ISBN 978-981-14-1684-2
Proceedings of 2019 the 9th International Workshop on Computer Science and Engineering
(WCSE 2019 SUMMER)
Hong Kong, 15-17 June, 2019, pp. 643-651
doi: 10.18178/wcse.2019.06.095

Research on Colorectal Cancer Prediction and Survival Analysis with Data Fusion Based on Deep Learning

Shiqi Li [1], Jun Zheng [1 +] and Shuxun Wei [2]

[1] School of Computer Science and Software Engineering
East China Normal University, China
[2] Department of General Surgery of Changzheng Hospital
the Second Military Medical University, China

Abstract. Colorectal cancer is a highly aggressive type of cancer. Accurate prognosis prediction of colorectal cancer can help patients and doctors choose the best treatment and avoid unnecessary costs. Nevertheless, most of previous work relied mostly on selected gene expression data to create a predictive model by traditional machine learning methods to reduce dimensions and predict cancer. Such methods cannot fully consider the correlation between samples to effectively extract feature information, and the cost of calculation is huge. In this paper, we propose a multi-modal neural network which integrates GCN and DNN to train multi-modal data for the prediction of colorectal cancer and its prognosis. The novelty of the method lies in the design of our method's architecture and the fusion of multi-dimensional data, which can give full play to the performance of the neural network. The comprehensive performance evaluation results show that the proposed method achieves a better performance than the widely used prediction methods with single dimensional data and other existing approaches.

Keywords: Colorectal cancer, Cancer prognosis prediction, Graph convolution network, Multimodal deep neural network, Multi-modal data.

1. Introduction

Colorectal cancer (CRC), defined as cancer starting in the colon or the rectum, is among the most common types of cancer for both men and women. It is the third most common cancer and ranks fourth as a cause of cancer-related death globally [1]. Because of the concealment of colon cancer and atypical symptoms, it is difficult to diagnose accurately and the rate of misdiagnosis is high. Therefore, with the help of the widely used method of deep learning, a multimodal deep neural network is constructed to analyze multi-modal data, which can be used to predict the prognosis of colorectal cancer. Our method adopts and improves GCN and DNN structure. The main contributions of this paper can be summarized as follows:

(1) In order to avoid the loss of feature information caused by traditional feature selection methods, such as PCA, t-SNE, we take full account of the correlation between samples for feature extraction.

(2) Considering that the relationship between omics data is not isolated, we fused multi-modal data for the first time and used back-end fusion technology to predict the prognosis of colorectal cancer more accurately. This avoids the error caused by using traditional machine learning methods for single dimension data.

We conduct experiments based on widely-used TCGA datasets. The experimental results consistently demonstrate the superiorities and competitiveness of our proposed model.

+ Corresponding author. Tel.: + 86-18121085702.
E-mail address: jzheng@cc.ecnu.edu.cn.

The remainder of this paper is organized as follows. We first introduce related work in Section 2. In Section 3, we describe the details of our proposed model. In Section 4, we introduce the experimental settings and results. Finally, the conclusions and future work are presented in Section 5.

2. Related Work

During past few decades, there are a large number of efforts to effectively understand molecular mechanisms of cancer by employing gene expression profile data [2][3]. However, most studies are designed to work with a handful of gene expression profile data and under the assumption of independence between different genes in hypothesis testing. In fact, more than tens of thousands of gene expression profiles are recorded in TCGA and different genes may have potential associations between each other.

While it is difficult for a doctor to consider many factors, machine learning algorithms can efficiently find patterns in large datasets of patient data. Delen et al. used data mining methods to predict breast cancer survivability in 2004 [4]. Xu et al. propose an efficient feature selection method based on support vector machine (SVM) for breast cancer prognosis prediction and achieve a superior prediction performance than that of the widely used 70 gene signatures [5]. Nguyen et al. introduce a method for breast cancer prognosis prediction based on random forest (RF) combined with feature selection and achieve the highest classification accuracy than previous methods [6]. Khademi et al. propose a probabilistic graphical model (PGM) by integrating two independent models of microarray and clinical data for prognosis and diagnosis of breast cancer [7]. They first apply Principal Component Analysis (PCA) to reduce the dimensionality of microarray data and construct a deep belief network to extract feature representation of the data. Given the complexity and heterogeneity of diagnosis and survival prediction, a more practical strategy is needed to use clinical data and genetic predictive markers that may contain some additional information. Zhang et al. propose a multiple kernel machine learning method by fusing different types of data for the GBM prognosis prediction [8].

