Summer 7-25-2019

Biologically Interpretable, Integrative Deep Learning for Cancer Survival Analysis

Jie Hao

BIOLOGICALLY INTERPRETABLE, INTEGRATIVE DEEP LEARNING FOR CANCER SURVIVAL ANALYSIS

by

JIE HAO

Presented to the Faculty of the Graduate College of
Kennesaw State University in Partial Fulfillment
of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

Kennesaw State University
July 2019

Kennesaw State University
Graduate College

Thesis/Dissertation Defense Outcome

Name Jie Hao KSU ID 000665945

Email jhao2@students.kennesaw.edu Phone Number 423-737-6672

Program Ph.D. in Analytics and Data Science

Title

Biologically Interpretable, Integrative Deep Learning for Cancer Survival Analysis

Thesis/Dissertation Defense: Date 7/8/2019

☑ Passed ☐ Failed ☐ Passed With Revisions (attach revisions)

Signatures

DocuSigned by: Mingon Kang — July 13, 2019
Thesis/Dissertation Chair — Date

DocuSigned by: Donghyun Kim — July 14, 2019
Committee Member — Date

DocuSigned by: Jung Hun Oh — July 15, 2019
Committee Member — Date

DocuSigned by: Sherry Ni — July 15, 2019
Committee Member — Date

DocuSigned by: Herman Ray — July 17, 2019
Committee Member — Date

DocuSigned by: Sherrill Hayes — July 17, 2019
Program Director — Date

DocuSigned by: Jennifer Priestley — July 17, 2019
Department Chair — Date

M. Apr — 7/23/19
Graduate Dean — Date

Last Modified 05/26/16

Dedicated to my dearest grandfather Yunqi Li,
who passed away on 4th September 2016,
but who was always supporting me in
this monumental academic goal.

ABSTRACT

BIOLOGICALLY INTERPRETABLE, INTEGRATIVE DEEP LEARNING FOR CANCER SURVIVAL ANALYSIS

Jie Hao, Ph.D.

Kennesaw State University, 2019

Supervising Professor: Mingon Kang

Identifying complex biological processes associated to patients' survival time at the cellular and molecular level is critical not only for developing new treatments for patients but also for accurate survival prediction. However, highly nonlinear and high-dimension, low-sample size (HDLSS) data cause computational challenges in survival analysis. We developed a novel family of pathway-based, sparse deep neural networks (PASNet) for cancer survival analysis. PASNet family is a biologically interpretable neural network model where nodes in the network correspond to specific genes and pathways, while capturing nonlinear and hierarchical effects of biological pathways associated with certain clinical outcomes. Furthermore, integration of heterogeneous types of biological data from biospecimen holds promise of improving survival prediction and personalized therapies in cancer. Specifically, the integration of genomic data and histopathological images enhances survival predictions and personalized treatments in cancer study, while providing an in-depth understanding of genetic mechanisms and phenotypic patterns of cancer. Two proposed models will be introduced for integrating multi-omics data and pathological images, respectively. Each model in PASNet family was evaluated by comparing the performance of current cutting-edge models with The Cancer Genome Atlas (TCGA) cancer data. In

the extensive experiments, PASNet family outperformed the benchmarking methods, and the outstanding performance was statistically assessed. More importantly, PASNet family showed the capability to interpret a multi-layered biological system. A number of biological literature in GBM supported the biological interpretation of the proposed models. The open-source software of PASNet family in PyTorch is publicly available at `https://github.com/DataX-JieHao/`

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

CHAPTER 1

PATHWAY-ASSOCIATED SPARSE DEEP NEURAL NETWORK FOR PROGNOSIS PREDICTION FROM HIGH-THROUGHPUT DATA

1.1 Background

Predicting prognosis in patients from large-scale genomic data is a fundamentally challenging problem in genomic medicine [1, 2, 3]. Along with the rapid advances of high-throughput technologies and their effectivenesses, high-dimensional genomic data provides more accurate and richer biological descriptions of clinical phenotypes of interests than ever before. Therefore, translating large-scale genomic profiles to clinical outcomes not only improves predicting patient prognosis but also helps in identifying prognostic factors and biological processes.

The capabilities of high-level biological representation and interpretation of the prognosis are often more desired in biomedical research rather than merely improving predictive performance. Pathway-based analysis is an approach that a number of studies have been investigating to improve both predictive performance and biological interpretability [4, 5, 6]. In pathway-based analyses, the incorporation of biological pathway databases in a model takes advantage of leveraging prior biological knowledge so that potential prognostic factors of well-known biological functionality can be identified. Pathway-based analyses identify biological links between pathways and clinical outcomes and enable the interpretation of biological processes where their corresponding genes and proteins are involved. Thus, pathway-based interpretation and visualization provide an intuitive and comprehensive understanding of functionally-related molecular mechanisms.

Moreover, pathway-based approaches have shown more reproducible analysis results than gene expression data analysis alone [4, 7, 8, 9, 10]. High-level representations of gene co-expressions are considered in most pathway-based analyses; each

of which represents a biological pathway while preserving the original information. Thus, pathway-based analyses remedy the limitations of gene expression data, which are intrinsically sensitive to stochastic fluctuations and are often caused by multiple potential sources, such as inherent stochasticity of biochemical processes, environmental differences, and genetic mutation [11]. Pathway-based markers were proposed for classifying breast cancer metastasis and ovarian cancer survival time [5]. Cancer subtypes were discovered with pathway-based markers via Restricted Boltzmann Machine (RBM) [8]. A group LASSO-based approach associated genes with pathways and characterized them based on biological pathways [10]. Higher-order functional representation of pathway-based metabolic features provided reproducible biomarkers for breast cancer diagnosis [9].

However, reliable and accurate prognosis still remains poor in many diseases due to the following challenges: high-dimension, low-sample size data and complex nonlinear effects between biological components.

Genomic data are highly dimensional relative to their sample sizes. High-dimension, low-sample size (HDLSS) data often make prediction models sensitive to noise and false positive associations, which consequently make predicting accurate prognoses difficult. LASSO-based approaches have been mainly considered to estimate the effects of a gene set that are associated with various types of clinical outcomes on HDLSS data. The LASSO-based approaches embed sparse coding schemes into linear or logistic regression models for selecting few but greatly informative features among the high-dimensional data. For instance, a logistic regression with sparse regularization was applied for the prognostic model of mortality after acute myocardial infarction [12]. Random LASSO was proposed to enhance the LASSO solution by applying multiple bootstrapping and was applied to predict patients' survival times with glioblastoma gene expression data [13]. LASSO-based regression models as a prediction model were validated with multiple imputed data in chronic obstructive pulmonary disease patients [14].

Pathway-based analysis also helps to reduce data dimensionality. The number of biological pathways is relatively smaller than the number of genes, and a set of genes in the same pathway can be represented by the pathway's effect. Thus, pathways can be used as summary variables for the input of the predictive model instead of including all genes, which consequently reduces the model complexity.

Most association studies between a gene set and various clinical outcomes have considered linear or logistic regression models for identifying prognostic factors as well as understanding a biological mechanism of the progression of disease. However, nonlinear effects of genes or pathways may fail to be identified by linear-based approaches. As a solution, kernel-based models have been proposed to capture nonlinear effects of complex pathways [15, 16]. Multiple kernel learning models were introduced to aggregate complex effects from multiple pathways [17, 18]. Kernel Principle Component Analysis (KPCA) was applied to reduce the dimensionality of the feature space by using the correlation structure of the pathways [18].

Recently, several attempts to capture hierarchical effects of genes and pathways have been made. Inferences of multilayered hierarchical gene regulatory networks have been considered to understand how pathways regulate each other hierarchically. A bottom-up graphic Gaussian model [19] and a recursive random forest algorithm [20] were proposed to construct multilayered hierarchical gene regulatory networks. Moreover, complex biological networks were modeled by inferring the multiple hierarchical models (1) between gene expression and pathways and (2) within pathways [21]. However, complex hierarchical relationships between pathways have not been considered for prognostic studies yet, to the best of our knowledge, although hierarchical effects of pathways are prevalent in biological systems [22].

In this chapter, we propose a Pathway-Associated Sparse Deep Neural Network (PASNet) to achieve the goals:

- to predict prognosis in patients accurately by incorporating biological pathways,

- to provide a solution for hierarchical interpretation of nonlinear relationships between biological pathways of disease systematically, and
- to handle with computational problems, such as HDLSS data.

An innovative aspect of our model is biological interpretability; we achieved this with sparse coding and by constructing hidden layers with biological pathways, which oppose the *black box* nature of deep learning. Our new sparse deep learning architecture represents multiple molecular biological layers, such as a gene layer and a pathway layer, along with their hierarchical relationships, which use sparse regularization.

1.2 Related Works in Deep Learning

In recent years, deep learning has been spotlighted as the most active research field in various machine learning communities, such as image analysis, speech recognition, and natural language processing as its promising potential is being actively discussed in bioinformatics and biomedicine [23]. Most deep learning-based approaches have been developed for classification and association studies in bioinformatics. For instance, D-GEX infers the expression of target genes from landmark genes, capturing the nonlinear relationships by combining gene expression, DNA methylation, and miRNA expression data [24]. A convolutional neural network (CNN) was adapted to predict DNA-protein binding sites with Chromatin Immunoprecipitation sequencing (ChIP-seq) data [25]. Additionally, CNN-based DeepBind was proposed to predict whether a specific DNA/RNA binding protein will bind to a specific DNA sequence [26]. The functionality of non-coding variants was predicted by DeepSEA by employing a CNN model [27].

Although only a small subset of deep learning research has been reported in bioinformatics due to the difficulty of structure definition and interpretation, the future of deep learning in biology and medicine is promising [28]. First, since a neural network is inspired by the neurons in the human brain, a neuron network architecture is applicable to modeling a mechanism for a complex biological system. Specifically,

deep learning approaches take advantage of flexible representation of hierarchical structures from inputs to outputs. The representation of nonlinear effects of neurons in multiple layers in neural networks may be able to model hierarchical biological signals. DCell constructs a multi-layer neural network based on extensive prior biological knowledge to simulate the growth of a eukaryotic cell [29]. However, DCell's network architecture is entirely based on well-known prior biological knowledge, so the model was applied to relatively simple biological system of yeast. Moreover, deep learning captures nonlinear effects of variables with high-level feature representation, which allows deep learning to outperform other state-of-the-art methods.

However, training deep neural networks with HDLSS data poses a computational problem. A large number of parameters are involved in deep neural networks, and it often makes the training infeasible or causes a model overfit on HDLSS data. Particularly, backpropagation gradients in neural networks are of high variance on HDLSS data, which consequently causes the model overfit [30]. In order to tackle the HDLSS problem, the leave-one-out approach was used to avoid the overfitting problem in backpropagation [31]. Regarding backpropagation, the risk of overfitting was examined with validation data by the leave-one-out approach and terminates the training early when overfitting occurs. For an alternative solution, an attempt to reduce the dimensionality of the input space to a feasible size has been made [32]. Dimension reduction techniques, such as subsampled randomized Hadamard transform (SRHT) and Count Sketch-base construction, were utilized to reduce the dimensional size of the input data. Then, the projected data into the lower space were introduced to a neural network for training.

For HDLSS data, feature selection is one of the conventional approaches. Deep Feature Selection (DFS) was developed to select a discriminative feature subset in a deep learning model [33]. Although DFS is not the optimal solution to low-sample size data, DFS shows that deep learning can detect informative and discriminative features of nonlinearity effects through multiple layers with high-dimensional data.

Then, Deep Neural Pursuit (DNP) improved the solution of the feature selection in deep learning, taking the HDLSS data problem into account [30]. DNP iteratively augments features in the input layer by performing multiple dropouts. The multiple dropouts grant the ability to train a small-sized sub-network at a time and to compute gradients with low variance for alleviating the overfitting problem.

1.3 Methods

Pathway-Associated Sparse Deep Neural Network (PASNet) identifies a subset of genes and pathways involved in a disease as prognostic biomarkers, as well as their interactions. PASNet models a multilayered, hierarchical biological system of genes and pathways on a disease, while leveraging the strengths of deep learning for competitive predictive performance. The sparsity of PASNet allows one to interpret the model, which is what conventional fully-connected networks lack. The architecture of PASNet and the strategies for training a sparse neural network model with HDLSS and imbalanced data are described.

1.3.1 The Architecture of PASNet

PASNet incorporates biological pathways and the concept of sparse modeling based on Deep Neural Network (DNN). The neural network architecture of PASNet consists of a gene layer (an input layer), a pathway layer that represents the biological pathways linked with input genes, a hidden layer that represents hierarchical relationships among biological pathways, and an output layer that corresponds with clinical outcomes, e.g. binary classes of long-term and short-term survival, stages of cancer (see Figure 1.1).

In PASNet, sparse coding is considered on the connections between layers for model interpretability. Sparse coding provides a solution to capture significant components of a biological mechanism in the model, since biological processes may involve

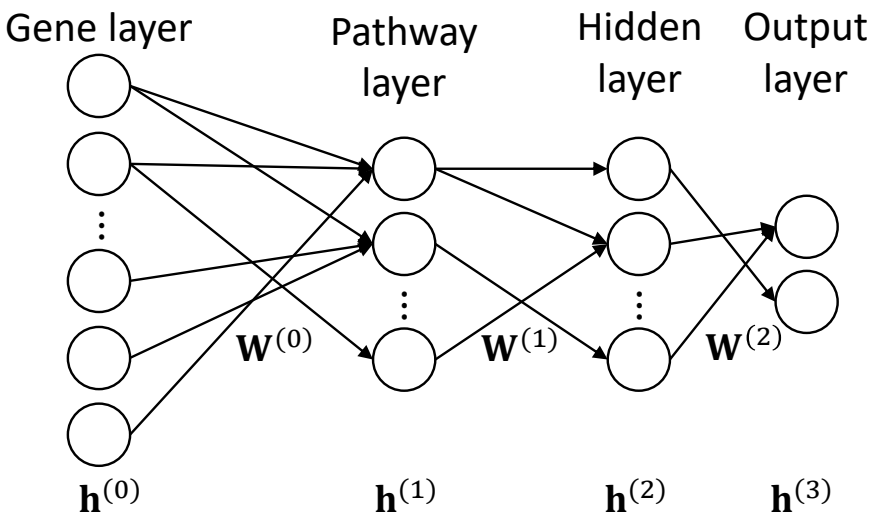

Figure 1.1: Architecture of PASNet. The structure of PASNet is constructed by a gene layer (an input layer), a pathway layer that represents the biological pathways linked with input genes, a hidden layer that represents hierarchical relationships among biological pathways, and an output layer that corresponds with clinical outcomes, e.g. a binary class that has long-term survival and short-term survival, stages of cancer.

only a few biological components. On the other hand, conventional fully-connected networks lack to represent biological mechanisms.

Gene Layer

The gene layer (as an input layer) corresponds to gene expression data. A patient sample of m gene expressions is formed as a column vector, which is denoted by $\mathbf{x} = \{x_1, x_2, ..., x_m\}$. Each input node represents one gene.

Pathway Layer

The pathway layer represents biological pathways, where each node indicates an individual pathway. The connections between the gene layer and the pathway layer are established by well-known pathway databases (e.g., Reactome and KEGG). Pathway databases contain associations between pathways and genes; each of which

provides a set of gene components. Therefore, the pathway layer makes it possible to interpret the model as a pathway-based analysis.

To begin with initializing the connections between the gene layer and the pathway layer, we consider a binary bi-adjacency matrix ($\mathbf{A}$) from biological pathway databases. The bi-adjacency matrix can be defined as $\mathbf{A} \in \mathbb{B}^{n \times m}$, where n is number of pathways and m is number of genes. Then, an element of $\mathbf{A}$, i.e., a_{ij}, is set to one if gene j belongs to pathway i; otherwise, zero. Sparse coding is applied based on the matrix $\mathbf{A}$ to represent the relationships between genes and pathways in the model.

1.3.1.1 Hidden Layer

Biological components may cooperate with others instead of functioning alone. A biological system involves multiple pathways which have interactions together, whereas a node in the pathway layer indicates a biological pathway. The associative interactions between pathways can be represented in the hidden layer. In PASNet, the hidden layer represents biological nonlinear associations between the pathways to outputs.

Sparse coding between the pathway and the hidden layers enables one to interpret these relationships. Although we consider only a single hidden layer in this study for simplicity's sake, multiple hidden layers can be used for deeper hierarchical representations of pathways. For example, if there are two hidden layers, the second hidden layer will represent deeper hierarchical associations of the nodes of the first hidden layer, which are association effects of pathways.

1.3.1.2 Output Layer

The output layer shows clinical outcomes for which nodes compute the posterior probabilities. In this layer, sparse coding allows to distinguish hierarchical groups of pathways (which are detected from hidden layers) to predict clinical outcomes. In

PASNet, more than two clinical outcomes can be easily represented with multiple nodes in the output layer.

Consequently, PASNet can dissect biological processes of hierarchical nonlinear relationships and associations of genes and pathways to predict clinical outcomes. This *generative* model-based approach would be useful to predict prognosis accurately with complex HDLSS data. Furthermore, the integration of the biological structures and prior knowledge to the model would produce a robust solution.

1.3.2 Overall Description of PASNet Training

The main challenge in training PASNet is to reduce both risk of overfitting and computational complexity of training on HDLSS data. The related works that have handled the HDLSS data problem are discussed in Section Related Work in Deep Learning. To unravel the problems, PASNet optimizes a small sub-network, which involves feasible nodes and parameters to train instead of the whole network and then makes the sub-network sparse. Figure 1.2 illustrates the overall training flow of PASNet.

First, we initialize the connections between the gene layer and the pathway layer with prior biological knowledge of pathways (see Figure 1.2(a)). Active/inactive connections are determined by the bi-adjacency matrix, $\mathbf{A}$. The weights of active connections and biases are randomly initialized from standard normal distribution, while the weights of inactive connections are set to zero. The sparsity of the connections between the gene layer and the pathway layer is invariant over the entire training. The remaining layers are fully interconnected as the initial.

In the training phase, we repeat training sub-networks and applying sparse coding on the sub-networks until convergence (Figure 1.2(b) – (c)). A sub-network is selected by a dropout technique, where neurons are randomly dropped in the intermediate layers. In Figure 1.2(b), a small sub-network is shown with bold solid circles and lines. Then, the small sub-network is trained by feed-forward and backpropaga-

tion. Note that only weights and biases of the sub-network are trained. Upon the completion of the sub-network's training, sparse coding is applied to the sub-network by trimming the connections that do not contribute to minimize the loss. In Figure 1.2(c), the dropped connections and nodes are marked as bold, dashed lines. The details of the training are elucidated in the following sections.

1.3.3 Sparse Coding

Once the small sub-network is trained with the HDLSS data, the sub-network is imposed to be sparse for the model interpretation. The sparsity of the sub-network is determined by the mask matrix $\mathbf{M}$ on each layer as:

$$\mathbf{h}^{(\ell+1)} = a\Big((\mathbf{W}^{(\ell)} \star \mathbf{M}^{(\ell)})\mathbf{h}^{(\ell)} + \mathbf{b}^{(\ell)}\Big), \tag{1.1}$$

where $\star$ denotes element-wise multiplication, and $a(\cdot)$ is an activation function. $\mathbf{h}^{(\ell)}$ denotes an output vector on the ℓ-th layer, and $\mathbf{W}^{(\ell)}$ and $\mathbf{b}^{(\ell)}$ are a weight matrix and a bias vector, respectively. An element value of $\mathbf{M}$ is either one or zero, which determines whether the associated weights are dropped in the current epoch.

The mask matrix $\mathbf{M}$ is generated with respect to a sparsity level (S) that indicates the proportion of weights to be dropped in a single layer. S is a value between 0 to 100 (e.g. [0, 10, ..., 100)), where zero creates a fully-connected layer while 100 causes no connection. The optimal S^* is approximated on each layer individually in the sub-network, while most related methods consider a single hyper-parameter for the sparsity of all layers [34, 35]. The individual setting of the sparsity on each layer shows different levels of biological associations on the genes and pathways.

We obtain the optimal sparsity level S^* that minimizes the cost score. For efficient computation, the cost scores are computed with a small number of finite sparsity levels. Then, the optimal sparsity level is estimated by applying a cubic-spline interpolation to the cost scores with the assumption that the cost function, with respect to the sparsity level, is continuous.

In particular, an element of $\mathbf{M}$ is set to one if the absolute value of the corresponding weight is greater than threshold Q; otherwise, the element is zero, where Q is an S-th percentile of absolute values of $\mathbf{W}$. Note that the mask between the gene layer and the pathway layer, i.e. $\mathbf{M}^{(0)}$, is determined by the bi-adjacency matrix $\mathbf{A}$ of biological pathways. Thus, the mask matrices are formulated as

$$\mathbf{M}^{(\ell)} = \begin{cases} \mathbb{1}(|\mathbf{W}^{(\ell)}| \geq Q^{(\ell)}), & \text{if } \ell \neq 0 \\ \mathbf{A}, & \text{if } \ell = 0 \end{cases} \tag{1.2}$$

where $Q^{(\ell)}$ is the S-th percentile of $|\mathbf{W}^{(\ell)}|$ if $\ell \neq 0$.

1.3.4 Cost-sensitive Learning for Imbalanced Data

We refine the cost function and the backpropagation for cost-sensitive learning, since imbalanced data causes bias of the predictions towards the majority class. We adapt the Mean False Error (MFE) method [36], which penalizes the errors of the majority class.

Let K be the number of clinical outcomes. The normalized cost is computed separately for each class by:

$$\mathcal{L} = \sum_{k=1}^{K} \mathcal{C}_k + \frac{1}{2}\lambda \|\mathbf{W}\|_2, \tag{1.3}$$

$$\mathcal{C}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} c(\mathbf{y}_i, \tilde{\mathbf{y}}_i), \tag{1.4}$$

where $\mathcal{C}_k$ denotes mean error on the class k, and n_k is the number of samples in the class k. $\mathbf{y}_i$ is a vectorized ground truth class label of the i-th sample, and $\tilde{\mathbf{y}}_i$ is its vectorized prediction. $c(\cdot)$ denotes a cost function (e.g., cross-entropy loss), and $\mathcal{L}$ is the total cost. $\|\mathbf{W}\|_2$ denotes a L^2-norm of $\mathbf{W}$, and $\lambda > 0$ is a regularization hyperparameter.

In the backpropagation phrase, the gradient is also computed separately for each class. Hence, the weights and biases on the ℓ-th layer are updated by:

$$\mathbf{W}^{(\ell)} \leftarrow (1-\eta\lambda)\mathbf{W}^{(\ell)} - \eta \sum_{k=1}^{K} \frac{\partial \mathcal{C}_k}{\partial \mathbf{W}^{(\ell)}}, \tag{1.5}$$

$$\mathbf{b}^{(\ell)} \leftarrow \mathbf{b}^{(\ell)} - \eta \sum_{k=1}^{K} \frac{\partial \mathcal{C}_k}{\partial \mathbf{b}^{(\ell)}}, \tag{1.6}$$

where η is a learning rate. The algorithm of PASNet is briefly described in Algorithm 1.

Algorithm 1 Training of PASNet

1: Initialize weights $\mathbf{W}^{(\ell)}$ and biases $\mathbf{b}^{(\ell)}$
2: $\mathbf{W}^{(0)} \leftarrow \mathbf{W}^{(0)} \star \mathbf{M}^{(0)}$
3: **repeat**
4: Select a small sub-network via dropout
5: Train the sub-network by Eq. (1.5) and Eq. (1.6)
6: Sparse coding with the optimal $\mathbf{M}^{(\ell)}$ by Eq. (1.2)
7: $\mathbf{W}^{(\ell)} \leftarrow \mathbf{W}^{(\ell)} \star \mathbf{M}^{(\ell)}$
8: **until** convergence

1.4 Results

We conducted experiments to evaluate PASNet's predictive performance for long-term survival prediction in Glioblastoma multiforme (GBM). The capability of the prediction was assessed by comparing our model with the classifiers that have been used for long-term survival prediction. Furthermore, we will describe how PASNet can represent the biological system of GBM in the following sections.

