A Cancer Survival Prediction Method Based on Graph Convolutional Network

Chunyu Wang, Junling Guo, Ning Zhao, Yang Liu, Xiaoyan Liu, Guojun Liu, and Maozu Guo

***Abstract*— *Background and objective:* Cancer, as the most challenging part in the human disease history, has always been one of the main threats to human life and health. The high mortality of cancer is largely due to the complexity of cancer and the significant differences in clinical outcomes. Therefore, it will be significant to improve accuracy of cancer survival prediction, which has become one of the main fields of cancer research. Many calculation models for cancer survival prediction have been proposed at present, but most of them generate prediction models only by using single genomic data or clinical data. Multiple genomic data and clinical data have not been integrated yet to take a comprehensive consideration of cancers and predict their survival. *Method:* In order to effectively integrate multiple genomic data (including genetic expression, copy number alteration, DNA methylation and exon expression) and clinical data and apply them to predictive studies on cancer survival, similar network fusion algorithm (SNF) was proposed in this paper to integrate multiple genomic data and clinical data so as to generate sample similarity matrix, min-redundancy and max-relevance algorithm (mRMR) was used to conduct feature selection of multiple genomic data and clinical data of cancer samples and generate sample feature matrix, and finally two matrixes were used for semi-supervised training through graph convolutional network (GCN) so as to obtain a cancer survival prediction method integrating multiple genomic data and clinical data based on graph convolutional network (GCGCN). *Result:* Performance indexes of GCGCN model indicate that both multiple genomic data and clinical data play significant roles in the accurate survival time prediction of cancer patients. It is compared with existing survival prediction methods, and results show that cancer survival prediction method GCGCN which integrates multiple genomic data and clinical data has obviously superior prediction effect than existing survival prediction methods. *Conclusion:* All study results in this paper have verified effectiveness and superiority of GCGCN in the aspect of cancer survival prediction.**

***Index Terms*— Graph convolution network, multiple genomic data, clinical data, similarity network fusion, minimum redundancy maximum relevance feature selection, cancer survival prediction.**

I. Introduction

AS MORBIDITY and mortality rates gradually rise, cancer is becoming the main death cause in the globe and one of important public health problems. According to the global cancer report, additional 12.7 million cancer cases occurred in the globe with death toll reaching 7.6 million in 2008; the two numbers reached 14.10 million and 8.20 million respectively in 2012. Morbidity and mortality rates of lung cancer are the highest and those of gastric cancer, oesophageal cancer and hepatic cancer grow with each passing day. Meanwhile, cancers have become common among young people in recent years. Cancer morbidity and mortality rates in the whole world have been rising. Cancer has become the main death cause since 2010 and it becomes one of the main public health problems in the whole world [1]. Therefore, it's urgent to design a more accurate calculation method for cancer survival prediction, which can contribute to development of individual-based treatment and management. Hence, this will also be conductive to lowering total death rate of cancers and further improving living quality of cancer patients.

The heterogenous disease, cancer, has different molecular features, clinical behaviors, morphological appearances and different reactions to therapies [2]–[5]. In addition, complexity of invasive cancers and their clinical outcomes presenting significant changes result in extreme difficulties in prediction and treatment [6], [7]. Therefore, a more accurate prediction of cancer prognosis can not only help cancer patients to understand their expected life and can also help clinicians to make wise decisions and give proper therapeutic guidance. Meanwhile, prognosis plays a significant role in clinical work of all clinicians, especially those clinicians working with patients with short survival. Being able to estimate prognosis reasonably and accurately, clinicians generally make clinical decisions with the help of prognostic prediction knowledge [8], confirm that patients accept therapeutic schemes [9] and design and analyze qualifications for clinical test. In addition, when a patient is predicted as short-survival patient, the clinician can provide him/her with an opportunity to consider whether he/she wants to be cared, take time to take actual measures and make preparations for the death [10].

Manuscript received April 10, 2019; revised July 10, 2019; accepted August 13, 2019. Date of publication August 21, 2019; date of current version December 31, 2019. This work was supported in part by the Natural Science Foundation of China under Grant 91735306, Grant 61872114, Grant 61671189, Grant 61671188, Grant 61571163, and Grant 61871020 and in part by the National Key Research and Development Plan Task of China under Grant 2016YFC0901902. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Funding was provided by C. Wang., Y. Liu, X. Liu, and M. Guo; ideas were initiated by Y. Liu and M. Guo; the manuscript was drafted by C. Wang and J. Guo; the manuscript was reviewed and edited by Y. Liu and X. Liu; the experiments were implemented by J. Guo, N. Zhao, and G. Liu; the paper was finalized by C. Wang; and all authors have read and approved the final manuscript. The authors declare no conflict of interest. *(Corresponding author: Maozu Guo.)*

C. Wang, J. Guo, N. Zhao, Y. Liu, X. Liu, and G. Liu are with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China.

M. Guo is with the School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing 100044, China, and also with the Beijing Key Laboratory of Intelligent Processing for Building Big Data, Beijing University of Civil Engineering and Architecture, Beijing 100044, China (e-mail: guomaozu@bucea.edu.cn).

Digital Object Identifier 10.1109/TNB.2019.2936398

In order to realize the goal, many researches have adopted microarray technique to study genetic expression profiling of cancers in the past years, but only some of them have displayed definite prognostic significance [11], [12]. For instance, Van's Veer et al., used DNA microarray to analyze 117 primary breast cancer patients and used supervised classification method to identify prognostic features of 70 genes [13]. Moreover, they tested these previously applicable prognostic markers among 295 breast cancer patients, and results indicated that prognostic features of 70 genes had great significances [14]. Wang *et al.* [15] revealed prognostic features of 76 genes and clustered genetic expression profiles and associated them with prognostic values, which could accurately predict tumor recurrence in the later phase. Microarray markers used, some machine learning classification methods such as support vector machine (SVM) [9], [16], [17], Naive Bayesian Classifier (NB) [18] and random forest (RF) [19] are also used to predict cancer survival. For instance, Brenton *et al.* [20] proposed a breast cancer diagnosis and prediction method based on random forest classifier and feature selection technique, and the result was superior to previously reported results.

In view of complexity and heterogeneity of cancer survival prediction, Brenton *et al.* [20] put forward a more pragmatic strategy and used clinical data and genetic prediction markers which might contain some supplementary information. In addition, with rapid development of new technologies in the medical field, a large number of clinical cancer data have been generated and collected. Clinical data and microarray markers combined, different computing methods have been developed to accurately predict cancer survival [8], [21]. For example, Gevaert et al., developed Bayesian network [14], integrated clinical data and information of 70 genes through three different strategies (including complete, decision-making or partial integration) and proved that combination of clinical data and microarray data had better or considerable performance than single use of clinical data or microarray data. Khademi et al., [22] put forward an interesting strategy, reduced dimensionalities of microarray data using manifold learning and deep belief network and integrated clinical data through the probabilistic graph model. Through a large quantity of experiments, compared with traditional classification methods, this method has more excellent effects. Besides microarray and clinical information, human reference protein interaction network has also been used to predict survival of breast cancer. Das et al., [23] designed a method based on elastic network and named it Encapp, and then combined protein network and genetic expression dataset in order to accurately predict survival time of human breast cancer. However, limitation of Encapp lies in that accuracy of survival prediction highly depends on quality of genetic expression datasets. As cancer is a very complicated disease, combination of genetic expression profiling data and clinical data may improve accuracy of prognosis and diagnosis by the prediction model [22]. Based on 70 genetic expression features, Sun et al., further identified a hybrid feature through prediction of genetic features of breast cancer prognosis and combination of clinical markers [8]. Therefore, it is still of considerable space to improve prediction performance of cancer survival by combining more cancer-related information.