However, most of these methods directly combine different types of data, and ignore that the features from different modalities (e.g. gene signature and clinical) may have different representations. Accordingly, based on the accelerated development of multiple omics data, there are urgent needs to develop efficient computational methods for accurately predicting prognosis of human colorectal cancer. Meanwhile, no matter what kind of data combination method, effective feature selection method is needed. There are many methods of gene selection using machine learning. Ding et al. [9] uses MRMR to select genes that are critical for the classification of diseases such as leukemia and colon cancer. Wang et al. [10] selected genes by filter, CFS and wrapper, and used these genes to construct classifiers to obtain better classification performance on leukemia data sets. Diaz et al. [11] proposed a gene selection method based on random forest.

In view of the above two points, in this study, we proposed a novel multimodal neural network by integrating GCN (Graph Convolutional Networks) model and DNN (Deep Neural Networks) model to utilize multi-dimensional data better for colorectal cancer prediction and survival analysis. This way is an efficient method to integrate multidimensional data including gene expression profile, DNA methylation, copy number alteration (CNA) profile and clinical data with a score level fusion at the final prediction results. This method considers the heterogeneity among different data types and makes full use of abstract high-level representation from each data source. In this method, GCN is applied to multidimensional data to predict whether an individual has cancer or not. In this process, feature extraction is performed on each modal data to effectively capture the features affecting cancer. These features will be used in DNN training by back-end fusion to analyze the survival time of human cancer. This method avoids the problem of explosive growth of parameters and local optimum when only using DNN, and also gives full play to the performance of DNN and GCN.

3. Our Methodology

3.1. Graph convolutional networks

There are many Non-Euclidean structures data in the real world, such as molecular structure, protein composition, recommender system and so on. In order to obtain the relationship between samples better and

extract effective features, it is a good choice to adopt a method like convolution neural network similar to image. However, the traditional discrete convolution can't maintain translation invariance on the data of Non-Euclidean structure. It is commonly understood that the number of adjacent vertices of each vertex in a topological graph may be different, so of course, it is impossible to use a convolution kernel of the same size for convolution operations.

The key of GCN is spectral decomposition based on Laplacian matrix. The spectral graph convolution operation of a sample x $\in \mathbb{R}^N$ on g_θ convolution kernel is defined as [12]:

$$g_\theta^* x = U g_\theta U^T x \tag{1}$$

U is the eigenvalue matrix of the Laplacian matrix of a graph. The Laplacian matrix of a graph is L = $I_N - D^{-1/2} A D^{-1/2}$. A is an adjacency matrix, $D_{ii} = \sum_j A_{ij}$.

In (1), x can obtain the information of neighbour nodes, because the Laplacian matrix L contains the global information of the graph. These theories constitute the framework of most graph convolution networks. Based on this, Kipf et al. [13] proposed a high-efficient graph convolution architecture using a first-order approximation of Chebyshev polynomials and obtained propagation rules:

$$H^{(l+1)} = \sigma[\hat{D}^{-1/2} \hat{A} \hat{D}^{-1/2} H^{(l)} W^{(l)}] \tag{3}$$

$\hat{A} = A + I$, I is a unit matrix. l is the number of layers. D that is diagonal, is the node degree matrix of adjacency matrix A. σ *is* ReLU activation function. $H^{(l+1)}$ means output to next layer or result. Where $H^{(l)} \in R^{n \times d^{(l)}}$ represent node-level expression of the graph in the l^{th} layer. N is the number of nodes. $d^{(l)}$ is the dimension expressed by the nodes of l^{th} layer. $W^{(l)}$ is the weights of l^{th} layer.

