TLSurv: Integrating Multi-Omics Data by Multi-Stage Transfer Learning for Cancer Survival Prediction

Yixing Jiang
National University of Singapore
Singapore
jiang@u.nus.edu

Kristen Alford
Georgia Institute of Technology
Atlanta, Georgia
kalford7@gatech.edu

Frank Ketchum
Georgia Institute of Technology
Atlanta, Georgia
fketchum3@gatech.edu

Li Tong
Georgia Institute of Technology and
Emory University
Atlanta, Georgia
ltong@gatech.edu

May D. Wang
Georgia Institute of Technology and
Emory University
Atlanta, Georgia
maywang@bme.gatech.edu

ABSTRACT

Lung cancer is one of the leading cancers, but survival models have not been explored to the extent of other cancers like breast cancer. In this study, we develop a super-hybrid network called TLSurv to integrate Copy Number Variation, DNA methylation, mRNA expression, and miRNA expression data for TCGA-LUAD datasets. The modularity of this super-hybrid network allows the integration of multiple -omics modalities with tremendous dimensional differences. Additionally, a novel training scheme called multi-stage transfer learning is used to train this super-hybrid network incrementally. This allows for training of a large network with many subnetworks using a relatively small data sets. At each stage, a shallow subnetwork is trained and these networks are combined to form a powerful prediction network. The results show the combination of DNA methylation data with either mRNA or miRNA expression data has produced promising performances with C-indexes of around 0.7. This performance is better than previous studies. Interpretability analysis confirms the clinical significance of some biomarkers identified. In addition, some novel biomarkers are suggested for future medical research. These findings reveal the potential of super-hybrid network for integrating multiple data modalities and the potential of multi-stage transfer learning for addressing the "curse of dimensionality."

KEYWORDS

multi-omics data integration, cancer survival analysis, multi-stage transfer learning, super-hybrid network

BCB '20, September 21–24, 2020, Virtual Event, USA

ACM ISBN 978-1-4503-7964-9/20/09…$15.00
https://doi.org/10.1145/3388440.3412422

ACM Reference Format:
Yixing Jiang, Kristen Alford, Frank Ketchum, Li Tong, and May D. Wang. 2020. TLSurv: Integrating Multi-Omics Data by Multi-Stage Transfer Learning for Cancer Survival Prediction. In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '20), September 21–24, 2020, Virtual Event, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3388440.3412422

1 BACKGROUND

1.1 Lung Cancer

Lung cancer is one of the most common cancers. In 2017, it caused more deaths than breast, prostate, colorectal, and brain cancers combined[37]. Even though there has been a long-term decline in the death rate associated with lung cancer, it is still projected to lead to 135,720 deaths in the United States in 2020[37]. As such, lung cancer is a commonly studied disease in oncology. However, in terms of deep-learning based prognosis predictions, lung cancer has not been explored to the extent of other cancers like breast cancer.

1.2 Survival Analysis and Related Works

Time-to-event prediction is a statistical term that is used to measure the duration or time it takes for an event to happen. Survival analysis is one type of time-to-event prediction with patients' death as the event, and it is widely used in medicine to predict patients' prognosis[19]. Prognosis results provide clinicians with guidance for treatment decisions, and these results are important indicators for risk stratification[15]. Furthermore, significant biomarkers may be identified during survival analysis, which assists in further understanding of the molecular mechanisms of cancer development and treatment resistance[9][29].

There are two main functions of interest in survival analysis: the survival function $S(t)$ and the hazard function $h(t)$. The survival function $S(t)$ defines the probability of an individual surviving past time t, while the hazard function $h(t)$ defines the probability of death occurring at time t, provided the patient has not died as of time t. Mathematically, given T as the random variable denoting survival duration,

$$S(t) = Pr(T > t) \tag{1}$$

$$h(t) = \lim_{\Delta t \to 0} Pr(t \leq T \leq t + \Delta t | T > t) \tag{2}$$

As a result, $h(t) = \frac{-S'(t)}{S(t)}$, so a hazard function can be derived from a survival function and vice versa. Each function characterizes a patient's survival. The hazard function is estimated most often in practice.

Machine learning methods have been employed in predicting survival outcomes in many types of cancers. One standard survival analysis model is the Cox proportional hazards (CoxPH) model, which assumes that a log-partial hazard is a linear combination of k covariates[5]. Mathematically, the hazard function at time t given covariate vector $\vec{x}$ and baseline hazard $b_0(t)$ is:

$$h(t|\vec{x}) = b_0(t)\exp(\sum_{i=1}^{k} \beta_i x_i) \tag{3}$$

In this equation, $exp(\sum_{i=1}^{k} \beta_i x_i)$ is the partial hazard and $\sum_{i=1}^{k} \beta_i x_i$ is the log-partial hazard. The model uses the training data to estimate those k parameters ($\beta_1, ..., \beta_k$). This model allows prediction of survival time based on covariates such as treatments and cancer stage and has been widely used in tools to aid physicians in proposing treatments for cancer patients[12].

The learning capabilities of proportional hazard models are limited by their dependence on linear relationships between inputs and the log-partial hazard. With the rise of -omics data collected from human cancer cells, other machine learning techniques have been employed to capture nonlinearities in these -omics data and give more accurate predictions of patient survival. Among these methods, one commonly used shallow learner for survival analysis is the random survival forest (RSF) model[13]. As an ensemble method, RSF grows many individual decision trees to enhance performance. It is essential to keep individual tress highly uncorrelated, and therefore, RSF utilizes randomization to build trees with low variance for ensemble. The randomization includes using bootstrapped data and random feature selection when splitting tree nodes. The final decision is made by combining decisions by individual decision trees.

Other than RSF, traditional classification models such as support vector machines (SVM)[3] have been used to learn and categorize -omics data in attempts to distinguish between "high-risk" and "low-risk" populations.

Since -omics data usually have large numbers of dimensions, unsupervised dimensionality reduction methods such as principal component analysis (PCA)[1], t-distributed stochastic neighbor embedding (t-SNE)[25], or autoencoders are utilized to decrease the dimensionality of these data. This serves to limit the number of learnable parameters in the subsequent survival models.

1.3 Deep Learning

The promise of deep neural networks has been demonstrated by recent studies, especially in the fields of computer vision and natural language processing[20]. Unlike traditional shallow learners, deep neural networks are able to capture sophisticated, non-linear data via activation functions and multiple levels of abstraction[20][36]. Previous studies have shown that deep learning-based survival models have superior performance compared to shallow learners such as Cox-PH and RSF[4][11][15].

Since 2018, many deep learning methods; Cox-nnet[4], DeepSurv[15], AECOX[11] and Nnet-survival[7]; have been proposed. The first three adopt the CoxPH assumption by using a Cox regression layer as its output layer, while the last one supports non-proportional hazards. Based on previous studies[7][11], Cox-nnet has achieved the best performance among the four and it is suitable for small datasets. The Cox-nnet architecture is used in the embedding and survival section in this study.

