1-score. The accuracy is denoted as

$$Acc = \frac{1}{c}\sum_{i}^{c}\frac{TP_i + TN_i}{N_i}, \tag{24}$$

where c denotes the number of the class, TP_i denotes the true positive samples of class i, TN_i denotes the true negative of class i, and N_i represents the total samples of class i. Precision metrics express the ability of the classifier to predict the accuracy of positive samples correctly, and it is denoted as

$$Precision = \frac{1}{c}\sum_{i}^{c}\frac{TP_i}{TP_i + FP_i}, \tag{25}$$

where FP_i denotes the false positive samples of class i. The recall metric reflects the ability of the classifier to correctly predict the fullness of positive samples, and it is denoted as

$$Recall = \frac{1}{c}\sum_{i}^{c}\frac{TP_i}{TP_i + FN_i}, \tag{26}$$

where FN_i denotes the false negative samples of class i. Finally, the F1-score denotes the harmonic mean of precision and recall, and it is expressed as

$$F1_{score} = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \tag{27}$$

In the cancer subtyping phase, we use two evaluation criteria i.e., $-log10$ P-values and the number of significant clinical parameters, which is used by the previous study [34]. First, the differential survival $-log10$ P-values are measured between the obtained clusters using the log-rank test [35]. This metric assumes that the subtypes are biologically meaningful if they have significantly different survival. Then the number of significant clinical parameters is evaluated for the enrichment of clinical labels in the clusters. we use six clinical labels to test the enrichment, which are age at diagnosis, gender, pathologic stage, pathologic T, pathologic N, and pathologic M. The age at diagnosis parameter is the numeric parameter and is carried out by the Kruskal–Wallis test [34]. The last four parameters are discrete pathological parameters and are calculated using $\mathcal{X}^2$ test.

III. Results and Discussion

A. Performance on Survival Analysis

We select 14 representative cancer types and analyze the performance of our method in comparison with two state-of-the-art models: Cox-nnet [12] and VAE-Cox [14]. Fig. 5 shows the Concordance Index on 14 cancer types. To better compare the Concordance Index of these methods, we list the mean Concordance Index of these methods on 14 cancer types in Table I. Compared with Cox-nnet and the VAECox, the Gene-MOE outperforms the state-of-the-art models on 13 cancer types, with a higher mean Concordance Index, which shows that Gene-MOE carries out a more accurate survival analysis compared with these two models. We find that the performance of Gene-MOE on the STAD dataset is inferior to some state-of-the-art methods. By examining the STAD dataset, we discover a significant imbalance between positive and negative samples, indicating that our proposed model is insufficient in learning from the imbalanced dataset.

We further employ Gene-MOE for survival analysis predictions. Specifically, we select the test datasets of the 12 cancer types indicated in Fig. 5. Patients are divided into high-risk and low-risk groups based on the medium predictions of Gene-MOE. Subsequently, we plot Kaplan–Meier (KM) survival curves and conduct log-rank tests. We also carry out the same survival analysis for VAE-Cox. Fig. 6 displays the KM curve results for 12 cancer types. We find that Gene-MOE outperforms VAE-Cox significantly for these 12 cancer types with a lower logP value, indicating the effective capability of Gene-MOE to split patients into high-risk and low-risk groups. These results demonstrate that Gene-MOE can effectively predict the risk of hazards to patients.

B. Performance on Cancer Type Classification

We evaluate the precision, recall, and F1-score of the classification model employing Gene-MOE. The performance results are depicted in Fig. 7. As illustrated in this figure, our model demonstrates significant performance on 32 cancer types except rectum adenocarcinoma (READ), with a total accuracy of 95.8%.