Besides successes achieved through the abovementioned methods, some new methods are proposed, Multi-dimensional data are integrated and applied to human cancer-related prediction fields. Hayes et al., determined related microRNA and mRNA features of patients with high and low risks of glioblastoma (GBM) [24]. Zhang et al., put forward a multi-core machine learning method integrating different types of data for prognostic prediction of GBM [25]. In consideration of multiple different kinds of data features and when these features are extensively applied to prognostic prediction of cancers, it's not surprising that favorable effect is achieved. However, most of these methods directly integrate different types of data into model generation while neglecting that features from different patterns (like genomic signature and clinic) may have different representations. With the latest progress of the new generation of sequencing technique, multi-omics cancer diagnosis and prognosis based on genetic expression profile, clinical information and DNA copy number alteration have enjoyed broad development [26]–[28]. Therefore, based on accelerated development of multi-omics data, it's urgent to develop an effective computing method to accurate predict cancer prognosis.

In order to solve these problems and enlightened by successful application of the present deep learning methods and great contributions made by multi-dimensional data to cancer prognosis prediction, a cancer survival prediction method integrating multiple genomic data (including gene expression, copy number alteration, DNA methylation and exon expression) and clinical data based on graph convolutional network namely GCGCN was proposed in this study. The method considered heterogeneity between different data types while taking full advantages of abstract high-level representations of different data sources. In order to verify effectiveness of multiple genomic data and clinical data, GCGCN was compared with different independent models which only used multiple genomic data or clinical data. Results indicated that both multiple genomic data and clinical data could improve the prediction performance of cancer survival, and this meant that both multiple genomic data and clinical data reflected cancer survival from different aspects. In addition, the proposed GCGCN method was compared with present popular classification methods. In the aspect of cancer survival prediction, GCGCN achieved the optimal effect in multiple evaluation indexes, thus proving the feasibility of integrating multiple genomic data and clinical data as well as the significance of GCGCN to cancer survival prediction.

II. Materials and Methods

Fig. 1 shows experimental framework of the proposed GCGCN, which is divided into three steps: (1) generating sample similarity matrix; (2) generating sample feature matrix; (3) obtaining cancer survival classifier through training. Specifically speaking, the whole process is divided into three steps. First of all, similarity network fusion algorithm (SNF) was used to integrate multiple genomic data and clinical data to obtain a sample similarity matrix A, and then min-redundancy

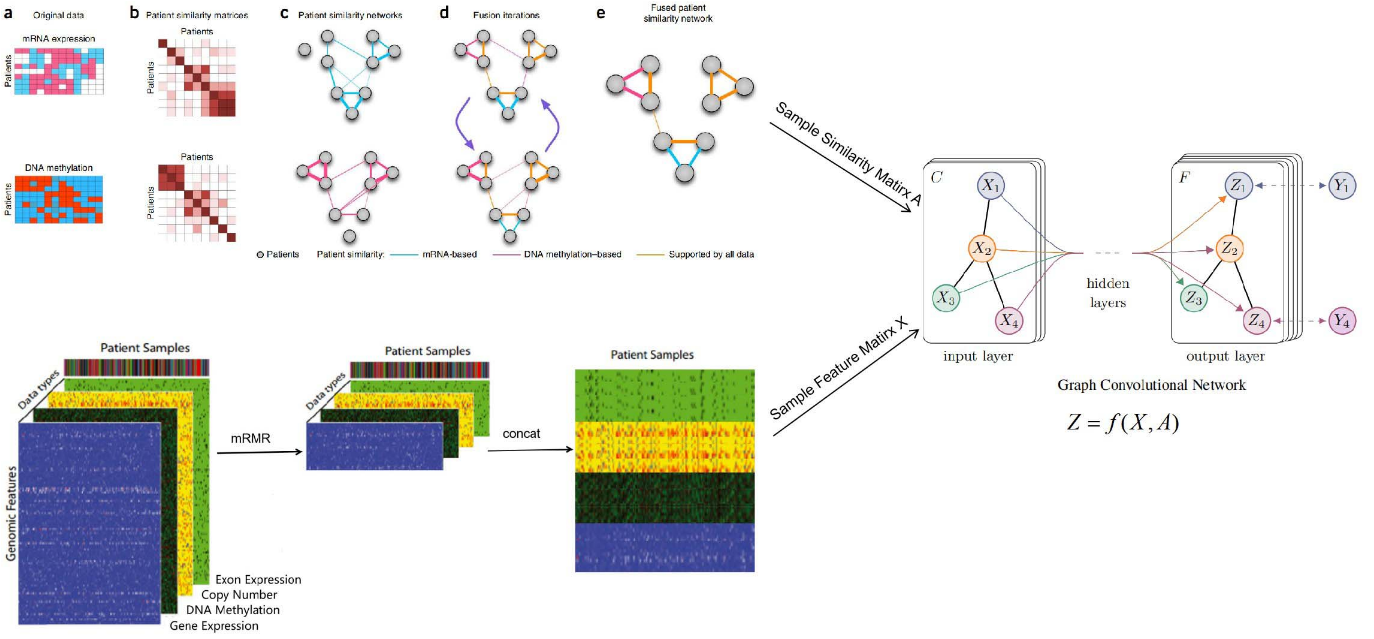

Fig. 1. GCGCN experimental framework [29], [34].

max-relevance feature selection algorithm (mRMR) was used to conduct feature selection of multiple genomic data and clinical data to obtain the optimal feature combination. According to these optimal features, a sample feature matrix X could be established. X and A were placed in the graph convolutional network (GCN) for classified training and prediction, and finally a cancer survival prediction model was built based on graph convolutional network.

A. Experimental Data

Breast cancer (BRCA) and Lung squamous cell cancer (LUSC) were taken as study objects, where cancer genomic data and clinical data mainly came from TCGA database. As the largest cancer genetic information database at present, TCGA is comprehensive in aspect of numerous cancer types but more in aspect of genomic data record with high confidence level, including gene expression, copy number alteration, DNA methylation, exon expression, clinical information, etc. For patient samples used in this chapter, their multiple genomic data and clinical data were obtained from TCGA library. According to ID matching of patient samples, 249 BRCA patient samples and 220 LUSC patient samples were finally obtained, and each sample covered detailed information of four types of genomic data and clinical data. 5 years and 3 years were taken as thresholds to divide two types of patients with two cancer types. Samples were divided into long-survival and short-survival patients according to the thresholds, and meanwhile, classification label of short-survival samples was set as 1 and that of long-survival samples was set as 0, and concrete information is seen in Table I.

TABLE I
Information Summary of BRCA and LUSC Patients

Cancer type	BRCA	LUSC
Total number of samples	249	220
Threshold (year)	5	3
Long-survival patients	70	109
Short-survival patients	179	111
Average age (years old)	57.19	68.29

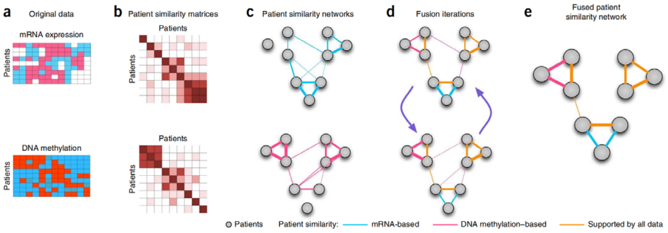

Fig. 2. Similarity network fusion algorithm [29].