1.4 Data Integration

The popularity of multi-omics data for cancer survival prediction is growing[26][32][33][41][42][45]. Multi-view representation learning, the use of multiple parallel input data modalities, is an attractive feature of these multi-omics data sets. It is primarily based on two basic principles: uniqueness and consensus. On one hand, each data modality is assumed to have some unique information which is lacking in other modalities. On the other hand, it is assumed that the integrated representation is consistent with each modality. As a result, information from different data modalities can complement each other. The resulting model is able to learn more comprehensive representations by learning from various views[21].

Variational Autoencoders (VAE)[27] and Multi-view Factorization AutoEncoders (MAE)[24] have recently been used to integrate -omics data and have proven good results. Each was implemented as a fusion section in this study.

Variational autoencoders (VAE) are generative adversarial neural networks composed of an encoder and a decoder. The encoder attempts to map input data in $\mathbb{R}^p$ to a latent representation of the data in $\mathbb{R}^q$, where $q << p$. The decoder attempts to recreate the original data from the latent representation produced by the encoder, which validates that the latent representation is representative of the original data set. Simidjevski et. al[27] propose many VAE architectures for integration of two data modalities. The two best performing architectures, the hierarchical and X-shaped VAEs, produced similar results in integration of breast cancer -omics data. The hierarchical VAE was much more complex than the X-shaped VAE, and in the context of this complicated model which we propose, a higher computational cost is not desirable. As such, the X-shaped VAE architecture was incorporated into the proposed model as the fusion network.

The Multi-view Factorization AutoEncoder (MAE)[24] uses an autoencoder to allow complex non-linear factorization. Unlike VAE learning a joint representation, MAE imposes an autoencoder for each data modality and fuses factor matrices into one fused view. In order to ensure the consistency of different views, a regularization term was added based on fused patient similarity network using affinity network fusion[23].

Machine learning models which implement two standalone models and combine them are often referred to as "hybrid models." In order to use the fusion network for survival analysis, a fusion network and survival network should be incorporated. Traditional unsupervised dimensionality reduction techniques do not capture information brought by survival outcomes, so supervised embedding can be used instead. The combination of many individual networks finally leads to a "super-hybrid" network.

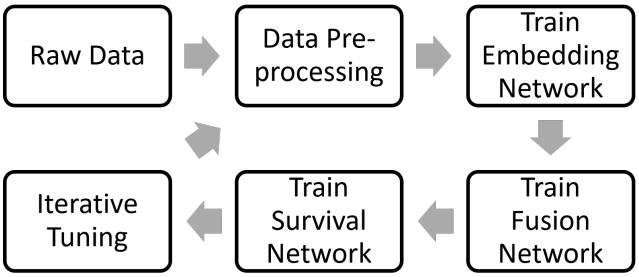

Figure 1: Overall Pipeline for Development of TLSurv. Raw data were first preprocessed for quality control. The cleaned data were fed into training the embedding networks of each modality for dimensionality reduction. The fusion network was then trained, followed by the survival network for predictions. The entire training pipeline was iterated to optimize the performance.

1.5 Transfer Learning

Transfer learning is traditionally used for knowledge transfer when dealing with insufficient training data, and the underlying assumption is that training data for the new task may have different distributions from ones for the original task[30]. However, the learned representation by an original task can be passed to the next stage of training for different purposes with the same set of training data. As a result, multi-stage transfer learning allows knowledge learned during each training stage to be accumulated so that a powerful model can be obtained at the end.

In the super-hybrid network mentioned in Section 1.4, the large number of trainable parameters poses a problem for convergence and overfitting of the model. To solve this issue, multi-stage transfer learning is used for training the super-hybrid network so that each subnetwork is trained individually. This way, the number of trainable parameters is reduced for each training stage.

2 MATERIALS AND METHODS

2.1 Overall Pipeline

As sample sizes of biomedical data are usually limited, training a large-scale deep neural network is very challenging due to the "curse of dimensionality" [24] and vanishing gradients [18]. In order to solve these problems, TLSurv is a super-hybrid network comprised of four individual shallow neural networks. There are in total three sections: embedding, fusion, and survival analysis, as shown in **Figure 3B and 3C**. The individual sections are explained in detail in Section 2.4. In order to train this network, multi-stage transfer learning was used. At each training stage, only the parameters of a small neural network were learned, which effectively addressed the "curse of dimensionality" by limiting the number of learnable parameters respective to the number of data points.

Figure 1 shows the overview of pipeline for development of TLSurv. Raw data were preprocessed to remove noisy features and normalize the ranges. Embedding networks were trained on the cleaned data, followed by a fusion network trained on the embedded data. Finally, a survival network was trained and the entire network was evaluated. The hyper-parameters were then tuned repeatedly based on evaluations. At this point, the whole network can be used for prediction of survival time.

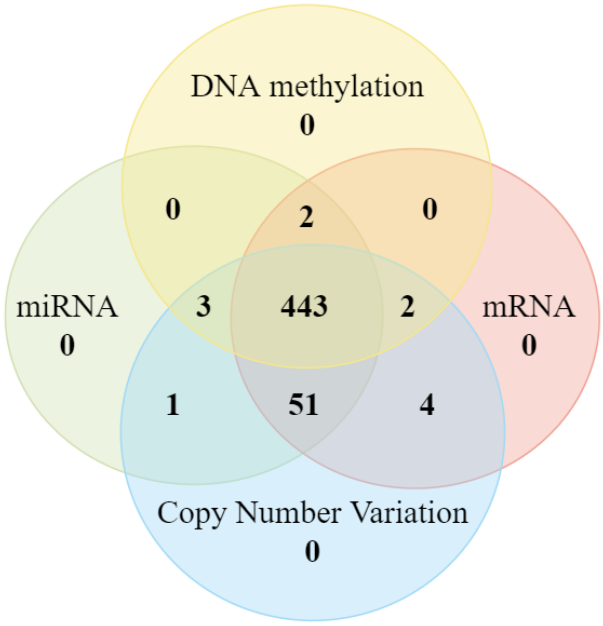

Figure 2: Venn Diagram of Numbers of Patients With Different Modalities Available. Only patients with survival outcomes are presented.

This project was built with Python 3.8 under Anaconda 3. PyTorch 1.4 package with cuda 10.1.243 and cudnn 7.6.3 was used to implement neural networks, and PyCox package[19] was used for survival analysis training and evaluation. The server used for training was equipped with Intel Xeon E5-2690 v3 CPUs and Tesla K80 GPUs. The source codes and the conda environment configuration are available at https://github.com/kylejyx/TLSurv.