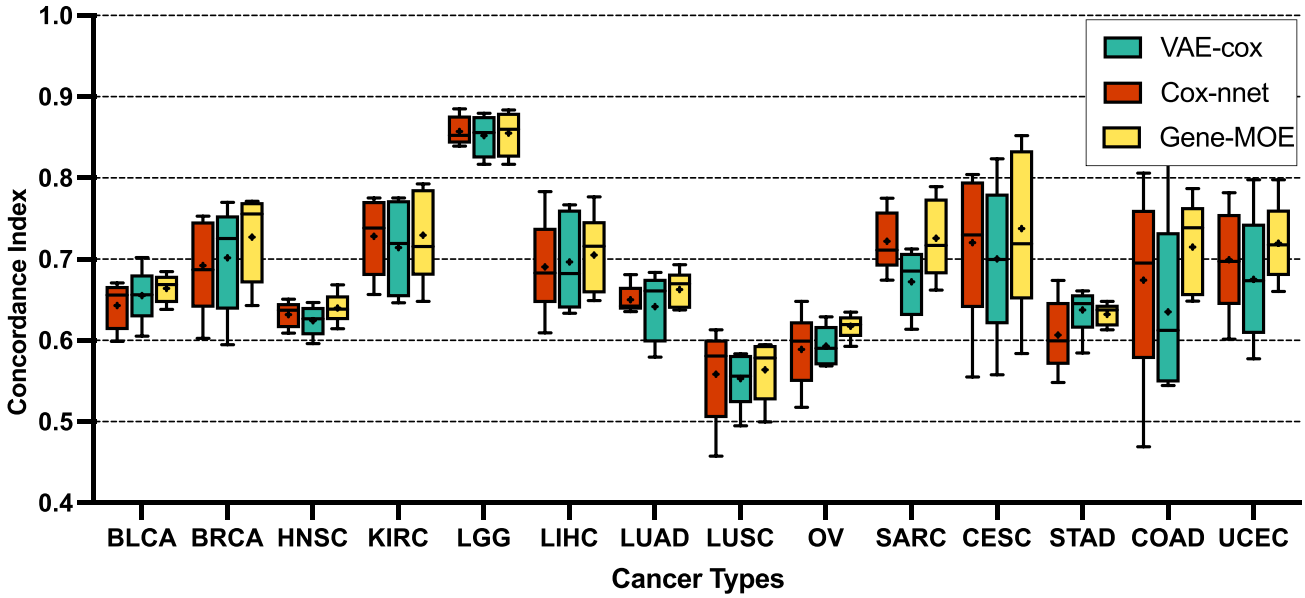
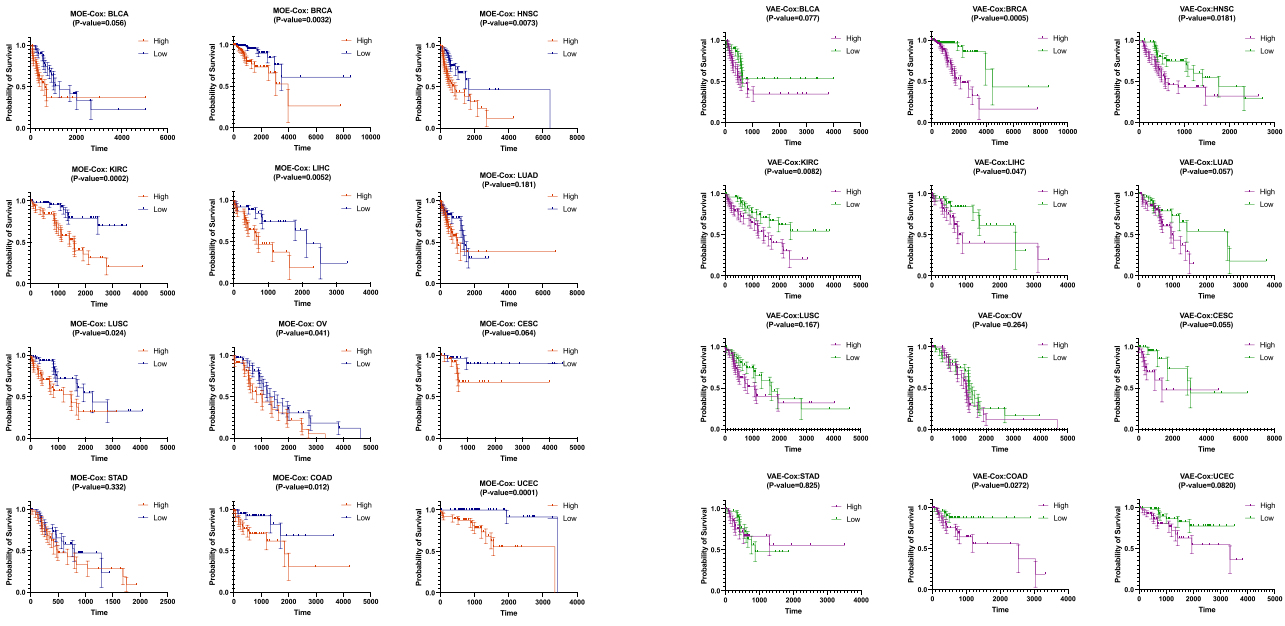


Fig. 5. Concordance index comparison of survival analysis on 14 cancer types. "+" of each box represents the mean concordance index.

TABLE I
COMPARISON OF MEAN CONCORDANCE INDEX ON 14 CANCER TYPES

	BLCA	BRCA	HNSC	KIRC	LGG	LIHC	LUAD	LUSC	OV	SARC	CESC	STAD	COAD	UCEC
Cox-nnet	0.643	0.692	0.632	0.728	0.847	0.691	0.649	0.558	0.590	0.722	0.720	0.607	0.674	0.699
VAE-cox	0.655	0.702	0.624	0.714	0.852	0.697	0.642	0.553	0.593	0.672	0.700	0.638	0.635	0.675
Gene-MOE	0.664	0.727	0.640	0.730	0.855	0.705	0.662	0.564	0.617	0.726	0.738	0.632	0.715	0.720



(a) Gene-MOE (b) VAE-Cox

Fig. 6. Kaplan–Meier survival curves using MOE-Cox and VAE-cox on 12 cancer types. The Gene-MOE model shows a lower logP value on 10 cancer data sets. It shows that the Gene-MOE can predict hazard ratios that divide into high-risk groups and low-risk groups more significantly compared with VAE-cox on 10 cancer datasets.

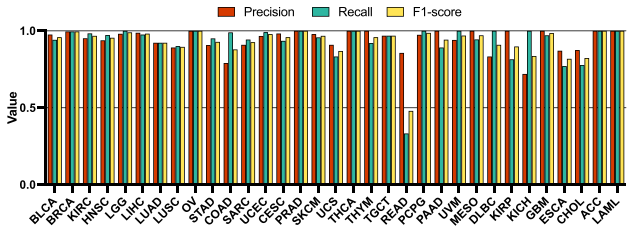


Fig. 7. Classification performance on 33 cancer types using Gene-MOE.

