Electronic Research Archive

ERA, 31(8): 4951–4967.
DOI: 10.3934/era.2023253
Received: 28 May 2023
Revised: 02 July 2023
Accepted: 05 July 2023
Published: 12 July 2023

http://www.aimspress.com/journal/ERA

Research article

Deciphering and identifying pan-cancer RAS pathway activation based on graph autoencoder and ClassifierChain

Jianting Gong[1], Yingwei Zhao[1], Xiantao Heng[1], Yongbing Chen[1], Pingping Sun[1,*], Fei He[1,*], Zhiqiang Ma[2,*] and Zilin Ren[1,3,*]

[1] School of Information Science and Technology, Northeast Normal University, Changchun 130117, China
[2] Department of Computer Science, College of Humanities & Sciences of Northeast Normal University, Changchun 130119, China
[3] Changchun Veterinary Research Institute, Chinese Academy of Agricultural Sciences, Changchun 130122, China

* **Correspondence:** Email: sunpp567@nenu.edu.cn, hef740@nenu.edu.cn, mazq@nenu.edu.cn, zilin.ren@outlook.com.

Abstract: The goal of precision oncology is to select more effective treatments or beneficial drugs for patients. The transcription of ‘‘hidden responders’’ which precision oncology often fails to identify for patients is important for revealing responsive molecular states. Recently, a RAS pathway activation detection method based on machine learning and a nature-inspired deep RAS activation pan-cancer has been proposed. However, we note that the activating gene variations found in KRAS, HRAS and NRAS vary substantially across cancers. Besides, the ability of a machine learning classifier to detect which KRAS, HRAS and NRAS gain of function mutations or copy number alterations causes the RAS pathway activation is not clear. Here, we proposed a deep neural network framework for deciphering and identifying pan-cancer RAS pathway activation (DIPRAS). DIPRAS brings a new insight into deciphering and identifying the pan-cancer RAS pathway activation from a deeper perspective. In addition, we further revealed the identification and characterization of RAS aberrant pathway activity through gene ontological enrichment and pathological analysis. The source code is available by the URL https://github.com/zhaoyw456/DIPRAS.

Keywords: pan-cancer; RAS pathway; multi-label classification; graph autoencoder

1. Introduction

Precision oncology is defined as the use of molecular profiling to formulate targeted therapies based on different patients, and it emerged to improve the efficacy of oncology treatments, reduce the side effects of drugs and avoid drug resistance [1]. Currently, significant success has been achieved in applying genomics-driven cancer therapies [2]. However, genomics-driven cancer therapy suffers from the limitation that genomic profiling information is not comprehensive and accurate enough, resulting in the inability to provide patients with precise medical regimens [3,4]. With the data in the largest uniformly processed cancer dataset increasing, the Cancer Genome Atlas (TCGA) PanCancerAtlas, transcriptomics-driven machine learning and deep learning approaches have promising to provide reference and basis for treatment decisions [5–7].

The RAS gene families KRAS, NRAS and HRAS are frequently found in tumors and are the most frequently activated proto-oncogenes [8]. Mutations in RAS genes or NF1 loss-of-function events are transmitted to proteins via transcription or translation, resulting in abnormal activation of the RAS pathway which leads to cell proliferation. RAS mutations are found in 30% of human cancer patients with KRAS gene mutations accounting for the largest proportion [9]. KRAS genes are frequently mutated in pancreatic cancer, colorectal cancer, rectal adenocarcinoma and lung adenocarcinoma, NRAS genes are frequently mutated in melanoma and thyroid cancer and HRAS genes are mutated less frequently overall [10]. Studies have pointed out that RAS gene mutation is an important cause of a variety of cancers [11] and it often occurs in the early stages of tumorigenesis. Therefore, it is particularly important to design and develop effective methods to detect RAS pathway activation. Recently, a RAS pathway activation detection method based on machine learning [5] and a nature-inspired deep RAS activation pan-cancer [6] have been proposed. Although they have achieved good performance, the similarity between different cancers or relationships within one type of cancer are ignored and underlying variants in the oncogenes KRAS, NRAS and HRAS are unclear.

Here, we propose a deep neural network framework for deciphering and identifying pan-cancer RAS pathway activation (DIPRAS). DIPRAS trains the low-dimensional feature vectors named embedding to stand for samples considering the similarity between cancers or relationships between types of cancer. Since the graph neural network [12] (GNN) has the ability to learn node relationships in a graph, we proposed a graph framework graph autoencoder [13] and linear model (GALM) with two loss functions to learn the characteristics of certain cancer types and relationships between two types of cancers. The model GALM demonstrated the ability to detect the activation of the pan-cancer RAS pathway. If we detect that the RAS pathway have been activated, further causal inference about which variants cause it is needed to explore. To further detect whether genes associated with RAS pathway activation such as KRAS, HRAS and NRAS have mutations, copy number gains (CNG) or both, we constructed a multi-label classification model (ensemble algorithm ClassifierChain, EACC) to study. We apply the method DIPRAS to RAS genes, demonstrating that our method can detect which RAS gene mutations or copy number gains in the RAS activation pathway. DIPRAS achieves to detect RAS pathway activation from new perspective.