B. Similarity Network Fusion

The flow of similarity network fusion is shown in Fig. 2, where Fig.2-a is mRNA expression and DNA methylation of patients of the same type; Fig.2-a is patient-patient similarity matrix of each data type; Fig.2-c is patient-patient similarity network, node represents patient, and edge represents similarity between two patients; Fig.2-d is network fusion process. Each network is upgraded iteratively through information from other networks using SNF algorithm so that each step is more similar; e is continuous iteration and fusion of networks until the final fusion network is obtained through convergence, and edge color expresses that data type has contributed to the given similarity [29].

For SNF model, assuming that there are m different data types. For the v(th) data type ($v = 1, 2, ..., m$), similarity matrix $P^{(v)}$ of all patients and K-nearest similarity matrix $S^{(v)}$

are respectively calculated,the calculated formula is shown as follows. Two data types are taken as the example namely m = 2.

$$\text{similarity matrix}: P(i,j) = \begin{cases} \dfrac{W(i,j)}{2\sum_{k\neq i} W(i,k)}, & j \neq i \\ 1/2, & j = i \end{cases} \tag{1}$$

$$\text{K-nearest similarity matrix}: S(i,j) = \begin{cases} \dfrac{W(i,j)}{\sum_{k\in N_i} W(i,k)}, & j \in N_i \\ 0, & otherwise \end{cases} \tag{2}$$

where $W(i, j)$ is a scaled exponential similarity matrix.

Step1: Calculate $P(i, j)$ and $S(i, j)$ of each data type
Calculate $P^{(1)}$ and $P^{(2)}$ through equation (1), and calculate and through equation (2),so that $P^{(1)}_{\mathrm{t=0}} = P^{(1)}$ and $P^{(2)}_{\mathrm{t=0}} = P^{(2)}$ represent two initial state matrixes under t=0(t is time).

Step2: Update the similarity matrix
Similarity matrix of each data type is iteratively updated as follows:

$$P^{(1)}_{t+1}(i,j) = \sum_{k\in N_i}\sum_{l\in N_j} S^{(1)}(i,k) \times S^{(1)}(j,l) \times P^{(2)}_t(k,l) \tag{3}$$

$$P^{(2)}_{t+1}(i,j) = \sum_{k\in N_i}\sum_{l\in N_j} S^{(2)}(i,k) \times S^{(2)}(j,l) \times P^{(1)}_t(k,l) \tag{4}$$

Step3: Standardization
$P^{(1)}_{\mathrm{t+1}}$ and $P^{(2)}_{\mathrm{t+1}}$ obtained in Step2 are substituted into formula (1) for standardization.

Step4: Result output
After t steps, the output overall state matrix can be calculated as: $P^{(c)} = \frac{P^{(1)}_t + P^{(2)}_t}{2}$

As a general rule, when m>2, the following formula (5) can be obtained according to formulas (3) and (4)

$$P^{(v)} = S^{(v)} \times (\frac{\sum_{k\neq v} P^{(k)}}{m-1}) \times (S^{(v)})^T, \quad v = 1, 2, \ldots, m \tag{5}$$

Through the above method, a patient network graph structure integrating multiple genomic data and clinical data can be finally obtained. SNF algorithm captures shared and supplementary information from different data sources. Information quantity of similarities observed between samples of each data type should be deeply understood. As it is based on sample network, useful information can be obtained even from a small quantity of samples, and meanwhile, it is of robustness to noise and data heterogeneity.

C. Min-Redundancy Max-Relevance Feature Selection

A common problem which will be encountered when high-throughput sequencing dataset is used to predict survival time of human cancers is so-called "curse of dimensionality". In our study, the sample size was limited while sample features were obtained by combining gene expression, copy number alteration, DNA methylation, exon expression and clinical data, so dimensionality of sample features was considerably enormous, and then sample feature dimensionality was far larger than sample size, which brought about certain difficulties to model learning and prediction. Because high dimensionality of dataset and small sample size can easily cause model overfitting and the model fits in with the training set too much, prediction effects are very poor for untrained validation set and test set. Hence, for problems involving a large number of features, feature selection plays a critical role in success of a learning algorithm.

A common feature selection method is maximizing the relevance between features and classification variables, namely selecting the first k variables with the highest relevance with classification variables. However, a single good feature combination can't improve the classifier performance in feature selection, because features may be highly relevant, which causes redundancy of feature variables. Hence, min-redundancy max-relevance feature selection algorithm comes into being, namely maximizing relevance between features and classification variables while minimizing relevance between features, and this is core idea of mRMR [30], which is a common dimension reduction algorithm enjoying extensive application [31]–[33]. Therefore, mRMR feature selection method was used to select features from the original dataset, dimensionality of the dataset was reduced while significant information loss was not caused, optimal combination features in each genomic part were selected, and feature dimensionality was reduced, thus improving generalization ability of the model.

In the specific experimental steps, we need to iterate to find the optimal feature, so this method mainly uses incremental search method to select features. For example, when we have obtained the optimal feature subset S_{m-1}, the next goal is to find the mth feature from the remaining feature set $X - S_{m-1}$, and maximize $\Phi(\cdot)$ by selecting features. The process of feature selection is shown in Figure 3-4. The incremental search algorithm finds the optimal feature by optimizing the following conditions:

$$\max_{x_j\in X-S_{m-1}}[I(x_j;c) - \frac{1}{m-1}\sum_{x_i\in S_{m-1}} I(x_j;x_i)] \tag{6}$$

D. Graph Convolutional Network

Since 2012, deep learning has achieved enormous success in two fields—computer vision and natural language processing. However, its study object is still restricted to Euclid field data. In the real world, many important datasets exist in forms of graph or grids. Graph is a data format. Graph convolutional network algorithm model on the graph structure was considered in this paper.

Spectral convolution of graph is defined [34] as the product of signal $x \in R^N$ and filter $g_\theta = diag(\theta)$:

$$g_\theta * x = U g_\theta U^T x \tag{7}$$

where U is a matrix consisting of eigenvectors of normalized graph Laplacian matrix. Graph Laplacian (matrix) is

$L = I_N - D^{-\frac{1}{2}}AD^{-\frac{1}{2}} = U\Lambda U^T$, Λ is diagonal matrix consisting of eigenvalues of L and $U^T x$ is graph Fourier transform of x. According to related literatures, Chebyshev polynomial $T_k(x)$ can be used for approaching, and the following formula is obtained:

$$g_{\theta'} * x \approx \sum_{k=0}^{k} \theta'_k T_k(\tilde{L})x \tag{8}$$

where $\tilde{L} = \frac{2}{\lambda_{\max}}L - I_N$.