TABLE II
COMPARISON OF CLASSIFICATION METRICS

	Accuracy	Precision	Recall	F1-Score
RandomForest	0.9010	0.8943	0.8338	0.8409
SVM	0.9498	0.9281	0.9114	0.9149
MLP	0.9250	0.8989	0.8756	0.8790
2D-Hybrid-CNN	0.9507	0.9300	0.9229	0.9232
Gene-MOE	**0.9580**	**0.9554**	**0.9374**	**0.9333**

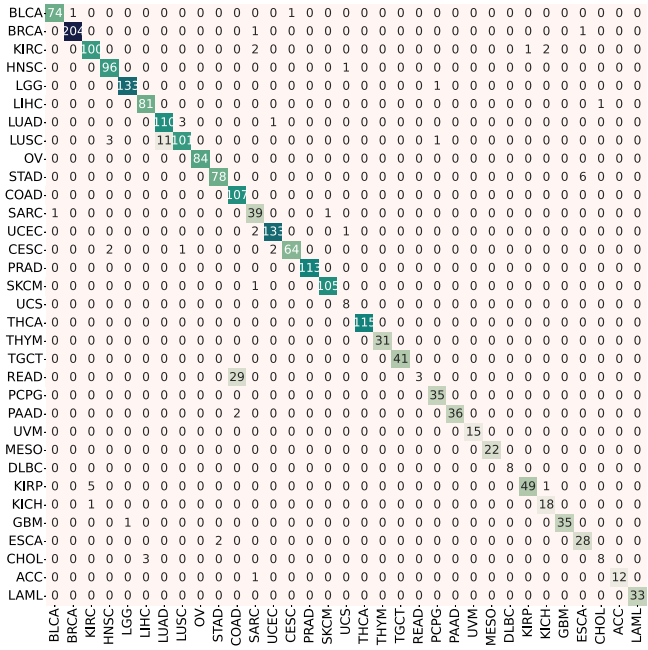


Fig. 8. Confusion matrix of test samples predicted by classification model using Gene-MOE with 33 cancer types.

To prove the novelty of proposed classifier, we select several state-of-the-art classifiers for comparison, including classifier models based on machine learning methods such as random forest [36] and SVM [36], [37], as well as deep learning based classifier models like MLP [38] and CNN [39]. To ensure the fairness of the comparative experiments, we conduct five independent processes and calculate the average of the classification results to avoid bias. Table II shows the classification result of these five models. According to Table II, We find that the classification results of the Gene-MOE model on the test dataset are improved by 0.1–0.5 compared with the other four methods.

We further analyze the classification performance by constructing a confusion matrix. Fig. 8 illustrates the confusion matrix constructed for 33 types of cancer. Our model demonstrates accurate classification of all 33 cancer types with a notably low misclassification rate according to the confusion matrix. Moreover, we observe that our model misclassifies 29 cases of READ as colon adenocarcinoma (COAD). This issue occurs because READ and COAD are genetically identical [40], and the samples in COAD samples are larger than READ, causing the model to misclassify some READ patients as COAD patients.

C. Performance on Cancer Subtyping

In this section, we conduct a cluster performance comparison of Gene-MOE against four state-of-the-art methods: Subtype-DCC, Subtype-GAN, NEMO, and SNF. We select nine representative cancer types and analyze the clustering performance of these models. According to the related works, these nine representative cancer types are identified as reasonable subtypes (i.e., BLCA with 5 subtypes, BRCA with 5 subtypes, KIRC with 4 subtypes, LUAD with 3 subtypes, STAD with 3 subtypes, and UCEC with 4 subtypes, SKCM with 4 subtypes, UVM with 4 subtypes, PAAD with 2) [29]. We then perform subtype identification and clustering based on the known subtype numbers. The performance results are listed in Table III. We use the -log10 P-values and the number of significant clinical parameters to evaluate the performance of the cancer subtyping. Compared with the other method, the Gene-MOE model achieves the best result on at least one metric over seven datasets. According to Table III, the best results for the two metrics performed by Gene-MOE are achieved on the LUAD and STAD datasets. On the LUAD dataset, the survival -log10 P-value reaches 2.07, and the number of significant clinical parameters is 5. On the STAD dataset, the survival -log10 P-value reaches 1.58, and the number of significant clinical parameters is 3. These results suggest that Lung Adenocarcinoma and Stomach Adenocarcinoma are well-subtyped compared to state-of-the-art methods.

D. Feature Analysis of Gene-MOE

1) Correlation Analysis: We conduct a correlation analysis between the low-dimensional features extracted by the Gene-MOE model and the high-dimensional genes of patients. Specifically, we use 1,000 patient samples from the test dataset to feed into the Gene-MOE model, and the model evaluates the low-dimensional features of each patient. We then select the top 20 features with the highest variance as the leading features and calculate their Pearson correlations with the original patient genetic information. Not that the normalization step is performed before selecting the top features to confirm the features are in the same scale. The results are shown in Fig. 9(a), which presents a heat map illustrating the correlation between these leading features and the genes of 1,000 patients, which unequivocally indicates a significant correlation between the low-dimensional features predicted by the Gene-MOE model and the original input features. Based on Fig. 9(a), we further refine our analysis by identifying genes with an average absolute correlation exceeding 0.4 with respect to the 20 leading features. Then, 29 strongly correlated genes are filtered, which is shown in Fig. 9(b). According to the result in Fig. 9(b), we observe that many genes with strong correlations to the leading features are cancer-related genes. For example, the TMPRSS4 gene is an emerging potential therapeutic target in cancer [41]. Moreover, tensin4 expression

TABLE III
PERFORMANCE COMPARISON (-LOG10 P-VALUES/NUMBER OF SIGNIFICANT CLINICAL PARAMETERS) OF GENE-MOE AND OTHER STATE-OF-THE-ART METHODS ON NINE DATASETS