2. Materials and methods

2.1. Methodology overview of DIPRAS

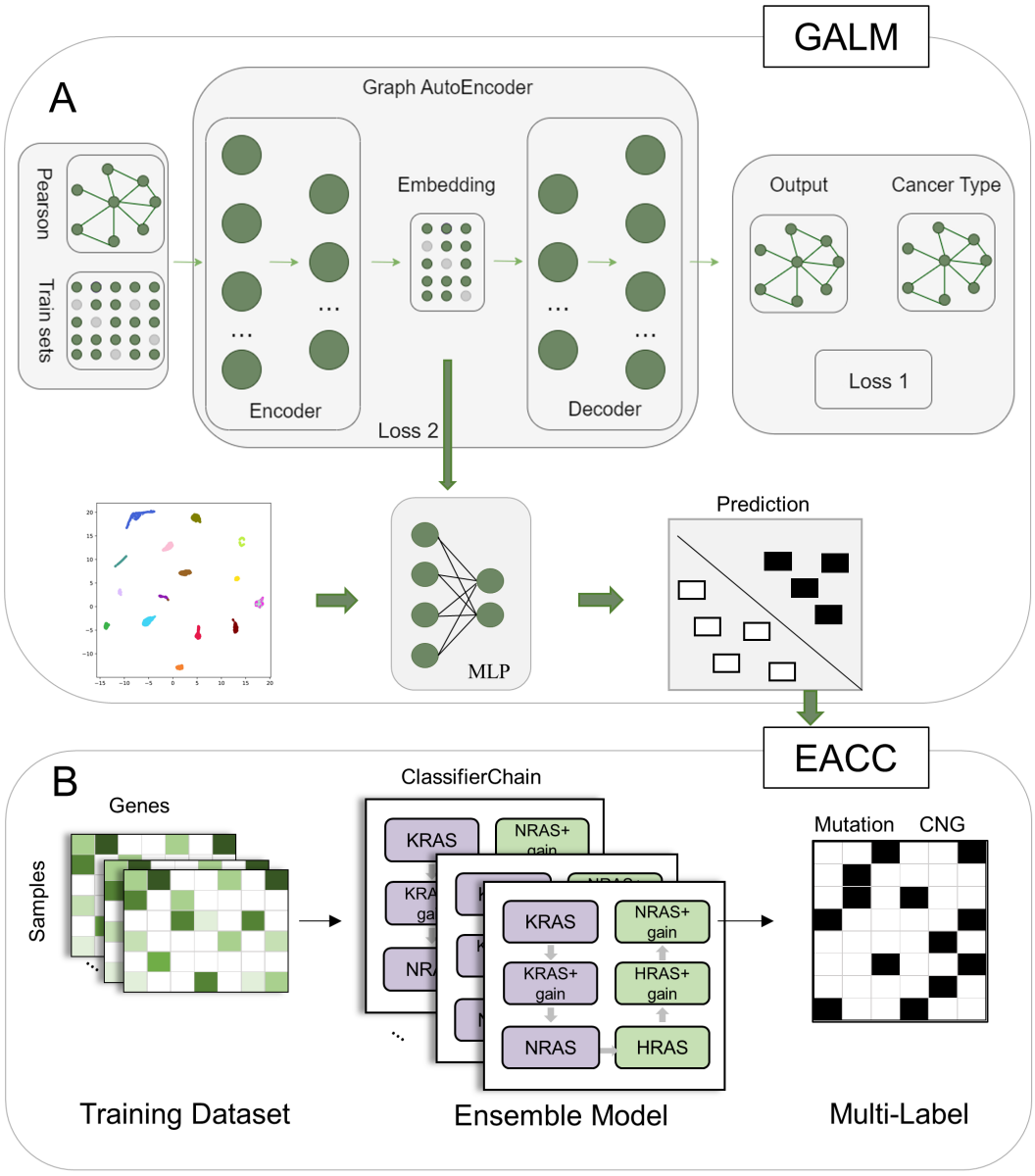

Figure 1. The framework of DIPRAS. A. The overview of GALM model. It is a transformed model of graph autoencoder in which there are two loss functions. Loss1 is used to learn the nodes' relationships such as cancer types and Loss2 is used to classify the variations related to RAS pathway activation. We trained five models and finally used the average strategy to output the results. B. The overview of EACC. It is based on the ensemble algorithm ClassifierChain. Inputs are the 'positive' samples from GALM. Output is the condition of genetic variations (mutations or CNG) in the RAS pathway.

We proposed a DIPRAS framework to further study which variants detected in KRAS, HRAS, and NRAS were related to the RAS activates pathway. Our proposed general framework is based on

graph autoencoder and linear model (GALM) and the ensemble algorithm ClassifierChain [14] (EACC) shown in Figure 1. To consider the relationship between every types of cancer and the similarity between cancers, a graph autoencoder model was proposed to get low-dimensional feature vectors to represent them (Figure 1A). Loss function Loss1 is used to increase cancer-type information while the loss function Loss2 is used to construct a binary classification model to identify the RAS pathway activation. Then the learned low-dimensional feature vectors were used to predict the RAS pathway activation. For the GALM classifier, non-silent somatic mutations and CNGs in the oncogenes KRAS, NRAS and HRAS can identify pathway activation. If detected the activated pathway, further causal investigation is needed to do. Here, EACC (Figure 1B) is used to decipher the relationships between mutations or CNG in KRAS, HRAS, NRAS and the RAS activates pathway, which is a multi-label classification task. The input is the detected samples with mutations or copy number gains in RAS activates pathway from GALM output.

2.2. Dataset construction

There are two datasets in our work. The first dataset contains 16 cancer types with a total of 4759 samples and 20,486 genes. The second dataset is an imbalanced dataset of 33 cancer types with 9074 samples and 20,486 genes. These datasets were obtained from the work by Li et al. [6]. These datasets were randomly split into a training set and a test set in a 9:1 ratio.

For each RAS-activated pan-cancer tumor, the activation state of the pan-cancer RAS pathway can be represented by the vector $y = \{y_1, y_2, \cdots, y_n\}$ and y_i represents the gene state of the i-th sample. The RAS is activated based on whether the gain of function mutations and copy number gains or not. It can be regarded as binary classification and multi-label classification tasks. For the binary classification task, $y_i \in \{0,1\}$, where 1 indicates that the activated RAS pathway is related to mutations or CNGs and 0 is not related. For the multi-label classification task, $y_i \in \{\mathrm{KRAS}_m, \mathrm{HRAS}_m, \mathrm{HRAS}_m, \mathrm{KRAS}_c, \mathrm{HRAS}_c, \mathrm{HRAS}_c\}$, where the KRAS_m, HRAS_m, HRAS_m stands for the mutations of KRAS, HRAS and NRAS, respectively and KRAS_c, HRAS_c, HRAS_c stands for copy number gains of KRAS, HRAS and NRAS, respectively.

2.3. Initial feature selection and graph construction

DIPRAS takes the RNA-seq data as the input. After data filtering and quality control, genes are ranked by the standard deviation to get the top 2000 genes in variances as a vector for study. The 2000 most variably expressed genes were ranked using median absolute deviation (MAD) [5].

To maximize learning about the relationships between cancers and similarities between the cancers, two graphs were constructed. One is a similarity graph created by calculating the Pearson correlation coefficients (PCC) whose edge is defined as PCC greater than 0.8. Another is the cancer-type graph. The cancer type graph was constructed by the relationship of cancer. The edges between samples in one type of cancer were positive edges and the edges between samples in two types of cancer were negative edges.

2.4. Graph autoencoder and linear model

GALM consists of a graph autoencoder with two graph convolution layers and a model with two

linear layers. The core of GALM is the graph autoencoder, by which low-dimensional feature vectors to represent relationships between cancers can be learned. The low-dimensional feature vector (embedding) Z of graph autoencoder is defined as:

$$Z = GCN(X, A) \tag{1}$$

X is the input gene expression matrix and Z represents low-dimensional embedding. The decoder of graph autoencoder is defined as:

$$\hat{A} = \sigma(ZZ^T) \tag{2}$$

$\hat{A}$ is the reconstructed adjacency matrix by embedding Z. σ is the activate function.

A graph convolution network [15] (GCN) is defined as:

$$H^{l+1} = \sigma(\widetilde{D}^{-\frac{1}{2}}\tilde{A}\widetilde{D}^{-\frac{1}{2}}H^{(l)}W^l) \tag{3}$$

$\tilde{A} = A + I$ represent its adjacency matrix add a unit matrix and $\widetilde{D}$ represent its degree matrix. $H^{(l+1)}$ is the output of the $l + 1$ layer. H^0 is the gene expression matrix with the top 2000 genes. W is the parameter for learning.

The objective of training the graph autoencoder is to get relationships between cancer by achieving a maximum similarity between the cancer-type graph and the reconstructed graph. Therefore, the loss function (Loss1) is used to evaluate the maximum similarity between cancer-type graph $\bar{A}$ and reconstructed graph of the model.

$$Loss1 = -\frac{1}{N_{pos}}\sum_{i=1}^{N_{pos}} log(\hat{y}_i + \varepsilon) - \frac{1}{N_{neg}}\sum_{i=1}^{N_{neg}} log(1 - \hat{y}_i + \varepsilon) \tag{4}$$

where N_{pos} denotes the number of edges within one cancer, N_{neg} denotes the number of edges between cancers, $\hat{y}_i$ denotes the predicted probability of the ith sample and ε denotes a very small number used to avoid infinity when taking logarithms.

