Improving Cancer Survival Prediction via Graph Convolutional Neural Network Learning on Protein-Protein Interaction Networks

Hongmin Cai, *Senior Member, IEEE*, Yi Liao, Lei Zhu, Zhikang Wang, and Jiangning Song

***Abstract*—Cancer is one of the most challenging health problems worldwide. Accurate cancer survival prediction is vital for clinical decision making. Many deep learning methods have been proposed to understand the association between patients' genomic features and survival time. In most cases, the gene expression matrix is fed directly to the deep learning model. However, this approach completely ignores the interactions between biomolecules, and the resulting models can only learn the expression levels of genes to predict patient survival. In essence, the interaction between biomolecules is the key to determining the direction and function of biological processes. Proteins are the building blocks and principal undertakings of life activities, and as such, their complex interaction network is potentially informative for deep learning methods. Therefore, a more reliable approach is to have the neural network learn both gene expression data and protein interaction networks. We propose a new computational approach, termed CRESCENT, which is a protein-protein interaction (PPI) prior knowledge graph-based convolutional neural network (GCN) to improve cancer survival prediction. CRESCENT relies on the gene expression networks rather than gene expression levels to predict patient survival. The performance of CRESCENT is evaluated on a large-scale pan-cancer dataset consisting of 5991 patients from 16 different types of cancers. Extensive benchmarking experiments demonstrate that our proposed method is competitive in terms of the evaluation metric of the time-dependent concordance index(C^{td}) when compared with several existing state-of-the-art approaches. Experiments also show that incorporating the network structure between genomic features effectively improves cancer survival prediction.**

Manuscript received 28 April 2022; revised 31 May 2023 and 3 October 2023; accepted 6 November 2023. Date of publication 14 November 2023; date of current version 6 February 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFE0112200, in part by the National Natural Science Foundation of China under Grants U21A20520 and 62325204, in part by the Science and Technology Project of Guangdong Province under Grant 2022A0505050014, and in part by the Key-Area Research and Development Program of Guangzhou City under Grant 202206030009. *(Corresponding author: Jiangning Song.)*

Hongmin Cai, Yi Liao, and Lei Zhu are with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China (e-mail: hmcai@scut.edu.cn; csliaoy@mail.scut.edu.cn; 201921041587@mail.scut.edu.cn).

Zhikang Wang and Jiangning Song are with the Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Melbourne, VIC 3800, Australia (e-mail: zhikang.wang@monash.edu; jiangning.song@monash.edu).

Digital Object Identifier 10.1109/JBHI.2023.3332640

***Index Terms*—Survival analysis, protein-protein interaction, machine learning, graph convolutional network.**

I. Introduction

CANCER is a leading cause of death globally. There are on average 19 million new cancer patients, and nearly 10 million patients die from cancer annually. In addition, the number of cancer patients is expected to increase by half by 2040 [1]. Therefore, cancer prevention is a crucial part of efforts to control cancer, as it can help reduce both the incidence and mortality of cancer. Accurate prognostic estimation on cancer patients are of great value to patients' quality of life, physicians' clinical decision-making, and personalized treatment. However, clinicians' prognostic estimates can be inaccurate [2]. In cases where physicians overestimate survival rates, patients frequently receive aggressive treatment before death, resulting in poor quality of end-of-life care [3]. In this regard, machine learning methods can be applied to build models for cancer survival prediction with improved performance [4]. The emerging large-scale genomic databases of cancer patients and their detailed follow-up records provide a solid database foundation and rich resource to support the development of accurate automated prognostic prediction methods.

Survival time records correspond to the time from disease diagnosis to death. It is common to observe no event of interest or censored observations in the actual data records, which provides a lower bound on the patient's survival time. In the early days, classical machine learning approaches were applied to mine the associations between patient characteristics and survival time. The Cox proportional hazards (CPH) [5] model is a semi-parametric regression model and has been widely used to predict hazard ratios. CPH assumes that the effect of patient features on survival risk does not change over time, which is unreasonable because the survival curves of different patients can overlap. The random survival forest (RSF) [6] introduces an interpretable mortality measure to predict outcomes. Clusters of patients with different prognoses were classified by the gene signatures through a support vector machine classifier [7]. In addition, Bayesian networks [8] have also been used for cancer survival prediction.

Recently, deep learning (DL) techniques have been increasingly used in bioinformatics and computational biology, which

demonstrated a better performance than traditional methods. DL-based methods can effectively capture nonlinear associations between the sample features and survival time. Moreover, they do not require manual feature engineering but instead essentially rely on the raw data as the input [9]. DeepSurv [10] is an approach based on neural network (NN) for the prediction of hazard ratios. Cox-nnet [11] is also an NN-based method for predicting patient prognosis using the RNA-sequencing data from The Cancer Genome Atlas (TCGA) [12]. More recently, different deep network architectures have been developed and deployed to address the challenges of various types of cancer data. As a powerful deep learning technique, convolutional neural networks can capture the correlation between image data and survival time. For example, CXR-risk [13] uses chest radiographs to assess long-term mortality in cancer patients. LUNG-net [14] is a shallow convolutional neural network that predicts the prognosis of lung cancer patients. Bichindaritz et al. used a bidirectional long short-term memory (LSTM) network to predict the prognosis of breast cancer cases in combination with gene expression and methylation data [15]. Modeling on large-scale pan-cancer data can leverage inter-cancer similarities to improve model performance. For example, MultimodalPrognosis [16] illustrates that pan-cancer training can help predict the survival probability of each cancer site. Although the aforementioned approaches have made great strides in developing prognostic prediction models based on various cancer data types, most of them are based on proportional hazards. The assumption of proportional hazards models is that the effects of covariates on survival do not vary over time. However, similar gene expression levels in patients with two different gene expression networks may lead to different tumor development trajectory, thereby largely violating the proportional hazard assumption. Theoretically, the discrete-time survival models are more plausible [17]; however, previous pan-cancer studies [16], [18] have not considered group differences in cancer patients.

For the majority of deep learning methods, the original matrices of genomics data of the samples are directly fed into the deep networks after normalization. However, there is a caveat of directly feeding the genomic feature matrix to the neural networks, in that it fails to take into account the interaction network information between different genes within the genomes. For those methods, the decision of the neural networks only depends on the characteristics of the features whilst ignoring the interactions between biological molecules. Therefore, it is necessary to develop improved methods by incorporating the knowledge of the interaction networks of biomolecules. In this regard, ENCAPP [19] is an elastic network-based approach that can predict prognosis in different cancers by combining protein interaction networks with gene expression data. MKGI [20] leverages the gene interaction networks in biochemical pathways to drive survival analysis of ovarian cancer. GCGCN [21] utilizes the Graph Convolutional Network (GCN) to predict the prognosis of BRCA and LUSC, which uses the similarity matrix between sample features as the raw input. Despite the progress of these approaches, to date, little effort has been made to investigate pan-cancer survival prediction by applying the rich biomolecular interaction priors to GCNs.