1.4.1 Data

GBM is a primary brain cancer that shows poor prognosis performance. Comprising more than half of all brain tumors, GBM is the most prevailing and aggressive malignant type of primary astrocytomas [37]. Patients with GBM have a median survival time of approximately 15 months with intensive treatments [38]. Furthermore, long-term survival patients with GBM are rare as more than 90% of patients are deceased within three years of diagnosis. Although treatments in neurosurgery, chemotherapy, and radiotherapy have improved, the prognosis of GBM remains poor [39]. Hence, the advancement in understanding molecular mechanisms and related biological pathways of GBM is significant to accelerating the progress for new treatments [38].

We used the gene expression data of GBM patients, which is available at The Cancer Genome Atlas (TCGA, `http://cancergenome.nih.gov`). The dataset includes the gene expression data of 522 samples and 12,042 genes and provides survival time and status. We considered patients who survived past 24 months (regardless of survival status) as long-term survivals (LTS) and patients that deceased in less than 24 months as short-term survivals (non-LTS). Living patients with a survival time of less than 24 months were excluded in the experiments and considered censored data. Finally, we obtained 99 LTS and 376 non-LTS samples, where around 20% of the samples were LTS patients.

For pathway-based analysis, we utilized a biological pathway database from the Molecular Signatures Database (MSigDB) [40, 41, 42]. In MSigDB, we extracted the biological pathways of Reactome. Then, we excluded the pathways that include less than ten genes, because small pathways are often redundant with larger pathways [43]. As the input features, we considered the genes that belong to at least one pathway, since pathway annotations of genes are essential to construct the mask matrix $\mathbf{M}$ between the gene layer and the pathway layer. Finally, we considered 574 pathways

and 4,359 genes in the experiments. The gene expression data were standardized to a mean of zero and a standard deviation of one.

1.4.2 Experimental Design

We followed a typical design of conventional deep neural networks for PASNet. A sigmoid function and cross-entropy were considered for the activation and the cost function, respectively. A softmax function was used in the output layer so that the probabilities of output nodes add up to one. For the optimal tuning of PASNet's training, we empirically determined the hyper-parameters by random search before cross-validation experiments. The learning rate (η) was set to 1e$-$4, and L^2 regularization (λ) was set to 3e$-$4. Adaptive Moment Estimation (Adam) was performed as the stochastic optimizer [44]. The dropouts for two intermediate layers were also applied with a dropping probability of 0.8 and 0.7, respectively. PASNet was implemented by PyTorch, and the source code is available at `https://github.com/DataX-JieHao/PASNet`.

1.4.3 Comparison

We evaluated PASNet by comparing the performance with classifiers that have been used for prognosis prediction: Support Vector Machine (SVM), Random LASSO [13], LASSO Logistic Regression (LLR) [1], and neural network with dropout (Dropout NN).

Specifically, we used a SVM with a radial basis function (RBF) kernel ($\gamma = 2^{-16}$ and $C = 2^{3.9}$ by two-step grid search [45]). Random LASSO was trained so that every feature could be selected 20 times on average by bootstrapping, and the L^1 regularization parameter was determined by 10-fold cross-validation. The LASSO parameter for LLR was also selected by 10-fold cross-validation. The fully-connected Dropout NN was designed with the same numbers of intermediate layers and neurons as the proposed PASNet as well as the dropout probabilities. The learning rate was 0.01

and the L^2 regularization was 0.005. Note that PASNet has less number of weights to be trained in each epoch because of sparse coding, compared to Dropout NN. Hence, the optimal hyper-parameters of L^2 regularization and learning rate should be different between PASNet and Dropout NN. We empirically searched the optimal hyper-parameters for PASNet and Dropout NN separately through multiple experiments. Dropout NN was implemented by PyTorch (`https://pytorch.org/`).

The experiments were carried out by stratified 5-fold cross-validation for maintaining the same proportions of the imbalanced samples in the classes. The cross-validation experiments were repeated ten times for performance reproducibility. Data preprocessing, such as data normalization, was separately applied on each fold. The testing data on each fold was scaled with the mean and standard deviation of the training data of the same fold.

The predictive performances of the five models were evaluated with two metrics: Area Under the Curve (AUC) and F1-scores. The Receiver Operating Characteristic (ROC) curve (see Figure 1.3) was traced over the thresholds of scores to examine the trade-off between True Positive Rate ($TPR = TP/(TP + FN)$) and False Positive Rate ($FPR = FP/(FP + TN)$), where LTS was considered positive. An AUC was computed by the area under the ROC curve. An F1-score, an average of Positive Predicted Value ($PPV = TP/(TP + FP)$) and TPR, is calculated by $2(PPV \times TPR)/(PPV + TPR)$. The F1-score was computed for the LTS class.

The average AUC and the average F1-score of the five methods on the test datasets are shown in Table 1.1. PASNet outperformed others as both AUC and F1-score are relatively high. PASNet produced AUC of 0.6622±0.013 (mean±std) and F1-score of 0.3978±0.016. Following PASNet, Dropout NN produced AUC of 0.6408±0.014, and SVM produced AUC of 0.6337±0.015.

To statistically assess the performance of PASNet (AUC) as compared to others, we conducted the Wilcoxon signed-rank test: a non-parametric paired, two sided test for the null hypothesis that states the median difference in paired samples is

Table 1.1: Comparison of AUC and F1-score in over ten stratified 5-fold cross-validations

Model	AUC	F1-Score
Logistic LASSO	0.5899±0.020	0.3347±0.025
Random LASSO	0.6209±0.020	0.3370±0.020
SVM	0.6337±0.015	0.3446±0.015
Dropout NN	0.6408±0.014	0.2957±0.025
PASNet	**0.6622±0.013**	**0.3978±0.016**

zero. Specifically, the null hypothesis is that the benchmark classifier has equal or better performance than our proposed algorithm. Table 1.2 shows the performance of PASNet is significantly better than others, where the null hypotheses are rejected at the 5% significance level (p-value < 0.05). Hence, the outperformance of PASNet was statistically significant compared to the benchmark classifiers.

Table 1.2: The Wilcoxon signed-rank tests for comparing PASNet with the Benchmark Classifiers

	W Statistic	P-value
PASNet vs. Dropout NN	146.5	2.13e-06
PASNet vs. RBF-SVM	137.0	1.35e-06
PASNet vs. Random LASSO	45.0	1.06e-08
PASNet vs. Logistic LASSO	43.0	9.52e-09

SVM and Dropout NN showed a higher AUC than LASSO logistic regression and Random LASSO, probably because of their capability of capturing nonlinear effects of genes. Compared to Dropout NN, PASNet is a relatively thin network, where the connections between layers are very sparse. However, PASNet interestingly produced higher performance than Dropout NN. It shows that PASNet builds a robust

network model, which is simplified to represent the biological processes for prognosis prediction by incorporating biological prior knowledge.

1.5 Model Interpretation in GBM

Although PASNet yielded competitive predictive performance in the experiments, a more promising contribution of PASNet is in the model's interpretability. In this section, we demonstrate a plausible biological mechanism inferred by PASNet for long-term survival prediction in GBM. The graphical representations of the PASNet model are illustrated in Figures 1.4–1.6 in the top-down order. The heatmaps were generated by sorting the weights and node values of LTS, and positive and negative weight values are colored in red and blue, respectively.

First, Figure 1.4 manifests the posterior probability of the samples in the clinical outcomes. The dark block on the top shows the output node values ($-log_2$(node value)) of the LTS samples, while the remaining ones are non-LTS samples. The weight values of the connections from hidden nodes to the output nodes are depicted in Figure 1.5(a), where dropped connections are colored in white. The figure reveals distinct patterns of weights (opposite signs) to the two output neurons. Note that there are hidden nodes disconnected to the neurons in the output layer (colored in white) by sparse coding, which shows that the hidden nodes are insignificant.

The hidden node values of the samples are shown in Figure 1.5(b). The values of the hidden nodes indicate the intensity of the group effects on the pathways, which are connected to the hidden nodes. For instance, the first 16 hidden nodes in Figure 1.5(b) show distinguishable intensities on LTS and non-LTS patients. The LTS patients present significant intensities of the group effects of the 16 pathways while non-LTS patients show significant lower values.

The weights between the pathway nodes and the hidden nodes are exhibited in Figure 1.5(c), and the top-10 ranked pathways among them are zoomed in Figure 1.6(a). It appears that a small number of pathways mainly contribute to the

hidden nodes simultaneously, which implies that the cohort of the pathways may be candidates of prognostic biomarkers in long-term survival of GBM. The top-10 ranked pathways include signaling by GPCR, GPCR downstream signaling, innate immune system, adaptive immune system, metabolism of carbohydrates, transmembrane transport of small molecules, developmental biology, metabolism of proteins, class A/1 (rhodopsin-like receptors), and axon guidance. Most of the pathways are referred to as significant pathways in GBM in biological literature. The pathways and the references are listed in Table 1.3. Since the top-10 ranked pathways are all large (gene numbers > 200), we further explored small pathways as well. Class B/2 (Secretin family receptors) pathway which includes 88 genes is ranked 14th. One of the subgroups in Class B/2 family is categorized as brain-specific angiogenesis inhibitors that are growth suppressors of glioblastoma cells [46]. Hence, Class B/2 pathway may play an important role in inhibition of GBM.

The genes of the pathways are illustrated by the weight values in Figure 1.6(b). Since the connections between the gene layer and the pathway layer are given by pathway databases, e.g., Reactome, they are very sparse. It also shows that multiple pathways share genes in common. The genes, which are most frequently shown in the ten pathways, include CDC42, PRKCQ, RAC1, AKT1, AKT2, AKT3, C3, CREB1, GRB2, HRAS, KRAS, NRAS, PRKACA, PRKACB, PRKACG, RAF1, and YWHAB, where CDC42, PRKCQ, and RAC1 are shown in six pathways and others are in five pathways. Among them, several genes have been reported as biomarkers in GBM. For instance, AKT1, AKT2, and AKT3, belonging to the five pathways of signaling by GPCR, GPCR downstream signaling, innate immune system, adaptive immune system, and developmental biology, are three isoforms of AKT in PI3K/AKT pathway, which is an important drug target in many cancers including GBM [54]. In particular, AKT2 is a well-known proto-oncogene that promotes the growth of tumors and reduces the survival of patients in GBM [55, 56].

Table 1.3: Top-10 ranked pathways for survival prediction in GBM by PASNet

Pathway name	Pathway size	Reference	Top-5 ranked genes*
Signaling by GPCR	920	[47]	SHH, PTGFR, GNG5, CHRM5, LHB
GPCR downstream signaling	805	[48]	PTGFR, OR7C2, GNG5, OR10H3, MLNR
Innate immune system	933	[49]	CD79B, INPPL1, SRC, NUP85, DNM2
Adaptive immune system	539	[50]	CD79B, ASB6, PTEN, NCF4, FBXO2
Metabolism of carbohydrates	247	-	HS3ST3B1, NUP85, PFKFB3, LUM, SLC2A4
Transmembrane transport of small molecules	413	[51]	SLC9A7, ABCA7, GNG5, AQP8, HK3
Developmental biology	396	-	NRP2, FES, WNT10B, MYOD1, SLC2A4
Metabolism of proteins	518	-	EIF3G, CCT2, TIMM22, RPL3L, GMPPA
Class A/1 (rhodopsin-like receptors)	305	[52]	PTGFR, OPRD1, CHRM5, NPFF, NTSR2
Axon guidance	251	[53]	NRP2, NRTN, AGRN, FES, RPS6KA4

* The genes were ranked by absolute weights in the pathways.

Finally, we demonstrate a hierarchical representation of genes and pathways in PASNet. In Figure 1.7(a), PASNet is partially visualized, where positive and negative weights are colored in red and blue respectively. The pathways are represented by the corresponding genes in the pathway layer, and then the nonlinear effects of the pathways are described in the hidden layer. The hierarchical representations can be captured in the output layer, which produces a posterior probability for prognosis prediction. Although we considered a single hidden layer to simplify the model with HDLSS data in this study, multiple hidden layers may be able to capture the biological processes and their effects more accurately if a sufficient number of samples are available. Figure 1.7(b) – (c) illustrate distinctive representations of LTS and non-LTS samples in PASNet. The color of nodes in the figures shows the values computed with LTS/non-LTS samples in average. Note that node values between the pathway layer and the output layer are between zero and one. The node with a high value may be a potential prognostic biomarker in the group. Figure 1.7(b) shows that pathways including aquaporin-mediated transport, signaling by BMP, and cytokine signaling in immune system are activated with LTS samples. The second node in the hidden layer is triggered by the active pathways, and the hidden node activates the LTS node in the output layer. On the other hand, Figure 1.7(c) shows that additional pathways of signaling by GPCR and innate immune system are also activated for non-LTS samples. The other two hidden nodes take the active pathways into account, and they activate the non-LTS node in the output layer. Hence, the two pathways of signaling by GPCR and innate immune system may be potential prognostic biomarkers for predicting LTS/non-LTS. Pathway of signaling by GPCR has been investigated as a potential therapeutic target to inhibit the progression of glioblastomas. [47]. Activating the innate immune system, i.e. immunotherapy, is a promising strategy for the treatment of GBM [57]. Vascular endothelial growth factor (VEGF), a modulator of the innate immune system, is reported crucial for the tumor progression [49]. Moreover, aquaporin-mediated transport, signaling by BMP,

and cytokine signaling in immune system may play an important role in GBM, since they are shown in common as active in both LTS and non-LTS. Note that the activation/inactivation of a node in PASNet does not directly represent biological activation in the system, whereas it indicates different states of the biological components in the groups.

1.6 Conclusions

In this chapter, we proposed pathway-associated sparse deep neural network for prognosis predictions (long-term survivals in GBM in this study). PASNet builds a network model by leveraging prior biological knowledge of pathway databases and by taking hierarchical nonlinear relationships of biological processes into account. To improve the model interpretability, PASNet introduces sparse coding. Moreover, we developed a training strategy to avoid the overfitting problem with HDLSS data and the imbalanced problem.

To investigate the performance of PASNet, we used gene expression data of GBM patients in TCGA. PASNet was assessed by comparing the predictive performance with support vector machine, random LASSO, LASSO logistic Regression, and neural network with dropout that have been widely used for prognosis prediction. PASNet outperformed them with respect to both AUC and F1-score in the multiple stratified 5-fold cross-validation experiments. Furthermore, we discussed how PASNet can describe the biological system of GBM.

PASNet is the first deep neural network-based model that represents hierarchical representations of genes and pathways and their nonlinear effects, to the best of our knowledge. Additionally, PASNet would be promising due to its flexible model representation and interpretability, embodying the strengths of deep learning.

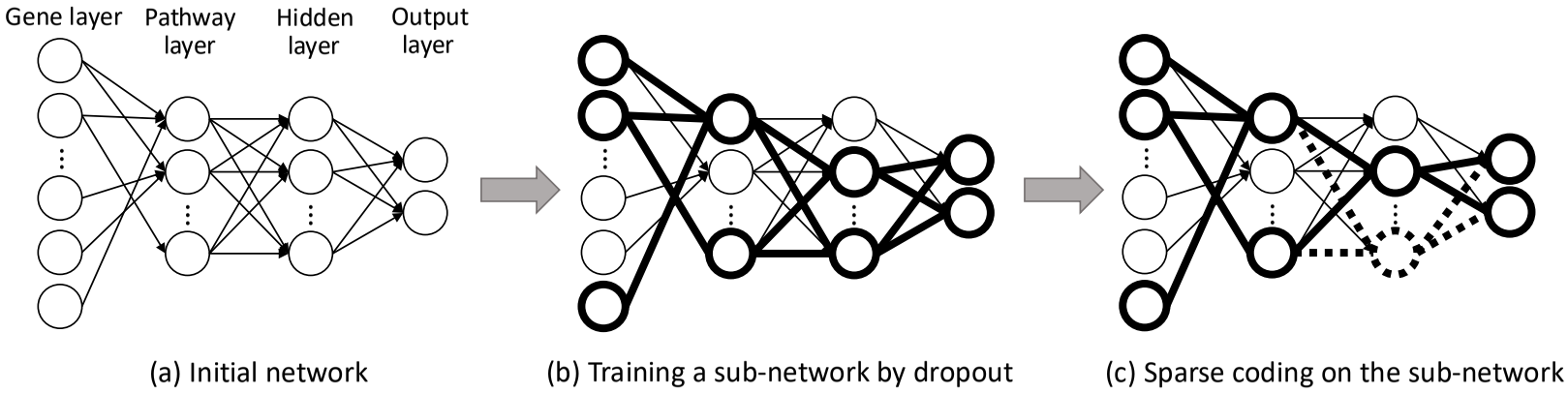

Figure 1.2: Training of PASNet. (a) Weights and biases are randomly initialized. Connections between the gene layer and the pathway layer are determined by biological pathway databases, and the remaining layers are considered as fully-connected in this step. (b) A sub-network is randomly selected using a dropout technique and trained. (c) Sparse coding optimizes the sparsity of connections in the sub-network.

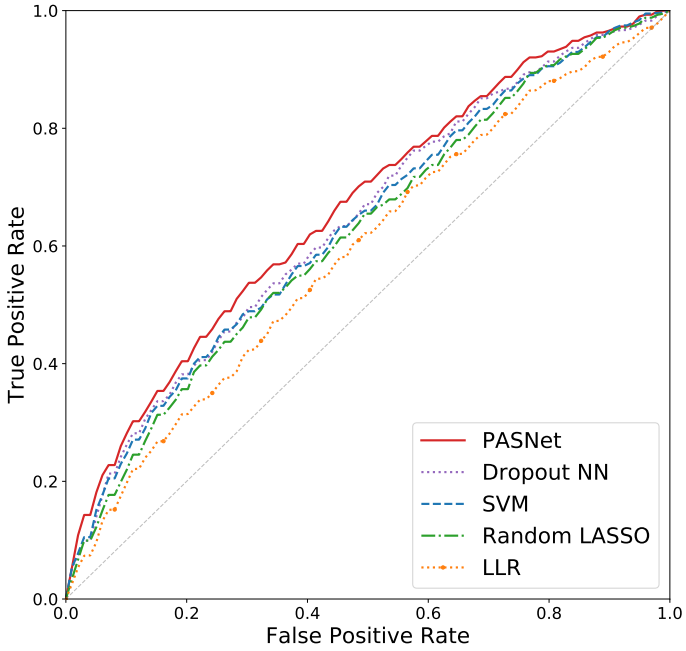

Figure 1.3: ROC Curves. PASNet produces the highest AUC of 0.6622 while the AUC of Dropout NN, SVM, random LASSO, and LLR is 0.6408, 0.6337, 0.6209, and 0.5899, respectively.

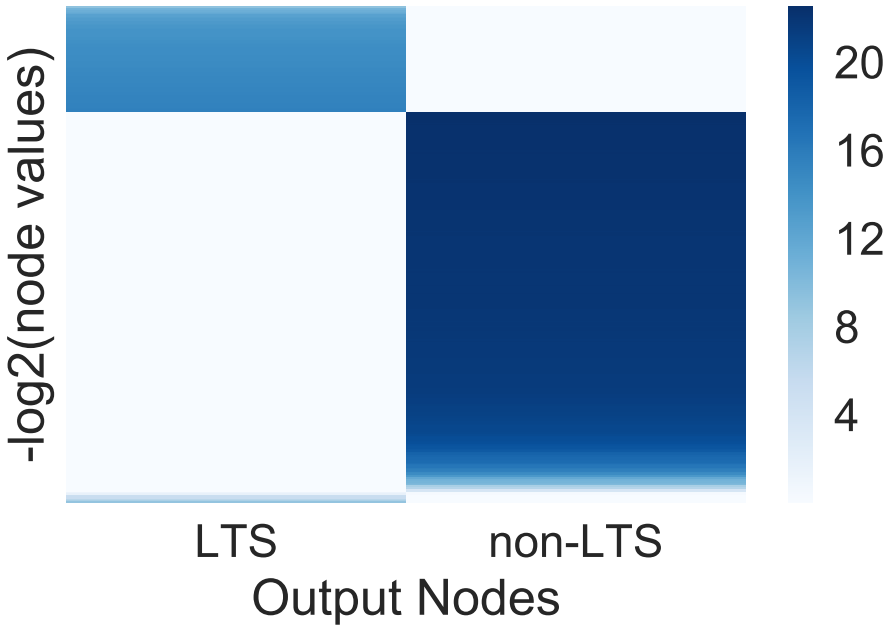

Figure 1.4: Graphical representation of the output node values over the samples by PASNet. LTS samples obtain higher node values in LTS node than non-LTS samples. Similarly, non-LTS samples obtain higher node values in non-LTS node than LTS samples.

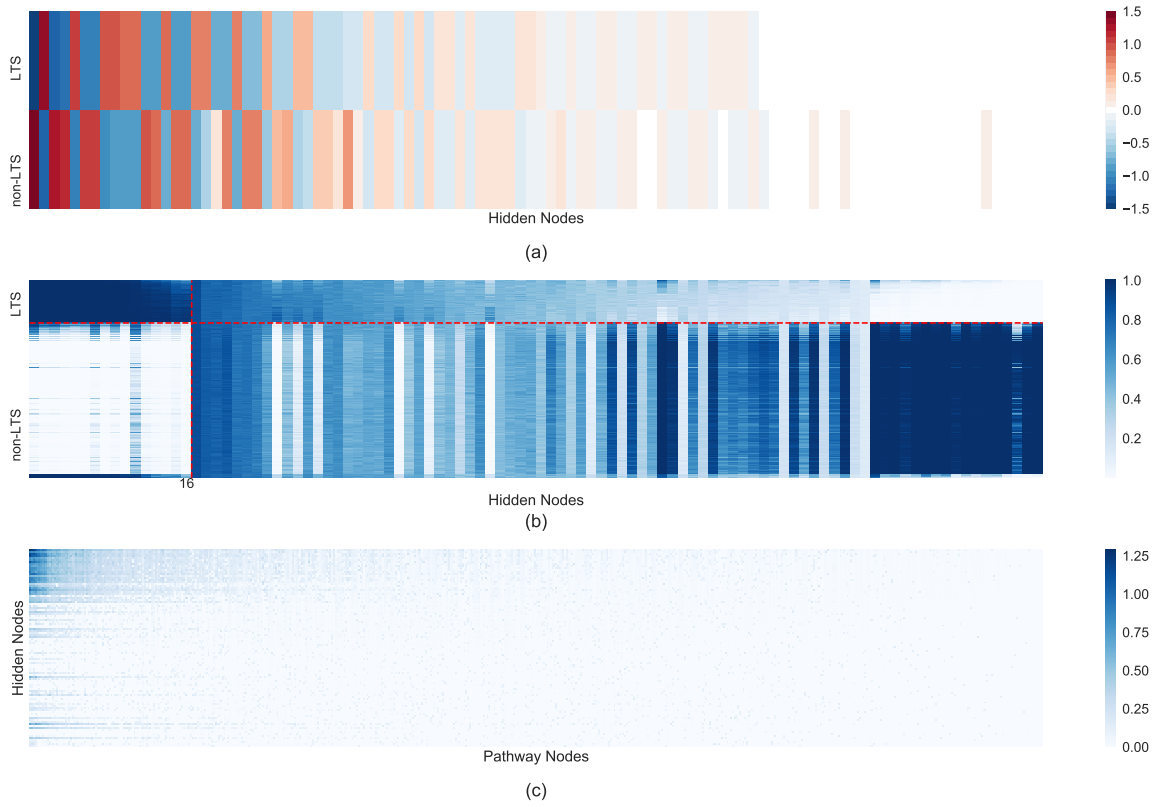

Figure 1.5: Graphical representation among the output layer, hidden layer, and pathway layer in PASNet. (a) The weights between the hidden layer and the output layer. Hidden nodes are sorted in a descending order. (b) The node values in the hidden layer. The horizontal dotted lines indicates LTS/non-LTS samples. The vertical dotted lines indicates LTS/non-LTS samples are significantly distinguished by top 16 pathways. (c) The absolute weights between the pathway layer and the hidden layer.