Graph convolutional neural network model can be constituted by stacking multiple convolutional layers in the form of formula (8). Hereby number of convolutional layers is restricted as k=1, and meanwhile, approximate $\lambda_{\max} = 2$ is taken, and then:

$$g_{\theta'} * x \approx \theta'_0 x + \theta'_1(L - I_N)x = \theta'_0 x - \theta'_1 D^{-\frac{1}{2}}AD^{-\frac{1}{2}}x \tag{9}$$

Furthermore, overfitting can be avoided by constraining number of parameters while operation frequency at each layer is minimized, and the following formula is obtained:

$$g_{\theta} * x \approx \theta(I_N + D^{-\frac{1}{2}}AD^{-\frac{1}{2}})x \tag{10}$$

$\theta = \theta_0^{'} = -\theta_1^{'}$ is set in (9). But eigenvalue interval of $I_N + D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ is [0,2]. In the deep neural network model, repetitive operation of this operation may result in numerical instability and gradient explosion/vanishing. In order to solve this problem, "re-normalization" is introduced: $I_N + D^{-\frac{1}{2}}AD^{-\frac{1}{2}} \rightarrow \tilde{D}^{-\frac{1}{2}}A\tilde{D}^{-\frac{1}{2}}$, where$\tilde{A} = A + I_N, \tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$. This definition is promoted to signal $X \in R^{N\times C}$ (N is number of nodes and C is dimensionality of node eigenvector) and F filters or feature map with C channels:

$$Z = \tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}X\Theta \tag{11}$$

where $\Theta \in R^{C\times F}$ is filter parameter matrix and $Z \in R^{N\times F}$ is signal matrix after convolution. Complexity of this filter operation is $O(|\varepsilon|FC)$. $\tilde{A}X$ can be regarded as the product of a sparse matrix and a dense matrix.

E. Cancer Survival Prediction Model GCGCN

Two problems exist in the cancer survival prediction namely too small sample size and high sample feature dimensionality. When the sample size is too small, different training samples will result in great differences in trained classification models and then give rise to a large variance in the model classification result. How to reasonably utilize all samples and guarantee a certain stability of the trained classification model has become the difficulty in the cancer survival prediction. Graph convolutional network not only considers features and labels of training samples but also will continuously acquire related features from adjacent points (namely samples) through adjacent matrixes or similarity matrixes of samples, so it will take full consideration features of all samples and labels of training samples in the training process, thus ensuring stability of the trained graph convolutional model under different training samples, and meanwhile, the graph convolutional model can obtain considerable or better effect under a small sample size.

Therefore, graph convolutional network was selected in this study to establish a cancer survival prediction model and establish sample similarity matrix or sample adjacent matrix and sample feature matrix needed by the graph convolutional model. It's necessary to integrate multiple genomic data (including gene expression data, copy number alteration, DNA methylation and exon expression) and clinical data in this study, but similarity network fusion (SNF) establishes the network of samples of each available data type (like patient or gene) and takes this sample network as the basis for integrating multi-omics data, aiming at utilizing complementation in data. Relative to analytical methods established for single data, SNF algorithm has been verified having considerable advantages in aspects of identifying tumor subtype and prediction. Hence, multiple genomic data and clinical data can be integrated through SNF algorithm to finally obtain a sample similarity matrix. As multiple genomic data and clinical data have high feature dimensionalities, mRMR algorithm was used for feature selection, and some optimal features were taken respectively for splicing to obtain a sample feature matrix. Sample similarity matrix and sample feature matrix integrating multiple genomic data and clinical data were used to establish the graph convolution-based cancer survival prediction model GCGCN.

F. Comparison With Other Cancer Survival Prediction Methods

In order to verify effectiveness of the proposed graph convolutional network-based cancer survival prediction method integrating multiple genomic data and clinical data, GCGCN was compared with other six models namely CGCN (clinical data), GGCN (multiple genomic data integrated), Gene-Expr (gene expression), DNAmethy (DNA methylation), CNA (copy number alteration) and ExonExpr (exon expression). Main differences of these methods from GCGCN lie in that they have not completely integrated multiple genomic data and clinical data, but instead, they only use multiple genomic data or data of other single types.

In order to verify effectiveness of the proposed cancer survival prediction method GCGCN, five commonly used classification methods were adopted for a comparison, respectively being naïve Bayesian classification (NB) [35], [36], K-nearest neighbor classification (KNN), logic regression (LR), decision tree (DT) and support vector machine (SVM) [37]. Meanwhile, in order to verify effectiveness of the proposed method integrating multiple genomic data and clinical data, dataset was divided into three groups—multiple genomic data, clinical data and multiple genomic data + clinical data—in this contrast experiment.

III. Experiment

A. Experimental Setting

In this paper, the total dataset (available from https://xenabrowser.net) was randomly divided into training set, Validation set and test set according to the proportion 7:1:2, and concrete data division is seen in Table II. In order to guarantee fairness and robustness of research methods,

TABLE II
DATA DIVISION OF TRAINING SET, VALIDATION SET AND TEST SET

Cancer	Data	Training	Validation	Test
BRCA	Long-survival	119	20	40
	Short-survival	40	10	20
	Total sample	159	30	60
LUSC	Long-survival	79	10	20
	Short-survival	81	10	20
	Total sample	160	20	40

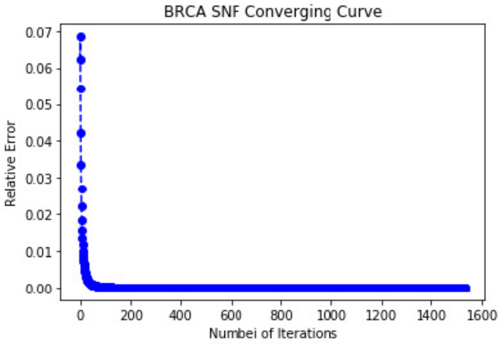

Fig. 3. BRCA SNF converging curve.

datasets were randomly divided for 5 times, experiment on each research method was carried out for 5 times, and final evaluation indexes were average values of 5 experiments.

In this experiment, hyper-parameters k=20 and $\mu = 0.5$ are taken in SNF algorithm and threshold value is taken as, namely when relative error in the iteration process satisfies, the algorithm can stop the iteration. In the GCN algorithm model, a three-layer graph convolutional network is uniformly adopted, respectively being input layer, hidden layer containing 40 nodes and output layer. In mRMR feature selection, 50 optimal features of gene expression profile, copy number alteration, DNA methylation and exon expression are respectively selected, two features—survival time and survival state—in clinical information are deleted, residual features are retained, and all of these features are combined to obtain the sample feature matrix X. Concrete results are seen in Table III.

B. Experimental Results

As shown in Fig. 3 and Fig. 4, in the similarity network fusion process of two cancers, rapid convergence can be realized, but in order to satisfy the condition for iterating termination, 1,500 times of iterations are needed.

GCGCN was compared with other six models respectively, namely CGCN (clinical data), GGCN(multiple genomic data integrated), GeneExpr (gene expression), DNAmethy (DNA methylation), CNA (copy number alteration) and ExonExpr (exon expression). Results corresponding to these models were compared using ROC curve and AUC value. Fig. 5 and Fig. 6 show average ROC curve chart and AUC mean value when the experiment was randomly repeated for 5 times. When multiplegenomic data and clinical data were integrated in the model, the obtained prediction performance was obviously superior to any other model. For BRCA, the AUC value of the used model integrating multiple genomic data and clinical data was 0.93, while the value was 0.58, 0.65, 0.56 and 0.62 respectively in gene expression model, DNA methylation model, copy number alteration model and exon expression model. Meanwhile, AUC value when only clinical data were used was 0.84. When the method integrated four types of genomic data, model AUC value could reach 0.78 which was obviously superior to the model established based on single genomic data. As for LUSC, AUC value of the proposed model integrating multiple genomic data and clinical data was 0.81, while the value was 0.57, 0.69, 0.52 and 0.65 respectively in gene expression model, DNA methylation model, copy number alteration model and exon expression model. Meanwhile, AUC value when only clinical data were used was 0.73. When the method integrated four types of genomic data, model AUC value could reach 0.78 which was obviously superior to the model established based on single genomic data. The above two results indicated that the proposed integration of multiple genomic data and clinical data reflected survival time of cancer

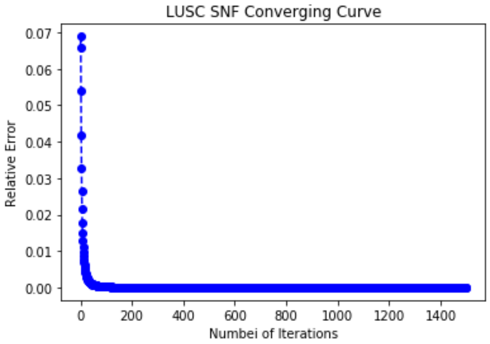

Fig. 4. LUSC SNF converging curve.