Based on the above considerations, in the present study, we develop an improved cancer survival prediction approach based on GCNs, termed Cancer suRvival prEdiction uSing Convolutional nEural NeTworks (CRESCENT). The structure of its gene expression network is derived from a rich protein-protein interaction (PPI) network, whose node features correspond to the gene expression profiles of patients. Importantly, GCN is capable of effectively learning the gene expression networks of patient genes and generating survival probabilities at different times. The original discrete-time survival model and ranking loss are combined into a hybrid loss function in our model. Furthermore, a new ranking loss function is designed and harnessed to allow the neural network to compare gene expression networks of different patterns across patients, thereby greatly facilitating GCN to identify survival-related sub-networks in pan-cancer patients.

The major contributions of the proposed work are summarized as follows:

1) This paper proposed a new GCN-based cancer survival prediction model, which mimics the natural framework of organisms. By incorporating the PPI prior knowledge networks, we transform the patient survival prediction as a graph classification problem.
2) Two survival loss functions are designed to guide the graph learning. The first loss function on individual survival error is used to ensure the predicted survival matches with the clinical record, while the second loss function on inter-patient survival ranking is to ensure consistent survival rankings. The two loss functions operate synergistically to improve the pan-cancer survival prediction accuracy.
3) To enhance the model interpretation underlying the decision of the survival prediction, the Shapley value was employed and used to interpret model predictions. Accordingly, we identified the genes and biological processes that the model relied on to make short- and long-term survival probability decisions for patients.

The rest of this paper is organized as follows: our proposed CRESCENT[1] method is introduced in Section II. Benchmarking experiments are conducted to investigate the performance of CRESCENT in comparison with different existing methods in Section III. Finally, Section IV draws a conclusion.

II. Materials and Methods

A. Datasets

Herein we introduce three PPI datasets and TCGA Pan-cancer dataset. These datasets are used to construct an undirected weighted graph for each cancer patient, as illustrated in Fig. 1. In such graph, each gene in the expression profiles represents a node in the graph, where the edges are established according to the structural relationships provided by the PPI networks.

1) Graph Construction From PPI Datasets: Three PPI public databases provide the constructed gene adjacency matrices for gene expression networks, which are ConsensusPathDB [22],

[1][Online]. Available: https://github.com/Ringoio/CRESCENT

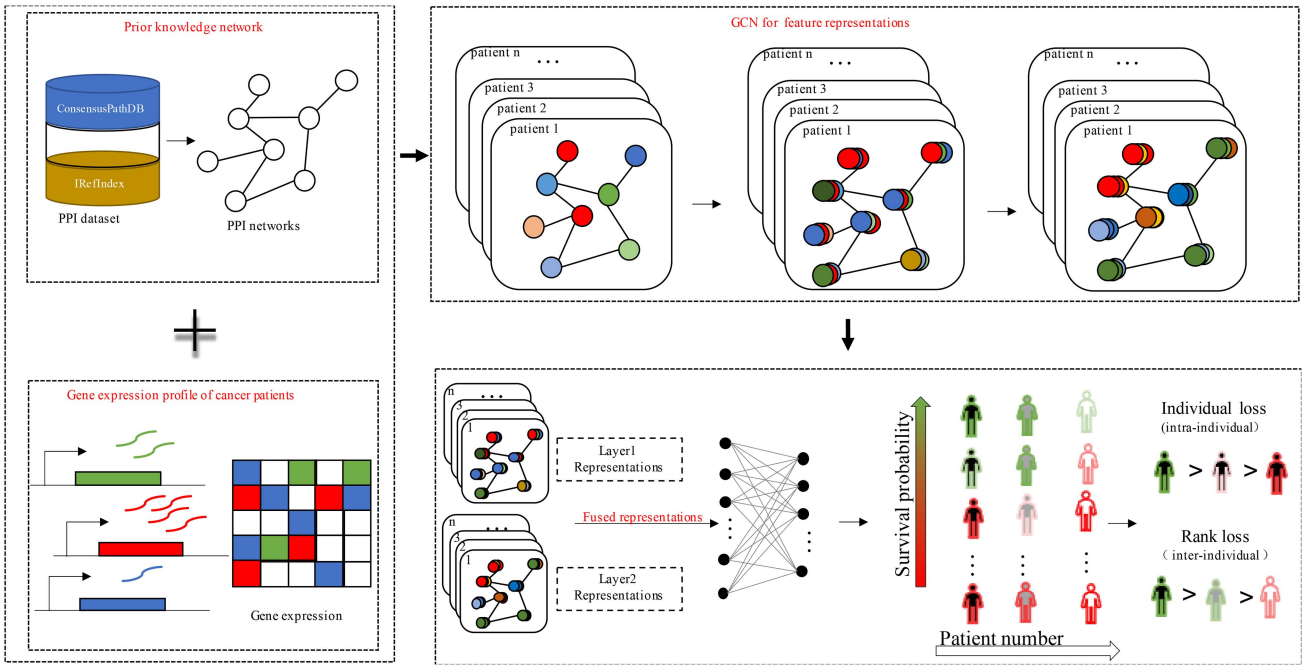

Fig. 1. Overview of the proposed CRESCENT method. Its development involves several major steps: First, the gene expression networks of patients were constructed using the PPI network data and the gene expression profiles of pan-cancer patients in TCGA. Then, the GCN model learns the features of the patients in the gene expression networks. Finally, the features learned by GCN are further fused to predict patient survival through a fully connected network. Furthermore, our CRESCENT method is able to predict the patient survival probability outputs through an individual loss function and an inter-individual ranking loss function.

TABLE I
STATISTICAL SUMMARY OF THREE PPI DATASETS CURATED IN THIS STUDY

	Number of nodes	Number of interactions	Number of high-confidence pairwise interactions
ConsensusPathDB	14092	521150	344862
STRING-db	12751	11759454	643574
IRefIndex	17837	500325	500325

STRING-db [23] and IRefIndex [24]. Table I provides a statistical summary of the included proteins and their interactions in the three PPI databases.

We filtered the complex protein interactions with more than two interacting partners in the ConsensusPathDB database and only selected those interactions with a confidence level of greater than 0.5. Likewise, we only retained high-confidence interactions with an interaction score of greater than 850 in STRING-db. For the IRefIndex database, we downloaded and used the version v.16.0, further removed the redundant interaction information and only considered the interactions between any two human proteins. We applied the python package mygene [25] to map the UniProt gene names to the corresponding Ensembl IDs, thereby simplifying the alignment of the adjacency and gene expression matrix.