The second loss uses the binary cross-entropy loss function [16] which is defined as follows:

$$L = -yL_+ - (1 - y)L_- \tag{5}$$

$$p_m = \max(p - m, 0) \tag{6}$$

$$Loss2 = \begin{cases} L_+ = (1 - p)^{\gamma_+}\log(p) \\ L_- = (p_m)^{\gamma_-}\log(1 - p_m) \end{cases} \tag{7}$$

where p denotes the probability that the sample is predicted to be a positive class. p_m represents the low probability of directly discarding the predicted negative samples and m is the hyperparameter of 0.05. γ is the concentration parameter when $\gamma_- > \gamma_+$, in processing easy samples, negative samples are more suppressed and positive samples are less suppressed. In the text γ_- is 10, γ_+ is 2.

The final loss is defined as:

$$Loss = Loss1 + \lambda Loss2 \tag{8}$$

To be able to minimize the loss function process we used the Adam optimizer [17].

2.5. Multi-label classification model EACC

EACC is a multi-label classification model based on the algorithm ClassifierChain for further study of non-silent somatic mutations and high copy gains in the oncogenes KRAS, NRAS and HRAS. ClassifierChain is a multi-label classification method that can model label correlations while maintaining acceptable computational complexity suitable for studying mutations and copy number gains of KRAS, HRAS and NRAS. In the Figure 1B, the model was constructed by specified sequential chains KRAS, KRAS_gain, NRAS, HRAS, HRAS_gain and NRAS_gain (mutations and copy number gains of KRAS, HRAS and NRAS). The training dataset of model EACC is the data of RAS pathway activation. Parameters are default values. The input to EACC is the samples with RAS pathway activation predicted by GALM.

2.6. Performance evaluation parameters

In the research of pan-cancer RAS pathway activation detection, the GALM is a binary classification model in which there are four cases: 1) True positive (TP) represents predicted mutations and CNG, same with labels. 2) False negative (FN) represents predicted neither mutation nor CNG, but labels are mutations and CNG. 3) False positive (FP) represents predicted mutations and CNG, but different with labels. 4) True negative (TN) represents prediction and true labels are neither mutation nor CNG. Two evaluation metrics Acc and Pre are defined as follows:

$$Acc = \frac{TP+TN}{N} \tag{9}$$

$$Pre = \frac{TP}{TP+FP} \tag{10}$$

where N represent the number of samples. Besides, we used the area under the receiver operating characteristics (AUROC) and the area under the precision-recall curve (AUPRC) to evaluate our GALM model.

As another multi-label classification works, the evaluation metrics are accuracy (ACC), Hamming loss (HL), precision (PRE) and F1-Measure (F1):

$$ACC = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i \cap \hat{y}_i}{y_i \cup \hat{y}_i}\right| \tag{11}$$

$$HL = \frac{1}{nL}\sum_{i=1}^{n}\sum_{j=1}^{L} I(y_i^j \neq \hat{y}_i^j) \tag{12}$$

$$PRE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i \cap \hat{y}_i|}{|\hat{y}_i|} \tag{13}$$

$$F1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2|y_i \cap \hat{y}_i|}{|y_i|+|\hat{y}_i|} \tag{14}$$

where y_i is the true label, $\hat{y}_i$ is the predicted label. Overall accuracy (ACC) is the average across all instances. Hamming loss considers the prediction and missing error normalized over the total number of classes and total number of examples.

3. Results and discussion

3.1. RAS benchmarking analyses

To further detect the mutations and gains of copy number, we proposed a multi-label classification task to study the RAS activate pathway based on predicting the variations successfully in the RAS activate pathway. For the first GALM model, we classified pathway activity in tumors like [5] by one of the mutations or CNG that appeared in the RAS to activate the pathway. If the RAS pathways are activated, the specific variations of RAS family mutations or CNS will be further analyzed (Figure 2).

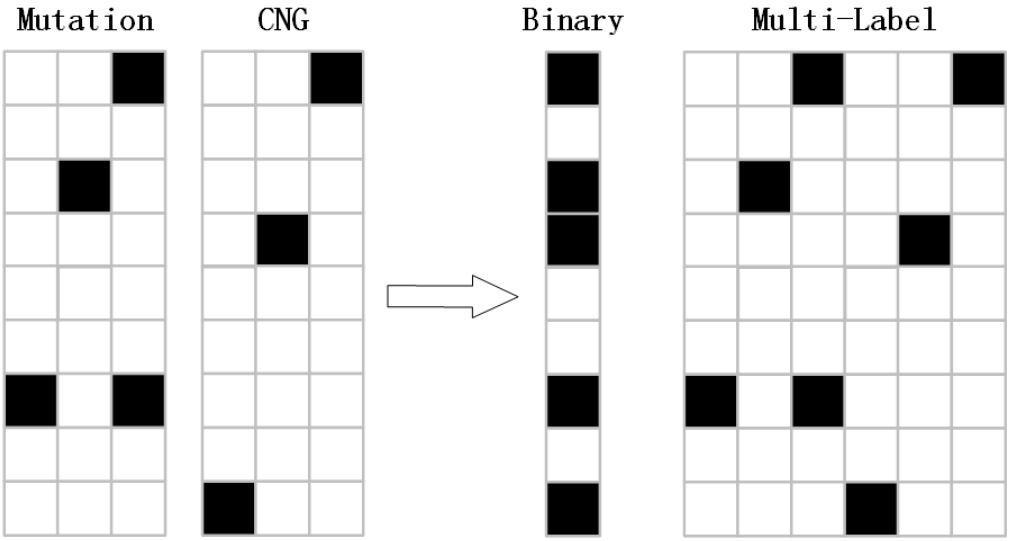

Figure 2. Identification of binary classification labels and multi-label classification labels.

Figure 3A shows the relationships between aberrant variations in the RAS activate pathway with cancer types. It simply shows how many cancer patients of each cancer type have RAS pathways activated. Figure 3B shows the association of patients with activated RAS pathways in all RAS pathways with RAS family mutations and CNG for each cancer type. Only nine RAS families' variations were analyzed in this study including 3 RAS family mutations (KRAS, HRAS, NRAS), 3 RAS family copy number gains (KRAS_gain, HRAS_gain, NRAS_gain), 3 copy number gains and RAS family mutations (KRAS-KRAS_gain, HRAS-HRAS_gain and NRAS-HRAS_gain). As we have seen, the activation of the RAS pathway in patients with cutaneous melanoma (SKCM) is mainly caused by HRAS or RAS-HRas_gain. Therefore, through the above analysis indicating possible differences in the signaling behavior of the mutant proteins that exploit the environment of specific cancer [18], we transformed the binary classification problem into a multi-label classification problem. On the basis of predicting the active state of the RAS pathway, the activation of the RAS pathway caused by RAS family mutations and copy number increases was further considered.