In our model, nodes $x_i = (x_{i1}, x_{i2}, \cdots, x_{in})$ represent samples. $(x_{i1}, x_{i2}, \cdots, x_{in})$ are the features set of x_i, in other words, it's the genome of sample x_i. Edges represent relationships between nodes. In this paper, the adjacency matrix is obtained by Minkowski distance:

$$d_{ij} = \sqrt[p]{\sum_{n=1}^{n} |x_{in} - x_{jn}|^p} \quad \Rightarrow \quad A_{ij} = \frac{1}{(x_i - x_j)^T (x_i - x_j) + 1} \tag{4}$$

p is a variable parameter, here set it to 2. The result is normalized to (0,1) for the convenience of calculation. The larger the A_{ij}, the greater the similarity between nodes i and j. Then we can easily get the node degree of matrix D:

$$D = \begin{pmatrix} D_1 & \cdots & \cdots \\ \vdots & \ddots & \vdots \\ \cdots & \cdots & D_n \end{pmatrix}, \quad D_i = \sum_{j=1}^{n} A_{ij} \tag{7}$$

After building the data graph, we feed the graph into a simple two layers GCN as in Kipf et al. [13]. Stochasticity in the training process is introduced via dropout. The second layer node has the same size as the labels set and are fed into a softmax classifier:

$$\mathrm{Z} = \mathrm{softmax}(\tilde{A}\mathrm{ReLU}(\tilde{A}(\mathrm{X}W_0)W_1)W_2) \tag{8}$$

where $\tilde{A} = D^{-1/2} A D^{-1/2}$ is the normalized symmetric adjacency matrix, and $\mathrm{softmax}(x_i) = \frac{1}{Z}\exp(x_i)$ with $\mathrm{z} = \sum_i \exp(x_i)$.The weight parameters W_0 , $W_1$and W_2 can be trained via gradient descent. In order to get the weight of each gene to achieve the goal of feature extraction, we added a hidden layer which uses backpropagation to update and get the weight of the gene. The loss function is defined as the cross-entropy error over all labeled data:

$$\mathcal{L} = -\sum_{d \epsilon Y_D} \sum_{f=1}^{F} Y_{df} ln Z_{df} \tag{9}$$

where Y_D is the set of data indices that have labels and F is the dimension of the output features, which is equal to the number of classes. Y is the label indicator matrix.

After sorting the weight of each gene obtained in GCN classification process, some gene features are selected according to the demand, and these gene data are input into the multi-mode deep neural network. In this way, we reduce the dimensions of datasets such as gene expression profile data.

3.2. Deep neural network prediction model

In this section, a deep neural network (DNN) is used for predicting the prognosis of human colorectal cancer. The DNN architectures build a hierarchy from the hidden layers. Higher level features are extracted implicitly by the combination of lower level features from each layer. Here, a DNN model is composed of an input layer, multiple hidden layers and an output layer. Units between layers are all fully connected. The input layer with an input vector x consists of one or multi-dimensional data. The output h_j^k for layer k including j units is calculated from the weighted sum of the outputs for the previous layer h^{k-1} (specially $h^0 = x$).

$$g^k = W^k h^{k-1} + b^k, 1 \le k \le N \tag{10}$$

$$h^k = f(g^k) = \sigma(g^k) \tag{11}$$

where W^k is the k^{th} weight matrix between $(k-1)^{th}$ layer and k^{th} layer. b^k is the bias vector for the k^{th} layer. N is the number of layers (here $N = 5$, including output layer) and ReLU activation function σ is used to hidden units, which naturally captures the non-linear relations within the data. The result of the last layer are fed into a softmax classifier.

Afterwards, in order to avoid the dead ReLU problem, we initialize the weights between each layer using MSRA [14] initialization method. The dead ReLU problem is a situation which the parameters because of the large learning rate are updated too much during the training process, which makes the network enter a state where some neurons may never be activated. We can also use tanh activation function and Xavier [15] initialization method to optimize the network. The distribution comparison of the two initialization methods is shown in Table 1.