Finally, after applying the aforementioned procedures, we obtained high-confidence PPI pairs. These high-confidence PPI pairs can be represented as an adjacency matrix $A \in R^{N \times N}$ of the graph, where N represents the number of genes. Next, we normalized the weights in the adjacency matrix in a row-wise manner. With the availability of the normalized adjacency matrix, we constructed a graph in which the nodes denote proteins, the edges represent the connections between proteins, while the normalized confidence scores denote the weights of the edges.

2) Preprocessing of the TCGA Pan-Cancer and Survival Data: Integration and standardization of the whole transcriptome sequencing (i.e., RNA-seq) data is challenging due to the differences across the samples, complicated data processing, and various sources. In the current study, the RNA-seq data were derived from a previous work [26]. This dataset employed a unified processing pipeline for quantile normalization and batch correction of the TCGA transcriptome data, thereby eliminating the potential study-specific bias. As the input of GCNs, the alignment of the PPI adjacency matrix and the gene expression matrix is required. The nodes (i.e., genes) present in both matrices would be retained, while the edges corresponding to the deleted nodes would also be deleted.

The TCGA pan-cancer dataset contained 5991 cancer patients belonging to 16 different types of cancer. Each patient had expression data of up to 12411 genes. Table II shows the numbers of

TABLE II
STATISTICAL SUMMARY OF THE THREE PPI NETWORKS SHARED WITH THE TCGA GENE EXPRESSION MATRIX

PPI networks	No. nodes	No. edges
ConsensusPathDB	9770	253210
STRING-db	8634	199610
IRefIndex	11219	288479

nodes and edges of the three PPI networks after aligning with the gene expression matrix. We used the R package TCGAbiolinks [27] to extract the clinical information of the patients, which included the survival status (death or alive), time of death, and follow-up of the patients.

B. Graph Convolutional Network

As illustrated in Fig. 1, we constructed an undirected weighted graph for each cancer patient. In such graph, each gene in the expression profiles represents a node in the graph, while the edges are constructed based on the structural relationships provided by the PPI network.

We constructed the graph for each cancer patient using the prior PPI knowledge and gene expression profiles. Specifically, for each graph $G(V, E)$, V is the set of nodes and E is the set of edges, respectively. Each node in the set V has a one-to-one correspondence with the proteins of the gene expression matrix. The gene expression matrix can be represented as $X \epsilon R^{N\times D}$, where N represents the number of gene nodes, while D represents the dimension of the gene expression data ($D = 1$), respectively. The adjacency matrix of the edge set E from the PPI database is denoted as $A \epsilon R^{N\times N}$.

GCNs [28] apply convolutional operations to the graphs, including spectral methods and spatial methods. The spectral decomposition method for collecting the node information depends on the decomposition of the graph Laplacian matrix. An approximate spectral graph convolution propagation rule in a GCN can be defined as:

$$h^{(l)} = \sigma\left(Lh^{(l-1)}W^{(l)}\right) \tag{1}$$

where the normalized graph Laplacian matrix L is defined as $L = D^{-1/2}\hat{A}D^{-1/2}$, and $\hat{A} = A + I$ in (1). $I\epsilon R^{N\times N}$ is the identity matrix and D is a degree matrix, thus $D_{ii} = \sum_j A_{ij}$. When $l = 1$, $h(l-1)$ is equal to gene expression matrix X, that is, $h^{(0)} = X$. Similar to a general neural network, W is a randomly initialized weight matrix, while σ represents a nonlinear activation function (relu).

We used (1) as the graph convolutional layer to extract the informative features from the gene expression networks of cancer patients. As our final goal is to output the probability of patient survival in discrete time periods, a pooling operation is required to complete this graph classification task. We decided to use the average pool to perform the pooling operation by allowing the network to take into account the majority of the genes. The pooling operation is a kind of down-sampling through which the feature extracted by GCN could be reduced to the survival vector. In addition, we also added a batch-norm layer between the GCN layer and the pooling layer to boost the generalization capability of the network during batch training.

In order to avoid the problem of over-smoothing in deep GCNs [29], we built a two-layer model ($l = 2$). Over-smoothing is a common issue in graph neural networks, as shown in several studies ([30], [29], [31]). To verify the relationship between the over-smoothing problem and number of the model's layers, we used different numbers of layers to model cancer survival prediction and found that a two-layer model achieved the best performance. Therefore, we applied this two-layer model to perform the follow-up experiments. In this model, each layer contained a graph convolutional layer, a batch-norm layer, and a global pooling operation. The features extracted by two GCN layers were subsequently fused and passed onto a fully connected layer to generate the survival probability of patients at different time intervals. In terms of the final output, an n-dimensional vector $y_{pred}(y_{pred}(1), \ldots, y_{pred}(n))$ represents the probability of survival of an individual patient over n time periods.

C. Joint Survival Loss Functions to Learn Individual Survival Probability and Consistent Survival Rank

Our experiments used $n+1$ time points to divide the n time intervals of the same length ($n = 19$). In particular, we divided the survival time of all patients into n time intervals $(t_1, t_2 \ldots t_n)$. The last n neuron outputs of the model corresponded to the survival probabilities at the respective n time intervals. We employed the discrete-time survival model [17] to construct a loss function ($L_{individual}$). For an individual sample, the loss function $L_{individual}$ requires two fitting labels ($surv_s$, $surv_f$) from the patient's known failure or censoring time. The length of the vector $surv_s$is n, which represents n time intervals of the individual survival. The vector $surv_f$ also has a length of n, which represents intervals of failure. In the time interval t_j, $surv_s\ (t_j) = 1$ indicates that survival status of the sample in the t_j is alive, otherwise the value is 0. The $surv_s(t_j)$ is also recorded as 1 if the censored patient survives for more than half of the time in the t_j. Moreover, $surv_f\ (t_j) = 1$ indicates that the uncensored patient died within this time interval t_j, otherwise the value is 0. For the output y_{pred} of each individual, it has the following negative log likelihood:

$$L_{individual} = -\sum_{i=t_1}^{t_n} \begin{array}{l} \ln\left(1 + surv_s\left(i\right)\cdot\left(y_{pred}\ \left(i\right) - 1\right)\right) \\ +\ln\left(1 - surv_f\left(i\right)\cdot y_{pred}\ \left(i\right)\right) \end{array} \tag{2}$$

where the first term encourages the probability of survival when the patient survives, while the second term suppresses the probability of survival when the patient dies.