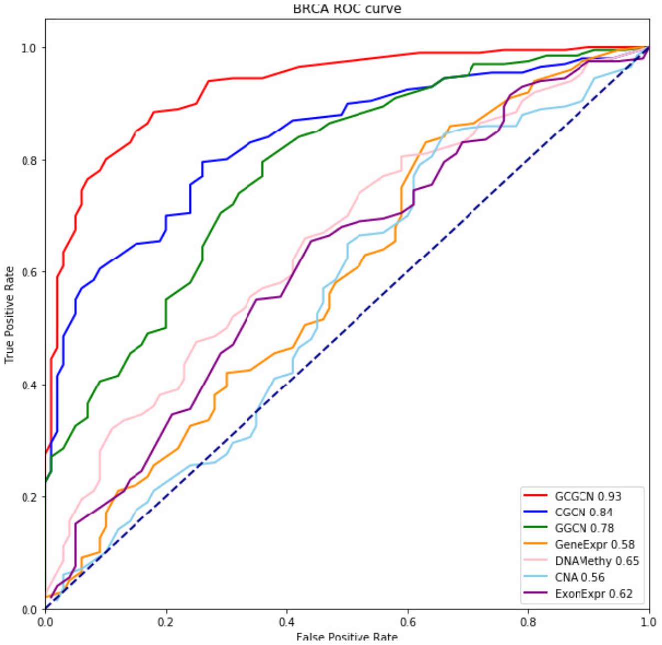

Fig. 5. Performance comparison of BRCA models.

TABLE III
NUMBERS OF OPTIMALLY SELECTED DIFFERENT FEATURES OF TWO CANCERS—BRCA AND LUSC

Cancer type	BRCA		LUSC	
Data type	num of original features	num of optimal features	num of original features	num of optimal features
Gene expression	20531	50	20531	50
Copy number alteration	24777	50	24777	50
DNA methylation	364737	50	365863	50
Exon expression	239323	50	239323	50
Clinical information	26	24	22	20

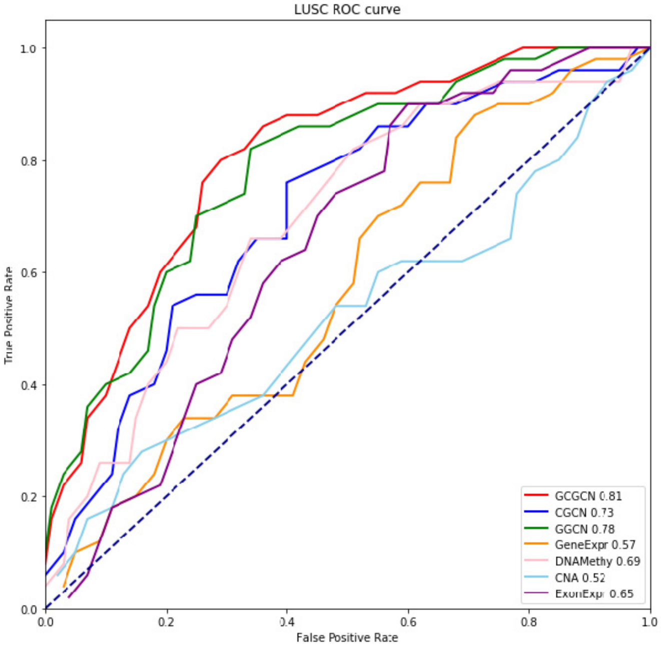

Fig. 6. Performance comparison of LUSC models.

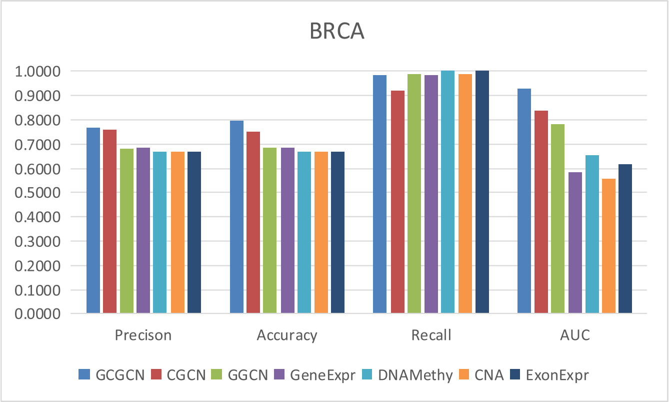

Fig. 7. Comparison histogram of multiple average performances of seven different BRCA models.

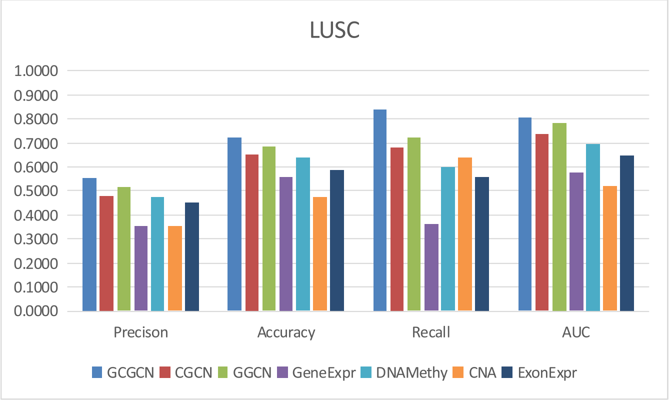

Fig. 8. Comparison histogram of multiple average performances of seven different LUSC models.

patients from different aspects, thus providing a certain help for medical personnel to formulate concrete medical methods.

In addition, in order to further verify reliability of the proposed model integrating multiple genomic data and clinical data. Four performance evaluation indexes—Precion, Accuracy, Recall and AUC—were respectively calculated for each model, where Precion, Accuracy and Recall were all measured under the threshold value 0.5.