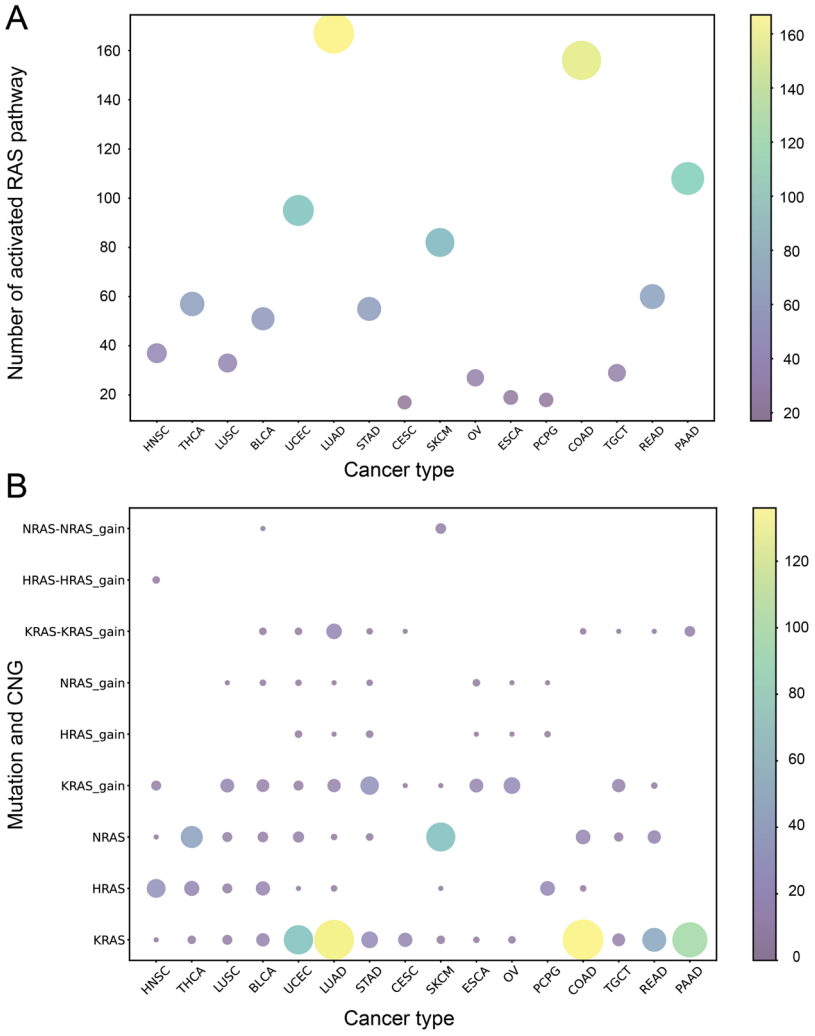

Figure 3. RAS pathway activation and cancers. A. Binary classification task label analyses; B. Multi-label classification task label analyses.

3.2. Parameter setting

To construct a strong and stable model and avoid the extreme best case of randomly selecting data at once we used 5-fold cross-validations on the training set. The GALM contains two GCN layers and one linear layer for graph convolution and normalization operations on the input features. The optimizer algorithm used Adam with a fixed learning rate of $5e^{-03}$. Epoch is set to 500 which means the number of training iterations. Patience is set to 25 which indicates that training will stop if there is no improvement on the validation set for a certain number of consecutive epochs. Batch_size is equal to the number of samples in the training set, indicating the number of samples used for each training iteration.

3.3. Analysis of features

In this work, we used the graph autoencoder method to learn the low-dimensional feature vectors

on the dataset. To verify the effectiveness of GALM and the ability to learn relationships between cancers we interpreted low-dimensional feature vectors by visualizing a UMAP plot [19]. Figure 4A shows the raw 2000 gene expressions which demonstrates that the raw features cannot distinguish the cancer type. Nevertheless, as shown in Figure 4B we can see that the embedding from GALM is clear in cancer types thus proving that GALM was a good choice in processing the data of RNA-seq and learning the relationship information.

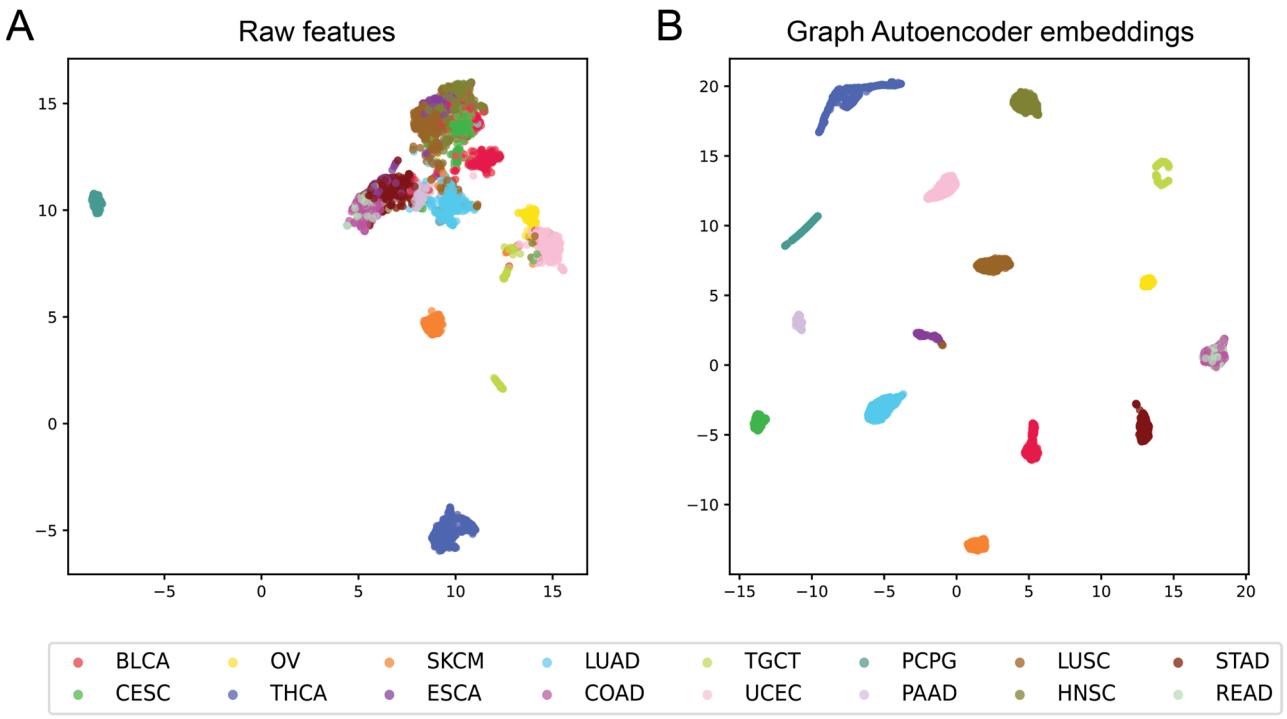

Figure 4. Visualization of raw features and embeddings from GALM. A. UMAP plot of raw features. B. UMAP plot of the embeddings from GALM.

3.4. Performance comparison of binary classification model GALM

To construct a strong and stable model and avoid the extreme best case of randomly selecting data at once we built five models. The average AUROC is 0.83, averaging AUPRC is 0.67, averaging Acc is 0.84 and averaging Pre is 0.62, respectively. We compared GALM with Way G et al. works (here define as LR) according to metrics AUROC and AUPRC which was an averaging result. As shown in Figure 5A,B, GALM is better than LR both in the balanced dataset and the imbalance dataset. The model GALM can be used as a basic model to distinguish whether the RAS pathway activation detected the mutations and the CNG.