	BLCA	BRCA	KIRC	LUAD	STAD	UCEC	UVM	PAAD	SKCM
SNF	1.31/6	0.93/5	8.19/6	2.23/4	0.72/2	5/1	2.77/0	**3.24**/3	5.27/4
NEMO	2.8/5	1.21/6	5.72/5	2.63/4	1.8/2	5.96/1	2.38/0	3.04/1	5.01/4
Subtype-GAN	1.45/4	1.28/**6**	7.77/**6**	2.83/3	0.39/2	7.4/1	2.62/0	1.65/1	0.1/2
Subtype-DCC	**2.35/6**	1.11/5	8.97/6	1.69/4	1.48/2	5.46/1	**2.77**/0	3.75/1	**5.94/4**
Gene-MOE	2.21/5	**1.33**/5	**11.99**/5	**2.07/5**	**1.85/3**	**4.08/1**	2.6/**2**	2.3/**3**	1.81/3

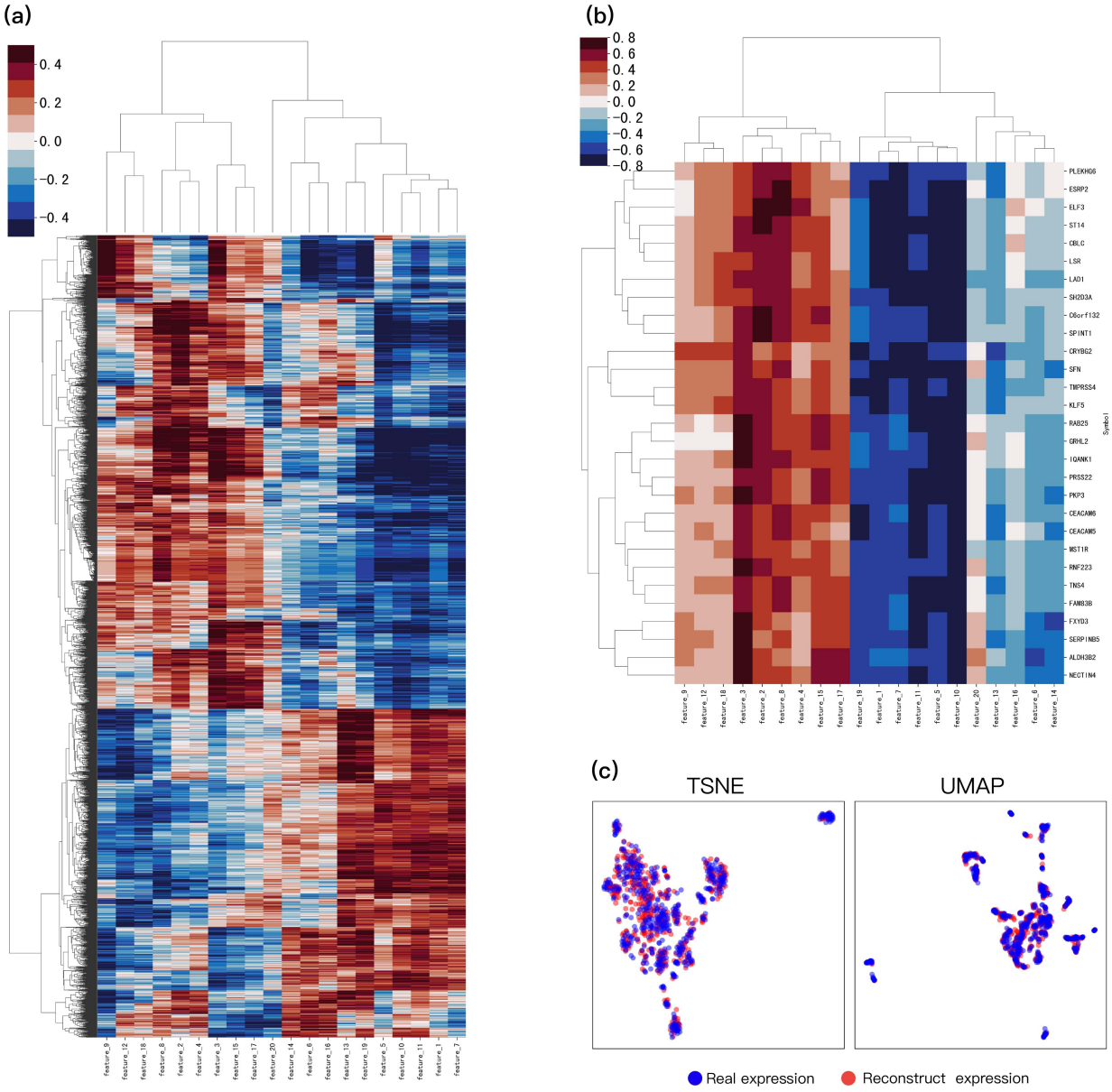


Fig. 9. Feature analysis of Gene-MOE. (a) Pearson correlation heat map between the leading feature and the patient genes. The features learned by Gene-MOE are strongly correlated with patient genes. (b) Pearson correlation heat map by selecting mean absolute coefficient greater than 0.5. (c) TSNE and UMAP results of real genes and genes reconstructed by Gene-MOE. According to the results, the reconstructed genes maintain the same distribution as the real genes.

showes prognostic relevance in gastric cancer [42]. Furthermore, E2F1-initiated transcription of PRSS22 promotes breast cancer metastasis by cleaving ANXA1 and activating the FPR2/ERK signaling pathway [43]. In addition, down-regulation of FXYD3 expression is observed in lung cancers [44]. Moreover, ST14 gene expression affects breast cancer [45], [46], [47]. In addition, RAB25 has been implicated in various cancers, with reports presenting it as both an oncogene and a tumor-suppressor gene [48], [49]. Long intergenic non-coding RNA 00324 promotes gastric cancer cell proliferation by binding with HuR and