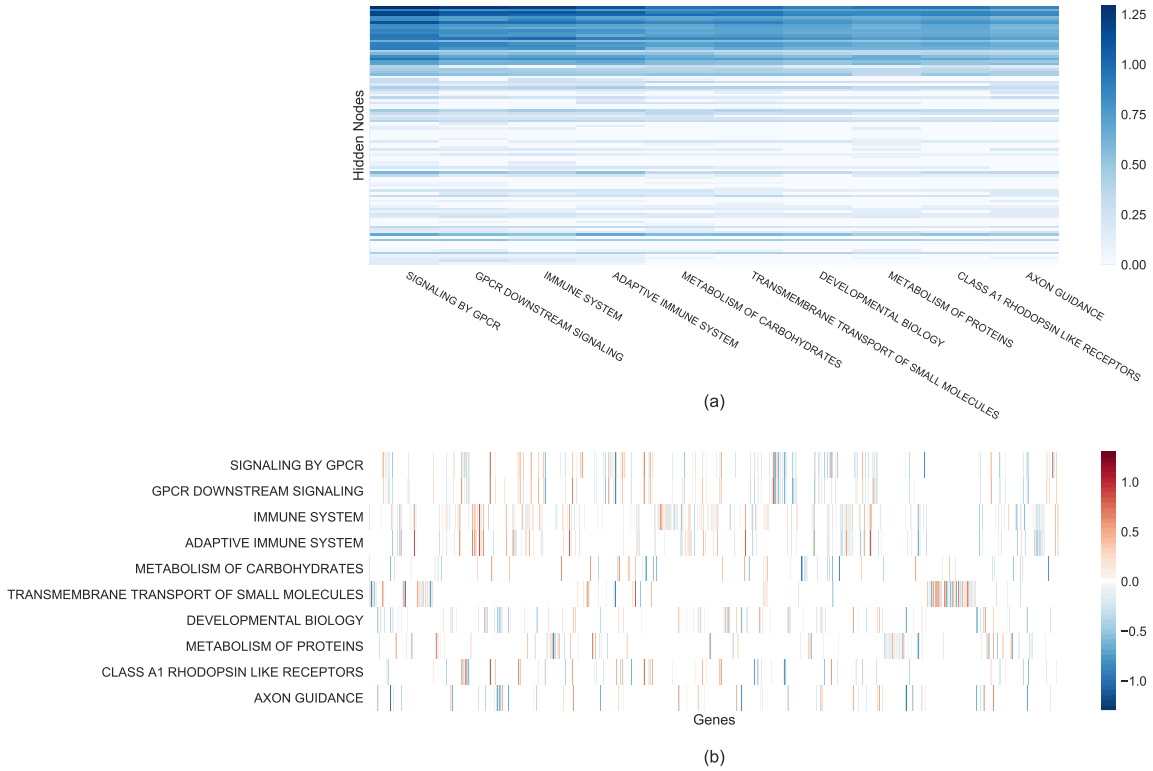

Figure 1.6: Graphical representation of the 10 top-ranked pathways by PASNet. (a) The absolute weights between the 10 top-ranked pathway nodes and the hidden layer. It is a zoom-in view of Figure 1.5(c). (b) Weights between the gene layer and the 10 top-ranked pathway nodes. The connections are determined by Reactome database.

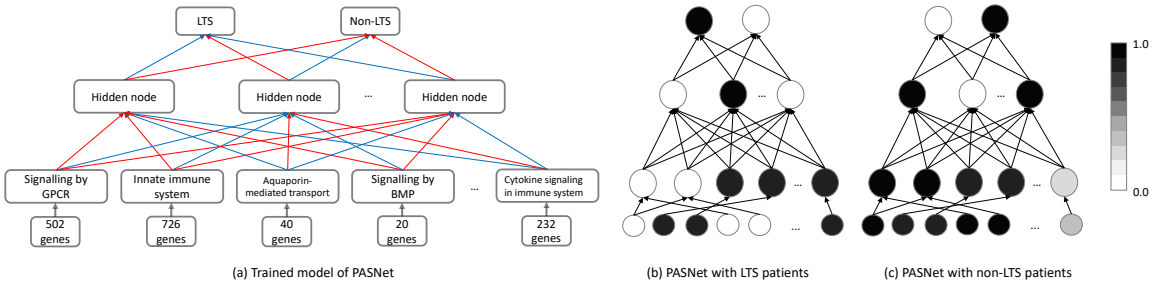

Figure 1.7: Hierarchical representation of pathways in PASNet. (a) PASNet is partially visualized showing the five pathways. Distinct neural network activations between LTS (b) and non-LTS (c) are shown via PASNet. The nodes of the neural network of (b) and (c) correspond to (a). For instance, the nodes in the pathway layer of (b) and (c) represent signaling by GPCR, innate immune system, aquaporin-mediated transport, signaling by BMP, and Cytokine signaling in immune system. The pathways of signaling by GPCR and innate immune system are inactive with LTS patients, whereas the both pathways are active with non-LTS patients.

CHAPTER 2

INTERPRETABLE DEEP NEURAL NETWORK FOR CANCER SURVIVAL ANALYSIS BY INTEGRATING GENOMIC AND CLINICAL DATA

2.1 Background

Dissecting complex biological processes associated to clinical outcomes (e.g., patients survival time) at the cellular and molecular level provides in-depth biological insights not only for developing new treatments for patients, but also for accurate prediction of clinical outcomes [58]. Advanced molecular high-throughput sequencing platforms produce high-dimensional genomic data (e.g., gene expression data) that can provide rich biological descriptions of molecular profiles of human diseases (e.g., cancer) as well as supporting clinical decision-making [59].

Survival analysis estimates survival distribution and investigates the effects of biological and clinical features on a patient's survival time, while handling censored data. The most widely used method for survival analysis is the Cox Proportional Hazards model (Cox-PH), a semi-parametric model that computes the effects of covariates on the risk of event [60, 61]. Cox-PH assumes that the linear combination of patients covariates may be associated with the hazard function (instantaneous rate of occurrence of the event).

However, traditional Cox-PH models have limitations: (1) analyzing high-dimension, low-sample size (HDLSS) data or (2) highly nonlinear data. Training models with HDLSS data is a challenging problem in bioinformatics, because most biological data have many more features (p) than the number of samples (n), i.e., $p >> n$. HDLSS data often make model training infeasible [62]. Thus, low dimensional data, such as clinical data (e.g., age, sex, and body-mass-index), have been analyzed with the Cox-PH model for survival analysis. However, recently, an increasing number of research studies have examined high-dimensional genomic data

to unveil the molecular mechanisms that cause different survival rates. To tackle the HDLSS problem on the Cox-PH model, feature selection techniques and regularization have been considered. Lasso (L1-norm) and elastic net penalizations were introduced in the Cox-PH model [63, 64, 65, 66], whereas in another study a feature selection approach was performed to reduce the number of covariates [67].

The relationship between genomic data and a patient's survival is often highly nonlinear in complex human diseases, whereas a hazard in the Cox-PH model assumes linear relationships between the predictors and a function of the outcome and time of the outcome. Kernel trick is a standard solution to convert nonlinear effects to linear, for linear learning algorithms. Kernel Cox-regression was proposed to capture nonlinear effects between gene expression data and survival data [68]. In the kernel Cox-regression model, regularized Cox-PH was considered in a reproducing kernel Hilbert space. Survival SVM model was developed with sparse regularization for high-dimensional and nonlinear data [69]. However, it is difficult to identify the optimal kernel function of the data, because a kernel function has to be specified in advance.

Lately, deep learning approaches have been successfully adapted due to the capability of modeling highly nonlinear systems and the flexibility of architecture design. In survival analysis, a number of deep learning approaches have been developed coupled with a Cox proportional hazards output layer. DeepSurv introduced a Cox proportional hazards function into a deep fully-connected feed-forward neural network for survival analysis and personalized treatment recommendation [70], and it showed competitive performance with Cox-PH and random survival forests. However, DeepSurv considered only low-dimensional clinical data, where only a small number of covariates ($p < 20$) were examined on simulation data and clinical data. Cox-nnet was constructed based on an artificial neural network with a Cox proportional hazards node in the output layer [71]. High-throughput transcriptomics data of RNA-Seq were introduced to Cox-nnet, and it produced better performance than

Cox-proportional hazards regressions, random survival forests, and CoxBoost. Cox-nnet reported that the high-level representations of gene expression at the top nodes of the hidden layer are correlated to survival rates, and each of the nodes in the hidden layer may implicitly reflect biological processes. SurvivalNet optimizes deep survival models via Bayesian optimization based on Cox-nnet for high-throughput different types of genomic data such as gene expressions, protein expressions, copy number variations, and mutations [72]. SurvivalNet automatically found the optimal network (e.g., numbers of layers and nodes), and the performance of SurvivalNet was slightly better than Cox elastic net (Cox-EN) and random survival forests when the dimension of the data is high. The risk backpropagation analysis enabled SurvivalNet to be interpretable by generating risk scores for each feature.

However, applying deep learning approaches to high-dimensional genomic data for survival analysis is still challenging due to (1) the problem of overfitting when training a deep learning model with HDLSS data and (2) lack of explicit model interpretation. Deep learning typically requires a large number of samples, since deep neural network models involve a number of parameters. Particularly, when training a deep learning model with HDLSS data, gradients tend to have high variance in backpropagation, which consequently causes model overfitting. Both Cox-nnet and SurvivalNet introduced only significant genomic data by feature selection approaches to avoid the overfitting problem. In order to tackle the HDLSS problem in deep learning, dimension reduction techniques were employed to reduce the dimension of the input data, and the results were introduced to a neural network [32]. Deep Feature Selection was developed to identify discriminative features in a deep learning model [33]. Deep Neural Pursuit trained a small-sized sub-network and computed gradients with low variance for feature selection [30].

Conventionally, deep neural networks consist of multiple fully-connected layers, which make it difficult to interpret. In survival analysis, model interpretation (e.g., identifying prognosis factors) is often more important than simply predicting patient

survival with high accuracy. However, fully-connected hidden layers lack to represent explicit biological components. Also, biological processes may involve only a small number of biological components rather than all input features. Thus, the capability of explicit model interpretation with sparse deep neural networks is highly desired in survival analysis.

Furthermore, high-level biological interpretation (e.g., hierarchical relationship between molecular pathways) has seldom been highlighted, whereas biological interpretation at low levels (e.g., gene expression level) has been often considered. Pathway-based model interpretation can provide better biological intuitive and interpretable solutions. Pathway-based analysis often produces significantly reproducible power in genomic study by incorporating well-known biological knowledge. For instance, higher-order functional representation of pathway-based metabolic features provided robust and highly reproducible biomarkers for breast cancer diagnosis [9].

Complex biological systems may involve hierarchical relationships between biological pathways. The hierarchical linkages of biological pathways may cause different survival rates. For instance, the hierarchical representation with receptor pathways and gene ontology was studied for antiviral signaling [73]. Therefore, the incorporation of the effects of inhibition and propagation of a pathway component to others in deep learning can allow the model to be interpretable.

Data integration of multiple types of data (e.g., multi-omics data or clinical data) in deep learning model is also challenging. A number of studies have reported that leveraging multi-omics and clinical data improves predictive performance in survival analysis [72, 1, 74]. A naive approach to integrate multi-omics data is to combine all types of data into a single matrix and perform survival analysis [75, 72]. The approach assumes that the heterogeneous data can be represented by an augmented matrix form. However, the augmented matrix causes problems: (1) generates much higher dimension of HDLSS data, (2) makes the sample size smaller due to missing values, and (3) ignores data types having a smaller number of covariates. Note that

multi-omics data on The Cancer Genome Atlas (TCGA) present substantial missing values; e.g., 160 samples of mRNA-Seq are available, while 595 clinical samples are in glioblastoma multiforme (GBM) dataset in TCGA.

In this chapter, we propose a novel method, Cox-PASNet, a pathway-based sparse deep neural network, for survival analysis integrating high-dimensional genomic data and clinical data. Our main contributions of Cox-PASNet for survival analysis are:

- to explicitly model nonlinear and hierarchical relationships in a biological pathway level,
- to enable one to interpret the model, where nodes in layers correspond to biological components of genes and pathways,
- to integrate clinical data in a deep learning model, and
- to provide an efficient solution to train the complex neural network model with HDLSS data without overfitting problem.

2.2 Methods

2.2.1 The Architecture of Cox-PASNet

We introduce our proposed model, Cox-PASNet, a pathway-based sparse deep neural network for survival analysis with genomic and clinical data. Cox-PASNet combines a Cox proportional hazards regression with a deep neural network, incorporating prior knowledge of biological pathways. The architecture of Cox-PASNet (see Figure 2.1) is comprised of (1) a gene layer, (2) a pathway layer, (3) multiple hidden layers, (4) a clinical layer, and (5) a Cox layer.

Gene Layer

The gene layer is an input layer of Cox-PASNet introducing gene expression data with n patient samples of p gene expressions. For pathway-based analysis, only the genes that belong to at least one pathway are considered in the gene layer.

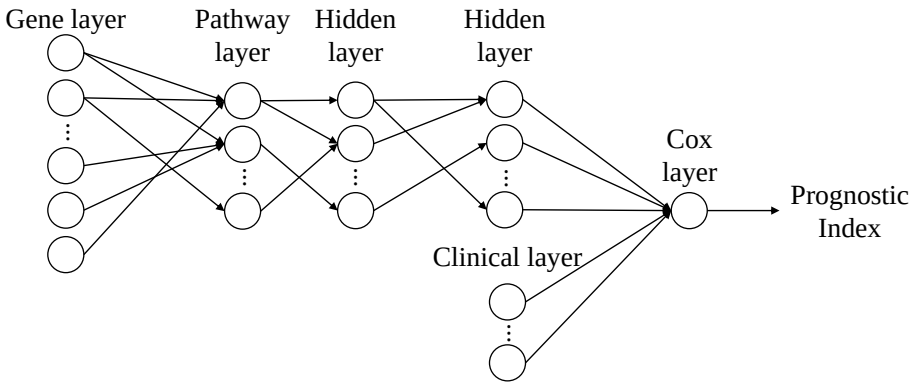

Figure 2.1: The architecture of Cox-PASNet. The structure of Cox-PASNet is constructed by a gene layer (an input layer), a pathway layer, multiple hidden layers, a clinical layer (additional input layer), and a Cox layer.

Pathway Layer

The pathway layer represents biological pathways where a node indicates a specific biological pathway. The pathway layer incorporates prior biological knowledge so that the model can be biologically interpretable. Pathway databases (e.g., KEGG and Reactome) contain a set of genes that are involved in a pathway, and each pathway characterizes a biological process. The knowledge of the given association between genes and pathways explicitly forms sparse connections between the gene layer and the pathway layer in Cox-PASNet, rather than fully-connecting the layers.

To implement the sparse connections between the gene and the pathway layers, we consider a binary bi-adjacency matrix. Given pathway databases containing pairs of m genes and n pathways, the binary bi-adjacency matrix ($\mathbf{A} \in \mathbb{B}^{n \times m}$) is constructed, where an element a_{ij} is one if gene j belongs to pathway i, otherwise zero, i.e., $\mathbf{A} = \{a_{ij} | 1 \leq i \leq n, 1 \leq j \leq m\}$ and $a_{ij} = \{0, 1\}$.

Hidden Layers

The hidden layers model the nonlinear and hierarchical effects of pathways. Node values in the pathway layer indicate the active/inactive status of a single path-

way in a biological system, whereas the hidden layers show the interactive effects of multiple pathways. The deeper hidden layer expresses the higher level representations of biological pathways.

Clinical Layer

The clinical layer introduces clinical data to the model separately from genomic data. The dimension of clinical data is usually much smaller than genomic data, so clinical data tend to be easily ignored if introducing them to the input layer with genomic data. In Cox-PASNet, the complex genetic effects of gene expression data are captured from the gene layer to the hidden layers, whereas the clinical data are directly introduced into the output layer along with the highest-level representation of genomic data (i.e., node values on the last hidden layer). Therefore, Cox-PASNet takes the effect of genomic data and clinical data separately into account in the neural network model.

Cox Layer

The Cox layer is the output layer that has only one node. The node value produces a linear predictor, a.k.a. Prognostic Index (PI), from both genomic and clinical data, which is introduced to a Cox-PH model. Note that the Cox layer has no bias node according to the design of the Cox model.

Furthermore, we introduce sparse coding so that the model can be biologically interpretable and mitigate overfitting. In a biological system, a few biological components are involved in biological processes. The sparse coding enables the model to include only significant components for better biological model interpretation. Sparse coding is applied to the connections from the gene layer to the last hidden layer by mask matrices. The details of sparse coding are described in Section 2.2.4.

2.2.2 Objective Function

In order to perform Cox proportional hazards regression on the Cox layer, Cox-PASNet defines the objective function using average negative log partial likelihood with L^2 regularization:

$$\ell(\boldsymbol{\Theta}) = -\frac{1}{n_E}\sum_{i\in E}\left(\mathbf{h}_i^I\boldsymbol{\beta} - \log\sum_{j\in R(T_i)}\exp(\mathbf{h}_j^I\boldsymbol{\beta})\right) + \lambda(\|\boldsymbol{\Theta}\|_2), \tag{2.1}$$

where $\boldsymbol{\Theta} = \{\boldsymbol{\beta}, \mathbf{W}\}$ is a set of parameters, $\boldsymbol{\beta}$ is the Cox proportional hazards coefficients (weights between the last hidden layer and the Cox layer), $\mathbf{W}$ is a union of the weight matrices on the layers before the Cox layer, and $\mathbf{h}^I$ is the integrative layer that integrates the second hidden layer's outputs and the clinical inputs from the clinical layer. E is a set of uncensored samples, n_E is the total number of uncensored samples, and $R(T_i) = \{i|T_i \geq t\}$ is a set of samples at risk of failure at time t. $\|\mathbf{W}\|_2$ and $\|\boldsymbol{\beta}\|_2$ are the L^2-norms of $\mathbf{W}$ and $\boldsymbol{\beta}$ respectively, and λ is a regularization hyperparameter to avoid overfitting ($\lambda > 0$).

2.2.3 Training Cox-PASNet

We propose an optimization strategy to train Cox-PASNet with HDLSS data along with L^2 regularization in the objective function. We optimize the model by partially training small sub-networks with sparse coding. Training a small sub-network guarantees the feasible optimization with a small set of parameters in each epoch.

The overall training flow of Cox-PASNet is illustrated in Figure 2.2. Layers are initially set to be fully connected, where weights and biases are randomly initialized. Particularly, the connections between the gene layer and the pathway layer are forced to be sparse by the bi-adjacency matrix, and the Cox layer includes no bias node.

A small sub-network is randomly chosen by a dropout technique in the hidden layers excluding the Cox layer (Figure 2.2a). Then, the weights and the biases of the sub-network are optimized by backpropagation. Once training of the sub-network completes, sparse coding is applied to the sub-network by trimming the connections

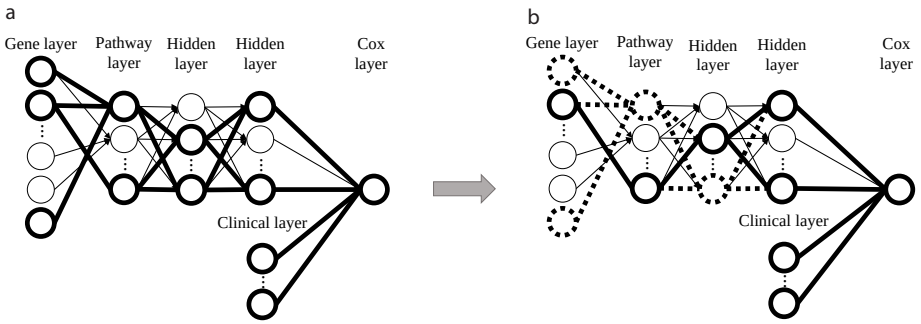

Figure 2.2: Training of Cox-PASNet with high-dimensional, low-sample size data. (a) A small sub-network is randomly chosen by a dropout technique in the hidden layers and trained. (b) Sparse coding optimizes the connections within the small network.

within the small network that do not contribute to minimizing the loss. In Figure 2.2b, the connections and the nodes dropped by sparse coding are marked with bold and dashed lines. The algorithm of Cox-PASNet is briefly described in Algorithm 2.

Algorithm 2 Training of Cox-PASNet

1: Initialize weights $\mathbf{W}^{(\ell)}$, biases $\mathbf{b}^{(\ell)}$, and $\boldsymbol{\beta}$
2: $\mathbf{W}^{(0)} \leftarrow \mathbf{W}^{(0)} \star \mathbf{M}^{(0)}$
3: **repeat**
4: Select a small sub-network via dropout
5: Train the sub-network
6: Sparse coding with the optimal $\mathbf{M}^{(\ell)}$ by Eq. (2.3)
7: Update weights
8: **until** convergence

2.2.4 Sparse Coding

Sparse coding is implemented by mask matrices. A binary mask matrix $\mathbf{M}$ controls a sparsity level of each layer on the sub-network, where an element indicates

whether the corresponding weight is dropped or not. Then, the outputs in the layer are computed by:

$$\mathbf{h}^{(\ell+1)} = a\Big((\mathbf{W}^{(\ell)} \star \mathbf{M}^{(\ell)})\mathbf{h}^{(\ell)} + \mathbf{b}^{(\ell)}\Big), \tag{2.2}$$

where $\star$ denotes an element-wise multiplication operator, and $a(\cdot)$ is a nonlinear activation function (e.g., sigmoid or Tanh). $\mathbf{h}^{(\ell)}$ is the outputs on the ℓ-th layer, and $\mathbf{W}^{(\ell)}$ and $\mathbf{b}^{(\ell)}$ are a weight matrix and a bias vector, respectively, with $1 \leq \ell \leq L-2$, where L is the number of layers.

In particular, an element of $\mathbf{M}$ is set to one if the absolute value of the corresponding weight is greater than threshold $s^{(\ell)}$, otherwise it is zero. Note that the mask between the gene layer and the pathway layer, i.e., $\mathbf{M}^{(0)}$, is determined by the bi-adjacency matrix $\mathbf{A}$ of biological pathways. Thus, the mask matrices are formulated as

$$\mathbf{M}^{(\ell)} = \begin{cases} \mathbb{1}(|\mathbf{W}^{(\ell)}| \geq s^{(\ell)}), & \text{if } \ell \neq 0 \\ \mathbf{A}, & \text{if } \ell = 0. \end{cases} \tag{2.3}$$

The optimal sparsity level ($s^{(\ell)}$) is estimated on each layer in the sub-network to generate the mask matrix. For efficient approximation of the optimal sparsity level, cost scores are computed with various finite sparsity levels in a range of $\mathbf{s} = [0, 100)$ where zero generates a fully-connected layer while 100 shows disconnected layers. Then, we approximate the cost function with respect to sparsity levels by applying a cubic-spline interpolation to the cost scores computed by the finite set of $\mathbf{s}$. Finally, the sparsity level that minimizes the cost score is considered for the optimal sparsity level. The optimal $s^{(\ell)}$ is approximated on each layer individually in the sub-network. The individual optimization of the sparsity on each layer represents different levels of biological associations on genes and pathways.

2.3 Results

2.3.1 Datasets

In this study, we considered GBM and ovarian serous cystadenocarcinoma (OV) to assess Cox-PASNet. GBM is the most aggressive malignant type of brain tumor, which shows poor prognosis [37]; OV is one of the most common cancer types among women in the world, and OV is usually diagnosed at a late stage [76]. Gene expression and clinical data of GBM and OV were obtained from the TCGA (http://cancergenome.nih.gov). The samples that lack survival time or survival status were filtered out.