To choose the suitable threshold value to construct a similarity graph we compared three GALMs' performances with 0.6, 0.7 and 0.8 as the threshold value. Since there are only 61 edges between samples when the threshold value is 0.9, we do not use 0.9 to construct a similarity graph. As shown in Figure 5C, after comparing the averaging values of the five models threshold 0.8 is better than others no matter on AUROC, AUPRC, Acc and Pre. Therefore, we used 0.8 as the threshold value to build GALM.

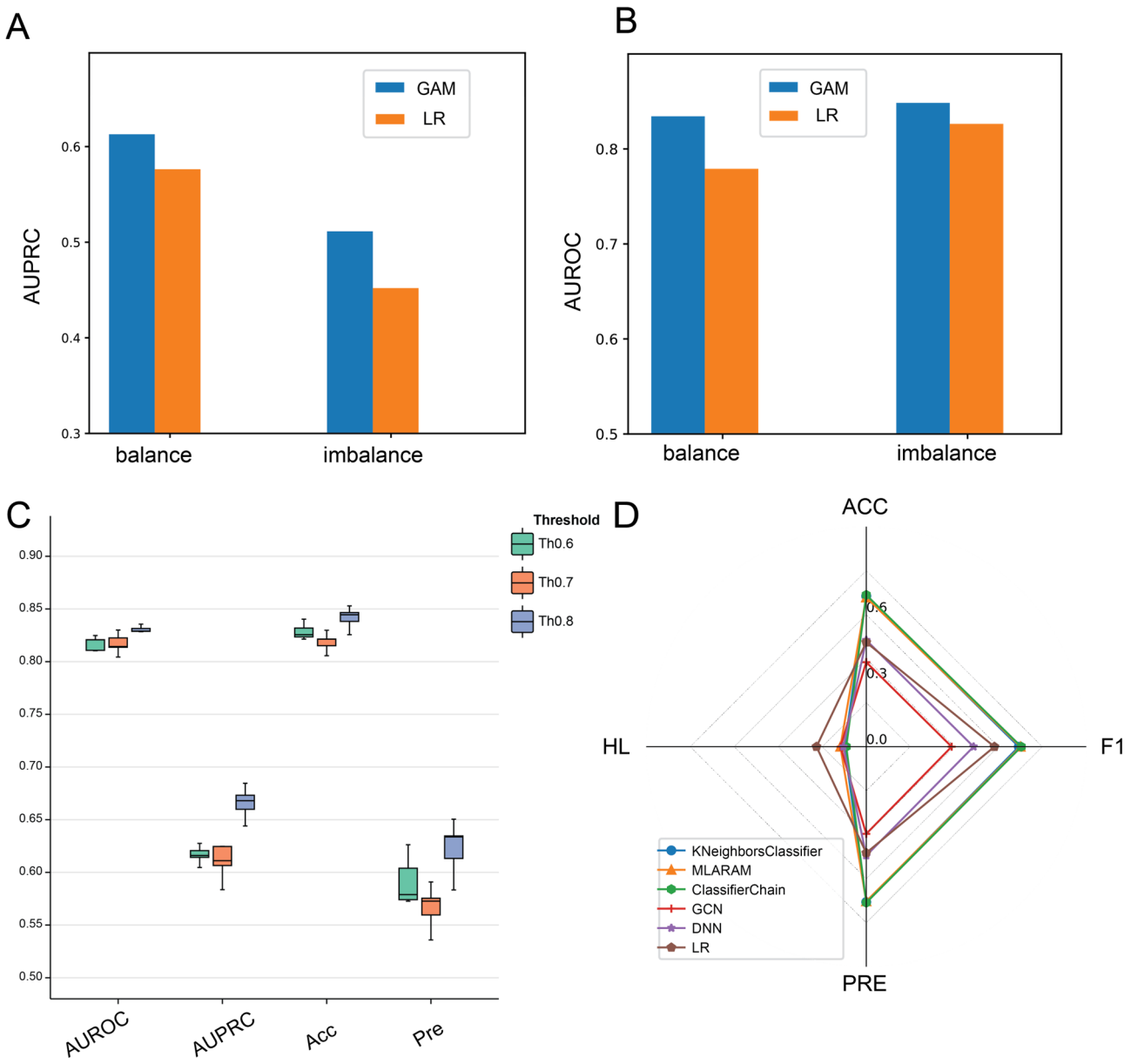

Figure 5. Performance of DIPRAS framework. A & B. The comparison of the performance to detect mutations or CNG in RAS pathway activation between the GALM model and prior tool LR. C. The comparison of threshold values to construct a similarity graph. D. The performance of different algorithms to multi-label classification to decipher whether it is a mutation, a CN or a combination of mutation and CNG in KRAS, HRAS and NRAS related to RAS pathway activation.

3.5. Comparison with different algorithms on the multi-label classification task

To validate the effectiveness of the multi-label classification model RAS pathway activation detection method we compared four multi-label classification algorithms and the existing tool LR. The multi-label classification algorithms are KNeighborsClassifier [20], MLARAM [21], GCN (Graph convolution network) and DNN (Deep neural network). The compared tool LR is from Way et al. [5].

Since LR is a binary classification model, we retrain six base binary classification models according to the labels. As shown in Table 1 and Figure 5D, the ensemble model ClassifierChain achieved the best performance with an ACC of 0.6898, HL of 0.0908, PRE of 0.7089 and F1 of 0.7046, respectively.

Table 1. Comparison with different algorithms on the multi-label classification task.

Methods	ACC	HL	PRE	F1
KNeighborsClassifier	0.6848	0.0987	0.7059	0.6934
MLARAM	0.6785	0.1175	0.7026	0.7013
GCN	0.3842	0.1168	0.396	0.3881
DNN	0.4842	0.104	0.495	0.4878
LR	0.4744	0.2261	0.4827	0.5825
ClassifierChain	**0.6898**	**0.0908**	**0.7089**	**0.7046**

To demonstrate the importance and correlation of features in the multi-label classification model EACC in the detection of pan-cancer RAS pathway activation we used the SHAP method to interpret our ClassifierChain model [22,23]. We chose the mean absolute SHAP value to gain insight into the impact of features and cancer types on pan-cancer RAS pathway activation detection. Figure 6 shows the mean absolute SHAP values of the top 20 features of the two testing samples. The two samples are from cancers colorectal adenocarcinoma (COAD) and pancreatic adenocarcinoma (PAAD) [18]. It was clear that IARS2 had the largest mean absolute SHAP value and had the greatest influence on the pan-cancer RAS pathway activation detection. Besides, the KRAS mutation showed a high proportion of feature importance in two PAAD and COAD samples. This conclusion was verified in Figure 3B, cancer PAAD and COAD showed a higher frequency of carrying the KRAS mutations. Therefore, it is necessary to further decipher the relationship between RAS pathway activation and tumor-specific in hopes of gaining insight into their complexity and the clinical consequences that drive them.