The loss function $L_{\text{individual}}$ is used to penalize the empirical survival errors for each individual sample. However, it does not consider the competitive relationships among the samples. In particular, the model does not know how survival differences would be observed between the pan-cancer patient groups, because the $L_{\text{individual}}$ only helps to predict a single patient's own probability distribution of survival. To address this, a second loss

function L_{rank} is accordingly designed to penalize the ranking difference between the samples.

For the sample m that failed at the interval t_{fail_m}, the cumulative survival probability of the sample m is $P_{sample_m}(t_{fail\_m}) = \sum_{i=t_1}^{t_{fail\_m}} y_{pred}(i)$. The cumulative survival ranking of the failed sample m at the interval t_{fail_m} should be lower than that of the surviving sample n at the same interval. Therefore, we define this ranking loss function L_{rank} as follows:

$$\sum_{m \neq n} Ind\,(m,n) \cdot \frac{exp\,(P_{sample_m}(t_{fail_m}) - P_{sample_n}(t_{fail_m}))}{\mu} \tag{3}$$

where μ is a small constant. The process of minimizing the exponential function guarantees the ranking of $P_{sample_m}(t_{fail\_m}) < P_{sample_n}(t_{fail\_m})$. $Ind(m,n)$ is an indicator function on the difference between a pair of the patients m and n, defined as follows:

$$Ind\ (m,n) = \begin{cases} 1 & if\ patient\ m\ dies\ before\ patient\ n \\ 0 & otherwise \end{cases} \tag{4}$$

where the survival time of the patient m is less than that of the patient n, $Ind_{mn} = 1$, otherwise $Ind_{mn} = 0$. By minimizing (3), the correct ordering of patient survival time can be improved. The two loss functions, i.e., the intra-individual loss $L_{individual}$ and inter-individual L_{rank} are used jointly to guide survival probability prediction:

$$L = L_{individual} + \beta L_{rank} \tag{5}$$

where β is a coefficient used to reconcile the penalties of the two loss functions. In a training batch, $L_{\text{individual}}$ penalizes each patient's erroneous distribution of survival probability predictions, encouraging them to conform to their records. L_{rank} penalizes incorrect ordering of pairs in this training batch, encouraging correct ordering of pairs. Incorporating L_{rank} into the total loss function would force the model to learn each pair of samples in the pan-cancer data in a comparative manner. The differences in the sample pairs allowed the model to understand the associations of gene expression networks with the differences in the survival among cancer patients.

We used two Nvidia GTX 1080Ti to train our GCNs. The dataset was divided into the training and test datasets with the ratio of 4:1. We performed five-fold cross-validation on the dataset. The hidden layer dimension of GCN was set to 20, and the learning rate of the neural network was set to 0.01, respectively. Deep learning was implemented using PyTorch v1.6.0, and the Adam algorithm [32] was used to optimize the hyperparameters during the training process.

An overview of our proposed CRESCENT approach is illustrated in Fig. 1. As can be seen, it consists of three major parts: The first part is acquisition of the adjacency matrix and gene expression matrix of cancer patients, which requires the preprocessing and feature alignment operations. After that, the PPI network adjacency matrix and gene expression matrix are combined as the input to a two-layer GCN. At the final step, the fused features from the GCNs are fed into the fully connected layer, and the resulting outputs represent the patient's survival probability during a certain period of time.

TABLE III
IMPACT OF INCORPORATING THE PPI NETWORK ON THE PREDICTIVE PERFORMANCE

Method	C^{td}
MLP	0.696
GCN (random network)	0.702
CRESCENT (ConsensusPathDB)	0.760

III. EXPERIMENTS AND RESULTS

A. Evaluation Criteria

The time-dependent concordance index (C^{td}) [33] is used as a primary metric of the performance of survival prediction. C^{td} measures the similarity between the actual survival time ordering and the predicted survival ordering. C^{td} can be accordingly defined based on the survival probability prediction as follows:

$$\frac{\sum_{m \neq n} Ind_{mn} \cdot sign(P_{sample_m}(t_{fail_m}) < P_{sample_n}(t_{fail_m}))}{\sum_{m \neq n} Ind_{mn}} \tag{6}$$

where the expression in the $sign$ function takes 1 if it is true, and zero otherwise. The maximum value of C^{td} is 1. The closer its value is to 1, the more accurate the model prediction result.

B. Performance Evaluation of CRESCENT

All evaluation experiments were performed on the datasets from [26]. There were 5991 patient samples enrolled in the current study. The average C^{td} of five experiments was used to measure the performance of the models trained using different PPI datasets and methods.

1) Impact of Incorporating the PPI Network on the Predictive Performance: We evaluated the impact of incorporating the PPI network on the predictive performanc of the final model in Table III. In the first attempt, we employed multi-layer perceptron (MLP) that directly utilized gene expression data as input (i.e., without using the network topology information) to train the model. Not surprisingly, it only achieved a poor prediction accuracy ($C^{td} = 0.695$). In our second attempt, we incorporated the network topology by training the model based on graph convolutional layers (GCN), where the graph structure was constructed by a random adjacency matrix. As a result, the model achieved a performance ($C^{td} = 0.702$) slightly better than that of MLP ($C^{td} = 0.695$), which may be attributed to use of the random graph structure that did not align well with the biological context. The CRESCENT model is our final proposed model by incorporating the PPI network from ConsensusPathDB described in the Section II (Table I). ConsensusPathDB -based CRESCENT achieved the best C^{td} reaching 0.760, which indicated the effectiveness of incorporating the aligned biological context as prior during the survival analysis process. We conducted 5-fold cross-validation test and calculated the mean of the C^{td} over 25 iterations of these three methods. Additionally, we

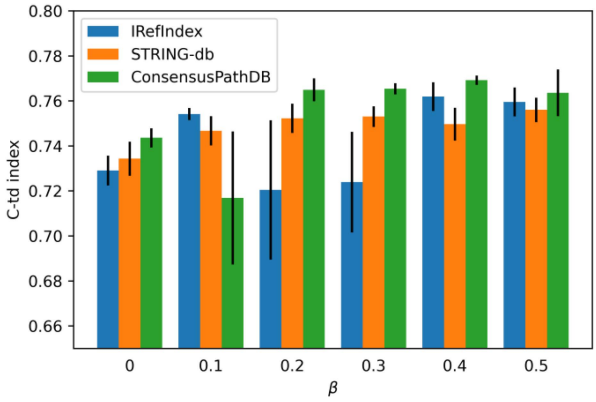

Fig. 2. Performance comparison in terms of the C^{td} index on the three PPI datasets, i.e., IRefindex, STRING-dab and consensusPathDB.