The prior knowledge of biological pathways was taken from the Molecular Signatures Database (MSigDB) [40, 41, 42], where KEGG and Reactome pathway databases were considered for the pathway-based analysis. We excluded small pathways (i.e., less than fifteen genes) and large pathways (i.e., over 300 genes), since small pathways are often redundant with other larger pathways and large pathways are related to general biological pathways rather than specific to a certain disease [43]. Moreover, only the genes that belong to at least one pathway were investigated.

For the integrative analysis, we included the clinical information of both GBM and OV patients. We incorporated only age in the clinical layer of Cox-PASNet, because age has been reported as a significant covariate for prognostic prediction in GBM [1] and most other clinical data have substantial missing values. Although Karnofsky Performance Score (KPS) is also reported as significant as well as age, KPS is highly correlated to age, and there are many missing values. Finally, we used 5,404 genes, 659 pathways, and clinical data of age from 523 GBM samples and 532 OV samples.

2.3.2 Experimental Design

Cox-PASNet was assessed by comparing the performance with Cox-EN [65], Cox-nnet [71], and SurvivalNet [72]. The performance of the four models was eval-

uated by C-index, which is a non-parametric metric that calculates concordance between predicted and actual survival curves. The value range of C-index is between zero and one, where one indicates a perfect model prediction and 0.5 means a random guess.

The dataset was randomly split into training (64%), validation (16%), and test (20%) data, while preserving the proportion of the censor status between censored and uncensored samples. The gene expression and clinical data in the training data were standardized to mean of zero and standard deviation of one. The validation and the test data were normalized with the mean and standard deviation from the training data. Each model was trained by the training data; the optimal hyper-parameters were obtained with the validation data; and the model performance was evaluated by the test data. The experiments were repeated over twenty times for reproducibility of model performance.

Cox-PASNet followed a modern deep learning design. We used the Tanh function as the activation function. Both dropout and L^2 regularizations were considered. Adaptive Moment Estimation (Adam) was performed for the optimization to approximate first-order gradients [44]. The optimal initial learning rate (η) and the L^2 regularization (λ) were estimated by the grid search technique. η and λ that minimize the cost function with validation data were selected as the optimal hyper-parameters. Dropout rates were empirically set to be 0.7 and 0.5 for the pathway layer and the following hidden layers, respectively. The open source code of Cox-PASNet implemented by PyTorch is available at `https://github.com/DataX-JieHao/Cox-PASNet`.

Cox-EN models were implemented using *Glmnet Vignette* package in Python [65]. The hyper-parameters of α and λ were optimized by grid search. We considered values of α between 0 and 1 with a step of 0.01 and 200 λ values. Then, Cox-EN was performed with the optimal hyper-parameters that minimize the cost function. Cox-nnet was conducted based on open source codes provided by the authors. The tuning setting of the model followed their recommendation. Grid search for L^2 was

applied. The optimal hyper-parameters of SurvivalNet were optimized by Bayesian Optimization technique, *BayesOpt* [77]. We also considered the hyper-parameters of L^1 and L^2 regularizations for the Bayesian optimization in addition to their default setting. SurvivalNet was carried out by its open source in GitHub.

For the data integration, both the clinical data of age and gene expression data were combined into an input matrix and introduced to Cox-EN, SurvivalNet, and Cox-nnet at the input level for the experiments, whereas Cox-PASNet introduced gene expression data into the gene layer and clinical data into the clinical layer separately.

2.3.3 Experimental Results

The experimental results with GBM and OV data are shown in Figure 2.3 and Table 2.1–2.3. Cox-PASNet showed the highest C-index of 0.6347 ± 0.0372 in GBM, whereas the second highest C-index of 0.5903 ± 0.0372 was shown in Cox-nnet (Figure 2.3a and Table 2.1). Cox-nnet is a simplified model of SurvivalNet that includes only a hidden layer. On the other hand, SurvivalNet is a generalized fully-connected neural network model for survival analysis with Cox-model, where the optimal architecture is determined by Bayesian optimization technique. Cox-nnet reported that the simple neural network architecture often produces better performance than deeper networks [71]. Cox-EN produced a C-index of 0.5151 ± 0.0336, which is close to a random guess. It may be due to the highly nonlinear HDLSS data of 5,404 features of 523 samples. The statistical significance of the performance was assessed by Wilcoxon rank-sum test. The distributions of C-index scores produced by Cox-PASNet were significantly higher than others in Table 2.2.

Moreover, we evaluated Cox-PASNet with OV data. Cox-PASNet showed the highest C-index of 0.6343 ± 0.0439 as well; Cox-nnet retained the second rank with C-index of 0.6095 ± 0.0356; Cox-EN was the last place with C-index of 0.5276 ± 0.0482 (Figure 2.3b and Table 2.3). The statistical testing of Wilcoxon rank-sum test showed that Cox-PASNet also statistically outperformed others in OV in Table 2.4.

Table 2.1: Comparison of C-index with GBM in over 20 experiments

Model	C-index
Cox-EN	0.5151 ± 0.0336
Cox-nnet	0.5903 ± 0.0372
SurvivalNet	0.5521 ± 0.0295
Cox-PASNet	**0.6347 ± 0.0372**

Table 2.2: Statistical assessment with GBM

	Wilcoxon rank-sum test (P-value)
Cox-PASNet vs. Cox-EN	8.85e-05*
Cox-PASNet vs. Cox-nnet	4.49e-4*
Cox-PASNet vs. SurvivalNet	1.40e-4*

* shows the statistical significance with significance level = 0.05.

Cox-PASNet shares the cost function of negative log partial likelihood with Cox-nnet and SurvivalNet. However, Cox-PASNet constructs the neural network based on prior knowledge of biological pathways, and the biologically inspired architecture produced better performance reducing noise that comes from the data complexity. Cox-PASNet also trains the model with sub-networks to avoid overfitting problem with HDLSS data. The outstanding performance supports the contributions of the new architecture Cox-PASNet and the training strategy.

2.4 Model Interpretation in GBM

For the biological model interpretation of Cox-PASNet, we re-trained the model with the optimal pair of hyper-parameters from 20 experiments using all available GBM samples. The samples were categorized into two groups of high-risk and low-risk groups by the median Prognostic Index (PI), which is the output value of Cox-PASNet. The node values of the two groups in the integrative layer (i.e., the second hidden layer (H2) and the clinical layer) and the pathway layer are illustrated in

Table 2.3: Comparison of C-index with OV in over 20 experiments

Model	C-index
Cox-EN	0.5276 ± 0.0482
Cox-nnet	0.6095 ± 0.0356
SurvivalNet	0.5614 ± 0.0524
Cox-PASNet	**0.6343 ± 0.0439**

Table 2.4: Statistical assessment with OV

	Wilcoxon rank-sum test (P-value)
Cox-PASNet vs. Cox-EN	1.03e-4*
Cox-PASNet vs. Cox-nnet	0.04*
Cox-PASNet vs. SurvivalNet	2.93e-4*

* shows the statistical significance with significance level = 0.05.

Figure 2.4 and Figure 2.5, respectively. In Figure 2.4a, the node values of 31 covariates (30 from the genomic data and age from the clinical data) were sorted by the average absolute partial derivatives with respect to the integrative layer. Age (the first column in Figure 2.4a) is shown as the most important covariate in Cox-PASNet with GBM data in terms of the partial derivatives.

The top ranked covariates show distinct distributions between high-risk and low-risk groups. For instance, the first three covariates in H2 (the 2nd, 3rd and 4th columns in Figure 2.4a) were activated in the high-risk group, but inactivated in the low-risk group. Moreover, we performed logrank test by grouping node values of the covariate into two groups individually again by their median. The -log10(p-values) computed by logrank test are depicted in the above panel aligning with the covariates in Figure 2.4a. The red triangle markers show significant covariates (-log10(p-value) > 1.3), whereas the blue markers show insignificant ones. The logrank tests revealed that the top ranked covariates by the absolute weight are associated to survival prediction. Figure 2.4b – 2.4c present Kaplan-Meier curves for the top two

covariates, where survivals between the two groups are significantly different. Thus, the top ranked covariates can be considered as prognostic factors.

In the same manner, the nodes in the pathway layer are partially illustrated in Figure 2.5. The heatmap in Figure 2.5a depicts the top 10 pathway node values of the high-risk and low-risk groups, where the pathway nodes are sorted by the average absolute partial derivatives with respect to the pathway layer. We also performed logrank tests on each pathway node, and 304 out of 659 pathways were statistically significant on the survival analysis. The two top-ranked pathways were further investigated by Kaplan-Meier analysis, shown in Figure 2.5b–2.5c. The Kaplan-Meier curves of the two top-ranked pathways imply the capability of the pathway nodes as prognostic factors.

The statistically significant nodes in the integrative layer and the top ten ranked pathway nodes are visualized by t-SNE [78] in Figure 2.6, respectively. The nonlinearity of the nodes associated with PI is illustrated. The integrative layer represents the hierarchical and nonlinear combinations of pathways. Thus, the more distinct associations with survivals are shown in the integrative layer than the pathway layer.

The ten top-ranked pathways by the partial derivatives are listed with related literature in Table 2.5. The p-values in the table were computed by logrank test with the pathway node values of the two groups of high and low risks. Among them, five pathways were reported as significant pathways in biological literature of GBM. Jak-STAT signaling pathway, which is usually called as an oncopathway, is activated for the tumor growth of many human cancers [79]. Inhibition of Jak-STAT signaling pathway was shown to reduce the malignant tumors using animal models of glioma. Neuroactive ligand-receptor interaction was explored as one of the most significant pathways in GBM [80]. PI3K cascade is also a well-known pathway that is highly involved in proliferation, invasion, and migration in GBM [81].

The ten top-ranked genes by partial derivatives with respect to each gene are listed with their p-values and related literature in Table 2.6. PRL is known as be-

ing associated with occurrence of neoplasms and central nervous system neoplasms, and the assessment with PRL expression in primary central nervous system tumors was investigated [86]. MAPK9 was identified as a novel potential therapeutic marker along with RRM2 and XIAP, which is associated with biological pathways involved in the carcinogenesis of GBM [87]. IL22 was reported to promote the malignant transformation of bone marrow-derived mesenchymal stem cells, which exhibit potent tumoritropic migratory properties in tumor treatment [88]. FGF5 contributes to the malignant progression of human astrocytic brain tumors as an oncogenic factor in GBM [89]. The activation of JUN along with HDAC3 and CEBPB may form resistance to chemotherapy and radiation therapy of hypoxic GBM, and the down-regulation of the genes appeared to inhibit temozolomide on hypoxic GBM cells [90]. Low expression of DRD5 was presented as being associated with relatively superior clinical outcomes in glioblastoma patients with ONC201 [91]. HTR7 involved in neuroactive ligandreceptor interaction and calcium signaling pathway was reported to contribute the development and progression of diffuse intrinsic pontine glioma [92].

It is worth noting that only IL22 and FGF5 are statistically significant (i.e., p-value < 0.05) by logrank test on each gene, which means that only the two genes can be identified as significant prognostic factors by conventional Cox-PH models. However, other genes such as PRL, MAPK9, JUN, DRD5, and HTR7 have been biologically identified as significant prognostic factors, even though significantly different distributions are not found on gene expression (i.e., p-value ≥ 0.05). The average absolute partial derivatives with respect to each gene measure contribution to patients' survival through the pathway and hidden layers in Cox-PASNet when gene expression varies on the gene. Therefore, the gene biomarker identification by Cox-PASNet allows one to capture significant genes nonlinearly associated to patients' survival.

Figure 2.7 illustrates the overall hierarchical representation of biological pathways in Cox-PASNet. A pathway node is represented by nonlinear effects of the associated gene nodes, and a hidden node expresses the high-level representation of

a set of pathways. The following hidden layers describe the hierarchical representation of the previous hidden nodes. Then, the last hidden nodes are introduced to a Cox-PH model with clinical data.

A pathway node value shows active or inactive status of the corresponding pathway, which may be associated to different survivals (e.g., Jak-STAT signaling pathway). The significance of the genes involved in the active pathway can be ranked by the average absolute partial derivatives with repect to the gene layer (e.g., AKT1 and AKT3). A set of the active pathways are represented in an active node in the following hidden layer, which improves the survival prediction. For instance, the Kaplan-Meier plots of *Node 19* and *PI* show more similar estimation of the survival than *Jak-STAT signaling pathway* in Figure 2.7.

2.5 Conclusion

We developed a pathway-based sparse deep neural network, Cox-PASNet, for survival analysis coupled with Cox-PH model on a deep neural network. Cox-PASNet builds a neural network model that can describe nonlinear and hierarchical effects of biological pathways and provide significant prognostic factors for accurate prediction of patients' survival. A new strategy to train the deep neural network model with HDLSS data was also introduced in the paper. Cox-PASNet outperformed the current cutting-edge survival methods such as Cox-nnet, SurvivalNet, and Cox-EN, and its predictive performance was statistically assessed.

Negative log-partial likelihood with a single node in the output layer is considered in Cox-PASNet as Cox-nnet and SurvivalNet also adapted. Using Cox log-partial likelihood function may raise several concerns with respect to the model assessment, which is commonly applied in conventional Cox-PH models. One concern is if there is multicollinearity in the last hidden layer's nodes and the clinical layer's node, which are the covariates in the Cox-PH model. The covariates are hierarchically derived high-level representations from gene expression data (inputs) rather than input data

introduced to the model directly. Therefore, we used partial derivatives with respect to inputs for the model assessment with neural network for identifying significant genes or pathways. However, testing the multicollinearity would be a potential solution to identify the optimal number of nodes. Another concern is how to assess Cox-PASNet model fit by residuals, such as Martingale residuals and deviance residuals. While the scope of our study is to develop neural network-based survival analysis by incorporating Cox log-partial likelihood function, testing the fitness of a Cox-PH model, eventually, would be beneficial to deep learning-based Cox-PH model.

Overall, Cox-PASNet constructs the neural network based on biological pathways with sparse coding. The genomic and clinical data are introduced to the model separately for model interpretation. Cox-PASNet integrates clinical data as well as genomic data. However, high-dimensional genomic data may cause bias in the integration due to the unbalanced size between genomic and clinical covariates. Furthermore, the incorporation of multi-omics data such as DNA mutation, copy number variation, DNA methylation, and mRNA expression is essential to describe complex human diseases involving a sequence of complex interactions in multiple biological processes. A solution of integration of complex heterogeneous data would be desired as a future work.

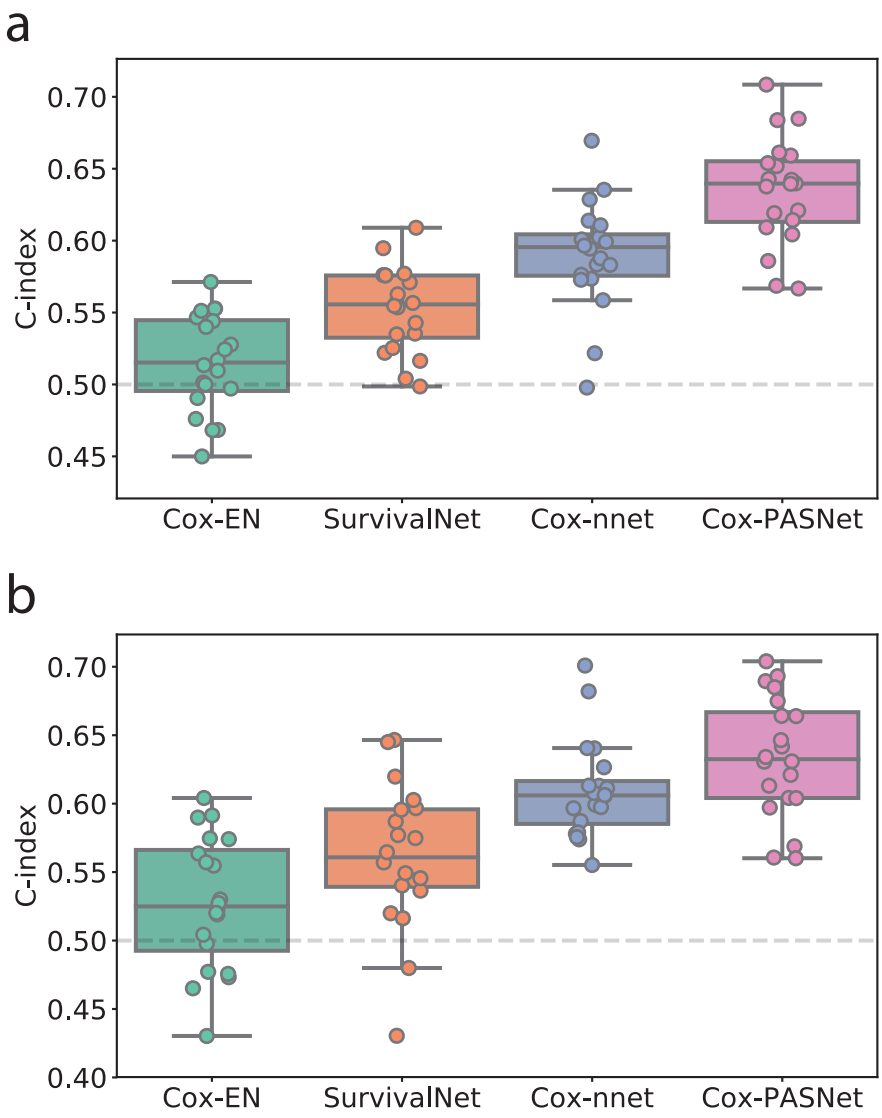

Figure 2.3: Experimental results with (a) GBM and (b) OV in C-index. Boxplots of the C-index of (a) TCGA GBM dataset and (b) TCGA OV dataset using Cox-EN, SurvivalNet, Cox-nnet, and Cox-PASNet. Each dataset was randomly split into training (64%), validation (16%), and test (20%) data, while preserving the proportion of the censor status between censored and uncensored samples. The experiments were repeated over twenty times.

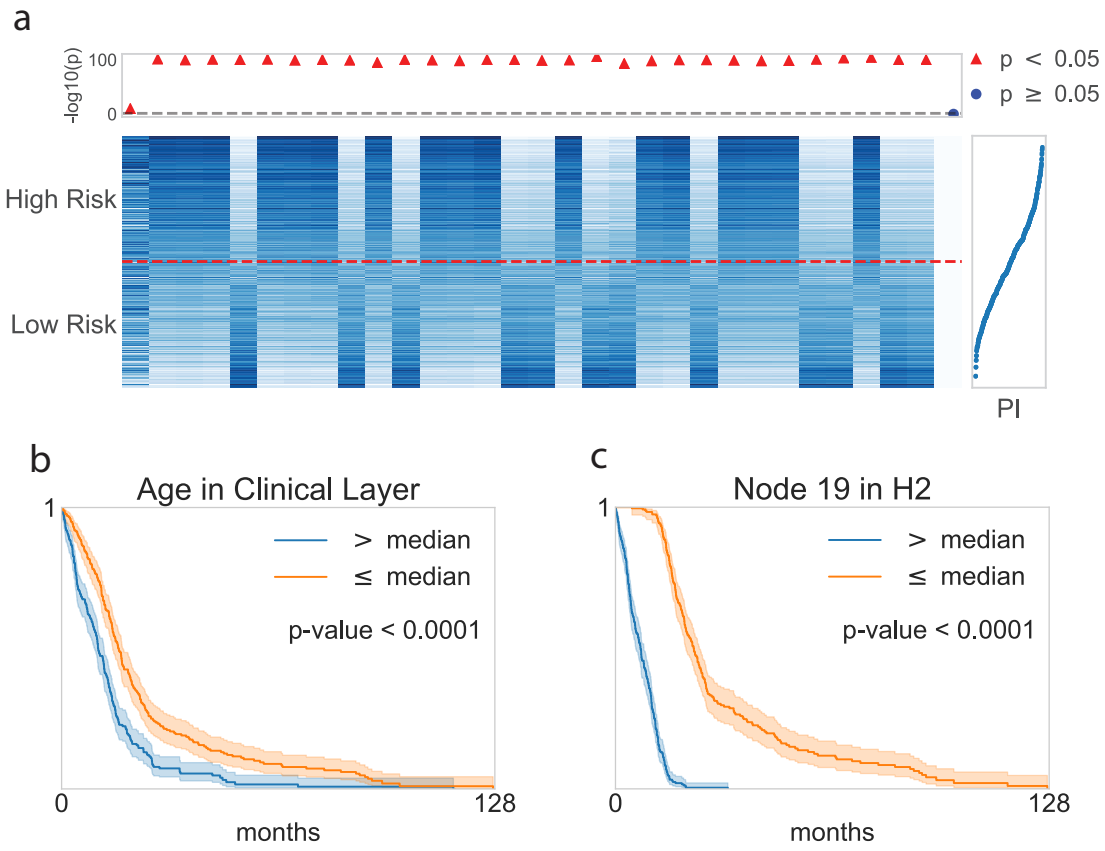

Figure 2.4: Graphical visualization of the node values in the second hidden layer (H2) and the clinical layer. (a) Heatmap of the 31 nodes (i.e., thirty H2 nodes and one clinical node). The horizontal dotted line indicates high-risk/low-risk samples. The upper dot plot shows -log10(p-values) of logrank test between high-risk/low-risk groups for each node. Red indicates statistical significance with logrank test, whereas blue shows insignificance. The curve in the right panel shows prognostic indices (PI) with the corresponding samples. (b) – (c) Kaplan-Meier plots for the two top-ranked nodes.

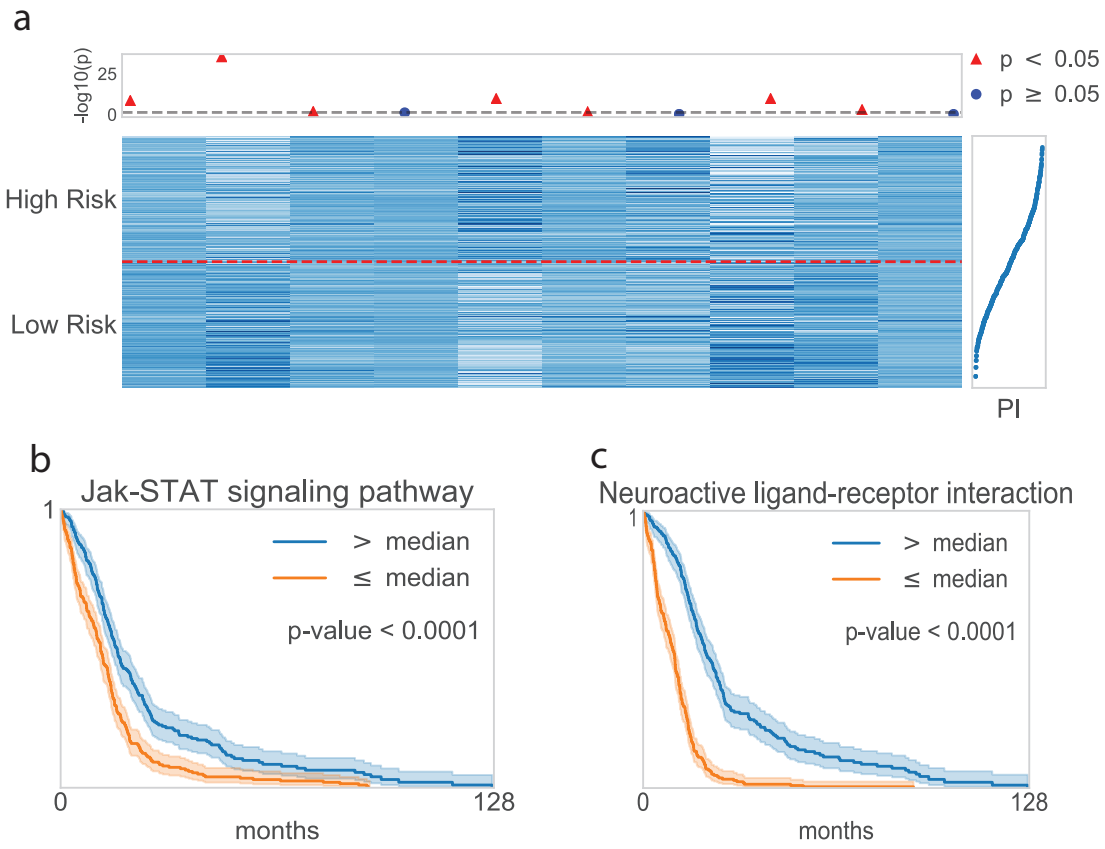

Figure 2.5: Graphical visualization of the node values in the pathway layer. (a) Heatmap of the ten top-ranked pathway nodes. The horizontal dotted line indicates high-risk/low-risk samples. The upper dot plot shows -log10(p-values) of logrank test between high-risk/low-risk groups for the top ten ranked pathway nodes. Red indicates statistical significance with logrank test, whereas blue shows insignificance. The curve in the right panel shows prognostic indices (PI) with the corresponding samples. (b) – (c) Kaplan-Meier plots for the top two ranked pathway nodes.