As shown in Table IV, Table V, Fig. 7 and Fig. 8, the proposed method integrating multiple genomic data and clinical data has higher prediction performance than other comparative models in all indexes. For instance, for BRCA, Precion, accuracy, recall and AUC value of the proposed method are 0.7698, 0.7933, 0.9850 and 0.9280 respectively with Precion, accuracy and AUC value higher than those of the method only using multiple genomic data by 0.0891, 0.1100 and 0.1475 respectively, and meanwhile their difference in recall is minor). Precion, accuracy, recall and AUC value are higher than those of the method only using clinical data by 0.0121, 0.0433, 0.0650 and 0.0930 respectively. For LUSC, Precion, accuracy, recall and AUC value of the proposed method are 0.5539, 0.7200, 0.8400 and 0.8050 respectively, which are higher than those of the method only using multiple genomic data by 0.0390, 0.0333, 0.1200 and 0.0210 respectively and higher than those of the method only using clinical data by 0.0739, 0.0666, 0.1600 and 0.0700 respectively. To sum up, the proposed model integrating multiple genomic data and clinical data has obviously superior prediction performance indexes than prediction methods only using single genomic data or clinical data in the aspect of cancer survival prediction, indicating that both clinical data and multiple genomic data can reflect their influences on concrete cancer survival time from different aspects.

Survival analysis expresses a statistical method considers both result and survival time. Moreover, it can give full play to incomplete information provided by censored data to describe distribution features of survival time and analyze main influ-

TABLE IV
COMPARISON CHART OF MULTIPLE PERFORMANCE EVALUATION INDEXES OF SEVEN BRCA MODELS

BRCA	Precison	Accuracy	Recall	AUC
GCGCN	**0.7698**	**0.7933**	**0.9850**	**0.9280**
CGCN	0.7577	0.7500	0.9200	0.8350
GGCN	0.6807	0.6833	0.9900	0.7805
GeneExpr	0.6841	0.6867	0.9850	0.5849
DNAMethy	0.6667	0.6667	1.0000	0.6537
CNA	0.6689	0.6667	0.9900	0.5555
ExonExpr	0.6667	0.6667	1.0000	0.6180

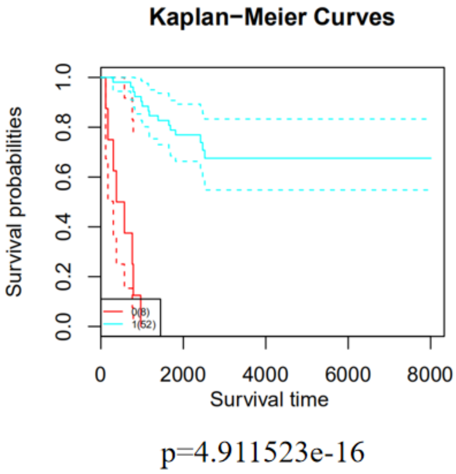

Fig. 9. KM survival curves of BRCA prediction category.

ence factors of survival time. Fig. 9 and Fig. 10 shows test sets of two cancers, which are divided into two groups according to classification results obtained through GCGCN, KM curves are drawn, and their P values are calculated according to the curves, where left picture is BRCA and right picture is LUSC. Both P values are smaller than 0.05. In addition, as shown in Fig. 11 and Fig. 12, we list the KM survival curves and P values of the original groups of two kinds cancer data to compare the results above. Therefore, the cancer survival classification results of test sets using the proposed GCGCN method have significant differences.

In order to verify favorable effect of the proposed method in cancer survival prediction, this method is hereby compared with five common classification methods. Concrete classification effects of the models are seen in Table VI and Table VII. For BRCA, when multiple genomic data + clinical data are used, average AUC value of GCGCN is 0.9280 while those of five classification methods—NB, KNN, LR, DT, SVM and DNN—are 0.8683, 0.7083, 0.9080, 0.7500, 0.8855 and 0.9080 respectively. Meanwhile, when only multiple genomic data or clinical data are used, average AUC value obtained through GCGCN is 0.7805 and 0.8350 respectively, which are also obviously higher than those of other methods, and meanwhile, the variance is small. For LUSC, when multiple genomic data + clinical data are used, average AUC value of GCGCN is 0.8050, and similarly, the proposed GCGCN method has higher average AUC value and smaller AUC variance than other methods. In a similar way, it can be

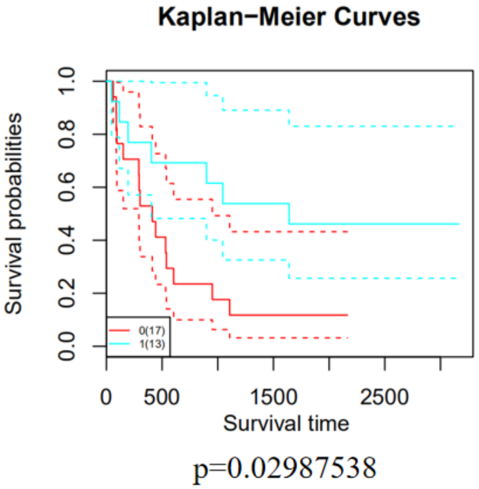

Fig. 10. KM survival curves of LUSC prediction category.

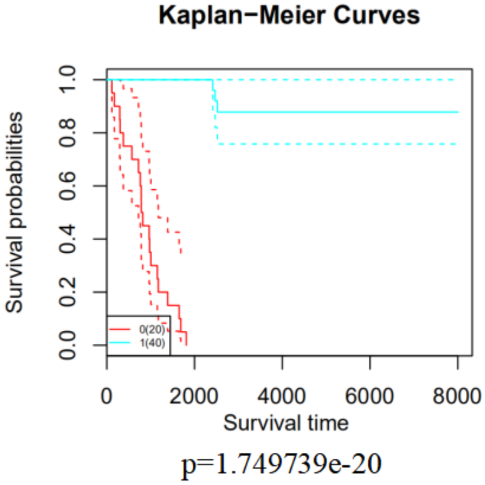

Fig. 11. KM survival curves of BRCA original category.

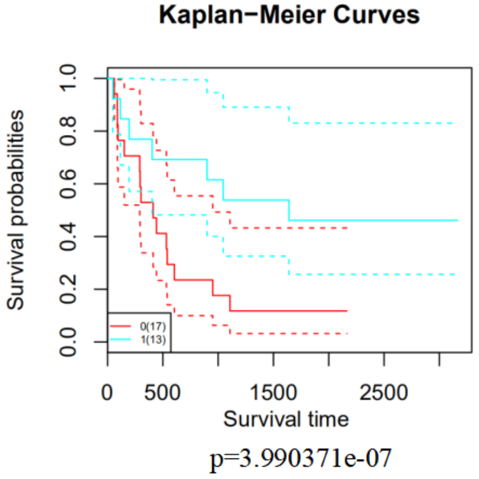

Fig. 12. KM survival curves of LUSC original category.

seen from Table VI and Table VII that for most classification methods, prediction performance based on integrated multiple genomic data and clinical data has been improved to a certain degree. For instance, for BRCA, when multiple genomic data + clinical data are used by NB algorithm, the obtained average AUC value increases by 0.1090 and 0.1008 respectively when

TABLE V
COMPARISON CHART OF MULTIPLE PERFORMANCE EVALUATION INDEXES OF SEVEN LUSC MODELS

LUSC	Precison	Accuracy	Recall	AUC
GCGCN	**0.5539**	**0.7200**	**0.8400**	**0.8050**
CGCN	0.4800	0.6534	0.6800	0.7350
GGCN	0.5149	0.6867	0.7200	0.7840
GeneExpr	0.3506	0.5600	0.3600	0.5750
DNAMethy	0.4746	0.6400	0.6000	0.6950
CNA	0.3516	0.4733	0.6400	0.5200
ExonExpr	0.4518	0.5867	0.5600	0.6480