TABLE IV
PERFORMANCE COMPARISON OF C^{td} AMONG FIVE DIFFERENT SURVIVAL PREDICTION METHODS

Method	C^{td}
CPH	0.731* (0.709-0.753)
RSF	0.720* (0.696-0.742)
DeepSurv	0.744* (0.721-0.766)
DeepHit	0.752* (0.728-0.775)
CRESCENT (ConsensusPathDB)	**0.760 (0.736-0.783)**

* indicates p-value<0.05
The bold values represent our proposed method. We use bold formatting to emphasize and distinguish it.

also made an effort to maintain the consistency of the parameters across different methods. Specifically, we set the number of linear layers = 3, matching the layer number used in the GCN and CRESCENT models. For CRESCENT, we designed two graph convolutional layers followed by a fully connected layer. While in the case of GCN, we followed the same structural configuration as CRESCENT but with different input graph structure between CRESCENT and GCN.

2) Performance Comparison on the Three PPI Datasets: We further expanded our analysis by incorporating several PPI networks from different sources as shown in Table I. We aimed to further validate the contribution of integrating PPI networks as the underlying network topology encompassing gene interactions. The performance results is illustrated in Fig. 2. In particular, we trained the networks with different values of β (0, 0.1, 0.2, 0.3, 04, 05) and subsequently compared the C^{td} resulting values on the test dataset. Fig. 2 shows the means of the C^{td} on the three PPI datasets. We can see that the GCN models trained on these three datasets achieved C^{td} of greater than 0.7, indicating that the GCN models could reliably predict cancer survival when being used in combination with the PPI networks. The results also indicate that L_{rank} could indeed improve the model performance in most cases, suggesting the effectiveness of this ranking loss function. The GCN model achieved the overall best performance on the ConsensusPathDB dataset, with a mean C^{td} value of 0.769 when $\beta = 0.4$. The adjacency matrix provided by STRING-db achieved a stable performance with $the\ C^{td}$ values ranging from 0.746 to 0.756. As a comparison, the GCN model trained on the IRefindex dataset reached a C^{td} value of 0.720 ($\beta = 0.2$), which was the worst. Thus, the PPI network from the IRefindex dataset did not improve the predictive performance, presumably due to the low-confidence protein-protein interactions of the dataset.

3) Performance Comparison With Previous Methods: According to the performance results from the aforementioned experiments, we used ConsensusPathDB for constructing the graph structure in CRESCENT, which has been demonstrated to lead to the most stable and accurate performance. In this section, four typical survival prediction methods were included in performance comparison. These included two traditional machine learning methods (CPH [5]) and RSF [6]) and two representative deep learning methods (DeepSurv [10] and DeepHit [34]). For the compared deep learning methods, the hyperparameters given in the corresponding papers were used. We provide the mean and 95% confidence interval for the C^{td} over 25 iterations of the 5-fold cross-validation in Table IV. As can be seen, of the five methods, three deep learning methods outperformed the other two ones, and in comparison, the RSF method achieved the worst performance. Amongst the five methods, our proposed CRESCENT method achieved the best performance with the mean C^{td} of 0.760 and p-value$<$0.05.

C. Visualization of Representations Generated by GCNs

To visualize the representations of patient data extracted by GCNs, we embedded the fused features learned from the two-layer GCNs into a two-dimensional space using the t- distributed stochastic neighborhood embedding (t-SNE) algorithm [35]. Fig. 3 shows the visualizations of the GCN output composed of three PPI adjacency matrices after the dimensionality reduction using t-SNE. As can be observed, the clusters of the patients with the same feature representation had a larger likelihood of belonging to the same cancer type.

We further analyzed the survival prediction outputs for four types of cancer patients on the ConsensusPathDB dataset. Fig. 4 illustrates the visualizations of feature embeddings for these four cancer types and their respective survival probability prediction curves. We colored six outliers from the four cancer clusters and their survival curves. It can be seen that the outliers of the lung squamous cell carcinoma (LUSC) cluster were close to the outliers of the uterine corpus endometrial carcinoma (UCEC) cluster, and both patients were located at the top of the curve of survival probability prediction. In addition, there also existed two outliers in the rectum adenocarcinoma (READ) cluster, and the survival probability prediction of the two patients was similar, both of which were positioned below the survival prediction curve. The two outliers of the kidney renal papillary cell carcinoma (KIRP) cluster also had similar survival probability predictions (Fig. 4).

D. Performance Comparison of Models Based on the Pan-Cancer and Cancer-Specific Training Data

In this section, we examined whether training on the pan-cancer data could help improve survival prediction compared with cancer-specific data. The seven cancer types with the most

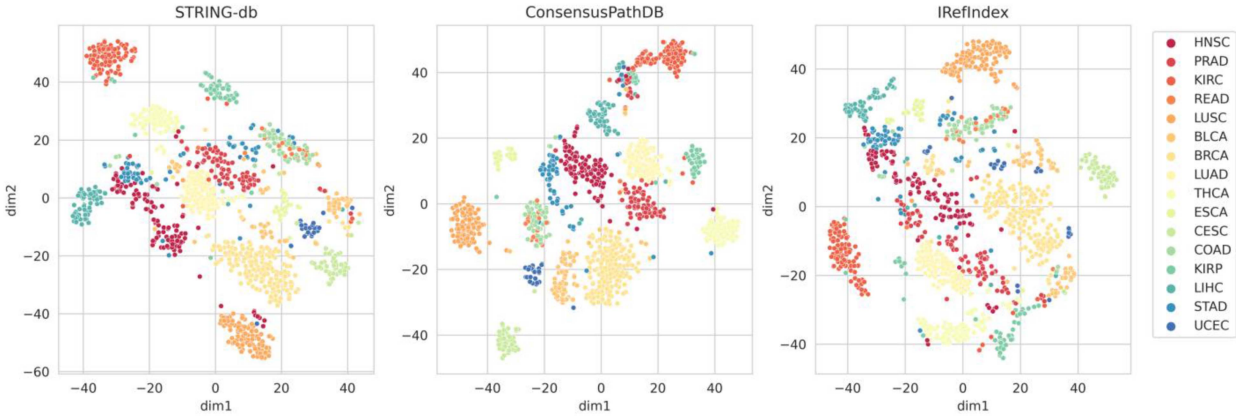

Fig. 3. Visualizations of feature representations generated by the GCN models using *t*-SNE.