TABLE IV
COMPARISON OF CONCORDANCE INDEX ON 14 CANCER TYPES USING FOUR DISTINCT MODELS

	BLCA	BRCA	HNSC	KIRC	LGG	LIHC	LUAD	LUSC	OV	SARC	CESC	STAD	COAD	UCEC
Baseline	0.643	0.692	0.632	0.728	0.847	0.691	0.649	0.558	0.590	0.722	0.720	0.607	0.674	0.699
MOE	0.670	0.711	**0.664**	0.729	0.848	0.700	0.663	0.596	0.616	0.733	0.720	0.641	0.705	0.712
MOAE	0.664	0.700	0.656	**0.730**	**0.850**	**0.713**	0.660	**0.599**	0.613	0.736	0.733	**0.728**	0.684	**0.719**
pre-train	**0.674**	**0.718**	0.649	**0.730**	**0.850**	0.708	**0.666**	0.590	**0.616**	**0.737**	**0.738**	0.630	**0.715**	0.716

TABLE V
PERFORMANCE COMPARISON (-LOG10 P-VALUES/NUMBER OF SIGNIFICANT CLINICAL PARAMETERS) USING FOUR DISTINCT MODELS ON NINE CANCER TYPES

	BLCA	BRCA	KIRC	LUAD	STAD	UCEC	UVM	PAAD	SKCM
Baseline	3.85/5	1.17/5	4.16/4	1.47/5	1.12/3	2.46/1	1.19/0	0.33/3	0.93/1
MOE	3.92/4	1.05/5	6.21/5	2.83/5	1.18/4	2.35/1	2.59/1	1.33/1	1.79/2
MOAE	3.92/4	1.05/5	6.22/5	2.85/5	1.18/**4**	2.36/1	2.6/2	1.97/1	1.80/1
pre-train	2.21/**5**	**1.33/5**	**11.99/5**	**2.07/5**	**1.85**/3	**4.08/1**	**2.6/2**	**2.3/3**	**1.81**/3

TABLE VI
COMPARISON OF CLASSIFICATION METRICS USING FOUR DISTINCT MODELS

	Accuracy	Precision	Recall	F1-Score
Baseline	0.9304	0.9235	0.8721	0.8799
MOE	0.9417	0.9214	0.9063	0.9098
MOAE	**0.9584**	0.9469	0.9349	**0.9392**
pre-train	0.9580	**0.9554**	**0.9374**	0.9333

stabilizing FAM83B expression [50]. CBLC expression is found to be higher in breast cancer tissues and cells than in normal tissues and cells [51]. Furthermore, Serpin B5 is shown to be a CEA-interacting biomarker for colorectal cancer [52]. PKP3 interactions with the MAPK-JNK-ERK1/2-mTOR pathway regulate autophagy and invasion in ovarian cancer [53]. In addition, LAD1 expression is associated with the metastatic potential of colorectal cancer cells [54]. ELF3 is found to be a negative regulator of epithelial–mesenchymal transition in ovarian cancer cells [55]. Moreover, the expression patterns of CEACAM5 and CEACAM6 are observed in primary and metastatic cancers [56]. Grhl2 determines the epithelial phenotype of breast cancers and promotes tumor progression [57]. KLF5 promotes breast cancer proliferation, migration, and invasion, in part by up-regulating the transcription of TNFAIP2 [58]. Finally, genetic predisposition to colon and rectal adenocarcinoma is found to be mediated by a super-enhancer polymorphism coactivating CD9 and PLEKHG6 [59]. According to the analysis in Fig. 9(a) and (b), Gene-MOE can extract rich cancer-related Genes, and each feature is strongly correlated with the specific genes, which indicates that the Gene-MOE can effectively learn the potential semantic correlation of genes.

2) Visualization Analysis: We further perform a visual analysis to measure the performance of Gene-MOE. We randomly select 1,000 patients from the test set to perform this evaluation. The real gene expression of test patients is first fed into Gene-MOE to generate the reconstruction expression. Then, we use TSNE and UMAP to perform the visualization and evaluate the similarity of these two distributions. Fig. 9(c) shows the visualization result using these two methods. According to Fig. 9(c), the real gene distribution perfectly coincides with the reconstructed gene distribution, which proves that the Gene-MOE model can perform reconstruction of the input genes more accurately based on the input real genes. Furthermore, Gene-MOE completes the reconstruction based on the low-dimensional feature obtained by the backbone model, which further reinforces that Gene-MOE can learn rich feature representation by the backbone network.

E. Ablation Studies

In this section, we extend our analysis by presenting ablation experiments to evaluate the performance of the model proposed in this paper. Four distinct models are constructed for this purpose: 1) the baseline model comprising two FCN layers, 2) the model incorporating MOE by replacing the FCN layers, 3) the Gene-MOE model, and 4) the pre-trained Gene-MOE model. Subsequently, survival analysis, cancer classification, and cancer subtyping tasks are performed using these four models. The results of these tasks are presented in Tables IV, V, and VI.

1) Performance of MOE: By comparing survival analysis results in Table IV, we find the model with the MOE layer could achieve better Concordance results than the baseline on 13 cancer types. By comparing cancer subtyping results in Table V, we find the model with the MOE layer outperforms the baseline on eight cancer types. Moreover, in Table VI, the model with MOE shows better accuracy and F1-score. These results indicate that using the MOE layer can increase the performance of genomic analysis, which proves our theory that using MOE can make a model learn rich features during the feature extraction process.