2.2 Datasets

The Cancer Genome Atlas (TCGA) is commonly used as it contains comprehensive multi-omics data from a moderate cohort. The GDC version of TCGA data was uniformly re-analyzed using the latest Human Genome Assembly hg38. Therefore, GDC TCGA Lung Adenocarcinoma (LUAD) dataset is used in this project. The data was downloaded from UCSC Xena platform[8].

After taking the intersection of samples across all five data modalities (copy number variation, DNA methylation, miRNA expression, mRNA expression and survival outcomes), there were in total 443 patients with all data modalities available, as shown in **Figure 2**. 35 percent of these patients had uncensored observations, and the median survival time was 651 days. Those patients were split into training and test sets with a ratio of eight to two, respectively. Random shuffling was conducted before splitting. The proportion of uncensored observations and the median survival time were measured post-split to ensure similar distributions between training set and test set.

Table 1: Summary of Processed Data

Modality	Dimension	Detailed Type
CNV	12476	GISTIC-focal score by gene
DNA methylation	38165	Illumina Human Methylation 450
miRNA	1417	miRNA Expression Quantification
mRNA	38708	HTSeq - FPKM-UQ
Survival	2	Time and event

2.3 Data Preprocessing

The pandas[31] package was utilized to process the downloaded tab-separated values(tsv) files. Firstly, duplicated samples were removed, and only one sample was kept for each patient. In order to improve model robustness and reduce irrelevant noise in the mRNA and copy number variation (CNV) data, genes with lowest 20% expression or aberration values and those within the lowest 20% variances across samples were removed. For miRNA data, features within the lowest 10% expression values and lowest 10% variances across samples were removed. For methylation data, features with a standard deviation smaller than 0.15 were removed. As a high occurrence of missing values negatively affects the data quality, features with many missing values were removed. The threshold used is 0.1 as suggested by literature; if at least 10% of the samples for any feature were not available, that feature would be removed[6]. For the remaining features, those missing values were filled with median values.

Empirical experiences have shown that neural networks work well with inputs with small ranges. In order to mitigate the significant differences in dynamic ranges across different modalities, for each feature in mRNA, CNV and miRNA data, MinMax scaling was used to normalize the data into ranges of [0, 1]. The transformation is given by:

$$X = \frac{X - min(X)}{max(X) - min(X)} \tag{4}$$

Table 1 summaries the processed data.

2.4 Neural Network Architecture

Figure 3 shows the architectures of the three neural networks: Cox-nnet, TLSurv(MAE) and TLSurv(VAE). Leftmost rectangular layers correspond to inputs from different data sources, and rightmost circular nodes indicate final prognostic indexes (log-partial hazards). Cox-nnet (**Figure 3A**) is the state-of-art deep-learning based survival network which serves as the baseline model.

Figure 3B and **Figure 3C** show examples of the TLSurv architecture for two modalities. It is an incrementally-trained super-hybrid network with four individual networks: two embedding networks, one fusion network and one survival network.

As -omics data usually have very high dimensionality, TLSurv begins with an embedding section for dimensionality reduction. Two networks with one hidden layer and one Cox regression layer are used. The architecture follows the Cox-nnet[4]. The hidden layer is densely connected with the input layer using a tanh activation. The Cox regression layer is densely connected with the hidden layer with no bias and linear activation.

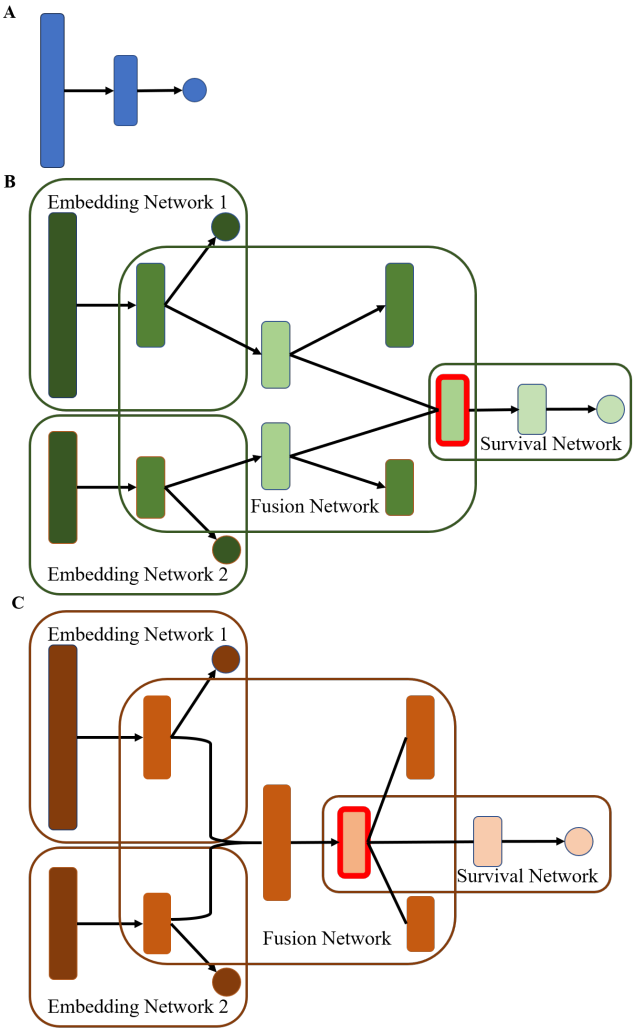

Figure 3: Architectures of Cox-nnet, TLSurv(MAE) and TLSurv(VAE). A. Cox-nnet with one hidden layer; B. TLSurv with MAE implementation for two modalities; C. TLSurv with VAE implementation for two modalities; For B and C, the four bounding boxes represent the four individual networks within the super-hybrid network. Leftmost rectangular layers correspond to inputs from different data sources. Rectangles with thick red outlines represent fused views. All seven circular output nodes represent log-partial hazard, while the four rectangular output nodes represent reconstructed embedding representations. Arrows represent dense connections between adjacent layers, and the right brace indicates concatenation followed by dense connections. The two lines in B refers to affinity network fusion, while the three lines in C refers to sampling from the learned distribution.

For the fusion section, two implementations (MAE and VAE) have been developed. They are provided in **Figure 3B** and **Figure 3C** respectively. For TLSurv(MAE), the output of the hidden layer of the embedding network is encoded into the latent embedding. A decoder follows for reconstruction of the input data set. The two

latent embeddings are then fused into one view, the layer with the red boundary outline, by taking the mean of the two embedded representations. A rectified linear unit(ReLU) activation is used. For the TLSurv with VAE implementation, outputs of the hidden layer of the two embedding networks are concatenated, and encoded into the Gaussian distributional representation with two nodes (mean and standard deviation). The fused view is then sampled from the distributional representation, and passed to two decoders for reconstruction. Exponential linear unit(ELU) activation was used. The fusion module serves as a denoise scheme in addition to feature reduction as -omics data are prone to both biological and technical noise.