Table 1: Comparison of the two initialization methods

Types/ Activation	**Distribution**	**Expression**
MSRA/ ReLU	Gaussian	$\mathrm{W} \sim \mathrm{G}\left[0, \sqrt{\frac{2}{N}}\right]$
Xavier/ tanh	Uniform	$\mathrm{W} \sim \mathrm{U}\left[-\frac{\sqrt{6}}{\sqrt{n_i + n_{i+1}}}, \frac{\sqrt{6}}{\sqrt{n_i + n_{i+1}}}\right]$

Given the fact that colorectal cancer prognosis prediction task in our study is a binary classification task (long-term survival and short-term survival), we define cross entropy loss as objective function for the DNN model in final output layer. In addition, to further prevent overfitting of our deep learning model, L2 regularization is also added into our loss function. Finally, our proposed DNN method aims at minimizing the loss function and is defined as:

$$\mathcal{L}(y_t, \hat{y}_t) = -\frac{1}{N}\sum_{i=0}^{N}[y_t(i) log \hat{y}_t(i) - (1 - y_t(i))\log(1 - \hat{y}_t(i))] + \frac{1}{2}\lambda \sum_{k=1}^{K}\sum_{j=1}^{n_k}\sum_{i=1}^{m_k} {w_{ij}^k}^2 \tag{9}$$

where $\mathcal{L}$ measures errors between predictive scores and the actual labels. $y_t(i)$ is the actual label for the i^{th} class, $\hat{y}_t(i)$ is the predictive scores obtained from the output layer of our method. N is the batch size. $w^k = \{w_{ij}^k\}_{m_k \times n_k}$ is the k^{th} weight matrix and K is the number of weight matrix in DNN model (here K=5).

A common issue in training a DNN model is named “internal-covariate-shift”, which is that input distributions change in each layer during training due to the update of parameters from previous layers. In 2015, a novel work called batch normalization [16] is proposed by Google to solve the aforementioned problem, which allows us to use higher learning rates and be less careful about weights initialization. As expected, the batch normalization is very significant to optimize our DNN model and obtains a good result. Finally, a DNN model employed in our work comprises one input layer, four hidden layers and an output layer.

3.3. Back-end fusion of multi-dimensional data

An important issue in our study is integrating multidimensional data including gene expression profile, CNA profile, DNA methylation and clinical data. One of the most straightforward approaches for

discriminative tasks is to train only one DNN model for all multi-dimensional data. However, different data may have different feature representation, and directly combining the three sources of data as an input of a DNN model may not be efficient [7]. We address this problem by proposing a multimodal DNN model which efficiently integrates multi-dimensional data. Figure 1 illustrates the structure of integration of multi-mode DNN and GCN method. Our proposed method conducts a score level fusion from each independent model. The final output of our model based on a weighted linear aggregation is calculated as:

$$R_{F-DNN} = aR_{gene-DNN} + bR_{CNA-DNN} + cR_{DNA-DNN} + dR_{clin-DNN}$$
$$\text{s.t.} \quad a + b + c + d = 1, a \geq 0, b \geq 0, c \geq 0, d \geq 0 \tag{13}$$

where the parameters a, b, c, d are four weight coefficients used to balance the contribution for each DNN model. In this study, our model chooses the optimal parameters for the parameters of different sub-DNN models, a, b, c, d according to the best prediction performance by using validation set (see Experimental Set). We screen different combinations of a, b, c, d by a step 0.1 and finally select a = 0.2, b = 0.1, c = 0.4 and d =0.3 for our dataset.