2) Performance of MOAE: By comparing the Concordance Index for the model with MOE and the Gene-MOE model in Table IV, we observe that Gene-MOE outperforms the model with MOE on nine cancer types. By comparing the cancer subtyping performance, we observe that Gene-MOE outperforms the model with MOE on six cancer types. Furthermore, in Table IV, we find that Gene-MOE performs better in accuracy, recall, precision, and F1-score. These findings prove that the MOAE model can further improve the accuracy of the model by improving the ability to learn deep semantic correlated features.

3) Performance of Pre-Training: By comparing the performance of the Gene-MOE and pre-trained Gene-MOE models

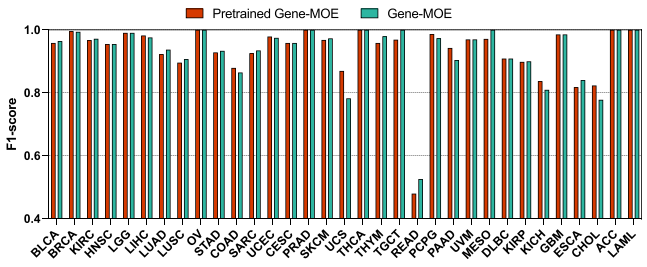


Fig. 10. Comparison of F1-score on 33 cancer types using pre-trained Gene-MOE and Gene-MOE.

in Table IV, we find that the pre-trained Gene-MOE model performs better in eight cancer types. At the same time, the Gene-MOE model demonstrates the same performance on KIRC and LGG datasets. By comparing the performance of Gene-MOE and the pre-trained Gene-MOE in Table V, we find the pre-trained Gene-MOE model performs better on at least one metric of the log10 P-values and the number of significant clinical on six cancers. These results reveal the effectiveness of pre-training, particularly in enhancing the performance of most survival analysis models. Further comparison of Gene-MOE and pre-trained Gene-MOE in Table VI shows that the pre-trained Gene-MOE model performs better in precision and recall, but the accuracy and F1-score are lower than that of Gene-MOE. By further comparing the F1-score of these two models in Fig. 10, we find that the pre-trained model performs less effectively on cancer types with few samples, such as the READ dataset. This discrepancy can be attributed to the fact that the pre-training stage primarily captures common cancer features, potentially missing certain characteristics for cancer types with small sample sizes. Consequently, during the fine-tuning stage, the model tends to favor learning patterns from cancer types with larger sample sizes. Although pre-training is unsuitable for cancer classification tasks, our proposed MOE module and MOAE module can further improve cancer classification, proving the Gene-MOE's novelty.

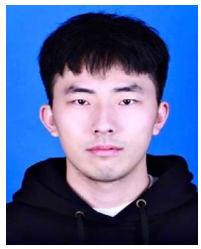
IV. Conclusion

In this work, we develop a sparsely gated model called Gene-MOE, which extensively leverages MOE layers to further deepen the ability to extract the deep correlation features of high-dimensional genes. Furthermore, we propose a novel MOAE module to explore the deep semantic associations between high-dimensional genetic features. Finally, we design novel pre-training strategies including data augmentation, self-supervised learning, and new loss functions to further improve the performance of Gene-MOE. The results show that Gene-MOE could achieve the best performance on cancer classification and survival analysis, indicating its strong potential for use in those applications. Currently, however, Gene-MOE has some limitations. During the pre-training stage, Gene-MOE focuses on cancer types with larger sample sizes, resulting in insufficient fitting of small sample datasets. Furthermore, the amount of existing data is insufficient, which leads to over-fitting issues in survival analysis and cancer classification tasks. In our future work, we aim to gather more genetic data for model training and to optimize the model training performance.