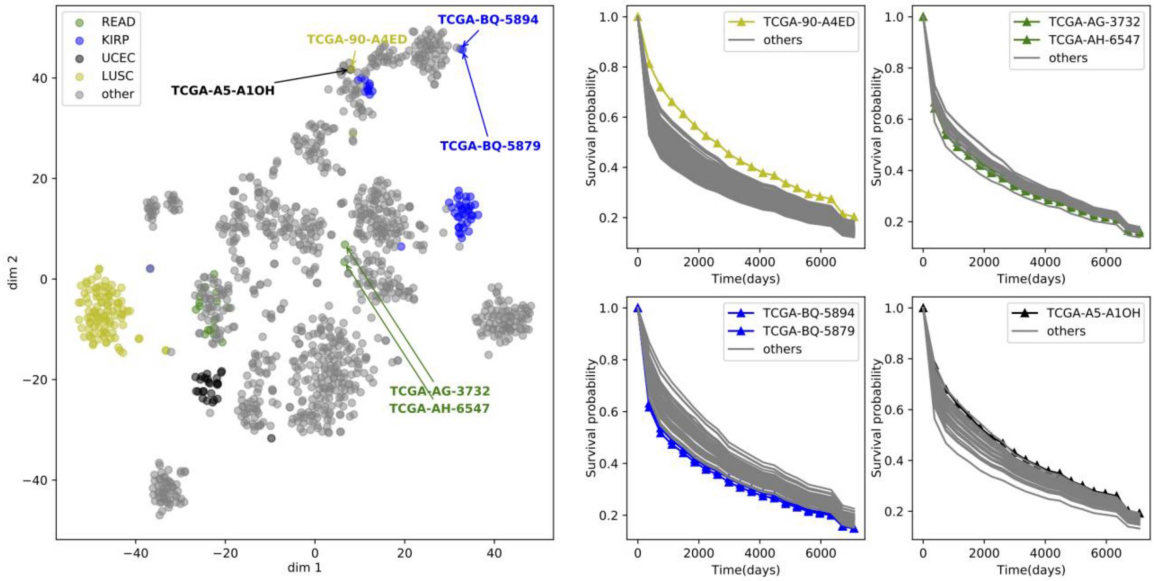

Fig. 4. Feature embeddings on the ConsensusPathDB dataset and prediction of survival probability for four different types of cancers, i.e., PRAD, READ, KIRP, and LUSC.

TABLE V
PERFORMANCE COMPARISON OF MODELS TRAINED ON CANCER-SPECIFIC DATA IN TERMS OF C^{td}

Cancer type	ConsensusPathDB	STRING-db	IRefIndex
BRCA	0.602	0.618	0.614
LUAD	0.632	0.671	0.647
LUSC	0.574	0.593	0.591
KIRC	0.658	0.689	0.695
HNSC	0.644	0.673	0.572
THCA	0.724	0.750	0.677
PRAD	0.696	0.663	0.627

significant number of patients were selected for respective training and testing. The results are provided in Table V. We can see that only the C^{td} value of thyroid carcinoma (THCA) exceeded 0.7. However, when being trained on the pan-cancer dataset, the performance of C^{td} could be further improved by up to 34% (Table V). These results suggest that training the models using the pan-cancer data could overcome the low accuracy issue, which is the case when training the model using the single cancer data. The results indicate that survival-related features across different types of cancers can be shared and leveraged to ultimately improve the pan-cancer survival prediction accuracy.

E. Functional Interpretation of Relevant Features

To illustrate the biological significance of the decision-making process of our method, we extracted the most relevant subsets of the gene network according to the Shapley values [36]. Shapley value is derived from coalitional game theory and has been widely used to interpret deep learning models in the fields of computer vision and natural language processing [36]. Here, we calculated the Shapley values of all input genes from the first neuron (short-term survival probability) and the last neuron (long-term survival probability) in the output layer of our model.

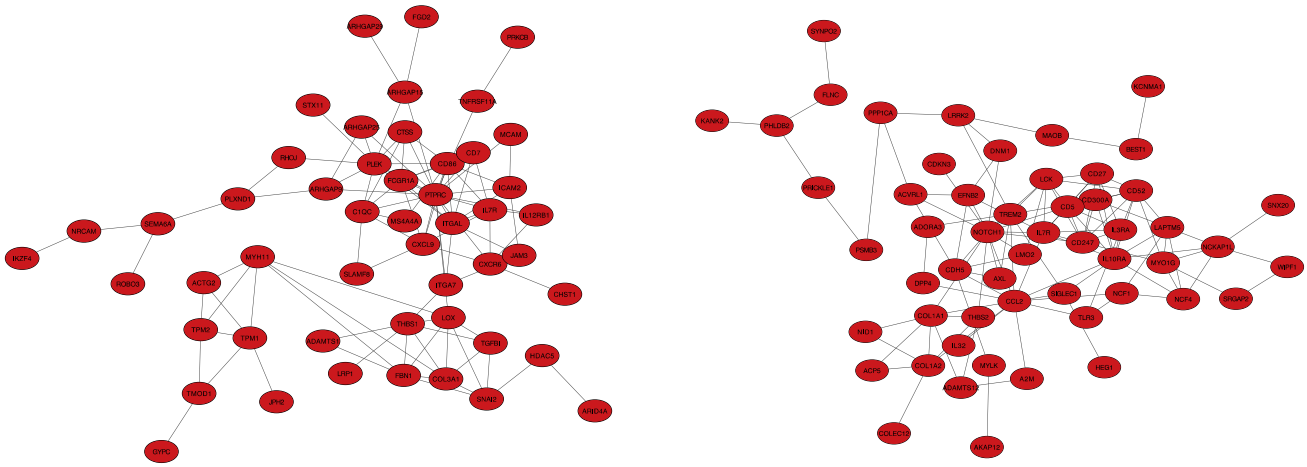

Fig. 5. Two exemplar PPI networks. The left network is constructed by genes that positively contribute to short-term survival, while the right network is assembled using genes that positively contribute to long-term survival.

We further identified the top 100 genes that made a positive contribution to the outputs of the two neurons, respectively.

First, we constructed PPI networks for the two core gene groups. For the PPIs between the query proteins extracted from the STRING-db database, we only retained the largest main network of the query results (Fig. 5). We observed that the genes in the network were closely related to proliferation and metastasis of cancer tissue, and many of them have been previously reported as potential therapeutic targets. For example, in the first network, TPM1 has been suggested to be able to inhibit cancer cell proliferation and migration in lung, prostate. Further, rectal cancers ([37], [38], [39], [40]) showed that silencing of the gene CTSS inhibited the migration and invasion of gastric cancer cells. PTPRC was characterized as a key gene for cervical and colon cancer metastasis ([41] and [42]). In the second network, COL1A1 has been shown to promote the metastasis of rectal, breast, and ovarian cancers ([43], [44], [45]). The study of [46] confirmed that downregulation of NOTCH1 contributes to the growth inhibition and apoptosis of pancreatic cancer cells. Another study suggested that knock-down of the AXL expression leads to reduced lung and breast cancer growth [47].