The survival section of the complete network utilizes the Cox-nnet architecture[4] with the exception that the fused view is used as the input layer. The hidden layer is densely connected with the input layer with a tanh activation. The Cox regression layer is densely connected with the hidden layer with no bias or activation. The final output is the prognostic index which is also called log-partial hazard.

2.5 Training Scheme

Since TLSurv is a super-hybrid network, it is trained using multi-stage transfer learning. There are three stages of learning in total, and each training stage has its own purpose. The neural network parameters are initialized by the default method provided by PyTorch package.

At the first stage, two individual embedding networks are trained to embed the predictive features and reduce the dimensional mismatch between different modalities. The Cox regression layer is used as the output node. Negative partial log-likelihood is used as the loss function[19]. Mathematically,

$$l(\boldsymbol{\beta}) = \sum_{i:D_i=1} (\log(\sum_{j:T_j \geq T_i} \exp(\boldsymbol{\beta}\mathbf{x_j} - \boldsymbol{\beta}\mathbf{x_i}))) \tag{5}$$

$\boldsymbol{\beta}$ is the previously discussed parameter of interest, $\mathbf{x_i}$ is the covariate vector of patient i, $D_i = 1$ denotes the occurrence of a death event for patient i, and T_i denotes the death time of patient i. The weights of embedding networks are saved as pt files after training for reuse later.

The second stage adopts multi-view representation learning for the information fusion network. It naturally incorporates complementary information from various data modalities, and renders those information consistent as a whole joint presentation. Before training, parameters for the embedding section are loaded from the two pt files obtained before, and are frozen during the second stage of training, as these parameters have already been optimized.

The loss function for the MAE is a combination of mean squared error (MSE) and regularization of patient view similarity. MSE is employed to verify the reconstructed data sets' similarity to the input data before encoding, and the regularization term is based on the fused patient cosine similarity network using affinity network fusion[23]. The model is trained using the Adam optimizer[17].

The loss function for the VAE is a combination of MSE and Kullback-Leibler divergence. Kullback-Leibler divergence is employed to ensure the embedded features fit a normal distribution with mean of zero and standard deviation of 1. The model is trained using the Adam optimizer[17] over 150 epochs. The parameters in fusion networks are saved to pt files after training for future use.

The final stage utilizes the integrative representation in latent space to train a survival model. Before training, parameters for the embedding networks and fusion network are loaded from pt files obtained before, and are frozen during the this stage of training. The loss function is the same as first stage: negative partial log-likelihood. The Adam optimizer[17] is once again utilized in this stage.

2.6 Parameter Selection

The initial hyper-parameters were taken from literature. The hyper-parameters for Cox-nnet were taken from the original publication[4]. The hyper-parameter search in the MAE and VAE was based on parameters tested in [24] and [27]. Some variations were added to those reported choices to optimize the performance.

Five-fold cross-validation was performed during each stage of training and was used to gauge models' performances relative to one another. For each training stage for every modality pair, the model with the highest mean cross-validation performance was passed to the next stage of training and finally tested using the test data set.

2.7 Performance Metrics

This study used concordance index (C-index)[2] as the evaluation metric. These metrics are widely used as indicators of performance of predictive models. The PyCox package was used for calculating these metrics.

The C-index is a standard measure in survival analysis that estimates how good the model is at ranking survival times. By calculating the probability of correctly ranking the event time of cases taken two at a time, it evaluates discrimination performance of models. It can be seen as a generalization of the area under the Receiver Operating Characteristic (ROC) curve to right-censored survival data[38], making it a good metric for survival time measurements. A C-index equal to 1 indicates perfect prediction whereas a C-index equal to 0.5 indicates a random prediction. Higher C-index indicates better capability to provide accurate survival predictions. State of the art survival models typically yield a C-index between 0.6 and 0.7[22].

2.8 Interpretability analysis

As noted in[16], interpretability of a machine learning model is essential in medical applications as it helps build trust between medical personnel and predictive models. However, getting precise and intuitive interpretations of deep neural networks is a challenge[36]. One common technique, feature visualization, generates images to maximize the activation of certain neurons in order to get an understanding of behavior of those neurons[28], but its application is limited to convolutional neural networks.

Other methods have been developed to determine feature importance in other models such as Integrated Gradient (IG) algorithm[40]. This algorithm integrates the model gradients along the path from a given input to a baseline, usually the zero vector. Effectively, this quantifies the output's dependence on each input feature. The input

features can then be ranked by importance to the output. Specifically, this allows the interpretation of the most important features to the log-partial hazard function output. The presence of many previously validated cancer biomarkers among the highest attributed input features suggests that the model is learning biologically relevant features. If these biomarkers are present in the most important features, physicians may have an easier time trusting the outputs of the model. The Captum package was used in this study to analyze input feature attributions using the IG algorithm.

2.9 Summary of novelty

Overall, our contributions can be summarized as follows. To our knowledge, this is the first deep-learning based integrative survival model for lung cancer.

In order to address the "curse of dimensionality" issue, we built TLSurv on a super-hybrid network architecture with four individual neural networks. Using a super-hybrid network to integrate multiple -omics data is new to the field, and TLSurv can be applied to other heterogeneous data modalities. As a modular network, many different embedding networks for different types of input data may be used. For example, a convolutional neural network can be trained to integrate clinical images alongside -omic data, or a recurrent neural network can be trained to integrate sequential data.

We also developed a novel multi-stage transfer learning scheme to train the super-hybrid network incrementally. Each training stage has its own purpose, and it inherits the weights from previously trained sections with the exception of the first stage. As a result, we are effectively training a shallow rather than deep neural network during each training stage, but a powerful deep neural network will be obtained at the end. Since only a shallow neural network is trained during each stage, it mitigates the issues of vanishing gradients[18] and exploding gradients[34].

Lastly, rather than using traditional Gene Set Enrichment Analysis (GSEA), we have used the primary attribution algorithm for interpretability analysis to identify those significant input features. In this way, novel biomarkers could be identified instead of verifying those well established pathways. Those findings can potentially provide new insights for medical and pharmaceutical professionals.

3 RESULTS

3.1 Predictive performance

Figure 4 shows the internal validation C-index based on the cross-validation on the training set for the three architectures and various combinations. Correct discrimination between patients based on predicted survival time yields a higher C-index. Different colours are used to indicate network architectures. It is observed that TLSurv(MAE) generally performs better than TLSurv(VAE).