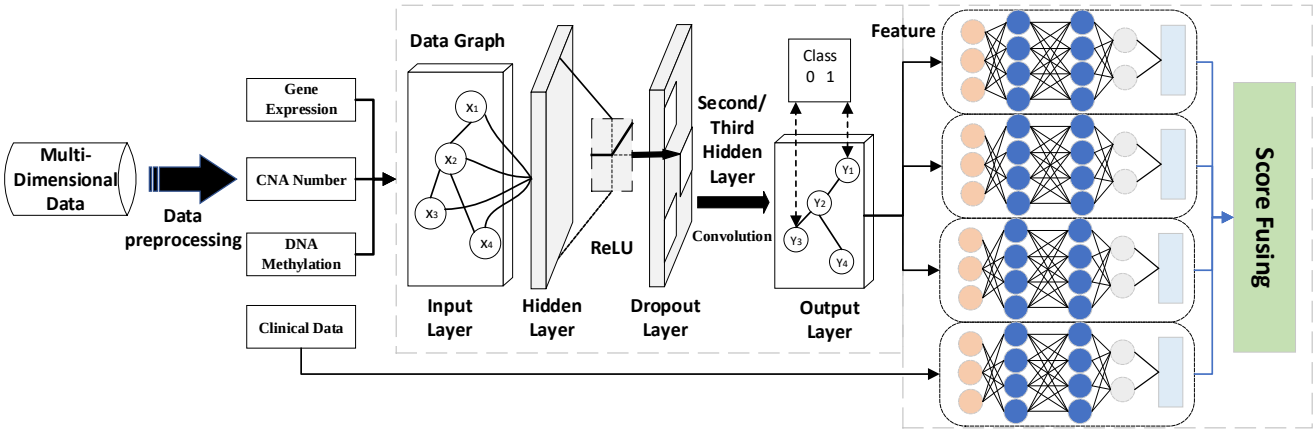

Fig. 1: The overall process of our model for the colorectal cancer prognosis prediction.

4. Experiment

Firstly, we preprocess multi-dimensional data of colorectal cancer, which includes four sub-data: gene expression, CNA, DNA methylation and clinical data. Secondly, we build data graph and use GCN model to select feature that can reduce the number of variables for gene expression, DNA methylation and CNA data. Thirdly, a multi-modal DNN is proposed to extract effectively information from different data, respectively. Therefore, we greedily train each DNN model corresponding to each sub-data.

4.1. Material and experimental settings

We use the TCGA dataset which is available at https://portal.gdc.cancer.gov/. We downloaded the latest high-quality dataset and obtained 460 valid patients' data for colorectal cancer including 68 long-term (greater than 5 years survival) and 392 short-term (less than 5 year's survival) survivors. The dataset contains 41 normal samples (They are labeled 1, patients are labeled 0) and multi-dimensional data among colorectal cancer such as gene expression profile, CNA profile, DNA methylation and clinical information. For classification labels in our work, the short-term patients are labeled as 0 and long-term patients are labeled as 1. For gene expression profile data and CNA profile data, missing values are estimated using a weighted nearest neighbors algorithm [17]. According to Z-score standardization and discretization, the gene expression features are normalized and further discretized into three categories: under-expression (-1), over-expression (1) and baseline (0). For copy number variation, all features are normalized in (-1,1). For clinical data, all features are normalized into the range [0, 1] by minmax normalization [2].

For GCN model, we randomly select 10% of training set as validation set for hyperparameter optimization (dropout rate for all layers, L2 regularization factor for the first GCN layer and number of hidden units). We do not use the validation set labels for training. We used two hidden layer size of 64 and 32 units and set the learning rate as 0.02, dropout rate as 0.5. Following Kipf et al. [13], we trained GCN for a maximum of 200 epochs using Adam [18] and stop training if the validation loss does not decrease for 10

consecutive epochs. We initialized weights using the Xavier [15] method and accordingly row normalize input feature vectors.

For DNN model, to comprehensively evaluate our proposed method, we used ten-fold cross validation experiment in consistent with previous existing studies of cancer prognosis prediction [8]. We use grid-search strategy to adjust optimal hyperparameter by using the AUC value as the criteria. Besides, A batch normalization is added to each hidden layer and a dropout is added before the output layer. Finally, we set 4 hidden layers with 1000, 500, 500 and 100 units. The size of minibatch and initial learning rate are set to 64 and 0.001. We initialized weights using the Xavier [15] method.