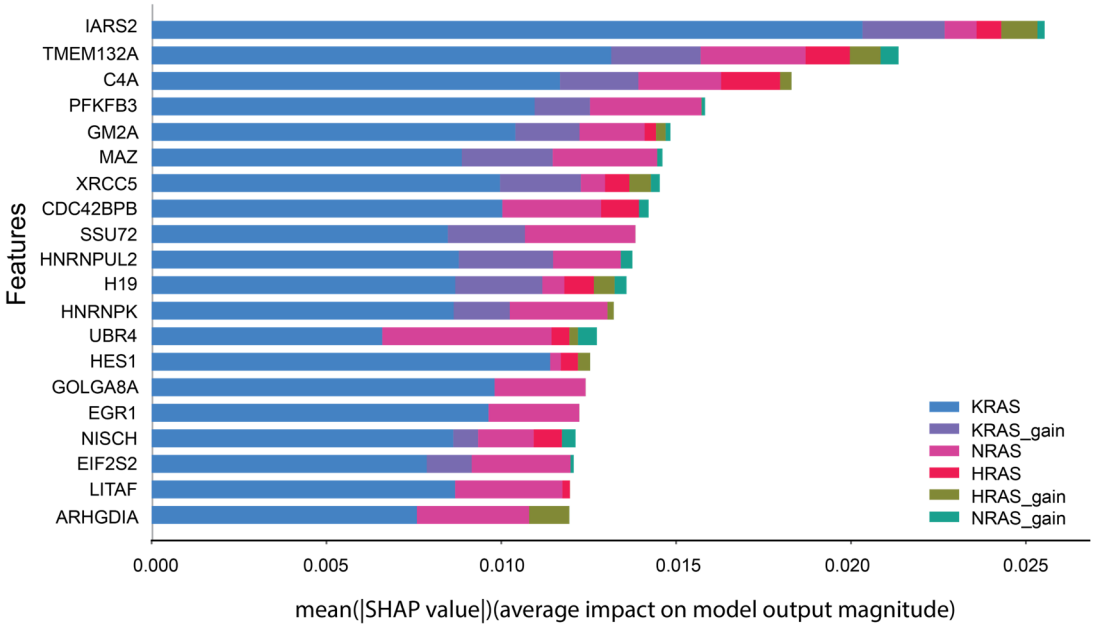

Figure 6. Histogram of the mean absolute SHAP values of ensemble model ClassifierChain in multi-label classification task.

3.6. GO and KEGG enrichment analysis

Based on the measurement of the feature importance, we ranked the importance of features using SHAP values in this section. We selected the top 200 genes and converted them into the Ensembl ID [24] and then we performed biological functional by GO enrichment analysis and KEGG [25] pathway enrichment analysis.

GO enrichment analysis of pan-oncogenic RAS pathway activation detection was performed on 200 genes and p-value < 0.05 was used as the filtering condition. The roles of these genes at biological process, cellular component and molecular function levels were derived. Due to the large number of results from GO enrichment analysis, only the top 20 ranked and significant ones were shown in Figure 7 for visualization.

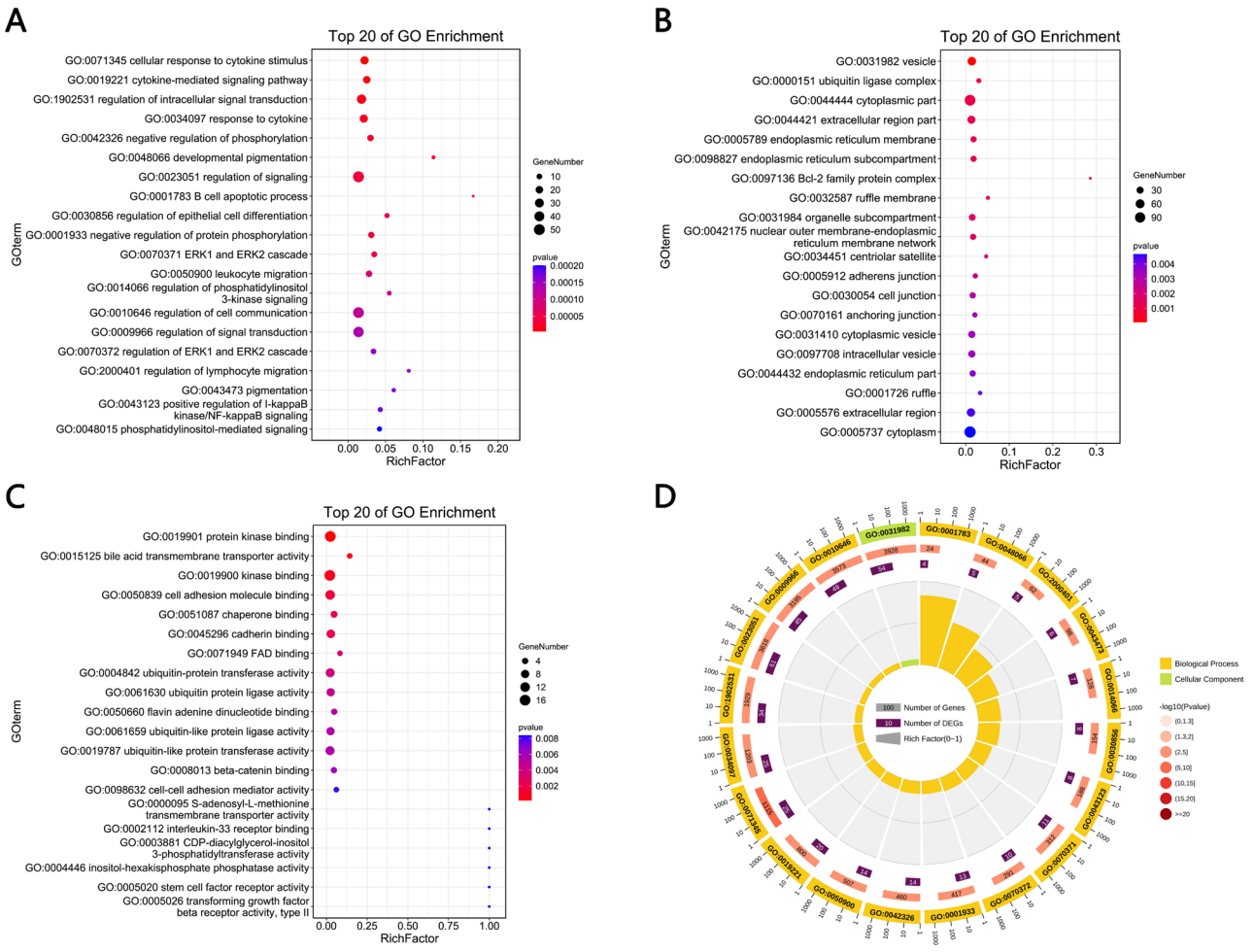

Figure 7. Gene GO enrichment analysis of pan-cancer RAS pathway activation detection. A. Bubble chart of biological processes. B. Bubble chart of cellular components. C. Bubble chart of molecular function. D. Enrichment circle chart.

From the bubble chart of biological processes in Figure 7, it can be seen that RAS pathway activation was closely related to biological processes such as phosphatidylinositol 3-kinase signaling [26,27], intracellular signal transduction [28] and epithelial cell differentiation. As seen in

the bubble chart of cellular components in Figure 7B, RAS pathway activation was involved in constituting cellular components such as vesicles, cytoplasm, endoplasmic reticulum and Bcl-2 family protein complexes. Figure 7C shows the molecular function of these genes and it showed that molecular functions such as protein kinase binding, kinase binding, cell adhesion molecule binding and cadherin binding were involved. The p-value ranking of the three functional categories can be seen in the enrichment circle chart in Figure 7D. The top 20 biological processes accounted for the most which may indicate the higher enrichment degree of biological processes.