Table 2.5: Ten top-ranked pathways in GBM by Cox-PASNet

Pathway name	# of genes	P-value	Ref.
Jak-STAT signaling pathway	155	< 0.0001	[79, 82, 83]
Neuroactive ligand-receptor interaction	272	< 0.0001	[80]
MAP kinase activation in TLR cascade	50	0.0176	–
NFκB and MAP kinases activation mediated by TLR4 signaling repertoire	72	0.0729	–
G alpha (i) signalling events	195	< 0.0001	–
PI3K cascade	71	0.0304	[81, 84]
Tyrosine metabolism	42	0.5671	–
Neuronal system	279	< 0.0001	[85]
Axon guidance	129	0.0012	[83]
Xenobiotics	16	0.6347	–

Table 2.6: Ten top-ranked genes in GBM by Cox-PASNet

Gene name	P-value	Ref.
PRL	0.1698	[86]
FGF22	0.4503	–
MAPK9	0.9580	[87]
IL22	0.0140	[88]
IFNA5	0.5401	–
FGF5	< 0.0001	[89]
AGTR1	0.1375	–
JUN	0.1798	[90]
DRD5	0.1288	[91]
HTR7	0.7751	[92]

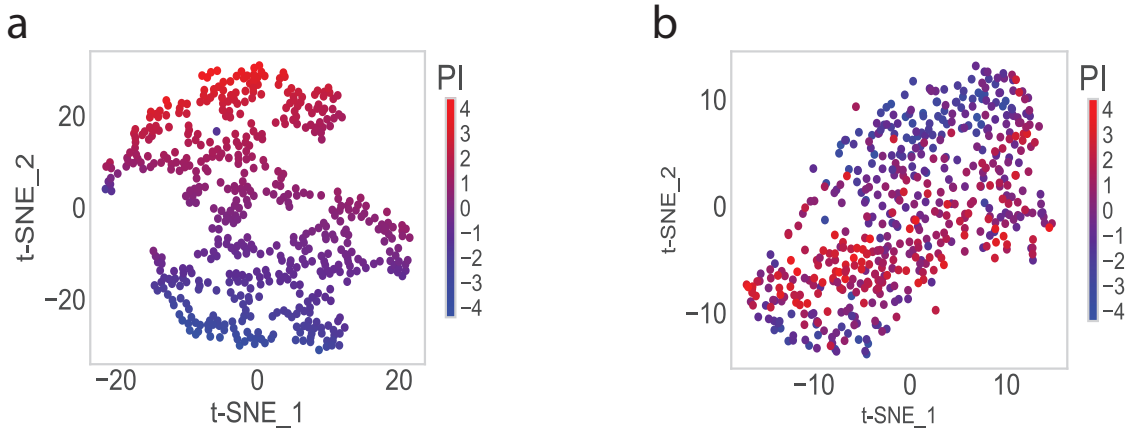

Figure 2.6: Visualization of the top-ranked nodes by Cox-PASNet. (a) t-SNE plot of the statistically significant nodes in the integrative layer (i.e. the second hidden layer (H2) and the clinical layer) and (b) t-SNE plot of the ten top-ranked nodes in the pathway layer.

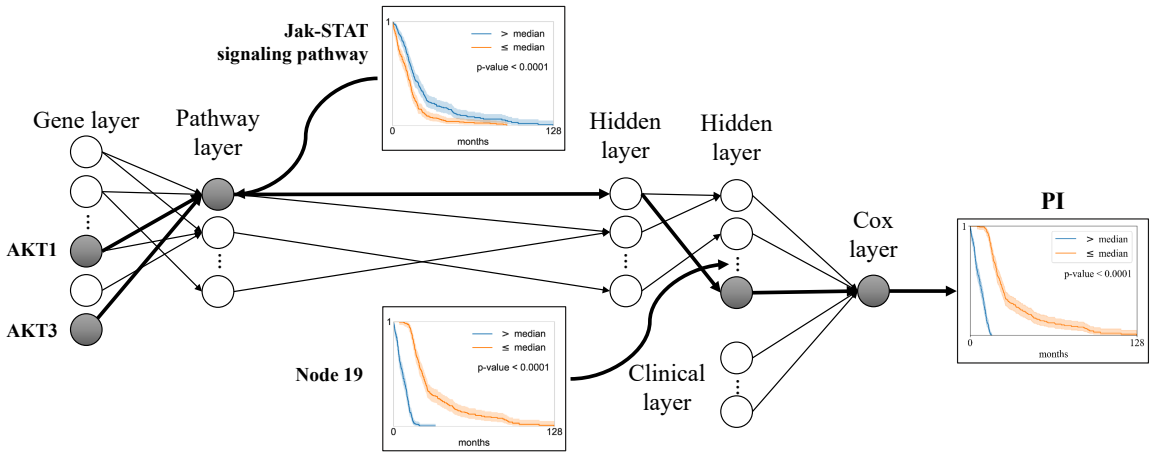

Figure 2.7: Hierarchical and associational feature representation in Cox-PASNet. For instance, Jak-STAT signaling pathway shows active status, which is associated to PI. The significance of the genes (i.e. AKT1 and AKT3) involved in the Jak-STAT signaling pathway can be ranked by the average absolute partial derivatives with respect to the gene layer. A set of the active pathways are represented in an active Node 19 in the following hidden layers, which improves the survival prediction. Note that the Kaplan-Meier plots of Node 19 and PI show more similar estimation of the survival than Jak-STAT signaling pathway.

CHAPTER 3

GENE- AND PATHWAY-BASED DEEP NEURAL NETWORK FOR MULTI-OMICS DATA INTEGRATION TO PREDICT CANCER SURVIVAL OUTCOMES

3.1 Introduction

Data integration of multi-platform based omics data (e.g., genomics, proteomics, and metabolomics) from biospecimens holds promise of improving survival prediction and personalized therapies in cancer [93, 94]. The importance of integrative studies has been increasingly emphasized along with the rapid development of various types of high-throughput multi-omics data. A large scale of multi-omics data sets have been generated in various cancer projects, such as The Cancer Genome Atlas (TCGA) and The Cancer Genome Project in Wellcome Trust Sanger Institute. In particular, TCGA provides various types of omics data of more than 33 cancers, including tissue exome sequencing, gene expression, Copy Number Alternation (CNA), DNA variation, DNA methylation, and microRNA, as well as clinical data such as race, tumor stage, and survival status and months of cancer patients.

Multi-omics data provide comprehensive descriptions of human genomes regulated by complex interactions of multiple biological processes such as genetic, epigenetic, and transcriptional regulation [95]. Thus, the integration of multi-omics data can be leveraged to decipher complex mechanisms of human diseases and to enhance cancer treatments based on genetic understanding of each patient in precision medicine. Specifically, genes are activated by sequential interactions of DNA variations, CNA, histone modifications, transcription factors, DNA methylation, and other genes in relevant pathways [96, 97]. CNA, which is a modified gene structure, often alters downstream pathways or regulatory networks, and DNA methylation often reduces gene expression in a nearby gene when the methyl groups are added to

the DNA. Hence, monozygotic twins discordance in disease is often caused due to different CNA, although they have nearly identical genetic variants [98, 99].

Recently, multi-omics data have been widely incorporated in an increasing number of research projects in survival analysis, rather than using a single type of genomic data that most genomic research traditionally has analyzed. Multi-omics data such as CNA, DNA methylation, and gene expresssion were integrated to identify knowledge-driven genomic interactions with clinical outcomes of interest in ovarian carcinoma [100]. The meta-dimensional models, which incorporate biological pathways with multi-omics data, enhanced the model interpretability in the biological pathway level. A multi-block bipartite graph was proposed not only to identify intra- and inter-block interaction effects of multi-omics data, but also to predict quantitative traits such as gene expression and survival time [101]. SurvivalNet integrated multi-omics data such as DNA mutation, CNA, protein, and mRNA along with clinical information into a deep neural network to improve survival prediction of patients in cancers [72]. Feature selection techniques were applied to each omics dataset separately, and selected features of the multi-omics data and clinical data were combined into a large augmented matrix in SurvivalNet. Another deep learning-based model integrated RNA-Seq, miRNA-Seq, and DNA methylation data to differentiate survival groups in hepatocellular carcinoma [102]. Furthermore, the differential subgroups identified several significant multi-omics features.

In this study, we propose a novel approach, called MiNet, to integrate multi-omics data and clinical data using a pathway-based deep neural network for survival analysis. Our previously published model, Cox-PASNet, which is a pathway-based deep neural network for predicting survival outcome, has considered only gene expression data as well as clinical data [103]. The main contributions of MiNet are as follows: (1) to introduce a multi-omics layer that represents gene-based interaction effects of multi-omics data and (2) to interpret the model in a biological pathway level.

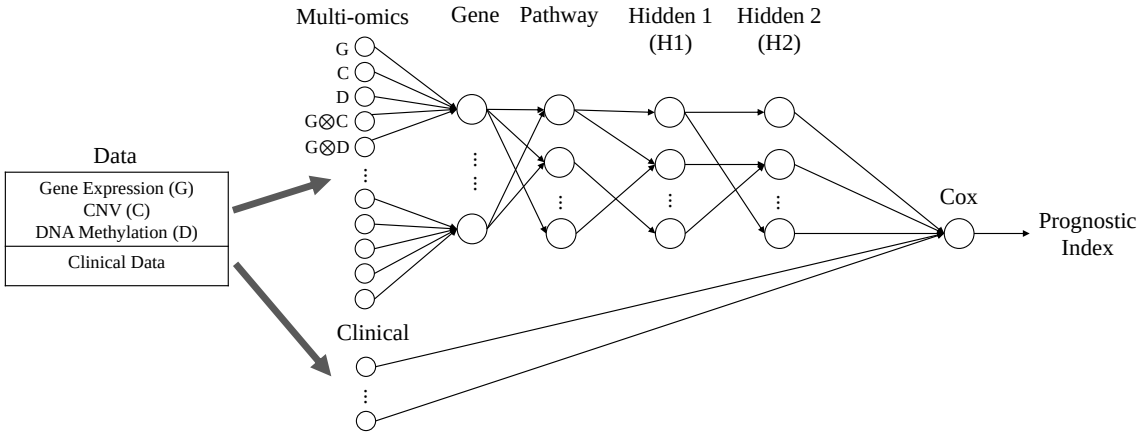

Figure 3.1: The architecture of MiNet

The rest of the paper is organized as follows. In Section 3.2, our proposed model is elaborated in detail. The experimental setting and results are demonstrated in Section 3.3. Section 3.4 discusses the model interpretation with biological findings, while Section 3.5 concludes the discussion.

3.2 Methods

We propose a gene- and pathway-based multi-omics integrative deep neural network (MiNet) to predict cancer survival outcomes. MiNet introduces a gene-based multi-omics layer to integrate multi-omics data, leveraging the advantages of the pathway-based neural network framework in Cox-PASNet [103]. The neural network structure of MiNet follows a biological system, which is multi-layered with multi-omics data and their interactions along with clinical features, by utilizing prior knowledge of biological pathways. The biologically inspired neural network architecture provides a rich interpretation of a biological system.

3.2.1 Multi-Omics Integration

Most studies have integrated multi-omics data by combining all types of omics data to a single matrix and performed analysis, e.g., survival analysis. However, the

consideration of the augmented multi-omics data as independent features lacks to represent interaction effects of genomic and epigenomic data with gene expressions. Note that CNA and DNA methylation often regulate transcriptional mechanisms of genes, so some genes may be down- or up-regulated caused by interaction effects of other omics data [104, 105].

We introduce a multi-omics layer that transfers gene-based interaction effects of multi-omics data to the pathway-based neural network of Cox-PASNet [103]. MiNet generates multi-omics features that include main and interaction effects of multi-omics data on each gene. Then, MiNet inputs the multi-omics features to the multi-omics layer followed by the gene layer that represents *canonical* gene expression level. Note that the gene layer of MiNet consists of *canonical* gene expressions which are high-level representations of gene-based multi-omics data, whereas Cox-PASNet introduces gene expression data directly into the gene layer.

We consider cis-regulatory interaction effects of CNA and DNA methylation to a nearest gene. Multi-omics feature vectors $\mathbf{x}_i$ are generated as:

$$\mathbf{x}_i = \begin{bmatrix} \mathbf{g}_i \\ \mathbf{c}_i \\ \mathbf{d}_i \\ \mathbf{g}_i \otimes \mathbf{c}_i \\ \mathbf{g}_i \otimes \mathbf{d}_i \end{bmatrix}^\top \begin{array}{l} \text{// Main effect of gene expression} \\ \text{// Main effect of CNA} \\ \text{// Main effect of DNA methlyation} \\ \text{// Interaction effect with CNA} \\ \text{// Interaction effect with DNA methlyation} \end{array}, \tag{3.1}$$

where $\mathbf{g}_i$, $\mathbf{c}_i$, and $\mathbf{d}_i$ are sample vectors of gene expression, CNA, and DNA methylation for the i-th gene, respectively. Note that we consider the genes that have at least a gene expression feature. Then, *canonical* gene expression ($\tilde{\boldsymbol{g}}_i$) for the i-th gene is expressed by:

$$\tilde{\boldsymbol{g}}_i = \sigma(\mathbf{x}_i \mathbf{w}_i), \tag{3.2}$$

where $\mathbf{w}_i$ is a weight vector, $\sigma(\cdot)$ is an activation function, and $\otimes$ is element-by-element multiplication. The main or interaction effects are ignored if there is no

CNA or DNA methylation associated to the i-th gene, so genes may have different numbers of multi-omics features.

3.2.2 The Architecture of MiNet

The architecture of MiNet is composed of a multi-omics layer, a gene layer, a pathway layer, multiple hidden layers, a clinical layer, and a Cox layer, as shown in Fig. 3.1. The multi-omics layer is an input layer, which introduces multi-omics features (see Eq. 3.1) from genomics (CNV), epigenomics (DNA methylation), and transcriptomics (gene expression) data into MiNet. The multi-omics layer contains multi-omics features of all genes, and the connections between multi-omics features and genes are implemented by a boolean mask matrix. Note that the associations of multi-omics features are determined with the nearest gene. Most databases often provide genes that CNV and DNA methlayion are mapped to. At the end, every multi-omics features are connected to only a node in the gene layer.

The gene layer represents *canonical* gene features computed by Eq. 3.2, where each node indicates a gene in a biological system. Since a set of genes are involved in biological pathways, genes in the gene layer transfer to corresponding pathway nodes in the pathway layer. Note that the connections between genes and pathways are given by pathway databases, so the number of nodes in the pathway layer is identical with the number of known biological pathways. Hidden layers show hierarchical representations of multiple pathways. A hidden node contains the interaction effect of a set of pathways. More hidden layers may capture more complex interactions of biological pathways.

The clinical layer is an additional input layer for clinical features (e.g., sex, age, and tumor stage). The clinical data are introduced to the output layer as additional features of the last hidden layer, rather than concatenating with the multi-omics layer. The independent clinical layer prevents a few input features from dominating others and makes the model interpretation effective in genomic level. Clinical features, such

as age, have often been shown as significant covariates in several cancer studies. The effects of clinical features may be suppressed by genomic features. Moreover, genomic data and clinical data should be separated for the model interpretation.

The output layer with one node is named as a Cox layer. A linear activation function without bias is applied to this layer to adopt Cox regression. The final outcome of MiNet is Prognostic Index (PI) which is a linear combination of covariates, and PI is introduced to the hazard function for the Cox proportional hazards model as:

$$\lambda(t|\mathbf{x}) = \lambda_0(t)\exp(\text{PI}), \tag{3.3}$$

where PI is an outcome of the Cox layer in MiNet.

3.2.3 Training MiNet with Sparse Coding

MiNet minimizes the average negative log partial likelihood with L^2 regularization. MiNet adapts the training strategy introduced in Cox-PASNet for effective training with high-dimensional, low-sample-size data, where small sub-networks are randomly selected and trained with sparse coding. For the parameter initialization, all layers are fully-connected with He's initialization strategy [106].

The connections between the multi-omics layer and the gene layer are masked by the given boolean mask matrix during the entire training process, similarly in the connections between the gene layer and the pathway layer. Note that the connections between the multi-omics layer and the pathway layer are defined by prior biological knowledge. Sparse coding is applied to the hidden layers following the pathway layer.

We apply sparse coding (L^1 regularization) individually on each layer pair, instead of entire weight matrix. Inspired by LASSO, a soft-thresholding strategy is applied to the connections on each layer pair. Thus, weight matrix is further optimized on each layer pair by:

$$\mathbf{W}^\star \leftarrow S(\mathbf{W}, Q_s), \tag{3.4}$$

where $S(\mathbf{W}, s) = sign(\mathbf{W})(|\mathbf{W}| - Q_s)_+$ is the soft-thresholding function and $sign(\mathbf{W})$ returns a sign of $\mathbf{W}$. $(|\mathbf{W}| - Q_s)_+$ returns $|\mathbf{W}| - Q_s$ if $|\mathbf{W}| - Q_s > 0$, otherwise, $(|\mathbf{W}| - Q_s)_+ = 0$. Q_s is the optimal threshold with respect to the optimal sparsity level s. The optimal sparsity level s is estimated with the strategy proposed in Cox-PASNet [103].

3.3 Experimental Results

In this paper, we conducted experiments with multi-omics data and clinical data in Glioblastoma Multiforme (GBM), which is the most invasive brain tumor. We downloaded multi-omics data including gene expressions, CNAs, and DNA methylations, and clinical data of GBM patients from The Cancer Genome Atlas (TCGA)[1]. We retrieved age, survival status (living or deceased), and survival months of the GBM patients. Age was used as a clinical feature, and both survival status and survival months were used for response variables. The other clinical features were not considered because of large missing values. We filtered out samples with missing values in survival information.

For pathway-based analysis, we downloaded KEGG and Reactome pathway databases from the Molecular Signatures Database (MSigDB) [40]. The pathway databases consist of gene sets of well-known biological pathways, which have molecular interactions in a cell that simultaneously lead to a certain biological process. Small pathways with less than 25 genes were excluded to avoid large redundancy with other pathways [43].

For the experiments, we considered genes that belong to at least one pathway. In particular, 5,481 genes were associated with 507 pathways in the dataset. We included CNAs and DNA methylations associated to the 5,481 genes. Missing values in CNV and DNA methylation features were imputed by 1-Nearest Neighbor (1-NN). Finally, we used 24,803 multi-omics features including interactions and one clinical

[1]https://cancergenome.nih.gov

feature (i.e. age) from 523 samples. The dataset for benchmark models has 14,142 multi-omics features and age from 523 samples, where interactions were excluded. Note that the benchmark methods considered much less numbers of input features than our model.

We compared the performance of MiNet with the current cutting-edge methods: Cox regression with elastic net regularization (Cox-EN) [65], SurvivalNet [72], and Cox-nnet [71]. Concordance index (C-index) was measured to evaluate the performance of the methods. C-index is commonly used to measure the predictive performance in survival analysis. We randomly split the entire data into three subsets of training (64%), validation (16%), and test data (20%) by stratified sampling with survival status, so that each subset preserves the same proportion of censored samples as the entire data. Then, all features were normalized to zero mean with variance of one. Validation and test data were normalized with the mean and variance obtained from training data. Validation data were used to perform early stopping and grid search for finding the optimal hyper-parameters. We repeated the experiments 20 times to show the reproducibility of the performance.

Our proposed method MiNet was implemented by PyTorch 1.0 with CUDA 10. We used ReLU for the activation function, and dropout and L^2 regularization were applied to avoid overfitting problems. Adaptive Moment Estimation (Adam) optimizer was performed to take advantage of a fast convergence and a reduced oscillation. The structure of MiNet was constructed with two hidden layers following multi-omics, gene, and pathway layers, as empirically showing better performance than with a single hidden layer. We considered 22 and 5 nodes in the two hidden layers (H1 and H2) respectively, following the rule of thumb that the number of hidden nodes is the square root of the number of input nodes [71]. Dropout rates were empirically set as 0.7 and 0.5 for pathway layer and hidden layer, respectively. The optimal initial learning rate (η) and L^2 regularization (λ) were determined by grid search that maximizes C-index in validation data on each experiment. All experiments were

Table 3.1: Performance comparison of MiNet with the benchmark methods using C-index in over 20 experiments

Model	C-index ($\mu \pm \sigma$)
Cox-EN [65]	0.5163 ± 0.0359
SurvivalNet [72]	0.5567 ± 0.0312
Cox-nnet [71]	0.5655 ± 0.0287
MiNet (proposed)	$\mathbf{0.6214 \pm 0.0352}$

Table 3.2: Statistical Assessment

	Wilcoxon rank-sum test
MiNet vs. Cox-EN	1e-4*
MiNet vs. Cox-nnet	2e-4*
MiNet vs. SurvivalNet	2e-4*

* shows the statistical significance with significance level = 0.05.

performed with four NVIDIA Tesla M40 (12GB memory) Graphics Processing Units (GPU). The source code of MiNet is publicly available in GitHub[2].

Experiments of SurvivalNet [72] and Cox-nnet [71] were performed by the Python packages published on GitHub[3] [4]. Bayesian optimization [77] was employed in SurvivalNet for the optimal neural network structure and hyper-parameters, such as number of layers, number of nodes, dropout rate, L^1 regularization, and L^2 regularization. For Cox-nnet, grid search strategy was applied for optimal regularization parameter (L^2). Cox-EN was implemented by the package *Glmnet Vignette* in Python [65]. The tunning hyperparamter λ and the elastic-net penalty term α ($\alpha \in [0, 1]$) were optimized by grid search. Kaplan-Meier analysis and log-rank test were performed by using the Python package *lifelines*.