To verify the benefit of multi-modal DNN by integrating multi-dimensional data, DNN based methods with single-dimensional data are examined for the prognosis prediction of colorectal cancer in this study. The main difference between these methods and our model is that they do not integrate multi-dimensional data and only use the data type in one type. For simplicity, the DNN based methods that use single-dimensional data of gene expression profile data, clinical data, DNA methylation data and CNA profile data are thereafter termed as gene-DNN, clin-DNN, DNA-DNN, and CNA-DNN, respectively. In order to show the effectiveness of multimodal deep learning method in prognosis prediction of colorectal cancer, we employ three widely used methods as classifiers，including support vector machines (SVM), random forest (RF) and logistic regression (LR) for comparison. In these algorithms, multi-dimensional data that we used are regarded as feature vector to train the model. The performance is also evaluated by ten-fold cross validation process.

The evaluation metric, besides of AUC value, Sensitivity (Sn), Specificity (Sp), Accuracy (Acc), Precision (Pre) and Matthew's correlation coefficient (Mcc) are also used for performance evaluation.

4.2. Result

In order to confirm the effectiveness of multi-dimensional data, we first adopt deep learning method on different single data type to predict colorectal cancer prognosis. We compare the performance of F-DNN(our method) with gene-DNN, clin-DNN, DNA-DNN and CNA-DNN.

Table 2: Comparison of Acc, Pre, Sn and Mcc between multi-DNN, SVM, RF and LR

method	**AUC**	**Acc**	**Pre**	**Sn**	**Mcc**	**Acc**	**Pre**	**Sn**	**Mcc**
		Sp = 99.0% (threshold: 0.587)				Sp = 95.0% (threshold: 0.441)			
CNA-DNN	0.606	0.705	0.464	0.025	0.061	0.683	0.408	0.103	0.094
Gene-DNN	0.732	0.731	0.642	0.053	0.138	0.719	0.543	0.167	0.201
DNA-DNN	0.748	0.759	0.720	0.079	0.192	0.757	0.629	0.282	0.362
Clin-DNN	0.794	0.773	0.818	0.137	0.274	0.784	0.681	0.324	0.388
SVM	0.814	0.778	0.811	0.122	0.257	0.805	0.708	0.365	0.407
RF	0.801	0.769	0.778	0.097	0.223	0.791	**0.766**	0.226	0.337
LR	0.663	0.754	0.563	0.036	0.093	0.760	0.551	0.179	0.203
mRMR-DNN	0.835	0.794	0.875	0.191	0.343	0.822	0.749	**0.450**	0.486
(GCN)F-DNN	**0.851**	**0.806**	**0.882**	**0.204**	**0.366**	**0.834**	0.752	**0.450**	**0.489**

The ROC curves are plotted for five different methods at each specificity level and displayed in Figure 2. As shown in Figure 2, F-DNN achieves better overall performance (AUC =0.851) than those of the single-dimensional data methods. The AUC value (showed in Fig. 2) of F-DNN(0.851) is 5.7%, 10.3%, 11.9% and 24.5% higher than those of clin-DNN, DNA-DNN, gene-DNN and CAN-DNN, respectively As the same time, by following the study of Fan et al. [19], two stringency levels of medium (Sp = 95.0% with corresponding threshold of 0.441) and high (Sp = 99.0% with corresponding threshold of 0.587) specificity are applied to each method for measuring the predictive performance. Besides of, we compare the performance of F-DNN with three widely used methods for prognosis prediction of colorectal cancer: SVM, RF and LR. The comparison of AUC, Sn, Acc, Pre, and Mcc with different methods at the two stringency levels is listed in Table 2. In order to test the performance of GCN, we use a new method called max-

relevance and min redundancy (mRMR) [20] in feature extraction. The mRMR is one of the most common dimensionality reduction algorithm in a wide range of applications. It will be compared with GCN while ensuring that the DNN model is unchanged. It can be seen from table 2, our proposed method achieves a better performance than those other prediction methods in most of evaluations.