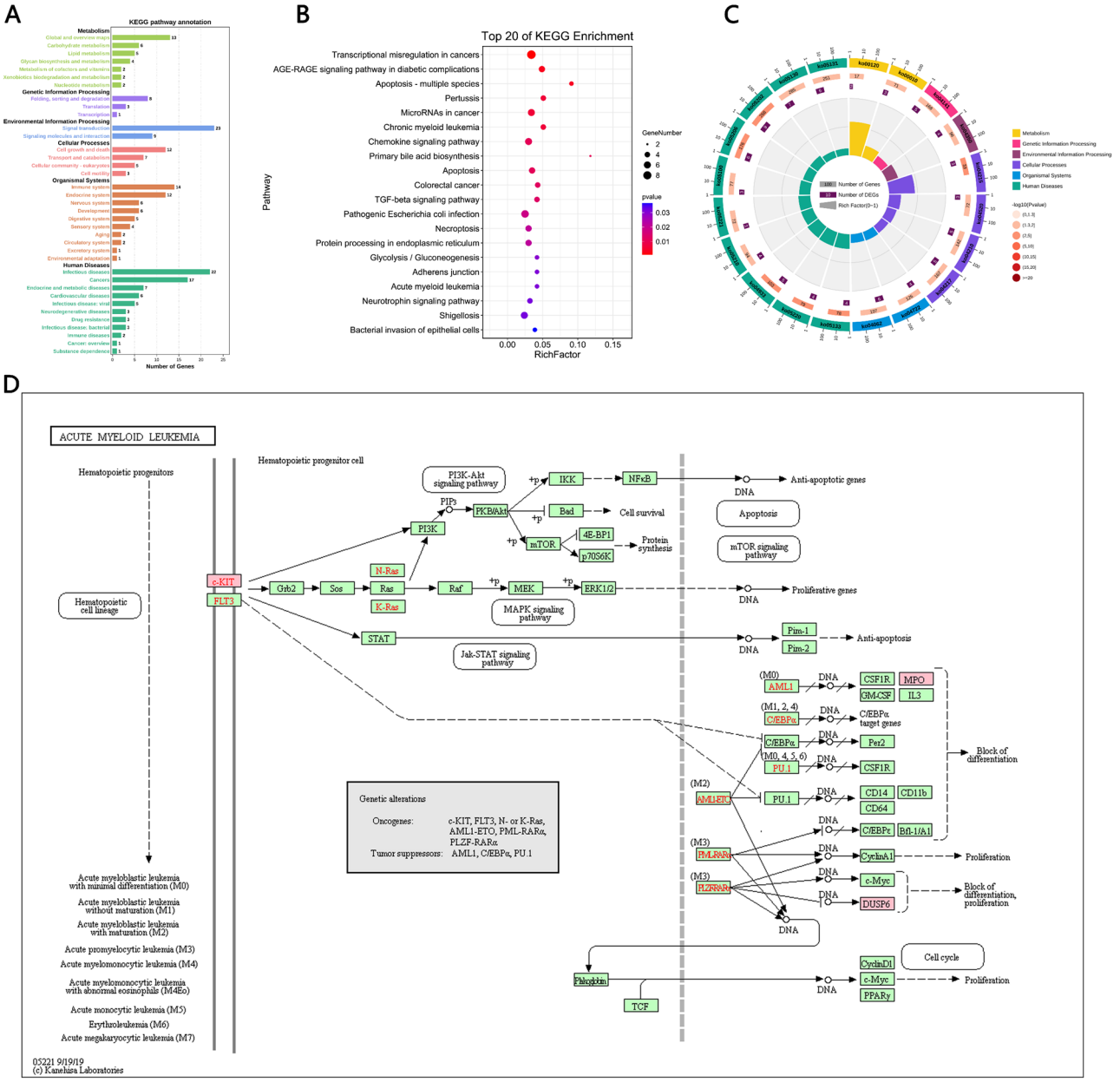

Figure 8. Gene KEGG enrichment analysis results of pan-cancer RAS pathway activation detection. A. Analysis chart of KEGG signal pathway. B. Bubble chart of KEGG signaling pathway. C. Enrichment circle chart of KEGG signaling pathway. D. Acute myeloid leukemia (hsa05221) related gene signaling pathway chart.

KEGG signaling pathway enrichment analysis of pan-oncogenic RAS pathway activation detection was performed on 200 genes. We found that genes were enriched to a variety of pathways, with the most enriched pathways mainly including signal transduction, infectious diseases, cancer and the immune system. The results were shown in Figure 8A. 20 KEGG signaling pathways were screened for significance by p-value < 0.05. The results showed that chronic myeloid leukemia, acute myeloid leukemia, MicroRNAs in cancer, apoptosis etc. are closely related to the RAS signaling pathway. Figure 8B,C showed the bubble and circle chart of the KEGG signaling pathway enrichment analysis. Moreover, RAS gene mutations often occur in patients with acute myeloid leukemia and RAS gene mutations may directly affect the treatment results of leukemia patients [29–31]. In addition, we used these 200 genes for molecular pathway analysis of KEGG (https://www.kegg.jp/). We found a total of 209 results from the KEGG Mapper tool. Among them, acute myeloid leukemia (hsa05221) contained 3 related genes as shown in Figure 8D.

4. Conclusions

Mutation in the RAS gene is an important cause that promotes the development of many types of cancer and often appears early in tumorigenesis. In this study, we proposed a deep learning framework DIPRAS which uses a varied graph autoencoder model detecting the activity of the pan-cancer RAS pathway from TCGA and use ClassifierChain to further detect specific mutations or CNG. The experiment results indicated that our DIPRAS could learn relationships between cancer and provide further analysis of RAS pathway activation-related mutations or CNG.

In the future, the results of our functional enrichment analysis will be obtained and we will jointly study clinical trials related to the RAS pathway with doctors to further verify that our RAS abnormal pathway activity detection can identify more patients with RAS activation providing a good idea for deep integration application in bioinformatics and medicine. It would be interesting to identify more RAS-activated patients by selecting genes for targeted RNA sequencing based on the results of our functional enrichment analysis. On the other hand, if we can apply the idea proposed to other mutant pathways it will be meaningful work.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work is supported by Natural Science Funds of Jilin Province (Grant No. 20200201158JC), the science and technology research project of "13th Five-Year" of the Education Department of Jilin Province (Grant No. JJKH20211297KJ), the National Natural Science Funds of China (No. 61802057), and the Science and Technology Development Plan of Jilin province. (Grant No. 20180414006GH).

GO enrichment analysis was performed using the OmicShare tools, a free online platform for data analysis (http://www.omicshare.com/tools).

Conflict of interest

The authors declare there is no conflict of interest.