TABLE VI
COMPARISON OF BRCA AUC VALUE OF GCGCN WITH EXISTING CLASSIFICATION METHODS

BRCA	Multiple genomic data	Clinical data	Multiple genomic data + clinical data
NB	0.7593±0.5318	0.7675±0.0494	0.8683±0.4549
KNN	0.6210±0.0904	0.7383±0.0660	0.7038±0.0848
LR	0.7663±0.0392	0.8065±0.0626	0.9080±0.0319
DT	0.6300±0.0509	0.6650±0.0365	0.7500±0.0622
SVM	0.7735±0.0503	0.8055±0.0628	0.8855±0.0226
DNN	0.7765±0.1021	0.8145±0.0920	0.9080±0.0510
GCGCN	**0.7805±0.0553**	**0.8350±0.0520**	**0.9280±0.0260**

TABLE VII
COMPARISON OF LUSC AUC VALUE OF GCGCN WITH EXISTING CLASSIFICATION METHODS

LUSC	Multiple genomic data	Clinical data	Multiple genomic data + clinical data
NB	0.6885±0.0796	0.7212±0.0622	0.7831±0.0797
KNN	0.6015±0.0900	0.5325±0.0797	0.6355±0.0409
LR	0.6985±0.1263	0.7330±0.0295	0.7660±0.0411
DT	0.5475±0.0406	0.5750±0.1140	0.5950±0.0400
SVM	0.7145±0.0236	0.7012±0.0764	0.7560±0.0471
DNN	0.7265±0.1302	0.7122±0.0902	0.7721±0.0890
GCGCN	**0.7840±0.0491**	**0.7350±0.0574**	**0.8050±0.0301**

compared with methods only using multiple genomic data or clinical data. When LR algorithm uses multiple genomic data + clinical data, average AUC value increases by 0.1417 and 0.1015 respectively when compared with methods only using multiple genomic data or clinical data. It can also be found that the proposed GCGCN has overall small variance, because in the training process of graph convolutional network model, the model considers both global information and sample relevance, and then a more stable prediction result can be obtained.

IV. DISCUSSION AND CONCLUSIONS

As cancer is one of the most common and malignant diseases in the world, this study was dedicated to improving prediction performance of cancer survival. A graph convolutional network-based cancer survival prediction method GCGCN integrating multiple genomic data and clinical data was proposed in this paper, where multiple genomic data included gene expression, copy number alteration, DNA methylation and exon expression. First of all, multiple genomic data and clinical data were integrated using the similarity network fusion algorithm, sample similarity matrix was obtained, cancer survival-related features were extracted using min-redundancy max-relevance mRMR feature selection algorithm, the influence of useless features was mitigated, and classification training and prediction were conducted through the graph convolutional network. In order to explore into effectiveness of multiple genomic data and clinical data in cancer survival prediction, compared with existing cancer survival prediction methods only using multiple genomic data or clinical data, indexes—ROC curve, AUC, Recall, Accuracy, Precion, etc.—were compared in this paper, and results indicated that both multiple genomic data and clinical data had a bearing on cancer survival in different aspects. Therefore, integration of multiple genomic data and clinical data could improve prediction effect on cancer survival. Meanwhile, five common classification methods—NB, KNN, LR, DT and SVM—were applied to cancer survival prediction. Through a comparison, the proposed GCGCN method had higher average AUC value and lower AUC standard deviation, so GCGCN showed obviously more excellent classification effect relative to other common classification methods.

Even though GCGCN has excellent classification effect, it still has certain limitations in the aspect of cancer survival. For example, this research work can realize extension and verification based on a larger sample size (cancer patients). The present sample size is restricted by availability of multiple genomic data and clinical data. It's estimated that when more cancer samples are available in the future, performance of this method can be significantly improved. In addition, it's believed that if cancers are classified by their subtypes and GCGCN model is established for each subtype, this will be more significant for cancer researchers and its performance may be further improved, because cancer survival prediction is greatly influenced by cancer subtypes [38]–[40]. Unfortunately, for each subtype of cancer patients, available data size is very small. Therefore, when there are more available samples in the future, the concrete analysis of cancer subtypes will be our extension direction. Another future research direction is to integrate more genomic data like protein expression, miRNA expression [41]–[44] and other genomic data [45]–[47]. Meanwhile, pathological image features of cancer patients may be considered in the future research work. The final goal is to establish a multitask learning system for different cancer researches, including cancer susceptibility prediction, cancer recurrence and cancer therapy.

REFERENCES

[1] L. A. Torre, F. Bray, R. L. Siegel, J. Ferlay, J. Lortet-Tieulent, and A. Jemal, "Global cancer statistics," *CA, Cancer J. Clinicians*, vol. 65, no. 2, pp. 87–108, 2015.

[2] E. A. Rakha *et al.*, "Breast cancer prognostic classification in the molecular era: The role of histological grade," *Breast Cancer Res.*, vol. 12, no. 4, 2010, Art. no. 207.

[3] O. Balacescu *et al.*, "Blood genome-wide transcriptional profiles of HER2 negative breast cancers patients," *Mediators Inflammation*, vol. 2016, Dec. 2015, Art. no. 3239167.

[4] Z. Liao, D. Li, X. Wang, L. Li, and Q. Zou, "Cancer diagnosis through isomiR expression with machine learning method," *Current Bioinf.*, vol. 13, no. 1, pp. 57–63, 2018.

[5] L. Yu, J. Huang, Z. Ma, J. Zhang, Y. Zou, and L. Gao, "Inferring drug-disease associations based on known protein complexes," *BMC Med. Genomics*, vol. 8, no. 2, 2015, Art. no. S2.

[6] L. R. Martin, S. L. Williams, K. B. Haskard, and M. R. DiMatteo, "The challenge of patient adherence," *Therapeutics Clin. Risk Manage.*, vol. 1, no. 3, pp. 189–199, 2005.

[7] L. Yu, X. Ma, L. Zhang, J. Zhang, and L. Gao, "Prediction of new drug indications based on clinical data and network modularity," *Sci. Rep.*, vol. 6, Sep. 2016, Art. no. 32530.

[8] Y. Sun, S. Goodison, J. Li, L. Liu, and W. Farmerie, "Improved breast cancer prognosis through the combination of clinical and genetic markers," *Bioinformatics*, vol. 23, no. 1, pp. 30–37, 2006.

[9] X. Xu, Y. Zhang, L. Zou, M. Wang, and A. Li, "A gene signature for breast cancer prognosis using support vector machine," in *Proc. 5th Int. Conf. BioMed. Eng. Inform.*, Oct. 2012, pp. 928–931.

[10] P. C. Stone and S. Lund, "Predicting prognosis in patients with advanced cancer," *Annals Oncol.*, vol. 18, no. 6, pp. 971–976, 2006.

[11] L. J. van 't Veer, *et al.*, "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, pp. 530–536, Jan. 2002.

[12] L. Yu, J. Zhao, and L. Gao, "Predicting potential drugs for breast cancer based on miRNA and tissue specificity," *Int. J. Biol. Sci.*, vol. 14, no. 8, pp. 971–980, 2018.

[13] D. M. A. El-Rehim *et al.*, "High-throughput protein expression analysis using tissue microarray technology of a large well-characterised series identifies biologically distinct classes of breast cancer confirming recent cDNA expression analyses," *Int. J. Cancer*, vol. 116, no. 3, pp. 340–350, Art. no. 2005.

[14] M. J. van de Vijver *et al.*, "A gene-expression signature as a predictor of survival in breast cancer," *New England J. Med.*, vol. 347, no. 25, pp. 1999–2009, 2002.