Figure 5 shows the external validation C-index on test set versus the mean internal validation C-index on training set for the three architectures and various combinations. Two dash lines are drawn to show performance of random guesses. Points far away from the diagonal line are cases of over-fitting. Colours are used to differentiate network architectures, and marker shapes indicate various modality combinations. There is significant overlapping among the three CNV-related models for TLSurv(MAE).

Table 2: Top 20 Most Important Features in mRNA and Methylation Data Sets

Number	mRNA Gene Name	DNA Methylated Genes
1	TNK2-AS1	SIPA1L3
2	AC009299.5	SUGCT
3	RP11-96K19.5	cg0495658*
4	RP11-303E16.3	NFATC1
5	RP11-443B7.1	SHISA6
6	RP11-214N9.1	RP11-444E17.6
7	EXOC3L4	CCDC154
8	CTB-41I6.1	LINC01019
9	RP11-490B18.6	HNF4A
10	FTH1P22	BRINP1
11	CTD-2588E21.1	ABR
12	RP11-712B9.4	F2RL3
13	FGF14-AS2	ADAMTS17
14	RP1-90G24.10	LINC01234
15	RP11-47L3.1	INO80E
16	RNU4-47P	SNHG14
17	LINC00881	AATK
18	GRIN3B	VGLL4
19	AC034220.3	OR9I1
20	NCMAP	ILKAP

*: selected gene is unnamed; Methylation ID listed instead

3.2 Sensitivity to hyper-parameter change

In order to demonstrate the robustness of models towards hyper-parameter changes, heat maps were generated for TLSurv(MAE) of DNA methylation and mRNA, as shown in **Figure 6**. C-indexes here were calculated on the test set using models trained with different hyper-parameter combinations. When changing two of hyper-parameters at one time, the rest hyper-parameters were kept the same as the optimal hyper-parameter combination.

3.3 Running time comparison

One significant concern with deep learning is the computational cost of training models. Especially in the use of -omics data, large data sets with thousands of features result in very high computational cost and therefore require more time to execute. In using a multi-stage network with several discreet training phases, this concern becomes more substantial. The computational cost of the networks was quantified by measuring the amount of time passed during each stage of network training. As seen in **Figure 7**, in some cases, training of the complete mixed-modal pipeline exhibited a significant increase in computational cost over the simpler single modality models.

3.4 Biological relevance of selected features

Based on the predictive performances of the model using the test set, the TLSurv network combining methylation and mRNA expression data was used for interpretability analysis. **Table 2** shows the top 20 input features for TLSurv(MAE) of methylation+mrna combination. Among the top 20 most contributing mRNA and methylation sites, many have been identified by previous studies as significant

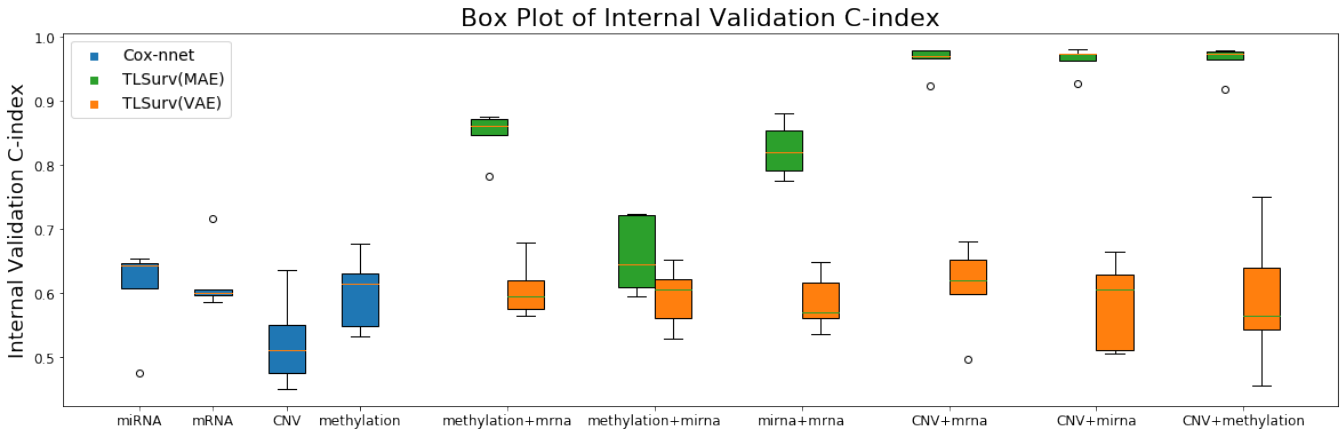

Figure 4: Box plot of internal validation C-index on training set for the three architectures and various modality combinations. The training set were split into five folds for cross validation. Different colours are used to indicate network architectures for comparison.

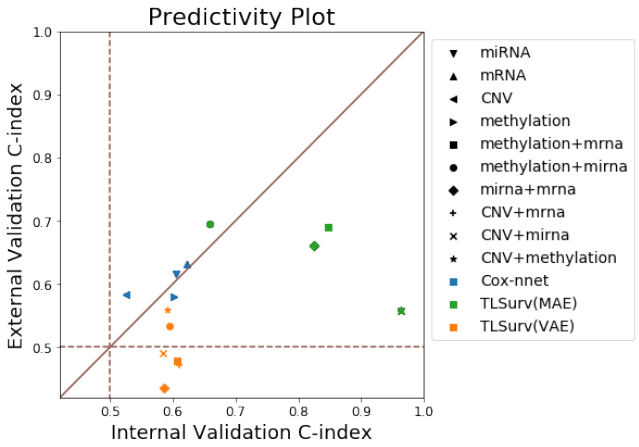

Figure 5: Predictivity plot for C-index on test set versus mean cross-validated C-index on training set for the three architectures and various combinations. Colors reflect network architectures, while marker shapes represent modality combinations. There is significant overlapping among the three CNV-related models for TLSurv(MAE).

biomarkers for lung cancer. The top positively contributing mRNA gene is TNK2-AS1. This gene is upregulated in non-small-cell lung carcinoma, and its over expression is correlated with poor survival outcomes[43]. GRIN3B has been shown to be over expressed in lung carcinoma as well[39]. Other top contributing genes have been identified as biomarkers or tumor related genes in other cancer types including EXOC3L4 in colorectal cancer and FGF14-AS2 in breast cancer[47]. Also, there are many pseudogenes identified as top contributing mRNA genes. Although they are not well studied, pseudogenes are highly specific markers of cell identity which indicate prognosis predictions as noted in[35]. SIPA1L3 is the top positively contributing methylated gene; low methylation of this gene could promote its overall expression which activates the Rap1 signalling pathway for cell adhesion and migration and is correlated with poor prognosis[48]. F2RL3 hypomethylation was also associated with lung cancer occurrences and low survival[49]. Other high ranking methylation features are current proposed biomarkers for other cancer forms, typically dealing with immune signaling pathways[10][44]. Therefore, these top mRNA and methylation genes indicate our model has learned biologically relevant features.