References

[1] G. Liu, C. Dong, and L. Liu, "Integrated multiple "-omics" data reveal subtypes of hepatocellular carcinoma," *PLoS One*, vol. 11, no. 11, 2016, Art. no. e0165457.
[2] M. J. Barry, "Prostate-specific–antigen testing for early diagnosis of prostate cancer," *New England J. Med.*, vol. 344, no. 18, pp. 1373–1377, 2001.
[3] G. Brett, "Earlier diagnosis and survival in lung cancer," *Brit. Med. J.*, vol. 4, no. 5678, pp. 260–262, 1969.
[4] J. C. Venter et al., "The sequence of the human genome," *Science*, vol. 291, no. 5507, pp. 1304–1351, 2001.
[5] K. Y. Yeung and W. L. Ruzzo, "Principal component analysis for clustering gene expression data," *Bioinformatics*, vol. 17, no. 9, pp. 763–774, 2001.
[6] Y. Lai, B. Wu, L. Chen, and H. Zhao, "A statistical method for identifying differential gene–gene co-expression patterns," *Bioinformatics*, vol. 20, no. 17, pp. 3146–3155, 2004.
[7] Y. Choi and C. Kendziorski, "Statistical methods for gene set co-expression analysis," *Bioinformatics*, vol. 25, no. 21, pp. 2780–2786, 2009.
[8] D. Y. Lin and L.-J. Wei, "The robust inference for the Cox proportional hazards model," *J. Amer. Stat. Assoc.*, vol. 84, no. 408, pp. 1074–1078, 1989.
[9] D. Capper et al., "DNA methylation-based classification of central nervous system tumours," *Nature*, vol. 555, no. 7697, pp. 469–474, 2018.
[10] R. Díaz-Uriarte and S. Alvarez de Andrés, "Gene selection and classification of microarray data using random forest," *BMC Bioinf.*, vol. 7, pp. 1–13, 2006.
[11] K.-B. Duan, J. C. Rajapakse, H. Wang, and F. Azuaje, "Multiple SVM-RFE for gene selection in cancer classification with expression data," *IEEE Trans. Nanobiosci.*, vol. 4, no. 3, pp. 228–234, Sep. 2005.
[12] T. Ching et al., "Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data," *PLoS Comput. Biol.*, vol. 14, no. 4, 2018, Art. no. e1006076.
[13] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, "DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network," *BMC Med. Res. Methodol.*, vol. 18, no. 1, pp. 1–12, 2018.
[14] S. Kim et al., "Improved survival analysis by learning shared genomic information from pan-cancer data," *Bioinformatics*, vol. 36, no. Supplement_1, pp. i389–i398, 2020.
[15] F. Gao et al., "DeepCC: A novel deep learning-based framework for cancer molecular subtype classification," *Oncogenesis*, vol. 8, no. 9, 2019, Art. no. 44.
[16] Z. Zeng et al., "Deep learning for cancer type classification and driver gene identification," *BMC Bioinf.*, vol. 22, no. 4, pp. 1–13, 2021.
[17] R. Ramirez et al., "Classification of cancer types using graph convolutional neural networks," *Front. Phys.*, vol. 8, 2020, Art. no. 00203.
[18] R. Ramirez et al., "Prediction and interpretation of cancer survival using graph convolution neural networks," *Methods*, vol. 192, pp. 120–130, 2021.
[19] X. Li et al., "MARPPI: Boosting prediction of protein–protein interactions with multi-scale architecture residual network," *Brief. Bioinf.*, vol. 24, no. 1, 2023, Art. no. bbac524.
[20] X. Wang et al., "TransFusionNet: Semantic and spatial features fusion framework for liver tumor and vessel segmentation under JetsonTX2," *IEEE J. Biomed. Health Informat.*, vol. 27, no. 3, pp. 1173–1184, Mar. 2023.
[21] N. Shazeer et al., "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," in *Proc. Int. Conf. Learn. Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=B1ckMDqlg
[22] D. Lepikhin et al., "GShard: Scaling giant models with conditional computation and automatic sharding," in *Proc. Int. Conf. Learn. Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=qrwe7XHTmYb
[23] J. Lee et al., "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.
[24] Y. Ji et al., "DNABERT: Pre-trained bidirectional encoder representations from transformers model for DNA-language in genome," *Bioinformatics*, vol. 37, no. 15, pp. 2112–2120, 2021.