Next, we performed enrichment analysis of Gene Ontology (GO) biological process terms of the core genes. Accordingly, we identified 145 enriched GO biological process terms for both sets of the core genes, with the Benjamini-corrected p-value $<$ 0.05. Fig. 6 shows a comparison of the enriched GO biological processes of the two core gene sets. Amongst these, vascular development is most pronounced in the first neuron. This is understandable because the growth and progression of large tumor masses depend on tumor angiogenesis [48], and the formation of tumor blood vessels and lymphatic vessels contributes to the spread of the primary tumor [49]. As such, the first neuron harness the vascular development to assess tumor expansion of the patients. Amongst the enriched biological process terms of the last neuron, cell adhesion is the most remarkable biological process term. Changes in cell adhesion state can affect both the signal transduction state of cells and change in the polarity of cells [50]. In addition, decreased cell adhesion in cancer tissue is a hallmark of malignancy [51]. In this context, the last neuron

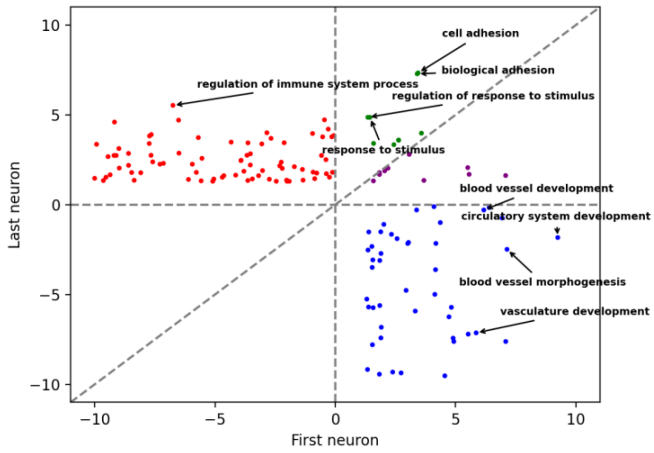

Fig. 6. Enriched GO biological processes identified from the positively contributing genes to the neuronal decision-making. The vertical and horizontal axes represent the evaluation scores (i.e., the negative logarithm of the p-value). Blue is the term corresponding to the first neuron, red is the term corresponding to the last neuron, and green and purple are the terms shared by both neurons.

may judge the status of cancer development by sensing the gene network related to cell adhesion status.

IV. Conclusion

In this study, we have proposed a new GCN-based cancer survival prediction approach, termed CRESCENT. It combines the PPI networks and gene expression profiles to construct an expanded graph of tissue gene expression to dictate each individual patient. The performance of our proposed method and four other methods was benchmarked and compared using the pan-cancer datasets. The performance benchmarking results demonstrated that our method outperformed the representative DNN structure-based methods. In addition, the GCN model trained using the pan-cancer data outperformed that trained using cancer-specific data. The main contribution of this work is manifested by the integration of prior knowledge in the form of PPI networks with the GCN models to improve cancer survival prediction.

By leveraging the outstanding learning ability of GCNs from structured graph data, the constructed GCNs could effectively learn the relationships between the genetic features of cancer patients. Nevertheless, this method still has some limitations, which should be addressed in the future work. For example, in our method, the average pooling was used to capture the effect of each graph extracted by GCN layers. Considering the differential degree of the influence of different genes on cancer development, pooling with node-selective selectivity might be more effective. In this study, only gene expression data that corresponded to the PPI networks were considered and utilized; however, there might exist other types of biomolecular interaction networks worthy of further interrogation. In the future work, we plan to develop effective strategies to incorporate a diverse collection of multi-omics data and their interaction networks to enable precise cancer survival prediction.

REFERENCES

[1] H. Sung et al., "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA:Cancer J. Clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[2] S. Cheon et al., "The accuracy of clinicians' predictions of survival in advanced cancer: A review," *Ann. Palliat. Med.*, vol. 5, no. 1, pp. 22–29, 2016.

[3] K. Sborov et al., "Impact of accuracy of survival predictions on quality of end-of-life care among patients with metastatic cancer who receive radiation therapy," *J. Oncol. Pract.*, vol. 15, no. 3, pp. e262–e270, 2019.

[4] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 8–17, 2015.

[5] D. R. Cox, "Regression models and life-tables," *J. Roy. Stat. Soc.: Ser. B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972.

[6] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random survival forests," *Ann. Appl. Statist.*, vol. 2, no. 3, pp. 841–860, 2008.

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

[8] O. Gevaert, F. D. Smet, D. Timmerman, Y. Moreau, and B. D. Moor, "Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks," *Bioinformatics*, vol. 22, no. 14, pp. e184–e190, 2006.

[9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[10] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, "DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network," *Biomed. Central Med. Res. Methodol.*, vol. 18, no. 1, pp. 1–12, 2018.

[11] T. Ching, X. Zhu, and L. X. Garmire, "Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data," *PLoS Comput. Biol.*, vol. 14, no. 4, 2018, Art. no. e1006076.

[12] J. N. Weinstein et al., "The cancer genome atlas pan-cancer analysis project," *Nature Genet.*, vol. 45, no. 10, pp. 1113–1120, 2013.

[13] M. T. Lu, A. Ivanov, T. Mayrhofer, A. Hosny, H. J. Aerts, and U. Hoffmann, "Deep learning to assess long-term mortality from chest radiographs," *J. Amer. Med. Assoc. Netw. Open*, vol. 2, no. 7, 2019, Art. no. e197416.

[14] P. Mukherjee et al., "A shallow convolutional neural network predicts prognosis of lung cancer patients in multi-institutional computed tomography image datasets," *Nature Mach. Intell.*, vol. 2, no. 5, pp. 274–282, 2020.

[15] I. Bichindaritz, G. Liu, and C. Bartlett, "Integrative survival analysis of breast cancer with gene expression and DNA methylation data," *Bioinformatics*, vol. 37, no. 17, pp. 2601–2608, 2021.

[16] A. Cheerla and O. Gevaert, "Deep learning with multimodal representation for pancancer prognosis prediction," *Bioinformatics*, vol. 35, no. 14, pp. i446–i454, 2019.

[17] M. F. Gensheimer and B. Narasimhan, "A scalable discrete-time survival model for neural networks," *PeerJ*, vol. 7, 2019, Art. no. e6257.

[18] L. A. Vale-Silva and K. Rohr, "Long-term cancer survival prediction using multimodal deep learning," *Sci. Rep.*, vol. 11, no. 1, pp. 1–12, 2021.

[19] 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," *Biomed. Central Genomic.*, vol. 16, no. 1, pp. 1–13, 2015.