4 DISCUSSION

The state-of-art deep learning-based survival model for using RNA-seq data is Cox-nnet, and the reported performance has a median C-index of around 0.63 for the TCGA-LUAD dataset[4]. The performance of Cox-nnet in this study agree with the reported one. The performance of the TLSurv(MAE) using the DNA methylation and mRNA data sets provided the best results in predicting the survival of lung cancer patients, with a external validated C-index of nearly 0.7, which is better than previous studies in lung cancer models. For future directions, the robustness of models can be further verified by cross-dataset validation once more data become available. Those data can be used to further refine models as well.

Furthermore, in studying the model itself, we have identified some most important features for predicting lung cancer survival, and they coincide with many previously identified biomarkers for cancer. It becomes apparent that the model is learning relevant features, thereby increasing its validity. Given the feature relevance and the promising performance, the power of the proposed TLSurv is clearly demonstrated. It may be of value to further study those other identified features as well, especially those pseudogenes since their roles in human cancer were only recently revealed[14][35][46].

While there is a large increase in computational cost associated with the use of this compound model, the complete training time is still under a minute for all models. Considering the implications of the trained models' performance, we believe that the benefits of this increase in complexity outweigh the increased computational cost.

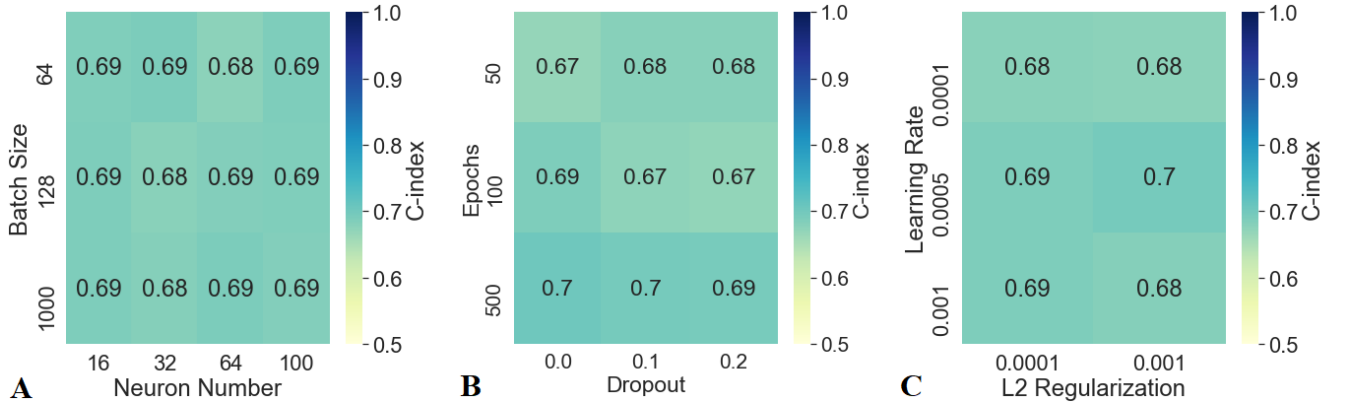

Figure 6: Heat map of external validation C-index on test set for various hyper-parameter sets. A. C-index for various batch sizes and neuron numbers; B. C-index for various training epochs and dropouts; C. C-index for various learning rates and L2 regularization

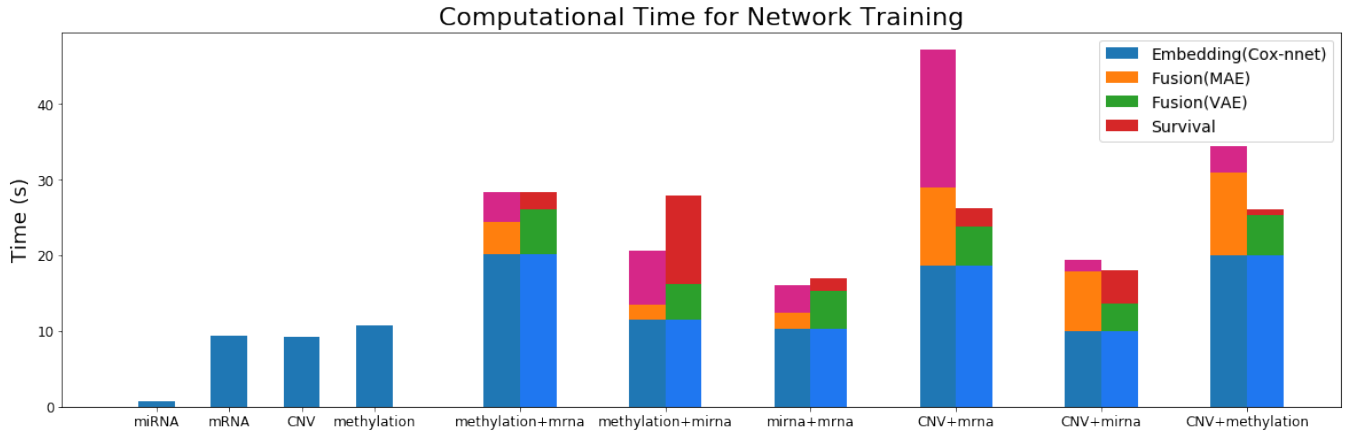

Figure 7: Bar plot of training time for each section of the three architectures and various combinations. For single modalities, computational time of survival models is equivalent to embedding models used in integrative modalities. For integrative modalities, the left column shows the total computational time of TLSurv(MAE), and the right column shows the total computational time of TLSurv(VAE). The order of items in the legend reflects the training order.

The disparity between the cross-validation performance and the external validation performance in the models integrating the copy number variation data set is of note. All three TLSurv(MAE) models incorporating the copy number variation (CNV) data had cross-validation C-indices greater than 0.9, meaning that during training, the model was able to learn the data so well that it correctly discriminated between two patients more than 90% of the time, which is an unprecedented performance. However, in the test set, these models achieved average C-indices around 0.55, which indicates relatively poor performance. These results imply that the model was overfitted to the training data, leading to poor performance when presented with data it had not seen before. Interestingly, these results were consistent across all combinations involving the CNV data set and this behavior was insensitive to differing hyper-parameters. The models were built with dropout layers, which turn off nodes randomly during training to combat overfitting to the training data. Even given a large probability of dropout in these layers and a small number of training epochs, both of which discourage overfitting, this problem persisted. One possible explanation is the categorical nature of the CNV data. Unlike the other data modalities as numerical data, copy number variations are expressed as categorical data for under-expression, normal expression, and over-expression. Additionally, the sequential training of the models may have contributed to overfitting. Further investigations should be conducted to check whether the method of cross-validation requires any modification when applying to multi-stage learning.