[25] F. Yang et al., "scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data," *Nature Mach. Intell.*, vol. 4, no. 10, pp. 852–866, 2022.
[26] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim, "A gene-coexpression network for global discovery of conserved genetic modules," *Science*, vol. 302, no. 5643, pp. 249–255, 2003.
[27] S. Horvath and J. Dong, "Geometric interpretation of gene coexpression network analysis," *PLoS Comput. Biol.*, vol. 4, no. 8, 2008, Art. no. e1000117.
[28] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.
[29] J. Zhao et al., "Subtype-DCC: Decoupled contrastive clustering method for cancer subtype identification based on multi-omics data," *Brief. Bioinf.*, vol. 24, no. 2, 2023, Art. no. bbad025.
[30] H. Yang, R. Chen, D. Li, and Z. Wang, "Subtype-GAN: A deep learning approach for integrative cancer subtyping of multi-omics data," *Bioinformatics*, vol. 37, no. 16, pp. 2231–2237, 2021.
[31] Y. Li, P. Hu, Z. Liu, D. Peng, J. T. Zhou, and X. Peng, "Contrastive clustering," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 8547–8555.
[32] C.-H. Yeh, C.-Y. Hong, Y.-C. Hsu, T.-L. Liu, Y. Chen, and Y. LeCun, "Decoupled contrastive learning," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2022, pp. 668–684.
[33] H. Steck et al., "On ranking in survival analysis: Bounds on the concordance index," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 1209–1216.
[34] S. Ge, J. Liu, Y. Cheng, X. Meng, and X. Wang, "Multi-view spectral clustering with latent representation learning for applications on multi-omics cancer subtyping," *Brief. Bioinf.*, vol. 24, no. 1, 2023, Art. no. bbac500.
[35] P. Mukhopadhyay et al., "Log-rank test vs MaxCombo and difference in restricted mean survival time tests for comparing survival under nonproportional hazards in immuno-oncology trials: A systematic review and meta-analysis," *JAMA Oncol.*, vol. 8, no. 9, pp. 1294–1300, 2022.
[36] A. Statnikov et al., "A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification," *BMC Bioinf.*, vol. 9, no. 1, pp. 1–10, 2008.
[37] K. Kourou et al., "Machine learning applications in cancer prognosis and prediction," *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 8–17, 2015.
[38] T. Ahn et al., "Deep learning-based identification of cancer or normal tissue using gene expression data," in *Proc. 2018 IEEE Int. Conf. Bioinf. Biomed.*, 2018, pp. 1748–1752.
[39] M. Mostavi et al., "Convolutional neural network models for cancer type prediction based on gene expression," *BMC Med. Genomic.*, vol. 13, pp. 1–13, 2020.
[40] The Cancer Genome Atlas Network, "Comprehensive molecular characterization of human colon and rectal cancer," *Nature*, vol. 487, no. 7407, pp. 330–337, 2012.
[41] A. De Aberasturi and A. Calvo, "TMPRSS4: An emerging potential therapeutic target in cancer," *Brit. J. Cancer*, vol. 112, no. 1, pp. 4–8, 2015.
[42] K. Sakashita et al., "Prognostic relevance of Tensin4 expression in human gastric cancer," *Ann. Surg. Oncol.*, vol. 15, pp. 2606–2613, 2008.
[43] L. Song et al., "E2F1-initiated transcription of PRSS22 promotes breast cancer metastasis by cleaving ANXA1 and activating FPR2/ERK signaling pathway," *Cell Death Dis.*, vol. 13, no. 11, 2022, Art. no. 982.
[44] K. Okudela et al., "Down-regulation of FXYD3 expression in human lung cancers: Its mechanism and potential role in carcinogenesis," *Amer. J. Pathol.*, vol. 175, no. 6, pp. 2646–2656, 2009.
[45] J. M. Kauppinen et al., "ST14 gene variant and decreased matriptase protein expression predict poor breast cancer survival," *Cancer Epidemiol. Biomarkers Prevention*, vol. 19, no. 9, pp. 2133–2142, 2010.
[46] Y. Wang et al., "SR14 (suppression of tumorigenicity 14) gene is a target for MIR-27B, and the inhibitory effect of ST14 on cell growth is independent of MIR-27B regulation," *J. Biol. Chem.*, vol. 284, no. 34, pp. 23094–23106, 2009.
[47] Y.-H. Dai et al., "Gene-associated methylation status of ST14 as a predictor of survival and hormone receptor positivity in breast cancer," *BMC Cancer*, vol. 21, no. 1, pp. 1–14, 2021.
[48] S. Mitra et al., "RAB25 in cancer: A brief update," *Biochem. Soc. Trans.*, vol. 40, no. 6, pp. 1404–1408, 2012.
[49] S. Wang et al., "RAB25 gtpase: Functional roles in cancer," *Oncotarget*, vol. 8, no. 38, 2017, Art. no. 64591.
[50] Z. Zou et al., "Long intergenic non-coding RNA 00324 promotes gastric cancer cell proliferation via binding with HuR and stabilizing FAM83B expression," *Cell Death Dis.*, vol. 9, no. 7, 2018, Art. no. 717.
[51] W. Li et al., "CBLC inhibits the proliferation and metastasis of breast cancer cells via ubiquitination and degradation of CTTN," *J. Receptors Signal Transduct.*, vol. 42, no. 6, pp. 588–598, 2022.
[52] J. Y. Baek et al., "Serpin B5 is a CEA-interacting biomarker for colorectal cancer," *Int. J. Cancer*, vol. 134, no. 7, pp. 1595–1604, 2014.
[53] V. Lim et al., "PKP3 interactions with MAPK-JNK-ERK1/2-mTOR pathway regulates autophagy and invasion in ovarian cancer," *Biochem. Biophysical Res. Commun.*, vol. 508, no. 2, pp. 646–653, 2019.
[54] B. Moon et al., "LAD1 expression is associated with the metastatic potential of colorectal cancer cells," *BMC Cancer*, vol. 20, pp. 1–12, 2020.
[55] T.-L. Yeung et al., "ELF3 is a negative regulator of epithelial-mesenchymal transition in ovarian cancer cells," *Oncotarget*, vol. 8, no. 10, 2017, Art. no. 16951.
[56] R. D. Blumenthal et al., "Expression patterns of CEACAM5 and CEACAM6 in primary and metastatic cancers," *BMC Cancer*, vol. 7, pp. 1–15, 2007.
[57] X. Xiang et al., "GRHL2 determines the epithelial phenotype of breast cancers and promotes tumor progression," *PLoS One*, vol. 7, no. 12, 2012, Art. no. e50781.
[58] L. Jia et al., "KLF5 promotes breast cancer proliferation, migration and invasion in part by upregulating the transcription of TNFAIP2," *Oncogene*, vol. 35, no. 16, pp. 2040–2051, 2016.
[59] J. Ke et al., "Genetic predisposition to colon and rectal adenocarcinoma is mediated by a super-enhancer polymorphism coactivating CD9 and PLEKHG6," *Cancer Epidemiol. Biomarkers Prevention*, vol. 29, no. 4, pp. 850–859, 2020.



Xiangyu Meng is currently working toward the PhD degree with the College of Computer Science and Technology, China University of Petroleum, Qingdao, China. His research interests include bioinformatics, parallel computing, and computational materials science.

Xue Li is currently working toward the PhD degree with the College of Computer Science and Technology, China University of Petroleum, Qingdao, China. Her research interests include bioinformatics, and deep learning.

Qing Yang is currently working toward the PhD degree with the College of Computer Science and Technology, China University of Petroleum, Qingdao, China. Her research interests include bioinformatics, and deep learning.

Huanhuan Dai is currently working toward the PhD degree with the College of Computer Science and Technology, China University of Petroleum, Qingdao, China. Her research interests include bioinformatics, and deep learning.

Lian Qiao received the master's degree from the College of Computer Science and Technology, China University of Petroleum, Qingdao, China. Her research interests include bioinformatics, and deep learning.

Long Hao is currently working toward the master's degree with the College of Computer Science and Technology, China University of Petroleum, Qingdao, China. His research interests include bioinformatics, and deep learning.

Hongzhen Ding is currently working toward the master's degree with the College of Computer Science and Technology, China University of Petroleum, Qingdao, China. His research interests include bioinformatics, and deep learning.

Xun Wang received the PhD degree in social systems and management from Tsukuba University. She is currently working as a professor with the School of Computer Science and Technology of China University of Petroleum, Qingdao, China. Her research interests include bioinformatics, parallel computing, and computational materials science.