C-index scores obtained from Cox-EN, SurvivalNet, Cox-nnet, and MiNet over 20 experiments with GBM data are shown in Table 3.1. Our proposed method, MiNet, produced the highest C-index of 0.6214 ± 0.0352 among the benchmark meth-

[2]https://github.com/DataX-JieHao/MiNet

[3]https://github.com/CancerDataScience/SurvivalNet

[4]https://github.com/lanagarmire/cox-nnet

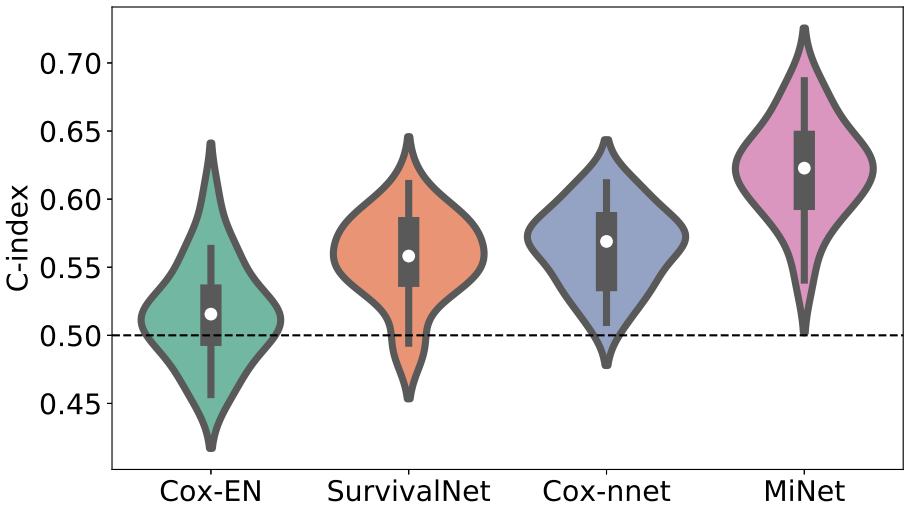

Figure 3.2: Distribution of C-index with 20 experiments

ods, whereas Cox-EN, SurvivalNet, and Cox-nnet showed 0.5163 ± 0.0359, 0.5567 ± 0.0312, and 0.5655 ± 0.0287, respectively. Fig. 3.2 depicts the distribution of C-index of the experiments. Moreover, we performed Wilcoxon rank-sum tests to assess the statistical significance of the model improvement. As shown at Table 3.2, the out-performance of MiNet against the other benchmarks was statistically assessed, i.e., p-values < 0.05.

3.4 Model Interpretation with GBM

For the model interpretation of MiNet with GBM data, we trained the model with the entire data again using the optimal hyper-parameters that have been selected most frequently over 20 experiments (i.e., $\lambda = 0.02$ and $\eta = 0.005$). In consequence, the C-index of the re-trained model was 0.91, which was not overfitted to the input data.

We first examined the six covariates, which are the input nodes to the Cox layer. Five covariates are in the last hidden layer (H2), and one covariate (age) is from the clinical layer. Fig. 3.3a illustrates the H2 and age node values, where the nodes are ranked by the partial derivatives with respect to the H2 layer and the clinical layer.

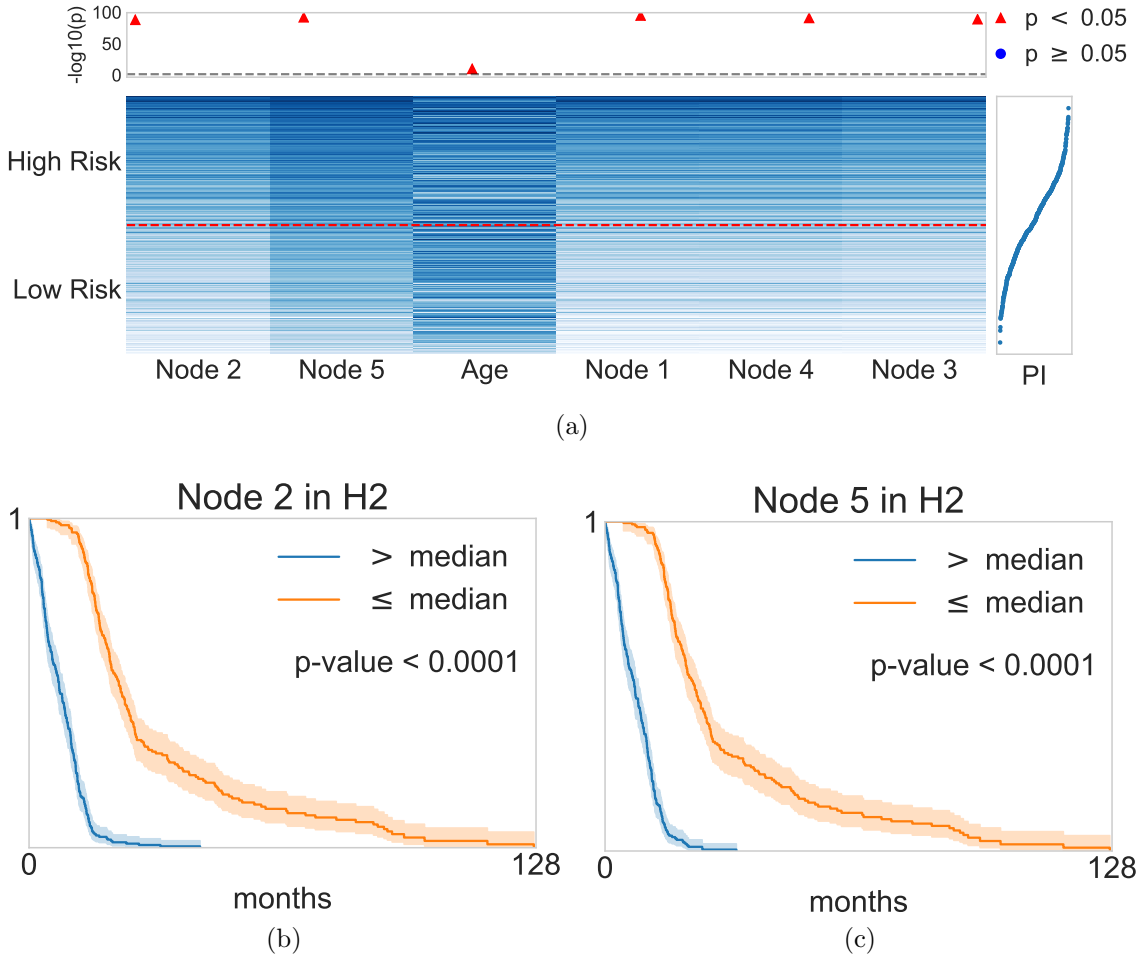

Figure 3.3: Graphical interpretation of the last hidden layer (H2) and the clinical layer. (a) Heatmap of the H2 and age node values. The horizontal dashed line separates high-risk and low-risk groups, which were separated by the median of PI. The upper dot plot shows -log10(p-values) from the logrank test between high-risk and low-risk groups for every single node. The right curve shows the distribution of PI with the corresponding samples on the heatmap. (b) – (c) Kaplan-Meier plots for the two top-ranked covariates.

Overall, the node values show high correlation with PI. Specifically, Node 2 in H2 (the first column in Fig. 3.3a) appeared as the most important covariate for predicting survival time in MiNet with GBM data. For evaluating each covariate, we separated the samples into two groups of high-risk and low-risk by the median of PI. Then, p-values were computed with logrank test. The p-values are shown in the upper plot

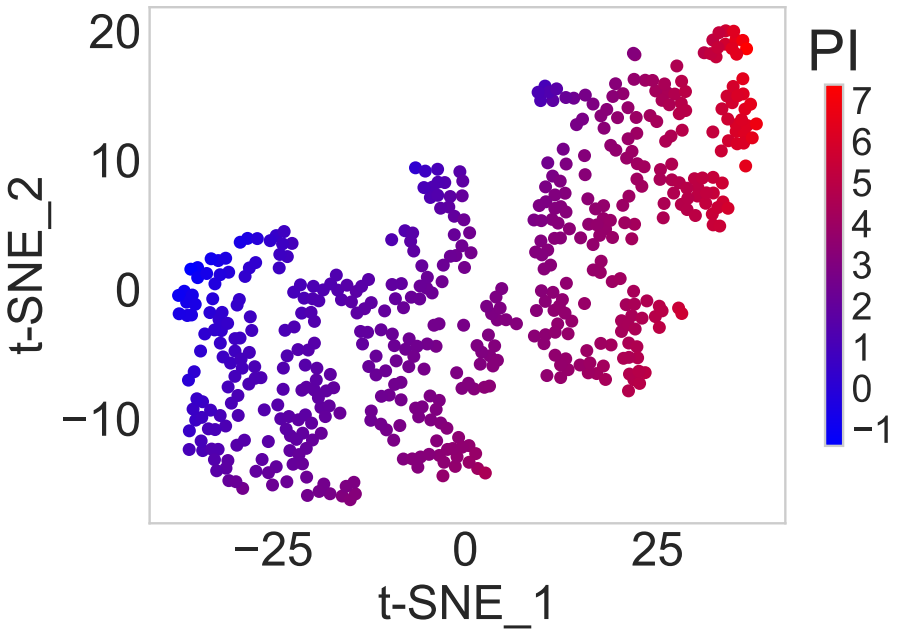

Figure 3.4: Visualization of the H2 and age nodes in MiNet using t-SNE.

in Fig. 3.3a, where all covariates were statistically significant (i.e., p-values < 0.05). Kaplan-Meier plots are depicted in Fig. 3.3b and Fig. 3.3c with the two top-ranked covariates, which demonstrates significantly distinct survival curves. Moreover, the six nodes are visualized by t-SNE in Fig. 3.4, which shows a highly linear correlation between the six covariates and the survival outcomes.

Table 3.3 shows five top-ranked pathways by MiNet, where pathway nodes are ranked by the partial derivatives with respect to the pathway layer. It was discovered that GnRH receptor is expressed in GBM [107]. Interestingly, GnRH signaling pathway was not identified with single omics data, but significantly enriched with multi-omics data [108]. MiNet accordingly ranked GnRH signaling pathway as a significant factor with multi-omics data. Furthermore, the other four pathways have been also recognized in GBM with several biological literature. The references are listed in Table 3.3.

Two genes of NRAS and PRKACA are identified as significant in GnRH signaling pathway (see Table 3.4). Then, we traced back to the multi-omics layer of

Table 3.3: Five top-ranked pathways by MiNet

Pathways	Size	Ref.
GnRH signaling pathway	101	[108]
Genes involved in RNA Polymerase I, RNA Polymerase III, and Mitochondrial Transcription	122	–
Genes involved in Response to elevated platelet cytosolic Ca2+	89	[109]
Melanogenesis	102	[110]
Genes involved in Extracellular matrix organization	87	[111]

Table 3.4: Two top-ranked genes in GnRH signaling pathway

Genes	Multi-omics	Ref.
NRAS	**G** (0.001829), **G** $\otimes$ **C** (0.000888), **D** (0.000791), **C** (0.000319), **G** $\otimes$ **M** (0.000037)	[112]
PRKACA	**C** (0.000774), **G** (0.000738), **G** $\otimes$ **C** (0.000698)	-

the genes. Somatic mutation of NRAS in GBM and its critical role in PI3K-AKT pathway were reported [112]. For NRAS, the main effects of gene expression was the most important factor, followed by the interaction effects of gene expression and CNA and the main effect of DNA methylation. The numbers in parenthesis show partial derivatives with respect to the input nodes, and the higher values indicate the more important factors. For PRKACA, the main effect of CNA and gene expression were highly ranked as the most important multi-omics factors and followed by the interaction effect of CNA, so CNA may play an important role in regulating PRKACA in GBM.

3.5 Conclusion

In this paper, we propose a gene- and pathway-based deep neural network for multi-omics data integration, named MiNet, to predict cancer survival outcomes. In MiNet, gene-based multi-omics features are generated by considering main and interaction effects of multi-omics data in the multi-omics layer. The multi-omics features produce *canonical* gene expression in the gene layer. The hierarchical representations of biological processes of multi-omics, genes, and pathways are captured in MiNet. MiNet showed the outstanding performance to predict cancer survival outcomes with GBM patients. More importantly, MiNet provides the capability to interpret a multi-layered biological system. A large number of biological literature supported our biological findings from MiNet.

The multi-omics layer of MiNet is designed as a neural network module for the integration of multi-omics data, and is compatible to the pathway-based neural network, Cox-PASNet. The high flexibility and expandability of the model architecture would allow one to take an advantage of utilizing the well-established pathway-based framework.

CHAPTER 4

INTERPRETABLE AND INTEGRATIVE DEEP LEARNING FOR SURVIVAL ANALYSIS USING HISTOPATHOLOGICAL IMAGES AND GENOMIC DATA

4.1 Introduction

Integration of histopathological images and genomic data has enhanced personalized treatments and survival predictions in cancer study, while providing an in-depth understanding of both phenotypic patterns and genetic mechanisms of cancer [113, 114]. Pathological images encompass rich phenotypic information with respect to tumor morphology, and high-throughput genomic data have unveiled molecular profiles of cancer [115]. Histopathology, as a clinical gold standard tool in diagnosis and prognosis for most cancers, allows clinicians to make decisions with precision on therapies [116]. Along with an advance of technology in microscopy, digital Whole Slide Imaging (WSI) enables pathologists to manage histopathological tissue slides efficiently. However, manual assessments with large-scale pathology images are highly time-consuming and subjective even by pathologists who have varying levels of experiences.

An increasing number of methods have been developed leveraging machine learning techniques for automatic classification of cancer subtypes, identification of metastases, and nuclei segmentation for pathological image analysis [117]. Deep learning techniques, especially convolutional neural networks, have shown tremendous potential in automatic pathological image analysis. A deep max-pooling convolutional neural network was applied for mitosis detection in breast cancer histological images [118]. A transfer learning-based deep convolutional activation features were extracted to classify glioma grades and to segment the presence of necrosis in GBM, where ImageNet was adopted for a pre-trained model [119]. An ensemble of CNN was developed for improving the predictive performance of tumor grades [120]. In

the ensemble, a CNN classified high- and low- grade glioma, and another CNN further differentiated the grade level in low-grade glioma only. An automatic recognition of nine important nuclear morphological characteristics in glioma pathological images were constructed by a semi-supervised CNN and a pre-trained CNN (i.e. VGG16) with SVM [121].

Survival analysis aims to estimate an expected survival time until a death event occurs. More importantly, a cancer survival model investigates prognostic factors associated to a cancer. The Cox proportional hazards model and its variants are the most commonly applied in medical research. However, the conventional Cox model assumes a linear relationship of covariates, which is barely applied to complex diseases without feature selection on high-dimensional data.

Deep learning-based Cox regressions with pathological images have been studied to tackle the problems of non-linearity and multicollinearity between covariates. Survival Convolutional Neural Networks (SCNNs) were developed to predict patient survival outcomes by high-power fields (HPFs) from Regions Of Interests (ROIs) that show morphological patterns with the representative tumor characteristics [122]. An Whole Slide Histopathological Images Survival Analysis framework (WSISA) was proposed to directly learn discriminative patches based on cluster-level Deep Convolutional Survival models for predicting patients' survivals [74]. The study introduced an aggregation strategy based on the weighted features evaluated by the performance in each cluster.

Recently, the integration of pathological and genomic data has been explored as a promising solution for predicting cancer survival outcomes. A lasso-regularized Cox proportional hazards model extracted pre-defined morphological features from digital WSIs and eigen-genes from gene coexpression data in clear cell renal cell carcinoma and outperformed the models with either morphoological features or eigen-genes individually [113]. A multiple kernel learning-based method was introduced to extract heterogeneous features from multiple types of genomic data and pathological images

in breast cancer [123]. Genomic Survival convolutional neural networks (GSCNN) integrated heterogeneous features from both pathological images and well-known genomic biomarkers for predicting patients survival with glioma [122].

Although the integrative models produced higher predictive performance than single data models for cancer survivals, most integrative models require intensive data preprocessing with manually annotated ROIs on pathological images and stringent feature selection to reduce numbers of input features, e.g., using well-known genomic biomarkers in biological literature. For instance, GSCNN integrated only two well-known genomic biomarkers with pretrained SCNN model for reducing a number of covariates and false negative prognostic factors [122]. Pre-defined image features of geometry, texture, and holistic statistics were extracted from Hematoxylin and Eosin (H&E) pathological slides, prior to integrating with gene expression data [114].

In this paper, we propose a biologically interpretable integrative deep learning model that integrates PAthological images and GEnomic data, called PAGE-Net, not only for improving survival predictive performance but also identifying genetic and pathological patterns that may cause different survivals between patients. The major methodological challenges are data heterogeneity and complexity, when integrating unstructured mega-pixel pathological images and structured genomic data. Our main contributions of PAGE-Net for cancer survival analysis are threefold:

- to integrate pathological images and genomic data in a biologically interpretable deep learning model,
- to identify survival-discriminative features without manually annotated ROIs, and
- to provide an aggregation strategy that aggregates patch-level features generated from multiple patches and produces image-level global features.

4.2 Methods

4.2.1 The Architecture of PAGE-Net

PAGE-Net consists of pathology-specific layers, genome-specific layers, and a demography-specific layer, each of which provides interpretability of biological mechanism and morphological phenotypic patterns associated to cancer survivals, as illustrated in Fig. 4.1. In order to tackle the integration challenge between an unstructured mega-pixel WSI and structured genomic data, we propose a novel patch-wise texture-based convolutional neural network with a patch aggregation strategy (described in Section 4.2.2 in detail) to extract survival-discriminative features without manually annotated ROIs for pathology-specific layers. First, survival-discriminative features are identified by a pre-trained deep learning model with uncensored data only. Then, the feature scores are aggregated from multiple patches of a whole slide image, which generates a structured vector data. For the genome-specific layers and the demography-specific layer, we adapt our previously proposed pathway-based sparse deep neural network, named Cox-PASNet [103]. Cox-PASNet is a cutting-edge deep learning model that interprets biological mechanism by incorporating gene expression data, clinical data, and prior biological knowledge of pathways, while holding outstanding predictive performance in patients' survival with high-dimension, low-sample size biological data. Finally, the high-level representations of pathological and genomic data along with clinical information are introduced to a shared layer that estimates Prognostic Index (PI) in a Cox proportional-hazards regression model.

4.2.2 Pathology-Specific Layers

In the pathology-specific layers, survival-discriminative features, which are identified in advance by a pre-trained CNN, are extracted from multiple patches of a pathological image. Then, the features are aggregated by a two-stage pooling strategy and introduced to Cox layer along with the last hidden layer of the genome-specific

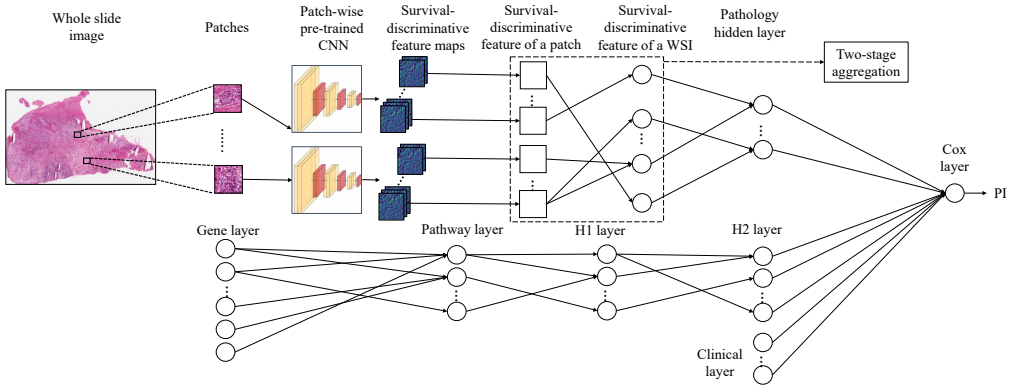

Figure 4.1: The architecture of PAGE-Net

layers and the clinical layer. We elucidate the pre-trained CNN model and the aggregation strategy in the following subsections.

4.2.2.1 Patch-Wise Pre-Trained CNN

We train a CNN model to identify survival-discriminative feature maps with patches from uncensored pathological images prior to the proposed integrative deep learning model. Morphological patterns of pathological images are captured by the pre-trained CNN with dilated convolutional layers. Dilated convolutional layers enlarges field-of-view (texture) without loss of spatial information [124]. The number of parameters does not increase with dilation, which makes model training computationally efficient. Moreover, dilated convolutional layers trade off computational time against context assimilation [125].

The CNN model is comprised of an input layer, three pairs of dilated convolutional layers (kernel size of 5×5, 50 feature maps, and dilation rate of 2) and a max-pooling layer of size of 2×2. The sequential layers are followed by a flatten layer and a fully connected layer. We use a linear model as the output layer, since the model is trained with only uncensored data. Finally, the 50 neurons in the last max-pooling layer are considered as survival-discriminative features in the integrative model.

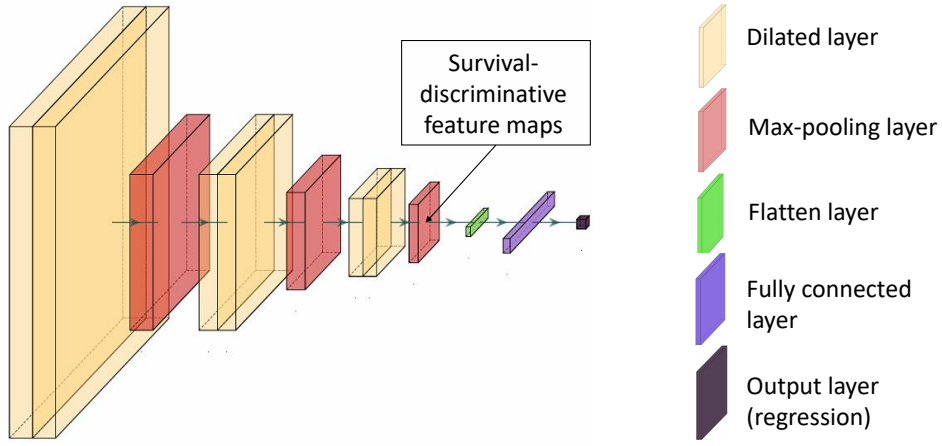

Figure 4.2: The architecture of the pre-trained CNN

4.2.2.2 Two-Stage Aggregation

Global survival-discriminative features for a WSI are generated by a two-stage pooling aggregation strategy. Each patch image produces N numbers of *local* survival-discriminative feature scores by the pre-trained CNN, and the scores of multiple patches from a WSI are aggregated. The aggregated scores are introduced to the last hidden layer in the pathology-specific layers.

We adapt a two-stage pooling approach [126] by computing 3-norm pooling so that only a highly-ranked subset of patches are considered [119, 127]. The first stage pooling ranks survival-discriminative features and identifies the most important features. Then, the second stage pooling forms global survival-discriminative features by aggregating only top-ranked patches.

The first stage pooling: Suppose that we have N survival-discriminative feature maps identified by the pre-trained CNN on each patch image (i.e., 50 neurons in the last max-pooling layer of the pre-trained CNN in this study). Let $\mathbf{X}$ denote

N survival-discriminative feature maps, where $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \ldots, \mathbf{X}_N]$. The i^{th} survival-discriminative feature map, $\mathbf{X}_i$ $(1 \leq i \leq N)$, can be represented as:

$$\mathbf{X}_i = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \ldots & x_{1w} \\ x_{21} & x_{22} & x_{23} & \ldots & x_{2w} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{h1} & x_{h2} & x_{h3} & \ldots & x_{hw} \end{bmatrix}, \tag{4.1}$$

where h and w are the height and the width of the feature map respectively (e.g., $h = w = 18$ in this study). Then, the flattened feature map becomes $\mathbf{X}_i^f = [x_{11}, x_{12}, x_{13}, \ldots, x_{hw}]$. After sorting the flattened feature map in the descending order, we consider top K_1 features as significant survival-discriminative feature map components, which is $\tilde{\mathbf{X}}_i^f = \tilde{x}_1, \tilde{x}_2, \tilde{x}_3, \ldots, \tilde{x}_{K_1}$. Then, a 3-norm pooling value on $\tilde{\mathbf{X}}_i^f$ is computed by:

$$f_i = \frac{1}{K_1}(\sum_{j=1}^{K_1}(\tilde{x}_j^f)^3)^{1/3}, \tag{4.2}$$

where f_i is an aggregated score for the i^{th} feature map on a patch.