References

1. S. H. Shin, A. M. Bode, Z. Dong, Addressing the challenges of applying precision oncology, *NPJ Precis. Oncol.*, **1** (2017), 28. http://doi.org/10.1038/s41698-017-0032-z
2. D. M. Hyman, B. S. Taylor, J. Baselga, Implementing genome-driven oncology, *Cell*, **168** (2017), 584–599. http://doi.org/10.1016/j.cell.2016.12.015
3. I. F. Tannock, J. A. Hickman, Limits to personalized cancer medicine, *N. Engl. J. Med.*, **375** (2016), 1289–1294. http://doi.org/10.1056/NEJMsb1607705
4. C. Kumar-Sinha, A. M. Chinnaiyan, Precision oncology in the age of integrative genomics, *Nat. Biotechnol.*, **36** (2018), 46–60. http://doi.org/10.1038/nbt.4017
5. G. P. Way, F. Sanchez-Vega, K. La, J. Armenia, W. K. Chatila, A. Luna, et al., Machine learning detects pan-cancer RAS pathway activation in the cancer genome atlas, *Cell Rep*, **23** (2018), 172–180. http://doi.org/10.1016/j.celrep.2018.03.046
6. X. Li, S. Li, Y. Wang, S. Zhang, K. C. Wong, Identification of pan-cancer RAS pathway activation with deep learning, *Briefings Bioinf.*, **22** (2021), bbaa258. http://doi.org/10.1093/bib/bbaa258
7. J. Zhang, Y. Zhang, Z. Ma, In silico prediction of human secretory proteins in plasma based on discrete firefly optimization and application to cancer biomarkers identification, *Front. Genet.*, **10** (2019), 542. http://doi.org/10.3389/fgene.2019.00542
8. R. Scharpf, G. Riely, M. Awad, M. Lenoue-Newton, B. Ricciuti, J. Rudolph, et al., Comprehensive pan-cancer analyses of RAS genomic diversity, *Cancer Res.*, **80** (2020), 1095. http://doi.org/10.1158/1538-7445.Am2020-1095
9. A. K. Murugan, M. Grieco, N. Tsuchida, RAS mutations in human cancers: Roles in precision medicine, in *Seminars in Cancer Biology*, Academic Press, (2019), 23–35. https://doi.org/10.1016/j.semcancer.2019.06.007
10. I. A. Prior, P. D. Lewis, C. Mattos, A comprehensive survey of RAS mutations in cancer, *Cancer Res.*, **72** (2012), 2457–2467. http://doi.org/10.1158/0008-5472.can-11-2612
11. W. Pao, T. Y. Wang, G. J. Riely, V. A. Miller, Q. Pan, M. Ladanyi, et al., KRAS mutations and primary resistance of lung adenocarcinomas to gefitinib or erlotinib, *PLoS Med.*, **2** (2005), e17. http://doi.org/10.1371/journal.pmed.0020017
12. Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, P. S. Yu, A comprehensive survey on graph neural networks, *IEEE Trans. Neural Networks Learn. Syst.*, **32** (2021), 4–24. http://doi.org/10.1109/TNNLS.2020.2978386
13. R. Hasibi, T. Michoel, A graph feature auto-encoder for the prediction of unobserved node features on biological networks, *BMC Bioinf.*, **22** (2021), 525. http://doi.org/10.1186/s12859-021-04447-3
14. J. Read, B. Pfahringer, G. Holmes, E. Frank, Classifier chains for multi-label classification, *Mach. Learn.*, **85** (2011), 333–359. http://doi.org/10.1007/s10994-011-5256-5
15. T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, *arXiv preprint*, (2016), arXiv:1609.02907. https://doi.org/10.48550/arXiv.1609.02907
16. P. T. De Boer, D. P. Kroese, S. Mannor, R. Y. Rubinstein, A tutorial on the cross-entropy method, *Ann. Oper. Res.*, **134** (2005), 19–67. http://doi.org/10.1007/s10479-005-5724-z

17. D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, *arXiv preprint*, (2014), arXiv:1412.6980. https://doi.org/10.48550/arXiv.1412.6980
18. J. H. Cook, G. E. Melloni, D. C. Gulhan, P. J. Park, K. M. Haigis, The origins and genetic interactions of KRAS mutations are allele-and tissue-specific, *Nat. Commun.*, **12** (2021), 1808. http://doi.org/10.1038/s41467-021-22125-z
19. E. Becht, L. Mcinnes, J. Healy, C. A. Dutertre, I. W. H. Kwok, L. G. Ng, et al., Dimensionality reduction for visualizing single-cell data using UMAP, *Nat. Biotechnol.*, **37** (2019), 38–44. http://doi.org/10.1038/nbt.4314
20. V. J. Pandya, Comparing handwritten character recognition by AdaBoostClassifier and KNeighborsClassifier, in *2016 8th International Conference on Computational Intelligence and Communication Networks*, IEEE, (2016), 271–274. http://doi.org/10.1109/Cicn.2016.59
21. Z. A. Huang, J. Zhang, Z. Zhu, E. Q. Wu, K. C. Tan, Identification of autistic risk candidate genes and toxic chemicals via multilabel learning, *IEEE Trans. Neural Networks Learn. Syst.*, **32** (2020), 3971–3984. http://doi.org/ 10.1109/TNNLS.2020.3016357
22. S. M. Lundberg, S. I. Lee, A unified approach to interpreting model predictions, in *Advances in Neural Information Processing Systems 30*, (2017).
23. A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, in *International Conference on Machine Learning*, PMLR, (2017), 3145–3153.
24. F. J. Martin, M. R. Amode, A. Aneja, O. Austine-Orimoloye, A. G. Azov, I. Barnes, et al., Ensembl 2023, *Nucleic Acids Res.*, **51** (2023), D933–D941. http://doi.org/10.1093/nar/gkac958
25. H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, M. Kanehisa, KEGG: Kyoto encyclopedia of genes and genomes, *Nucleic Acids Res.*, **27** (1999), 29–34. http://doi.org/10.1093/nar/27.1.29
26. T. Kodaki, R. Woscholski, B. Hallberg, P. Rodriguez-Viciana, J. Downward, P. J. Parker, The activation of phosphatidylinositol 3-kinase by RAS, *Curr. Biol.*, **4** (1994), 798–806. http://doi.org/10.1016/s0960-9822(00)00177-9
27. P. M. Campbell, A. L. Groehler, K. M. Lee, M. M. Ouellette, V. Khazak, C. J. Der, K-RAS promotes growth transformation and invasion of immortalized human pancreatic cells by Raf and phosphatidylinositol 3-kinase signaling, *Cancer Res.*, **67** (2007), 2098–2106. http://doi.org/10.1158/0008-5472.CAN-06-3752
28. M. Zenker, Clinical manifestations of mutations in RAS and related intracellular signal transduction factors, *Curr. Opin. Pediatr.*, **23** (2011), 443–451. http://doi.org/10.1097/MOP.0b013e32834881dd
29. J. L. Bos, M. Verlaan-De Vries, A. J. Van Der Eb, J. W. Janssen, R. Delwel, B. Lowenberg, et al., Mutations in N-RAS predominate in acute myeloid leukemia, *Blood*, **69** (1987), 1237–1241. https://doi.org/10.1182/blood.V69.4.1237.1237
30. J. D. Khoury, M. Tashakori, H. Yang, S. Loghavi, Y. Wang, J. Wang, et al., Pan-RAF inhibition shows anti-leukemic activity in RAS-mutant acute myeloid leukemia cells and potentiates the effect of sorafenib in cells with FLT3 mutation, *Cancers*, **12** (2020), 3511. http://doi.org/ARTN 351110.3390/cancers12123511

31. A. M. Akram, A. Chaudhary, H. Kausar, F. Althobaiti, A. S. Abbas, Z. Hussain, et al., Analysis of RAS gene mutations in cytogenetically normal de novo acute myeloid leukemia patients reveals some novel alterations, *Saudi J. Biol. Sci.*, **28** (2021), 3735–3740. http://doi.org/10.1016/j.sjbs.2021.04.089

AIMS Press