[15] Y. Wang *et al.*, "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer," *Lancet*, vol. 365, pp. 671–679, Feb. 2005.

[16] W. Chen H. Lv, F. Nie, and H. Lin, "i6mA-Pred: Identifying DNA N6-methyladenine sites in the Rice genome," *Bioinformatics*, vol. 35, no. 16, pp. 2796–2800, 2019.

[17] W. Chen, H. Yang, P. Feng, H. Ding, and H. Lin, "iDNA4mC: Identifying DNA N^4-methylcytosine sites based on nucleotide chemical properties," *Bioinformatics*, vol. 33, no. 22, pp. 3518–3523, 2017.

[18] O. Gevaert, F. De Smet, D. Timmerman, Y. Moreau, and B. De Moor, "Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks," *Bioinformatics*, vol. 22, pp. e184–e190, Jul. 2006.

[19] C. Y. W. Nguyen and H. N. Nguyen, "Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic," *J. Biomed. Sci. Eng.*, vol. 6, pp. 551–560, May 2013.

[20] J. D. Brenton, L. A. Carey, A. A. Ahmed, and C. Caldas, "Molecular classification and molecular forecasting of breast cancer: Ready for clinical application?" *J. Clin. Oncol.*, vol. 23, no. 29, pp. 60–7350, 2005.

[21] A.-L. Boulesteix, C. Porzelius, and M. Daumer, "Microarray-based classification and clinical predictors: On combined classifiers and additional predictive value," *Bioinformatics*, vol. 24, no. 15, pp. 1698–1706, 2008.

[22] M. Khademi and N. S. Nedialkov, "Probabilistic graphical models and deep belief networks for prognosis of breast cancer," in *Proc. IEEE 14th Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2015, pp. 727–732.

[23] J. Das, K. M. Gayvert, F. Bunea, M. H. Wegkamp, and H. Yu, "ENCAPP: Elastic-net-based prognosis prediction and biomarker discovery for human cancers," *BMC Genomics*, vol. 16, Apr. 2015, Art. no. 263.

[24] J. Hayes *et al.*, "Prediction of clinical outcome in glioblastoma using a biologically relevant nine-microRNA signature," *Mol. Oncol.*, vol. 9, no. 3, pp. 704–714, 2015.

[25] Y. Zhang, A. Li, C. Peng, and M. Wang, "Improve glioblastoma multiforme prognosis prediction by using feature selection and multiple kernel learning," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 13, no. 5, pp. 825–835, Sep. 2016.

[26] K. Tomczak, P. Czerwinska, and M. Wiznerowicz, "The cancer genome atlas (TCGA): An immeasurable source of knowledge," *Contemp. Oncol.*, vol. 19, no. 1A, pp. A68–A77, 2015.

[27] J. Gao *et al.*, "Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal," *Sci. Signaling*, vol. 6, no. 269, p. pl1, 2013.

[28] L. Yu, J. Zhao, and L. Gao, "Drug repositioning based on triangularly balanced structure for tissue-specific diseases in incomplete interactome," *Artif. Intell. Med.*, vol. 77, pp. 53–63, Mar. 2017.

[29] B. Wang *et al.*, "Similarity network fusion for aggregating data types on a genomic scale," *Nature Methods*, vol. 11, pp. 333–337, Jan. 2014.

[30] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[31] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *J. Bioinf. Comput. Biol.*, vol. 3, no. 2, pp. 185–205, 2005.

[32] F.-Y. Dao, H. Lv, F. Wang, C.-Q. Feng, H. Ding, W. Chen, and H. Lin, "Identify origin of replication in *Saccharomyces cerevisiae* using two-step feature selection technique," *Bioinformatics*, vol. 35, no. 12, pp. 2075–2083, 2018.

[33] H. Yang, H. Lv, H. Ding, W. Chen, and H. Lin, "iRNA-2OM: A sequence-based predictor for identifying 2'-O-methylation sites in homo sapiens," *J. Comput. Biol.*, vol. 25, no. 11, pp. 1266–1277, 2018.

[34] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. 5th Int. Conf. Learn. Represent. (ICLR)*. Toulon, France: Palais des Congrès Neptune, 2017, pp. 1–14.

[35] P.-M. Feng, H. Lin, and W. Chen, "Identification of antioxidants from sequence information using Naïve Bayes," *Comput. Math. Methods Med.*, vol. 2013, Jul. 2013, Art. no. 567529.

[36] P.-M. Feng, H. Ding, W. Chen, and H. Lin, "Naïve Bayes classifier with feature selection to identify phage virion proteins," *Comput. Math. Methods Med.*, vol. 2013, Apr. 2013, Art. no. 530696.

[37] C.-Q. Feng, Z.-Y. Zhang, X.-J. Zhu, Y. Lin, W. Chen, H. Tang, and H. Lin, "iTerm-PseKNC: A sequence-based tool for predicting bacterial transcriptional terminators," *Bioinformatics*, vol. 35, no. 9, pp. 1469–1477, 2019.

[38] Z. Liu, X.-S. Zhang, and S. Zhang, "Breast tumor subgroups reveal diverse clinical prognostic power," *Sci. Rep.*, vol. 4, Feb. 2014, Art. no. 4002.

[39] C. Desmedt *et al.*, "Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes," *Clin. Cancer Res.*, vol. 14, no. 16, pp. 5158–5165, 2008.

[40] L. Jiang, Y. Xiao, Y. Ding, J. Tang, and F. Guo, "Discovering cancer subtypes via an accurate fusion strategy on multiple profile data," *Frontiers Genet.*, vol. 10, p. 20, Feb. 2019.

[41] W. Tang, S. Wan, Z. Yang, A. E. Teschendorff, and Q. Zou, "Tumor origin detection with tissue-specific miRNA and DNA methylation markers," *Bioinformatics*, vol. 34, no. 3, pp. 398–406, Feb. 2018.

[42] Q. Wang, L. Wei, X. Guan, Y. Wu, Q. Zou, and Z. Ji, "Briefing in family characteristics of microRNAs and their applications in cancer research," *Biochimica Biophysica Acta-Proteins Proteomics*, vol. 1844, no. 1, pp. 191–197, 2014.

[43] L. Jiang, Y. Xiao, Y. Ding, J. Tang, and F. Guo, "FKL-Spa-LapRLS: An accurate method for identifying human microRNA-disease association," *BMC Genomics*, vol. 19, Dec. 2019, Art. no. 911.

[44] L. Jiang, Y. Ding, J. Tang, and F. Guo, "MDA-SKF: Similarity kernel fusion for accurately discovering miRNA-disease association," *Frontiers Genet.*, vol. 9, p. 618, Dec. 2018.

[45] X. Zhang, Q. Zou, A. Rodriguez-Paton, and X. Zeng, "Meta-path methods for prioritizing candidate disease miRNAs," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 1, pp. 283–291, Jan. 2019. doi: 10.1109/tcbb.2017.2776280.

[46] X. Zeng, N. Ding, A. Rodríguez-Patón, and Q. Zou, "Probability-based collaborative filtering model for predicting gene–disease associations," *BMC Med. Genomics*, vol. 10, no. 5, 2017, Art. no. 76.

[47] Y. Liu, X. Zeng, Z. He, and Q. Zou, "Inferring microRNA-disease associations by random walk on a heterogeneous network with multiple data sources," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 14, no. 4, pp. 905–915, Jul./Aug. 2016.