The second stage pooling: Suppose that M numbers of patches are available on a WSI. The aggregated feature maps of all patches after the first stage pooling can be represented as:

$$\boldsymbol{F} = \begin{bmatrix} f_{11} & f_{12} & f_{13} & \ldots & f_{1N} \\ f_{21} & f_{22} & f_{23} & \ldots & f_{2N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_{M1} & f_{M2} & f_{M3} & \ldots & f_{MN} \end{bmatrix}, \tag{4.3}$$

where f_{ij} is the j^{th} feature map of the i^{th} patch on a WSI, which is computed by Eq. (4.2). For each column of $\mathbf{F}$ (i.e. feature maps over M patches), we sort column-wise values in the descending order. The top K_2 number of values (i.e. important patches) in each column are truncated, i.e., $\tilde{f}_{ij}, 1 \leq i \leq K_2$. Then, another 3-norm pooling is performed on each column of the truncated $\mathbf{F}$. An aggregated score of top K_2 discriminative patches is obtained for a feature map. Therefore, a vector

of N aggregated survival-discriminative features represents a pathological WSI for a patient. In this study, we used $N = 50$, $M = 1000$, $K_1 = 65$, and $K_2 = 100$.

4.2.3 Genome- and demography-specific layers

The genome- and demography-specific layers are adapted from the pathway-based sparse deep neural network, Cox-PASNet [103]. The genome-specific layers include a gene layer, a pathway layer, and two hidden layers (H1 and H2). The gene layer is an input layer for gene expression data, where each node indicates a gene. The pathway layer embeds a prior biological knowledge using well-known biological pathway databases (e.g., KEGG) for biological interpretation. The connection between the gene layer and the pathway layer are sparsely established by given biological pathway databases where the relationships between genes and pathways are available. Hence, each pathway node explicitly represents a biological pathway. The following two hidden layers capture nonlinear and hierarchical relationships between pathways. Clinical data of a patient are directly introduced to the demography-specific layer and combined with genomic features from gene expressions and aggregated survival-discriminative features from a pathological image in the last hidden layer of the integrative model.

Overfitting is a critical issue to avoid when training a deep learning model with high-dimension, low-sample-size data. In order to prevent the overfitting problem, PAGE-Net applies the training technique that Cox-PASNet proposed [103]. Instead of training the whole network, small networks are randomly selected, and sparse coding was applied to make connection sparse for model interpretation. The training is repeated until it converges. Errors with the validation data was also traced for early stopping and preventing overfitting.

4.3 Experimental Results

We examined pathological images, gene expression data, and clinical data of Glioblastoma Multiforme (GBM) patients to assess our proposed model. The data were downloaded from The Cancer Imaging Archive (TCIA) [1] and The Cancer Genome Atlas (TCGA) that provide pathological images and genomics data from an identical set of patients. We considered only GBM patients' data where both gene expression and pathological image are available. We also filtered out the data without survival information. We included age only as a clinical feature for the demography-specific layer, i.e. clinical layer, since a large amount of missing values are shown in other clinical features.

KEGG and Reactome pathway databases, taken from the Molecular Signatures Database (MSigDB) [40, 41, 42], were used for biological pathways in the model, as a prior biological knowledge. Biological pathways that have either less than fifteen genes or over 300 genes were excluded [43]. Furthermore, only genes that belong to at least one pathway were considered as inputs to the model. Finally, we considered 5,404 genes of 447 GBM patients and 659 pathways were examined. For the pathological WSI, we considered WSIs of "top" frozen tissue sections with the 20X magnification. In the pre-training phase, 1,000 patches of size 256×256 were randomly sampled from the uncensored data for training the pre-train CNN. Note that only uncensored training and validation data were used for the pre-trained CNN on each experiment. In the integration phase, we sampled other 1,000 patches from a WSI for training and testing.

We compared the predictive performance of PAGE-Net with Cox-PASNet and Cox regression with elastic net regularization (Cox-EN) [65]. Cox-PASNet was applied to gene expressions and age, whereas aggregated survival-discriminative image features were trained by Cox-EN. Concordance index (C-index) with counted tied prediction pairs was measured to scale the performance of the models. The samples

[1] `https://www.cancerimagingarchive.net/`

were randomly split into training (80%), validation (10%), and test (10%) sets, by preserving the proportion between censored and uncensored statuses. The features in the training set were normalized to mean of zero and standard deviation of one. The validation and the test sets were normalized by the mean and standard deviation from the training set. We repeated the experiments twenty times to show the reproducibility of the performance.

PAGE-Net was implemented by PyTorch 1.0 with CUDA 10.0.130 and Keras 2.2.4 with TensorFlow 1.13.1 as backend. The model was optimized with the dilated kernel size of 5×5, dilated rate r of 2, and the max pooling size is 2×2. A dropout rate of 0.3 was applied for each dilated conventional layers and flatten layer. We used Adaptive Moment Estimation (Adam) optimizer and ReLU activation function. Mean squared error (MSE) was computed as the loss. A grid search was performed on each experiment to optimize a learning rate and a mini-batch size using validation data with a learning rate decay of 0.7 for every 5 epochs. An early stopping upon validation loss was applied.

In the integration phrase, Tanh function was used as the activation function between layers. We set 100, 30, and 30 nodes for H1, H2, and the pathology hidden layer, respectively. Dropout rates were empirically set as 0.7, 0.5, and 0.3 for the pathway layer, H1, and the global survival-discriminative feature layer, respectively. The optimal learning rate and L^2 regularization (λ) were automatically determined by grid search so as to maximize C-index with the validation data on each experiment. All experiments were performed with two NVIDIA Tesla M40 (8 cores, 12GB memory per each core) Graphics Processing Units (GPUs). The source code of PAGE-Net is accessible online via GitHub (`https://github.com/DataX-JieHao/PAGE-Net`). For the benchmark methods, Cox-PASNet was performed in the proposed manner in the paper. Cox-EN was implemented by the Python version of *Glmnet Vignette* [65]. 200 λs were considered for optimization. The regularization term α between zero and one was optimized by grid search with a step size of 0.01.

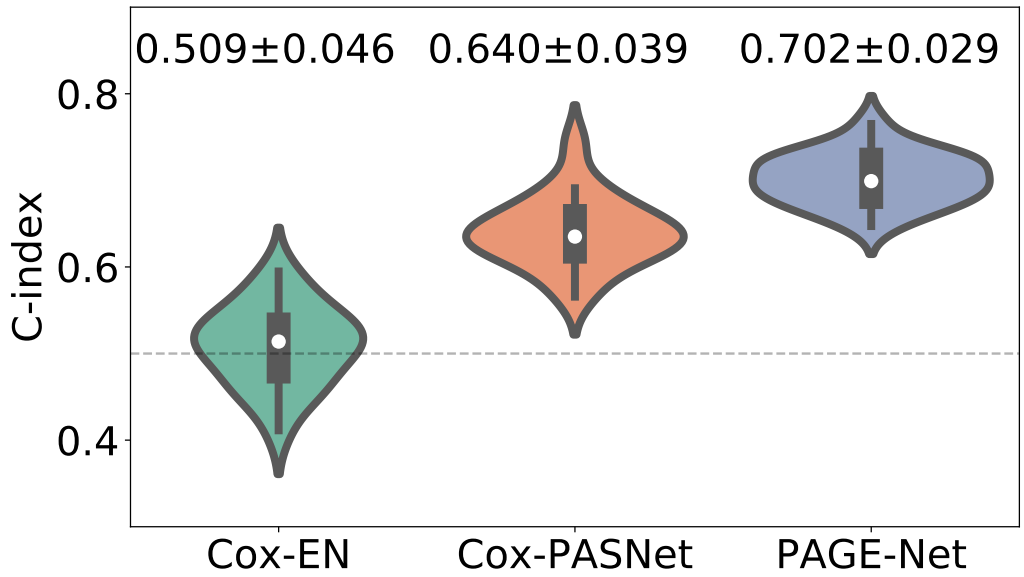

Figure 4.3: Performance comparison over 20 experiments with GBM in C-index

The experimental results with GBM data are shown in Fig. 4.3. Our proposed model, PAGE-Net, achieved the highest C-index of 0.702 ± 0.0294 (mean ± std) comparing to Cox-PASNet (with gene expressions and age) showing C-index of 0.6401 ± 0.00399, and Cox-EN (with aggregated image features) showing the lowest C-index of 0.5093 ± 0.0460. The highest C-index of PAGE-Net shows the increased power of the integrative model with pathological data and genomic data. Interestingly, a pathological WSI itself contributes little to the predictive performance. However, the experimental results show that the pathological WSI boosted the performance of survival analysis with genomic data in the proposed integrative model. The performances were assessed by Wilconxon rank-sum test, and PAGE-Net statistically outperformed Cox-EN with pathological images only and Cox-PASNet with genomic data only (both p-values are less than 0.0001).

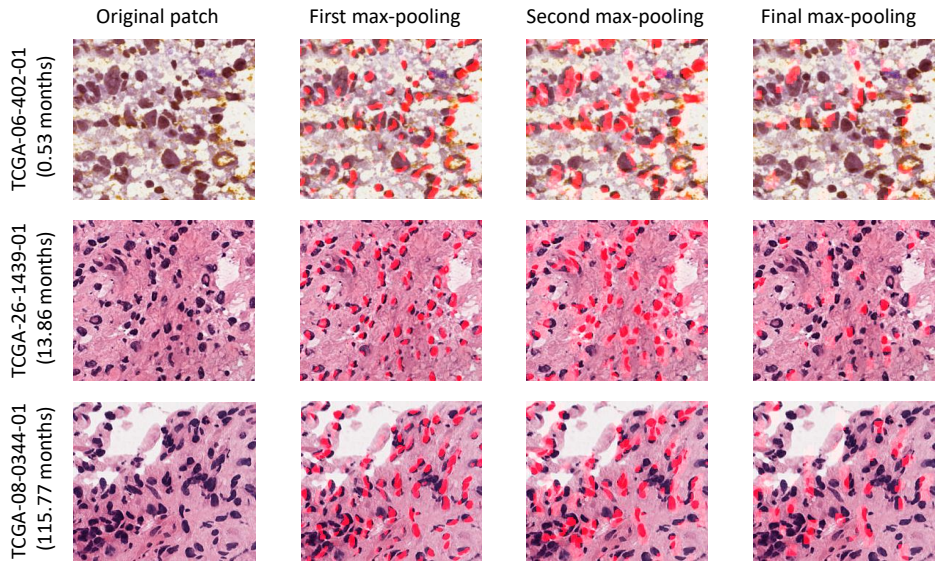

Figure 4.4: Survival-discriminative feature maps on the patches of three patients in various survivals

4.4 Model Interpretation

For the model interpretation of PAGE-Net, we re-trained the proposed models using the entire data and the optimal hyper-parameters that were most commonly used over the 20 experiments. We performed the analysis for biological interpretation with the pathology- and genome-specific layers. For the pathology-specific layers, we assessed pathological and morphological patterns of the survival-discriminative feature maps with a pathologist. For the genome-specific layers, we conducted the pathway-based interpretation by ranking the nodes with partial derivatives, as Cox-PASNet conducted [103].

Figure 4.4 exhibits pathological top-ranked patch images of three patients in a short (first row; *TCGA-06-402-01*; survival month = 0.53), median (second row; *TCGA-26-1439-01*; survival month = 13.85), and long-term (third row; *TCGA-08-0344-01*; survival month = 115.3) survivals and the survival-discriminative feature maps captured by the pre-trained CNN on the patches. The survival-discriminative feature map scores (higher than the median) are colored in red in the figures. Inter-

estingly, the survival-discriminative feature maps capture most nucleus and nuclear debris of interest on the patches. In GBM where boundaries between nuclei are not clearly shown, a distance between nuclei and a shape of nucleus are critical checkpoints on tissue readings. The feature maps show that the morphological patterns of pathologist's interest are also recognized by the proposed model. Moreover, nuclear debris implies necrosis of nucleus, and the relationship between nuclear debris and survival prognosis is known. The top-ranked patches were measured scores of nuclear pleomorphism (NP), cytoplasmic degeneration (CD), and brown pigment (BP) using three tiered scoring by a pathologist. The scores of NP, CD, and BP on *TCGA-06-402-01* were +3, +3, and +3, whereas the scores of *TCGA-26-1439-01* and *TCGA-08-0344-01* were +1, 0, and 0. The patch of the patient, *TCGA-06-402-01*, shows more severe scores on NP, CD, and BP than other two patients. It shows that PAGE-Net can also identify regions (patches) associated to patients' survival on a WSI.

Ten top-ranked pathways and genes in GBM are ranked with the genome-specific layers in PAGE-Net. The pathways and genes are listed in Table 4.1 and Table 4.2. Neuroactive ligand-receptor interaction pathway, ranked top one by PAGE-Net, is well known as one of the most associated pathways to GBM [80]. Survival models by both univariate and multivariate Cox regression analysis for the nine long noncoding RNAs (lncRNAs) in GBM identified neuroactive ligand-receptor interaction pathway as the most related pathway [128]. Axon guidance pathway harbored the top-ranked CNVs with respect to GBM [83]. The downregulation of endocytosis pathway was likely to be a common trait in glioma tumors [129]. For instance, the down-regulated differentially expressed genes (DEGs) assoiciated with the glioma gene expression profile GSE4290 were enriched in endocytosis pathway [130]. Collagen formation pathway enriched for the candidate genes identified by weighted gene co-expression network analysis with RNA sequencings of GBM patients from the Chinese Glioma Genome Atlas database [131]. 18 cytokines, which differentiated normal and GBM serum samples, were enriched in both of cytokine-cytokine receptor interac-

Table 4.1: Ten top-ranked pathways in GBM by PAGE-Net

Pathway name	# of genes	P-value	References
Neuroactive ligand-receptor interaction	272	< 0.0001	[80, 128]
Axon guidance	129	< 0.0001	[83]
Transmission across chemical synapses	186	< 0.0001	–
G alpha (s) signalling events	121	< 0.0001	–
Neuronal system	279	< 0.0001	–
Endocytosis	183	< 0.0001	[129, 130]
Tyrosine metabolism	42	0.3924	–
Collagen formation	58	0.1041	[131]
Neurotransmitter receptor binding and downstream transmission in the postsynaptic cell	137	< 0.0001	–
Cytokine-cytokine receptor interaction	267	< 0.0001	[132]

Table 4.2: Ten top-ranked genes in GBM by PAGE-Net

Gene name	P-value	Ref.
PTGER4	0.5679	[133]
NPY2R	0.0358	–
LHB	0.1379	–
GHRHR	0.0578	[134]
ADRB3	0.0217	–
ADORA2A	0.0064	[135]
MET	0.0066	[136]
FSHB	0.0330	–
HTR7	0.8468	[92]
GRM8	0.6673	[137]

tion and JAK-STAT pathways [132]. Furthermore, overexpressed ADORA2A is one of the evidences for high-grade gliomas by the World Health Organization (WTO) [135]. HTR7, enriched in neuroactive ligandreceptor interaction, was reported to contribute the diffuse intrinsic pontine glioma development and progression [92]. MET, well-known as an oncogene, has been revealed as a functional marker in Glioblastoma stem cells since it benefits glioma invasiveness and self-reconstruction [136].

Figure 4.5 shows hierarchical biological mechanisms on both pathological images and genomic data in PAGE-Net. In the pathology-specific layers, morphological patterns, which are associated to patients' survivals, are scored by the survival-discriminative features, and the global features are introduced to the model. The survival-discriminative feature maps substantially capture the nucleus and nuclear debris of interest on a WSI. In the genome-specific layers, activated genes including ADORA2A and ADORA2B trigger the neuroactive ligand-receptor interaction pathway, and the pathway contributes patient's survivals in a non-linear manner with other pathways in the hidden layers. The Kaplan-Meier plots of the pathway and Node 13 in the H1 layer shows the different survival distribution with the two groups

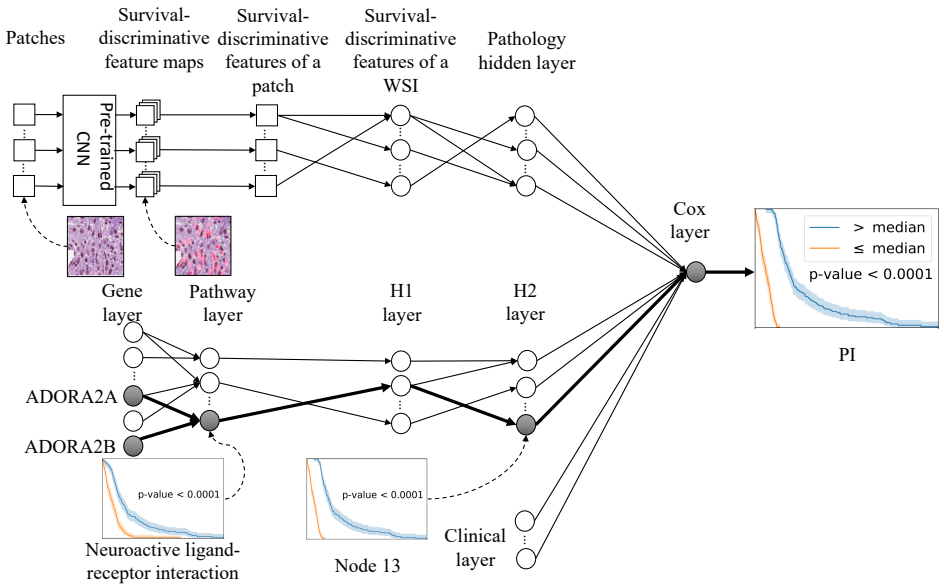

Figure 4.5: Overview of the model interpretation

separated by the median of the node values. The Node 13 values can be considered as a potential prognostic factor that can predict patient's survival.

4.5 Conclusion

In this paper, we propose an integrative deep learning model (PAGE-Net) that captures both morphological patterns on pathological WSIs and pathway-based genetic mechanisms of a complex human cancer, while predicting cancer survival outcomes with pathological images and genomic data. PAGE-Net produced the outstanding predictive performance and showed promising potential to identify genetic and pathological prognostic factors simultaneously associated with patients survival. The survival-discriminative features identified by the pre-trained CNN was asssessed by a pathologist that the features can identify nucleus and nuclear debris, which may be related to patient' survivals. The integrative deep learning model, PAGE-Net, also shows that the data integration of pathological images and genomic data is essential for enhancing patient's survival rather than analyses with a single data type.

REFERENCES

[1] J. Lu, M. C. Cowperthwaite, M. G. Burnett, and M. Shpak, "Molecular Predictors of Long-Term Survival in Glioblastoma Multiforme Patients," *PLOS ONE*, vol. 11, no. 4, p. e0154313, 2016.

[2] M. W. Onaitis *et al.*, "Prediction of Long-Term Survival After Lung Cancer Surgery for Elderly Patients in The Society of Thoracic Surgeons General Thoracic Surgery Database," *The Annals of Thoracic Surgery*, vol. 105, no. 1, pp. 309–316, 2018.

[3] Y. Cao *et al.*, "Prediction of longterm survival rates in patients undergoing curative resection for solitary hepatocellular carcinoma," *Oncology Letters*, vol. 15, no. 2, pp. 2574–2582, 2018.

[4] L. Jin *et al.*, "Pathway-based Analysis Tools for Complex Diseases: A Review," *Genomics, Proteomics & Bioinformatics*, vol. 12, no. 5, pp. 210–220, 2014.

[5] S. Kim, M. Kon, and C. DeLisi, "Pathway-based classification of cancer subtypes," *Biology Direct*, vol. 7, p. 21, 2012.

[6] E. Cirillo, L. D. Parnell, and C. T. Evelo, "A Review of Pathway-Based Analysis Tools That Visualize Genetic Variants," *Frontiers in Genetics*, vol. 8, p. 174, 2017.

[7] Y. Drier, M. Sheffer, and E. Domany, "Pathway-based personalized analysis of cancer," *Proceedings of the National Academy of Sciences*, vol. 110, no. 16, pp. 6388–6393, 2013.

[8] T. Mallavarapu, Y. Kim, J. H. Oh, and M. Kang, "R-PathCluster: Identifying cancer subtype of glioblastoma multiforme using pathway-based restricted boltzmann machine," in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Nov. 2017, pp. 1183–1188.

[9] S. Huang *et al.*, "Novel personalized pathway-based metabolomics models reveal key metabolic pathways for breast cancer diagnosis," *Genome Medicine*, vol. 8, p. 34, 2016.

[10] Y. Li, B. Nan, and J. Zhu, "Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure," *Biometrics*, vol. 71, no. 2, pp. 354–363, 2015.

[11] J. M. Raser and E. K. O'Shea, "Noise in Gene Expression: Origins, Consequences, and Control," *Science*, vol. 309, no. 5743, pp. 2010–2013, 2005.

[12] E. W. Steyerberg, M. J. C. Eijkemans, and J. D. F. Habbema, "Application of Shrinkage Techniques in Logistic Regression Analysis: A Case Study," *Statistica Neerlandica*, vol. 55, no. 1, pp. 76–88, 2001.

[13] S. Wang, B. Nan, S. Rosset, and J. Zhu, "Random lasso," *The Annals of Applied Statistics*, vol. 5, no. 1, pp. 468–485, 2011.

[14] J. Z. Musoro, A. H. Zwinderman, M. A. Puhan, G. ter Riet, and R. B. Geskus, "Validation of prediction models based on lasso regression with multiply imputed data," *BMC Medical Research Methodology*, vol. 14, no. 1, p. 116, 2014.

[15] D. Liu, X. Lin, and D. Ghosh, "Semiparametric Regression of Multidimensional Genetic Pathway Data: Least-Squares Kernel Machines and Linear Mixed Models," *Biometrics*, vol. 63, no. 4, pp. 1079–1088, 2007.

[16] D. Liu, D. Ghosh, and X. Lin, "Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models," *BMC Bioinformatics*, vol. 9, no. 1, p. 292, 2008.

[17] F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan, "Multiple Kernel Learning, Conic Duality, and the SMO Algorithm," in *Proceedings of the Twenty-first International Conference on Machine Learning*, ser. ICML '04, 2004, p. 6.

[18] J. A. Sinnott and T. Cai, "Pathway aggregation for survival prediction via multiple kernel learning," *Statistics in Medicine*, vol. 37, no. 16, pp. 2501–2515, 2018.

[19] S. Kumari *et al.*, "Bottom-up GGM algorithm for constructing multilayered hierarchical gene regulatory networks that govern biological pathways or processes," *BMC Bioinformatics*, vol. 17, no. 1, p. 132, 2016.

[20] W. Deng, K. Zhang, V. Busov, and H. Wei, "Recursive random forest algorithm for constructing multilayered hierarchical gene regulatory networks that govern biological pathways," *PLOS ONE*, vol. 12, no. 2, p. e0171532, 2017.

[21] L. M. Pham, L. Carvalho, S. Schaus, and E. D. Kolaczyk, "Perturbation Detection Through Modeling of Gene Expression on a Latent Biological Pathway Network: A Bayesian Hierarchical Approach," *Journal of the American Statistical Association*, vol. 111, no. 513, pp. 73–92, 2016.

[22] S. Kher, J. Peng, E. S. Wurtele, and J. Dickerson, "Hierarchical Biological Pathway Data Integration and Mining," in *Bioinformatics*, H. Pérez-Sánchez, Ed. Rijeka: IntechOpen, 2012, ch. 1. [Online]. Available: https://doi.org/10.5772/49974

[23] S. Min, B. Lee, and S. Yoon, "Deep learning in bioinformatics," *Briefings in Bioinformatics*, vol. 18, no. 5, pp. 851–869, 2016.

[24] M. Liang, Z. Li, T. Chen, and J. Zeng, "Integrative Data Analysis of Multi-Platform Cancer Data with a Multimodal Deep Learning Approach," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 12, no. 4, pp. 928–937, 2015.

[25] H. Zeng, M. D. Edwards, G. Liu, and D. K. Gifford, "Convolutional neural network architectures for predicting DNA-protein binding," *Bioinformatics*, vol. 32, no. 12, pp. i121–i127, 2016.