Compared with TLSurv(MAE), TLSurv with VAE implementation has a significantly lower performance. One explanation may be that VAE is trying to learn the latent representation in a continuous and highly structured manner so that every point sampled in the latent space is decoded to a valid output, and the small sample size hinders the VAE from projecting into the latent space well.

To the best of the authors' knowledge, this focus on deep learning-based multi-omics data integration for survival analysis is a new

development within lung cancer research and can be used to increase the level of personalized patient care. By giving physicians descriptive metrics, like predicted time until follow-up as presented in this study, physicians are able to provide more personalized care and make better informed decisions.

The great potential of multi-stage transfer learning was demonstrated in this study. By using it, powerful deep super-hybrid networks can be trained incrementally with shallow subnetworks. Since only a shallow neural network is trained during each stage, it can potentially alleviate the "curse of dimensionality" issue and vanishing gradients issue. The flexibility of embedding networks greatly enhances multi-stage learning's broad application. As various biomedical data modalities are becoming more and more available, this method can be applied to different contexts. For example, with the advancement of wearable technology, different sources like sleeping, sports and diet could be integrated to get a more comprehensive view of human health.

5 CONCLUSION

In this study, we have proposed a novel, multi-stage model called TLSurv for survival prediction in lung cancer using multi-view factorization autoencoders (MAE) for multi-omics data integration. Interpretability analysis has shown its biological relevance, and some novel biomarkers have been identified. The great potential of multi-stage transfer learning for super-hybrid networks is also demonstrated. Future investigations may be aimed at applying this framework to other diseases or other modalities of data.

ACKNOWLEDGMENTS

The results here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. We also want to thank UCSC Xena platform for hosting those data. The work was supported in part by grants from the National Science Foundation EAGER Award NSF1651360, Microsoft for Azure Cloud Grant Support, Georgia Cancer Coalition Distinguished Cancer Award, Amazon Research Award, Georgia Tech Petit Institute Faculty Fellow, and Carol Ann and David D. Flanagan Faculty Fellow Research Fund. This work was supported in part by the scholarship from China Scholarship Council (CSC) under the Grant CSC NO. 201406010343. The content of this article is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

REFERENCES

[1] Hervé Abdi and Lynne J Williams. 2010. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics* 2, 4 (2010), 433–459.
[2] Laura Antolini, Patrizia Boracchi, and Elia Biganzoli. 2005. A time-dependent discrimination index for survival data. *Statistics in medicine* 24, 24 (2005), 3927–3944.
[3] Kumardeep Chaudhary, Olivier B Poirion, Liangqun Lu, and Lana X Garmire. 2018. Deep learning–based multi-omics integration robustly predicts survival in liver cancer. *Clinical Cancer Research* 24, 6 (2018), 1248–1259.
[4] Travers Ching, Xun Zhu, and Lana X Garmire. 2018. Cox-nnet: an artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS computational biology* 14, 4 (2018), e1006076.
[5] David R Cox. 1972. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)* 34, 2 (1972), 187–202.
[6] Zijian Ding, Songpeng Zu, and Jin Gu. 2016. Evaluating the molecule-based prediction of clinical drug responses in cancer. *Bioinformatics* 32, 19 (2016), 2891–2895.
[7] Michael F Gensheimer and Balasubramanian Narasimhan. 2019. A scalable discrete-time survival model for neural networks. *PeerJ* 7 (2019), e6257.
[8] Mary J Goldman, Brian Craft, Mim Hastie, Kristupas Repečka, Fran McDade, Akhil Kamath, Ayan Banerjee, Yunhai Luo, Dave Rogers, Angela N Brooks, et al. 2020. Visualizing and interpreting cancer genomics data via the Xena platform. *Nature Biotechnology* (2020), 1–4.
[9] Balázs Győrffy, Pawel Surowiak, Jan Budczies, and András Lánczky. 2013. Online survival analysis software to assess the prognostic value of biomarkers using transcriptomic data in non-small-cell lung cancer. *PloS one* 8, 12 (2013).
[10] Christina Hausmann, Achim Temme, Nils Cordes, and Iris Eke. 2015. ILKAP, ILK and PINCH1 control cell survival of p53-wildtype glioblastoma cells after irradiation. *Oncotarget* 6, 33 (2015), 34592.
[11] Zhi Huang, Travis S Johnson, Zhi Han, Bryan Helm, Sha Cao, Chi Zhang, Paul Salama, Maher Rizkalla, Christina Y Yu, Jun Cheng, et al. 2020. Deep learning-based cancer survival prognosis from RNA-seq data: approaches and evaluations. *BMC medical genomics* 13 (2020), 1–12.
[12] Zhi Huang, Xiaohui Zhan, Shunian Xiang, Travis S Johnson, Bryan Helm, Christina Y Yu, Jie Zhang, Paul Salama, Maher Rizkalla, Zhi Han, et al. 2019. SALMON: Survival analysis learning with multi-omics neural networks on breast cancer. *Frontiers in genetics* 10 (2019), 166.
[13] Hemant Ishwaran and Min Lu. 2014. Random survival forests. *Wiley StatsRef: Statistics Reference Online* (2014), 1–13.
[14] Shanker Kalyana-Sundaram, Chandan Kumar-Sinha, Sunita Shankar, Dan R Robinson, Yi-Mi Wu, Xuhong Cao, Irfan A Asangani, Vishal Kothari, John R Prensner, Robert J Lonigro, et al. 2012. Expressed pseudogenes in the transcriptional landscape of human cancers. *Cell* 149, 7 (2012), 1622–1634.
[15] Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. 2018. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC medical research methodology* 18, 1 (2018), 24.
[16] Sujata Khedkar, Priyanka Gandhi, Gayatri Shinde, and Vignesh Subramanian. 2020. Deep Learning and Explainable AI in Healthcare Using EHR. In *Deep Learning Techniques for Biomedical and Health Informatics*. Springer, 129–148.
[17] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[18] Janusz Kolbusz, Pawel Rozycki, and Bogdan M Wilamowski. 2017. The study of architecture MLP with linear neurons in order to eliminate the "vanishing gradient" problem. In *International Conference on Artificial Intelligence and Soft Computing*. Springer, 97–106.
[19] Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. 2019. Time-to-event prediction with neural networks and Cox regression. *Journal of Machine Learning Research* 20, 129 (2019), 1–30.
[20] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436–444.
[21] Yingming Li, Ming Yang, and Zhongfei Zhang. 2018. A survey of multi-view representation learning. *IEEE transactions on knowledge and data engineering* 31, 10 (2018), 1863–1883.
[22] Margaux Luck, Tristan Sylvain, Héloïse Cardinal, Andrea Lodi, and Yoshua Bengio. 2017. Deep learning for patient-specific kidney graft survival analysis. *arXiv preprint arXiv:1705.10245* (2017).
[23] Tianle Ma and Aidong Zhang. 2017. Integrate multi-omic data using affinity network fusion (anf) for cancer patient clustering. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 398–403.
[24] Tianle Ma and Aidong Zhang. 2019. Integrate multi-omics data with biological interaction networks using Multi-view Factorization AutoEncoder (MAE). *BMC genomics* 20, 11 (2019), 1–11.
[25] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.
[26] Jonathan Mitchel, Kevin Chatlin, Li Tong, and May D Wang. 2019. A Translational Pipeline for Overall Survival Prediction of Breast Cancer Patients by Decision-Level Integration of Multi-Omics Data. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 1573–1580.
[27] Simidjievski Nikola, Bodnar Cristian, Tariq Ifrah, Scherer Paul, Helena Andres Terre, Zohreh Shams, Mateja Jamnik, and Pietro Liò. 2019. Variational Autoencoders for Cancer Data Integration: Design Principles and Computational Practice. *Frontiers in Genetics* 10, 1205 (2019). https://doi.org/10.3389/fgene.2019.01205
[28] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. 2017. Feature Visualization. *Distill* (2017). https://doi.org/10.23915/distill.00007 https://distill.pub/2017/feature-visualization.
[29] Michael Olivier, Reto Asmis, Gregory A Hawkins, Timothy D Howard, and Laura A Cox. 2019. The Need for Multi-Omics Biomarker Signatures in Precision Medicine. *International Journal of Molecular Sciences* 20, 19 (2019), 4781.
[30] Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 22, 10 (2009), 1345–1359.
[31] The pandas development team. 2020. *pandas-dev/pandas: Pandas*. https://doi.org/10.5281/zenodo.3509134
[32] John H Phan, Chang F Quo, Chihwen Cheng, and May Dongmei Wang. 2012. Multiscale integration of-omic, imaging, and clinical data in biomedical informatics.