To further demonstrate the predictive results of the multi-dimensional data in assessing the risk of developing distant metastases in colorectal cancer patients, survival data analyses of the proposed method is also performed according to previous studies [8] [21], the Kaplan-Meier curve is plotted and shown in Figure 3, for the aforementioned datasets. It suggests that there is a significant difference between the patients with short term survival time and the patients with long term survival time predicted by our predictive results (p-value = 1e-5).

In addition, after the first feature selection, the genes affecting cancer are obtained. After the second feature selection, the genes affecting survival are found. We compared some genes which are selected by our model, and we found that, such as genes ACVR1B, SMAD2, APC, GATA1 and TNF are highly expressed in each disease sample. But they cannot effectively distinguish the types of survival. However, such as gene BIRC5, the expression level of gene can distinguish the types of survival to some extent. After comparing the NCBI (at https://www.ncbi.nlm.nih.gov/) database, it was found that this gene was a regulatory gene significantly associated with most tumors and cancers. Figure 4 shows result of classification by tool of UCSC-xena (at https://xenabrowser.net/)

All above comparison results indicate that the value of our work, and confirm the tremendous benefits from integrating multidimensional data, multimodal deep neural network and multimodal fusion in the prognosis prediction of colorectal cancer. And our model plays an important and far-reaching role in the discovery and analysis of genes.

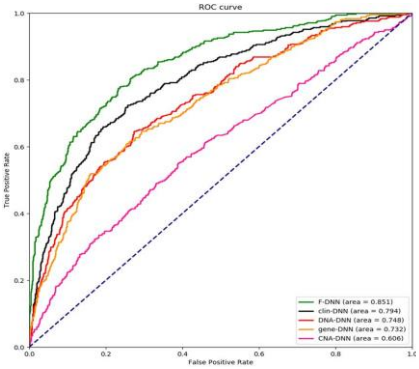

Fig. 2: Performance comparison in different ways

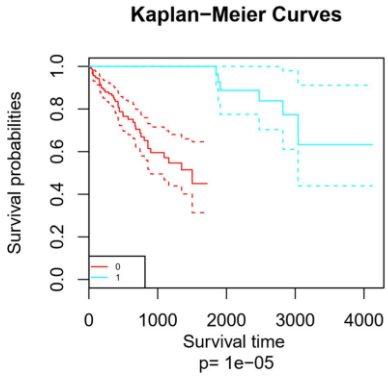

Fig. 3: KM-curve of colorectal cancer prognosis prediction.

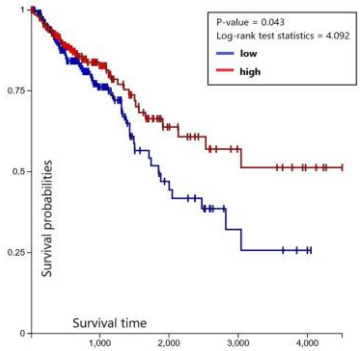

Fig. 4: the result of classification of Gene BIRC5

5. Conclusions and future work

In this work, we present a novel multimodal deep neural network by integrating multi-dimensional data named GCN-DNN to predict whether a sample has cancer and the survival time of human colorectal cancer. GCN not only considers the correlation between samples, but also utilizes unlabeled data (manual setting). Through convolution method similar to image, the nodes (samples) in the graph continuously obtain information from adjacent nodes. This approach maximizes the comprehensiveness of the information obtained. In this way, feature genes which will be input to DNN affecting cancer in samples can be selected. Then, to efficiently incorporate multidimensional data including gene expression profile, CNA, DNA methylation and clinical data in colorectal cancer, four independent DNN models are constructed to generate a final multimodal DNN model considering the heterogeneity of different types of data. Finally, a decision level multimodal fusion (score fusion) is used to integrate data of different types to predict colorectal cancer prognosis. On TCGA dataset, we compared our model with the widely used methods, and the results show that our model can achieve good performance. Due to the successful application of the multimodal deep learning method in our work, we believe that combining different data types is an efficient way to improve performance of human colorectal cancer prognosis prediction.