[26] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, "Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning," *Nature Biotechnology*, vol. 33, no. 8, pp. 831–838, 2015.

[27] J. Zhou and O. G. Troyanskaya, "Predicting effects of noncoding variants with deep learning-based sequence model," *Nature Methods*, vol. 12, pp. 931–934, 2015.

[28] T. Ching *et al.*, "Opportunities and obstacles for deep learning in biology and medicine," *Journal of The Royal Society Interface*, vol. 15, no. 141, 2018.

[29] J. Ma *et al.*, "Using deep learning to model the hierarchical structure and function of a cell," *Nature Methods*, vol. 15, pp. 290–298, 2018.

[30] B. Liu, Y. Wei, Y. Zhang, and Q. Yang, "Deep Neural Networks for High Dimension, Low Sample Size Data," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, ser. IJCAI'17, 2017, pp. 2287–2293. [Online]. Available: http://dl.acm.org/citation.cfm?id=3172077.3172206

[31] A. Pasini, "Artificial neural networks for small dataset analysis," *Journal of Thoracic Disease*, vol. 7, no. 5, 2015.

[32] P. I. Wójcik and M. Kurdziel, "Training neural networks on high-dimensional data using random projection," *Pattern Analysis and Applications*, 2018.

[33] Y. Li, C.-Y. Chen, and W. W. Wasserman, "Deep Feature Selection: Theory and Application to Identify Enhancers and Promoters," *Journal of Computational Biology*, vol. 23, no. 5, pp. 322–336, 2016.

[34] S. Han *et al.*, "DSD: Regularizing Deep Neural Networks with Dense-Sparse-Dense Training Flow," *CoRR*, vol. abs/1607.04381, 2016.

[35] B. Wang and D. Klabjan, "Regularization for Unsupervised Deep Neural Nets," *CoRR*, vol. abs/1608.04426, 2016.

[36] S. Wang *et al.*, "Training deep neural networks on imbalanced data sets," in *2016 International Joint Conference on Neural Networks (IJCNN)*, July 2016, pp. 4368–4374.

[37] F. Hanif, K. Muzaffar, k. Perveen, S. M. Malhi, and S. U. Simjee, "Glioblastoma multiforme: A review of its epidemiology and pathogenesis through clinical pre-

sentation and treatment," *Asian Pacific Journal of Cancer Prevention*, vol. 18, no. 1, pp. 3–9, 2017.

[38] M. E. Davis, "Glioblastoma: Overview of Disease and Treatment," *Clinical Journal of Oncology Nursing*, vol. 20, no. 5, pp. S2–S8, 2016.

[39] M. S. Walid, "Prognostic Factors for Long-Term Survival after Glioblastoma," *The Permanente journal*, vol. 12, no. 4, pp. 45–48, 2008.

[40] A. Subramanian *et al.*, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences*, vol. 102, no. 43, pp. 15 545–1 554 550, 2005.

[41] A. Liberzon *et al.*, "Molecular signatures database (MSigDB) 3.0," *Bioinformatics*, vol. 27, no. 12, pp. 1739–1740, 2011.

[42] ——, "The Molecular Signatures Database Hallmark Gene Set Collection," *Cell Systems*, vol. 1, no. 6, pp. 417–425, 2015.

[43] J. Reimand *et al.*, "Pathway enrichment analysis and visualization of omics data using g: Profiler, GSEA, Cytoscape and EnrichmentMap," *Nature Protocols*, vol. 14, no. 2, pp. 482–517, 2019.

[44] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *CoRR*, vol. abs/1412.6980, 2014.

[45] C.-W. Hsu, C.-C. Chang, and C.-J. Lin. (2016) A Practical Guide to Support Vector Classification. [Online]. Available: https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

[46] A. J. Harmar, "Family-B G-protein-coupled receptors," *Genome biology*, vol. 2, no. 12, p. reviews3013.10, Nov. 2001.

[47] A. E. Cherry and N. Stella, "G protein-coupled receptors as oncogenic signals in glioma: Emerging therapeutic avenues," *Neuroscience*, vol. 278, no. 1, pp. 222–236, 2014.

[48] J. Zhang, H. Feng, S. Xu, and P. Feng, "Hijacking gpcrs by viral pathogens and tumor," *Biochemical Pharmacology*, vol. 114, pp. 69–81, 2016.

[49] K. Turkowski *et al.*, "VEGF as a modulator of the innate immune response in glioblastoma," *GLIA*, vol. 66, no. 1, pp. 161–174, 2018.

[50] L. Feng *et al.*, "Heterogeneity of tumor-infiltrating lymphocytes ascribed to local immune status rather than neoantigens by multi-omics analysis of glioblastoma multiforme," *Scientific Reports*, vol. 7, no. 1, p. 6968, 2017.

[51] C. Zhou *et al.*, "Analysis of the gene-protein interaction network in glioma," *Genetics and Molecular Research*, vol. 14, no. 4, pp. 14 196–14 206, 2015.

[52] H. Y. Choi *et al.*, "G protein-coupled receptors in stem cell maintenance and somatic reprogramming to pluripotent or cancer stem cells," *BMB Reports*, vol. 48, no. 2, pp. 68–80, Feb. 2015.

[53] A. Chédotal, G. Kerjan, and C. Moreau-Fauvarque, "The brain within the tumor: New roles for axon guidance molecules in cancers," *Cell Death And Differentiation*, vol. 12, pp. 1044–1056, 2005.

[54] A. Joy *et al.*, "The role of AKT isoforms in glioblastoma: AKT3 delays tumor progression," *Journal of Neuro-Oncology*, vol. 130, no. 1, pp. 43–52, Oct. 2016.

[55] B. Hu *et al.*, "Astrocyte elevated gene-1 (AEG-1) interacts with Akt isoform 2 to control glioma growth, survival and pathogenesis," *Cancer Research*, vol. 74, no. 24, pp. 7321–7332, 2014.

[56] L. C. Hinske *et al.*, "Intronic miRNA-641 controls its host Gene's pathway PI3K/AKT and this relationship is dysfunctional in glioblastoma multiforme," *Biochemical and Biophysical Research Communications*, vol. 489, no. 4, pp. 477–483, 2017.

[57] M. Lim, Y. Xia, C. Bettegowda, and M. Weller, "Current state of immunotherapy for glioblastoma," *Nature Reviews Clinical Oncology*, vol. 15, no. 7, pp. 422–442, 2018.

[58] H. B. Burke, "Predicting Clinical Outcomes Using Molecular Biomarkers," *Biomarkers in Cancer*, vol. 8, no. 8, pp. 89–99, June 2016.

[59] G. Lightbody *et al.*, "Review of applications of high-throughput sequencing in personalized medicine: barriers and facilitators of future progress in research and clinical application," *Briefings in Bioinformatics*, p. bby051, 2018.

[60] F. E. Ahmed, P. W. Vos, and D. Holbert, "Modeling survival in colon cancer: A methodological review," *Molecular Cancer*, vol. 6, no. 1, p. 15, Feb. 2007.

[61] H.-C. Chen, R. L. Kodell, K. F. Cheng, and J. J. Chen, "Assessment of performance of survival prediction models for cancer prognosis," *BMC Medical Research Methodology*, vol. 12, no. 1, p. 102, July 2012.

[62] D. M. Witten and R. Tibshirani, "Survival analysis with high-dimensional covariates," *Statistical Methods in Medical Research*, vol. 19, no. 1, pp. 29–51, Feb. 2010.

[63] H. H. Zhang and W. Lu, "Adaptive Lasso for Cox's proportional hazards model," *Biometrika*, vol. 94, no. 3, pp. 691–703, Aug. 2007.

[64] R. J. Tibshirani, "Univariate Shrinkage in the Cox Model for High Dimensional Data," *Statistical Applications in Genetics and Molecular Biology*, vol. 8, no. 1, pp. 1–18, 2009.

[65] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent," *Journal of Statistical Software*, vol. 39, no. 5, pp. 1–13, 2011.

[66] J. Xu, "High-dimensional cox regression analysis in genetic studies with censored survival outcomes," *Journal of Probability and Statistics*, vol. 2012, 2012.

[67] J. Fan, Y. Feng, and Y. Wu, *High-dimensional variable selection for Cox's proportional hazards model*, ser. Collections. Beachwood, Ohio, USA: Institute of Mathematical Statistics, 2010, vol. 6, pp. 70–86.

[68] H. Li and Y. Luan, "Kernel Cox regression models for linking gene expression profiles to censored survival data," in *Pacific Symposium on Biocomputing 8*, 2003, pp. 65–76.

[69] L. Evers and C.-M. Messow, "Sparse kernel methods for high-dimensional survival data," *Bioinformatics*, vol. 24, no. 14, pp. 1632–1638, 2008.

[70] J. L. Katzman *et al.*, "DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network," *BMC Medical Research Methodology*, vol. 18, no. 1, p. 24, Feb. 2018.

[71] T. Ching, X. Zhu, and L. X. Garmire, "Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data," *PLOS Computational Biology*, vol. 14, no. 4, pp. 1–18, Apr. 2018.

[72] S. Yousefi *et al.*, "Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models," *Scientific Reports*, vol. 7, no. 1, p. 11707, 2017.

[73] P. Masson *et al.*, "An Integrated Ontology Resource to Explore and Study Host-Virus Relationships," *PLOS ONE*, vol. 9, no. 9, pp. 1–10, Sept. 2014.

[74] B. Zhu *et al.*, "Integrating Clinical and Multiple Omics Data for Prognostic Assessment across Human Cancers," *Scientific Reports*, vol. 7, no. 1, p. 16954, 2017.

[75] W. Zhang *et al.*, "Integrating Genomic, Epigenomic, and Transcriptomic Features Reveals Modular Signatures Underlying Poor Prognosis in Ovarian Cancer," *Cell Reports*, vol. 4, no. 3, pp. 542–553, 2013.

[76] B. M. Reid, J. B. Permuth, and T. A. Sellers, "Epidemiology of ovarian cancer: a review," *Cancer Biology & Medicine*, vol. 14, no. 1, pp. 9–32, 2017.

[77] M.-C. Ruben, "BayesOpt: A Bayesian Optimization Library for Nonlinear Optimization, Experimental Design and Bandits," *Journal of Machine Learning Research*, vol. 15, pp. 3915–3919, 2014. [Online]. Available: http://jmlr.org/papers/v15/martinezcantin14a.html

[78] L. J. P. van der Maaten and H. G. E., "Visualizing High-Dimensional Data Using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. Nov,

pp. 2579–2605, 2008. [Online]. Available: http://www.jmlr.org/papers/v9/vandermaaten08a.html

[79] G. P. Atkinson, S. E. Nozell, and E. T. N. Benveniste, "NF-κB and STAT3 signaling in glioma: targets for future therapies," *Expert review of neurotherapeutics*, vol. 10, no. 4, pp. 575–586, 2014.

[80] J. Pal *et al.*, "Abstract 2454: Genetic landscape of glioma reveals defective neuroactive ligand receptor interaction pathway as a poor prognosticator in glioblastoma patients," in *Proceedings of the American Association for Cancer Research Annual Meeting 2017*, vol. 77, no. 13 Supplement, Apr. 2017, pp. 2454–2454.

[81] G. L. Weber, M.-O. Parat, Z. A. Binder, G. L. Gallia, and G. J. Riggins, "Abrogation of PIK3CA or PIK3R1 reduces proliferation, migration, and invasion in glioblastoma multiforme cells," *Oncotarget*, vol. 2, no. 11, pp. 833–849, 2011.

[82] C. Senft *et al.*, "Inhibition of the JAK-2/STAT3 signaling pathway impedes the migratory and invasive potential of human glioblastoma cells," *Expert review of neurotherapeutics*, vol. 101, no. 3, pp. 393–403, 2011.

[83] M. Xiong *et al.*, "Genome-Wide Association Studies of Copy Number Variation in Glioblastoma," in *2010 4th International Conference on Bioinformatics and Biomedical Engineering*, June 2010, pp. 1–4.

[84] C. B. Chan and K. Ye, "Phosphoinositide 3-kinase enhancer (PIKE) in the brain: is it simply a phosphoinositide 3-kinase/Akt enhancer?" *Reviews in the neurosciences*, vol. 23, no. 2, pp. 153–161, 2013.

[85] D. K. Tanwar *et al.*, "Crosstalk between the mitochondrial fission protein, Drp1, and the cell cycle is identified across various cancer types and can impact survival of epithelial ovarian cancer patientss," *Oncotarget*, vol. 7, no. 37, pp. 60 021–60 037, 2016.

[86] G. A. Mendes *et al.*, "Prolactin gene expression in primary central nervous system tumors," *Journal of Negative Results in BioMedicine*, vol. 12, no. 1, p. 4, Jan. 2013.

[87] C. G. Brahm *et al.*, "Identification of novel therapeutic targets in glioblastoma with functional genomic mRNA profiling," *Journal of Clinical Oncology*, vol. 35, no. 15_suppl, pp. 2018–2018, 2017.

[88] X. Cui *et al.*, "IL22 furthers malignant transformation of rat mesenchymal stem cells, possibly in association with IL22RA1/STAT3 signaling," *Oncology reports*, vol. 41, no. 4, pp. 2148–2158, 2019.

[89] S. Allerstorfer *et al.*, "FGF5 as an oncogenic factor in human glioblastoma multiforme: autocrine and paracrine activities," *Oncogene*, vol. 27, no. 30, pp. 4180–4190, 2008.

[90] Y. Gao *et al.*, "Targeting JUN, CEBPB, and HDAC3: A Novel Strategy to Overcome Drug Resistance in Hypoxic Glioblastoma," *Frontiers in oncology*, vol. 9, p. 33, 2019.

[91] V. V. Prabhu *et al.*, "Dopamine Receptor D5 is a Modulator of Tumor Response to Dopamine Receptor D2 Antagonism," *Clinical Cancer Research*, vol. 25, no. 7, pp. 2305–2313, 2019.

[92] L. Deng *et al.*, "Bioinformatics analysis of the molecular mechanism of diffuse intrinsic pontine glioma," *Oncology letters*, vol. 12, no. 4, pp. 2524–2530, 2016.

[93] S. Huang, K. Chaudhary, and L. X. Garmire, "More Is Better: Recent Progress in Multi-Omics Data Integration Methods," *Frontiers in Genetics*, vol. 8, p. 84, 2017.

[94] R. Higdon *et al.*, "The Promise of Multi-Omics and Clinical Data Integration to Identify and Target Personalized Healthcare Approaches in Autism Spectrum Disorders," *OMICS: A Journal of Integrative Biology*, vol. 19, no. 4, pp. 197–208, 2015.

[95] V. N. Kristensen *et al.*, "Principles and methods of integrative genomic analyses in cancer," *Nature Reviews Cancer*, vol. 14, pp. 299–313, 2014.

[96] M. R. Aure *et al.*, "Individual and combined effects of DNA methylation and copy number alterations on miRNA expression in breast tumors," *Genome Biology*, vol. 14, no. 11, p. R126, 2013.

[97] J. R. Wagner *et al.*, "The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts," *Genome Biology*, vol. 15, no. 2, p. R37, 2014.

[98] G. Lyu *et al.*, "Genome and epigenome analysis of monozygotic twins discordant for congenital heart disease," *BMC Genomics*, vol. 19, no. 1, p. 428, 2018.

[99] C. E. Bruder *et al.*, "Phenotypically Concordant and Discordant Monozygotic Twins Display Different DNA Copy-Number-Variation Profiles," *The American Journal of Human Genetics*, vol. 82, no. 3, pp. 763–771, 2008.

[100] D. Kim *et al.*, "Using knowledge-driven genomic interactions for multi-omics data analysis: metadimensional models for predicting clinical outcomes in ovarian carcinoma," *Journal of the American Medical Informatics Association*, vol. 24, no. 3, pp. 577–587, 2017.

[101] M. Kang *et al.*, "Multi-Block Bipartite Graph for Integrative Genomic Analysis," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 14, no. 6, pp. 1350–1358, 2017.

[102] K. Chaudhary, O. B. Poirion, L. Lu, and L. X. Garmire, "Deep Learning-Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer," *Clinical Cancer Research*, vol. 24, no. 6, pp. 1248–1259, 2018.

[103] J. Hao, Y. Kim, T. Mallavarapu, J. H. Oh, and M. Kang, "Cox-PASNet: Pathway-based Sparse Deep Neural Network for Survival Analysis," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Dec 2018, pp. 381–386.

[104] S. Girirajan, C. D. Campbell, and E. E. Eichler, "Human Copy Number Variation and Complex Genetic Disease," *Annual Review of Genetics*, vol. 45, no. 1, pp. 203–226, 2011.

[105] L. D. Moore, T. Le, and G. Fan, "DNA Methylation and Its Basic Function," *Neuropsychopharmacology*, vol. 38, pp. 23–38, 2013.

[106] K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1026–1034.

[107] C. Gründker and G. Emons, "The Role of Gonadotropin-Releasing Hormone in Cancer Cell Proliferation and Metastasis," *Frontiers in Endocrinology*, vol. 8, p. 187, 2017.

[108] S. Jayaram, M. K. Gupta, R. Raju, P. Gautam, and R. Sirdeshmukh, "Multi-Omics Data Integration and Mapping of Altered Kinases to Pathways Reveal Gonadotropin Hormone Signaling in Glioblastoma," *OMICS: A Journal of Integrative Biology*, vol. 20, no. 12, pp. 736–746, 2016.

[109] L. Catacuzzeno and F. Franciolini, "Role of KCa3.1 Channels in Modulating Ca^{2+} Oscillations during Glioblastoma Cell Migration and Invasion," *International Journal of Molecular Sciences*, vol. 19, no. 10, p. 2970, 2018.

[110] D. D. J. Chi *et al.*, "Molecular Detection of Tumor-Associated Antigens Shared by Human Cutaneous Melanomas and Gliomas," *American Journal of Pathology*, vol. 150, no. 6, pp. 2143–2152, 1997. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/9176405

[111] T. A. Ulrich, E. M. de Juan Pardo, and S. Kumar, "The Mechanical Rigidity of the Extracellular Matrix Regulates the Structure, Motility, and Proliferation of Glioma Cells," *Cancer Research*, vol. 69, no. 10, pp. 4167–4174, 2009.

[112] F. E. Bleeker *et al.*, "Mutational profiling of kinases in glioblastoma," *BMC Cancer*, vol. 14, no. 1, p. 718, Sept. 2014.

[113] J. Cheng *et al.*, "Integrative Analysis of Histopathological Images and Genomic Data Predicts Clear Cell Renal Cell Carcinoma Prognosis," *Cancer Research*, vol. 77, no. 21, pp. e91–e100, 2017.

[114] X. Zhu *et al.*, "Lung cancer survival prediction from pathological images and genetic data - An integration study," in *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, 2016, pp. 1173–1176.

[115] V. Popovici *et al.*, "Joint analysis of histopathology image features and gene expression in breast cancer," *BMC Bioinformatics*, vol. 17, no. 1, p. 209, 2016.

[116] N. P. Group, "Histopathology is ripe for automation," *Nature Biomedical Engineering*, vol. 1, no. 12, p. 925, 2017.

[117] D. Komura and S. Ishikawa, "Machine Learning Methods for Histopathological Image Analysis," *Computational and Structural Biotechnology Journal*, vol. 16, pp. 34–42, 2018.

[118] D. C. Cireşan *et al.*, "Mitosis Detection in Breast Cancer Histology Images with Deep Neural Networks," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*, 2013, pp. 411–418.

[119] Y. Xu *et al.*, "DEEP CONVOLUTIONAL ACTIVATION FEATURES FOR LARGE SCALE BRAIN TUMOR HISTOPATHOLOGY IMAGE CLASSIFICATION AND SEGMENTATION," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 947–951.

[120] M. G. Ertosun and D. L. Rubin, "Automated Grading of Gliomas using Deep Learning in Digital Pathology Images: A modular approach with ensemble of convolutional neural networks," in *AMIA Annual Symposium Proceedings*, vol. 2015, 2015, pp. 1899–1908. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/26958289

[121] L. Hou *et al.*, "Automatic histopathology image analysis with CNNs," in *2016 New York Scientific Data Summit (NYSDS)*, 2016, pp. 1–6.

[122] P. Mobadersany *et al.*, "Predicting cancer outcomes from histology and genomics using convolutional networks," *Proceedings of the National Academy of Sciences*, vol. 115, no. 13, pp. E2970–E2979, 2018.

[123] D. Sun, L. Ao, T. bo, and W. Minghui, "Integrating genomic data and pathological images to effectively predict breast cancer clinical outcome," *Computer Methods and Programs in Biomedicine*, vol. 161, pp. 45–53, 2018.

[124] T. Hu, H. Turki, Phan, and Wang, "A 3D Atrous Convolutional Long Short-Term Memory Network for Background Subtraction," *IEEE Access*, vol. 6, pp. 43 450–43 459, 2018.

[125] T. Guan and H. Zhu, "Atrous Faster R-CNN for Small Scale Object Detection," in *2017 2nd International Conference on Multimedia and Image Processing (ICMIP)*, 2017, pp. 16–21.

[126] T. Zhi, L.-Y. Duan, Y. Wang, and T. Huang, "Two-stage pooling of deep convolutional features for image retrieval," in *2016 IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 2465–2469.

[127] Y.-L. Boureau, J. Ponce, and Y. Lecun, "A Theoretical Analysis of Feature Pooling in Visual Recognition," in *27th International Conference on Machine Learning (ICML 2010)*, 2010, pp. 111–118.

[128] B. Lei and ohters, "Prospective Series of Nine Long Noncoding RNAs Associated with Survival of Patients with Glioblastoma," *J Neurol Surg A Cent Eur Neurosurg*, vol. 79, no. 6, pp. 471–478, 2018.

[129] D. P. Buser *et al.*, "Quantitative proteomics reveals reduction of endocytic machinery components in gliomas," *EBioMedicine*, 2019.

[130] M. Liu *et al.*, "The Identification of Key Genes and Pathways in Glioma by Bioinformatics Analysis," *J Immunol Res*, vol. 2017, p. 1278081, 2017.

[131] ——, "Identification of survivalassociated key genes and long noncoding RNAs in glioblastoma multiforme by weighted gene coexpression network analysis,"

International Journal of Molecular Medicine, vol. 43, no. 4, pp. 1709–1722, 2019.

[132] M. B. Nijaguna *et al.*, "An Eighteen Serum Cytokine Signature for Discriminating Glioma from Normal Healthy Individuals," *PLoS One*, vol. 10, no. 9, p. e0137524, 2015.

[133] M.-E. Halatsch *et al.*, "Epidermal Growth Factor Receptor Pathway Gene Expressions and Biological Response of Glioblastoma Multiforme Cell Lines to Erlotinib," *Anticancer Res*, vol. 28, no. 6A, pp. 3725–3728, 2008.

[134] J. Guo, A. V. Schally, M. Zarandi, J. Varga, and P. C. Leung, "Antiproliferative effect of growth hormone-releasing hormone (GHRH) antagonist on ovarian cancer cells through the EGFR-Akt pathway," *Reproductive Biology and Endocrinology*, vol. 8, no. 1, p. 54, 2010.

[135] J. Huang *et al.*, "Differential Expression of Adenosine P1 Receptor ADORA1 and ADORA2A Associated with Glioma Development and Tumor-Associated Epilepsy," *Neurochem Res*, vol. 41, no. 7, pp. 1774–1783, 2016.

[136] C. Boccaccio and P. M. Comoglio, "The MET Oncogene in Glioblastoma Stem Cells: Implications as a Diagnostic Marker and a Therapeutic Target," *Cancer Research*, vol. 73, no. 11, pp. 3193–3199, 2013.

[137] D. Jantas *et al.*, "An endogenous and ectopic expression of metabotropic glutamate receptor 8 (mGluR8) inhibits proliferation and increases chemosensitivity of human neuroblastoma and glioma cells," *Cancer Letter*, vol. 432, pp. 1–16, 2018.