[20] D. Kim, R. Li, A. Lucas, S. S. Verma, S. M. Dudek, and M. D. Ritchie, "Using knowledge-driven genomic interactions for multi-omics data analysis: Metadimensional models for predicting clinical outcomes in ovarian carcinoma," *J. Amer. Med. Inform. Assoc.*, vol. 24, no. 3, pp. 577–587, 2017.

[21] C. Wang et al., "A cancer survival prediction method based on graph convolutional network," *IEEE Trans. Nanobiosci.*, vol. 19, no. 1, pp. 117–126, Jan. 2020.

[22] A. Kamburov, K. Pentchev, H. Galicka, C. Wierling, H. Lehrach, and R. Herwig, "ConsensusPathDB: Toward a more complete picture of cell biology," *Nucleic Acids Res.*, vol. 39, no. 1, pp. D712–D717, 2011.

[23] D. Szklarczyk et al., "STRING v11: Protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D607–D613, 2019.

[24] S. Razick, G. Magklaras, and I. M. Donaldson, "iRefIndex: A consolidated protein interaction database with provenance," *Biomed. Central Bioinf.*, vol. 9, no. 1, pp. 1–19, 2008.

[25] C. Wu, I. MacLeod, and A. I. Su, "BioGPS and MyGene. info: Organizing online, gene-centric information," *Nucleic Acids Res.*, vol. 41, no. D1, pp. D561–D565, 2013.

[26] Q. Wang et al., "Unifying cancer and normal RNA sequencing data from different sources," *Sci. Data*, vol. 5, no. 1, pp. 1–8, 2018.

[27] A. Colaprico et al., "TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data," *Nucleic Acids Res.*, vol. 44, no. 8, p. e71, 2016.

[28] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*.

[29] Q. Li, Z. Han, and X.-M. Wu, "Deeper insights into graph convolutional networks for semi-supervised learning," in *Proc. 32nd Assoc. Advance. Artif. Intell. Conf. Artif. Intell.*, 2018, pp. 3538–3545.

[30] D. Chen et al., "Measuring and relieving the over-smoothing problem for graph neural networks from the topological view," in *Proc. Assoc. Advance. Artif. Intell. Conf. Artif. Intell.*, 2020, pp. 3438–3445.

[31] J. Zhou et al., "Graph neural networks: A review of methods and applications," *AI Open*, vol. 1, pp. 57–81, 2020.

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[33] L. Antolini, P. Boracchi, and E. Biganzoli, "A time-dependent discrimination index for survival data," *Statist. Med.*, vol. 24, no. 24, pp. 3927–3944, 2005.

[34] C. Lee, W. Zame, J. Yoon, and M. Van Der Schaar, "DeepHit: A deep learning approach to survival analysis with competing risks," in *Proc. Assoc. Advance. Artif. Intell. Conf. Artif. Intell.*, 2018, pp. 2314–2321.

[35] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, 2008, pp. 2579–2605.

[36] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–7.

[37] Y. Mao, J. He, M. Zhu, Y. Dong, and J. He, "Circ0001320 inhibits lung cancer cell growth and invasion by regulating TNFAIP1 and TPM1 expression through sponging miR-558," *Hum. Cell*, vol. 34, no. 2, pp. 468–477, 2021.

[38] Y. Dai and X. Gao, "Inhibition of cancer cell-derived exosomal microRNA-183 suppresses cell growth and metastasis in prostate cancer by upregulating TPM1," *Cancer Cell Int.*, vol. 21, no. 1, pp. 1–15, 2021.

[39] W. Liang, J. Wu, and X. Qiu, "LINC01116 facilitates colorectal cancer cell proliferation and angiogenesis through targeting EZH2-regulated TPM1," *J. Transl. Med.*, vol. 19, no. 1, pp. 1–13, 2021.

[40] Y. Yixuan et al., "Cathepsin S mediates gastric cancer cell migration and invasion via a putative network of metastasis-associated proteins," *J. Proteome Res.*, vol. 9, no. 9, pp. 4767–4778, 2010.

[41] S. Chen, C. Gao, Y. Wu, and Z. Huang, "Identification of prognostic miRNA signature and lymph node metastasis-related key genes in cervical cancer," *Front. Pharmacol.*, vol. 11, 2020, Art. no. 544.

[42] S. Chu, H. Wang, and M. Yu, "A putative molecular network associated with colon cancer metastasis constructed from microarray data," *World J. Surg. Oncol.*, vol. 15, no. 1, pp. 1–9, 2017.

[43] Z. Zhang, Y. Wang, J. Zhang, J. Zhong, and R. Yang, "COL1A1 promotes metastasis in colorectal cancer by regulating the WNT/PCP pathway," *Mol. Med. Rep.*, vol. 17, no. 4, pp. 5037–5042, 2018.

[44] J. Liu et al., "Collagen 1A1 (COL1A1) promotes metastasis of breast cancer and is a potential therapeutic target," *Discov. Med.*, vol. 25, no. 139, pp. 211–223, 2018.
[45] M. Li et al., "Microenvironment remodeled by tumor and stromal cells elevates fibroblast-derived COL1A1 and facilitates ovarian cancer metastasis," *Exp. Cell Res.*, vol. 394, no. 1, 2020, Art. no. 112153.
[46] Z. Wang, Y. Zhang, Y. Li, S. Banerjee, J. Liao, and F. H. Sarkar, "Down-regulation of Notch-1 contributes to cell growth inhibition and apoptosis in pancreatic cancer cells," *Mol. Cancer Therapeutics*, vol. 5, no. 3, pp. 483–493, 2006.
[47] Y. Li et al., "Axl as a potential therapeutic target in cancer: Role of Axl in tumor growth, metastasis and angiogenesis," *Oncogene*, vol. 28, no. 39, pp. 3442–3455, 2009.
[48] T. P. Padera, E. F. Meijer, and L. L. Munn, "The lymphatic system in disease processes and cancer progression," *Annu. Rev. Biomed. Eng.*, vol. 18, pp. 125–158, 2016.
[49] S. Hirohashi and Y. Kanai, "Cell adhesion system and human cancer morphogenesis," *Cancer Sci.*, vol. 94, no. 7, pp. 575–581, 2003.
[50] U. Cavallaro and G. Christofori, "Cell adhesion and signalling by cadherins and Ig-CAMs in cancer," *Nature Rev. Cancer*, vol. 4, no. 2, pp. 118–132, 2004.
[51] R. H. Farnsworth, M. Lackmann, M. G. Achen, and S. A. Stacker, "Vascular remodeling in cancer," *Oncogene*, vol. 33, no. 27, pp. 3496–3505, 2014.