Despite the success application of our method, it still has some problem and point for improvement. Firstly, for clinical medicine it is difficult and expensive to obtain a large amount of complete clinical data. And it is unusable for researches where multiple omics data are unavailable or incomplete. Secondly, the dataset that we used are relatively small and may limit further analysis. It is expected that the performance of the proposed method would be enhanced when more samples become available in future. We also think that it will be more meaningful for cancer researchers if our model is built for each subtype of cancer which will make it to a multi-category task, and its performances may be further improved. And we also will try to integrate more omics data, such as image, miRNA expression.

6. References

[1] Ferlay, Jacques, et al. "Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012." *International journal of cancer* 136.5 (2015): E359-E386.

[2] Van De Vijver, Marc J., et al. "A gene-expression signature as a predictor of survival in breast cancer." *New England Journal of Medicine* 347.25 (2002): 1999-2009.

[3] Sun, Yijun, et al. "Improved breast cancer prognosis through the combination of clinical and genetic markers." *Bioinformatics* 23.1 (2006): 30-37.

[4] Delen, Dursun, Glenn Walker, and Amit Kadam. "Predicting breast cancer survivability: a comparison of three data mining methods." *Artificial intelligence in medicine* 34.2 (2005): 113-127.

[5] Xu, Xiaoyi, et al. "A gene signature for breast cancer prognosis using support vector machine." *2012 5th International Conference on BioMedical Engineering and Informatics*. IEEE, 2012.

[6] Nguyen, Cuong, Yong Wang, and Ha Nam Nguyen. "Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic." *Journal of Biomedical Science and Engineering* 6.05 (2013): 551.

[7] Khademi, Mahmoud, and Nedialko S. Nedialkov. "Probabilistic graphical models and deep belief networks for prognosis of breast cancer." *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2015.

[8] Zhang, Ya, et al. "Improve glioblastoma multiforme prognosis prediction by using feature selection and multiple kernel learning." *IEEE/ACM transactions on computational biology and bioinformatics* 13.5 (2016): 825-835.

[9] Ding, Chris, and Hanchuan Peng. "Minimum redundancy feature selection from microarray gene expression data." *Journal of bioinformatics and computational biology* 3.02 (2005): 185-205.

[10] Wang, Yu, et al. "Gene selection from microarray data for cancer classification—a machine learning approach." *Computational biology and chemistry* 29.1 (2005): 37-46.

[11] Díaz-Uriarte, Ramón, and Sara Alvarez De Andres. "Gene selection and classification of microarray data using random forest." *BMC bioinformatics* 7.1 (2006): 3.

[12] Hammond, David K., Pierre Vandergheynst, and Rémi Gribonval. "Wavelets on graphs via spectral graph theory." *Applied and Computational Harmonic Analysis* 30.2 (2011): 129-150.

[13] Kipf, Thomas N., and Max Welling. "Semi-supervised classification with graph convolutional networks." *arXiv preprint arXiv:1609.02907* (2016).

[14] He, Kaiming, et al. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification." *Proceedings of the IEEE international conference on computer vision*. 2015.

[15] Glorot, Xavier, and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks." *Proceedings of the thirteenth international conference on artificial intelligence and statistics.* 2010.

[16] Ioffe, Sergey, and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." *arXiv preprint arXiv:1502.03167* (2015).

[17] Troyanskaya, Olga, et al. "Missing value estimation methods for DNA microarrays." *Bioinformatics* 17.6 (2001): 520-525

[18] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980* (2014).

[19] Fan, Wenwen, et al. "Prediction of protein kinase-specific phosphorylation sites in hierarchical structure using

functional information and random forest." *Amino acids* 46.4 (2014): 1069-1078.

[20] Peng, Hanchuan, Fuhui Long, and Chris Ding. "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy." *IEEE Transactions on Pattern Analysis & Machine Intelligence 8* (2005): 1226-1238.

[21] Ranstam, J., and J. Cook. "Kaplan-Meier curve." *British Journal of Surgery 104.4* (2017).