IEEE reviews in biomedical engineering 5 (2012), 74–87.

[33] John H Phan, Chang-Feng Quo, and May D Wang. 2006. Functional genomics and proteomics in the clinical neurosciences: data mining and bioinformatics. *Progress in brain research* 158 (2006), 83–108.

[34] George Philipp, Dawn Song, and Jaime G Carbonell. 2017. The exploding gradient problem demystified-definition, prevalence, impact, origin, tradeoffs, and solutions. *arXiv preprint arXiv:1712.05577* (2017).

[35] Laura Poliseno, Andrea Marranci, and Pier Paolo Pandolfi. 2015. Pseudogenes in human cancer. *Frontiers in medicine* 2 (2015), 68.

[36] Daniele Ravì, Charence Wong, Fani Deligianni, Melissa Berthelot, Javier Andreu-Perez, Benny Lo, and Guang-Zhong Yang. 2016. Deep learning for health informatics. *IEEE journal of biomedical and health informatics* 21, 1 (2016), 4–21.

[37] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. 2020. Cancer statistics, 2020. *CA: A Cancer Journal for Clinicians* 70, 1 (2020), 7–30. https://doi.org/10.3322/caac.21590

[38] Harald Steck, Balaji Krishnapuram, Cary Dehing-Oberije, Philippe Lambin, and Vikas C Raykar. 2008. On ranking in survival analysis: Bounds on the concordance index. In *Advances in neural information processing systems*. 1209–1216.

[39] Andrzej Stepulak, Hella Luksch, Christine Gebhardt, Ortrud Uckermann, Jenny Marzahn, Marco Sifringer, Wojciech Rzeski, Christian Staufner, Katja S Brocke, Lechoslaw Turski, et al. 2009. Expression of glutamate receptor subunits in human cancers. *Histochemistry and cell biology* 132, 4 (2009), 435–445.

[40] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 3319–3328.

[41] Li Tong, Jonathan Mitchel, Kevin Chatlin, and May D Wang. in press. Deep Learning based Feature-Level Integration of Multi-Omics Data for Breast Cancer Patients Survival Analysis. (in press).

[42] Li Tong, Hang Wu, and May D Wang. in press. Integrating Multi-Omics Data by Learning Modality Invariant Representations for Improved Prediction of Overall Survival of Cancer. (in press).

[43] Yue Wang, Dongmei Han, Liming Pan, and Jing Sun. 2018. The positive feedback between lncRNA TNK2-AS1 and STAT3 enhances angiogenesis in non-small cell lung cancer. *Biochemical and biophysical research communications* 507, 1-4 (2018), 185–192.

[44] Christine Wolf, Angela Garding, Katharina Filarsky, Jasmin Bahlo, Sandra Robrecht, Natalia Becker, Manuela Zucknick, Arefeh Rouhi, Anja Weigel, Rainer Claus, et al. 2018. NFATC1 activation by DNA hypomethylation in chronic lymphocytic leukemia correlates with clinical staging and can be inhibited by ibrutinib. *International journal of cancer* 142, 2 (2018), 322–333.

[45] Po-Yen Wu, Raghu Chandramohan, John H Phan, William T Mahle, J William Gaynor, Kevin O Maher, and May D Wang. 2014. Cardiovascular transcriptomics and epigenomics using next-generation sequencing: challenges, progress, and opportunities. *Circulation: Cardiovascular Genetics* 7, 5 (2014), 701–710.

[46] Lu Xiao-Jie, Gao Ai-Mei, Ji Li-Juan, and Xu Jiang. 2015. Pseudogene in cancer: real functions and promising signature. *Journal of medical genetics* 52, 1 (2015), 17–24.

[47] Fan Yang, Ye-huan Liu, Si-yang Dong, Rui-ming Ma, Adheesh Bhandari, Xiao-hua Zhang, and Ou-chen Wang. 2016. A novel long non-coding RNA FGF14-AS2 is correlated with progression and prognosis in breast cancer. *Biochemical and biophysical research communications* 470, 3 (2016), 479–483.

[48] Ruyang Zhang, Linjing Lai, Xuesi Dong, Jieyu He, Dongfang You, Chao Chen, Lijuan Lin, Ying Zhu, Hui Huang, Sipeng Shen, et al. 2019. SIPA1L3 methylation modifies the benefit of smoking cessation on lung adenocarcinoma survival: an epigenomic–smoking interaction analysis. *Molecular oncology* 13, 5 (2019), 1235–1248.

[49] Yan Zhang, Ben Schöttker, José Ordóñez-Mena, Bernd Holleczek, Rongxi Yang, Barbara Burwinkel, Katja Butterbach, and Hermann Brenner. 2015. F2RL3 methylation, lung cancer incidence and mortality. *International journal of cancer* 137, 7 (2015), 1739–1748.