A Multitask Multiple Kernel Learning Algorithm for Survival Analysis with Application to Cancer Biology

Onur Dereli [1] **Ceyda Oğuz** [2] **Mehmet Gönen** [2 3 4]

Abstract

Predictive performance of machine learning algorithms on related problems can be improved using multitask learning approaches. Rather than performing survival analysis on each data set to predict survival times of cancer patients, we developed a novel multitask approach based on multiple kernel learning (MKL). Our multitask MKL algorithm both works on multiple cancer data sets and integrates cancer-related pathways/gene sets into survival analysis. We tested our algorithm, which is named as Path2MSurv, on the Cancer Genome Atlas data sets analyzing gene expression profiles of 7,655 patients from 20 cancer types together with cancer-specific pathway/gene set collections. Path2MSurv obtained better or comparable predictive performance when benchmarked against random survival forest, survival support vector machine, and single-task variant of our algorithm. Path2MSurv has the ability to identify key pathways/gene sets in predicting survival times of patients from different cancer types.

1. Introduction

Understanding the formation and progression mechanisms of the diseases plays a vital importance in treating them. To this aim, genomic characterizations have been used in answering various research problems. Survival analysis is one of these research problems that aims to predict survival times of patients. There are several machine learning algorithms developed to predict survival times using genomic characterizations and clinical information of patients (Cox, 1972; Cox & Oakes, 1984; Bakker et al., 2004; Shivaswamy et al., 2007; Evers & Messow, 2008; Ishwaran et al., 2008; Khan & Zubek, 2008; Van Belle et al., 2011a;b; Mogensen & Gerds, 2013; Kiaee et al., 2016; Wang et al., 2016; Yousefi et al., 2017). These existing algorithms consider the censored observations, but most of them cannot handle high-dimensional feature representations (e.g., genomic characterizations) effectively due to the limited number of training samples. These standard algorithms were recently shown to be more suitable for low-dimensional feature representations (i.e., clinical variables) (Yuan et al., 2014).

Pathways/gene sets are simply the sets of genes with roles in the same or similar biological mechanisms. Relating pathways/gene sets to clinical phenotypes helps us better understand the underlying mechanisms of diseases. That is why several machine learning algorithms were proposed to identify pathways/gene sets associated with disease-related phenotypes such as overall survival time after diagnosis. These algorithms either (i) identify survival-related molecular mechanisms using feature selection and learn a survival analysis model on selected features only or (ii) train a survival analysis model on each pathway/gene set separately and pick survival-related ones by comparing their predictive performances (Pang et al., 2012; Zhang et al., 2017; Pang et al., 2010; 2011). However, both approaches have drawbacks. The first approach might pick biologically unrelated genes due to highly correlated structure of genomic characterizations. The second approach might pick related or similar pathways/gene sets due to the fact that each pathway/gene set is analyzed separately. To eliminate these problems, pathway/gene set collections should be integrated into the model during the training step, so that learning algorithm can pick informative pathways/gene sets in a more robust manner.

Using high-dimensional genomic characterizations in machine learning algorithms is a challenging task due to their highly correlated structures. Kernel-based machine learning algorithms were used to address this problem in survival analysis (Shivaswamy et al., 2007; Evers & Messow, 2008; Khan & Zubek, 2008; Van Belle et al., 2011a;b). Kernel methods were also shown to be very successful in other cancer-related problems such as drug sensitivity predic-

[1]Graduate School of Sciences and Engineering, Koç University, İstanbul 34450, Turkey [2]Department of Industrial Engineering, College of Engineering, Koç University, İstanbul 34450, Turkey [3]School of Medicine, Koç University, İstanbul 34450, Turkey [4]Department of Biomedical Engineering, School of Medicine, Oregon Health & Science University, Portland, OR 97239, USA. Correspondence to: Mehmet Gönen <mehmetgonen@ku.edu.tr>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

tion (Costello et al., 2014) and gene essentiality prediction (Gönen et al., 2017). The success of kernel methods lies mainly in the fact that the number of model parameters optimized is proportional to the number of samples not to the number of features (Schölkopf & Smola, 2002).

The kernel function that defines a similarity measure between pairs of samples is the most important component of kernel methods. No single kernel function is the best one for different problems. That is why we can use a weighted combination of several kernel functions instead of using a single kernel, which is known as multiple kernel learning (MKL) (Gönen & Alpaydın, 2011). Following this idea, an MKL-based survival analysis algorithm can pick informative pathways/gene sets by assigning zero weight to uninformative ones during inference. This approach defines a kernel function on each pathway/gene set using genomic features of the genes included. MKL part then learns an optimized kernel function, which is used to predict survival times (Dereli et al., in press).

Multitask learning aims to mainly model related problems conjointly by exploiting commonalities between them (Caruana, 1997). This idea has also been applied in cancer studies to improve predictive performance (Costello et al., 2014; Gönen et al., 2017). Recently, studies modeling multiple cancer types simultaneously (i.e., pan-cancer studies) to capture common underlying biological mechanisms have attracted great attention (The Cancer Genome Atlas Research Network et al., 2013; Yang et al., 2014; Anaya et al., 2016). However, to the best of our knowledge, there is a limited number of multitask learning methods for survival analysis (Li et al., 2016; Wang et al., 2017).

In this study, we combined survival analysis, MKL (for pathway selection) and multitask learning (for modeling multiple cohorts) in a unified formulation for the first time. Our algorithm can identify survival-related biological pathways/gene sets using high dimensional genomic characterizations of patients from multiple cohorts.

2. Related Work

Random forest (RF) is a supervised machine learning algorithm originally developed for regression and classification (Breiman, 2001). It first creates multiple decision trees using randomly selected features from the input features or randomly selected samples from the training data. It then combines these decision trees to obtain more robust predictions. RF was also extended towards survival analysis and successfully used in many studies (Ishwaran et al., 2008).

Support vector machine (SVM) is another supervised machine learning algorithm originally developed for binary classification (Cortes & Vapnik, 1995). SVM was also extended towards censored regression problems, i.e., survival analysis (Shivaswamy et al., 2007; Khan & Zubek, 2008). Survival SVM can be formulated as follows:

$$
\begin{aligned}
\text{min.}\quad & \frac{1}{2}\boldsymbol{w}^\top\boldsymbol{w} + C\sum_{i=1}^{N}(\xi_i^+ + (1-\delta_i)\xi_i^-) \\
\text{w.r.t.}\quad & \boldsymbol{w}\in\mathbb{R}^D,\ \ \boldsymbol{\xi}^+\in\mathbb{R}^N,\ \ \boldsymbol{\xi}^-\in\mathbb{R}^N,\ \ b\in\mathbb{R} \\
\text{s.t.}\quad & \epsilon + \xi_i^+ \geq y_i - \boldsymbol{w}^\top\boldsymbol{x}_i - b \quad \forall i \\
& \epsilon + \xi_i^- \geq \boldsymbol{w}^\top\boldsymbol{x}_i + b - y_i \quad \forall i \\
& \xi_i^+ \geq 0 \quad \forall i \\
& \xi_i^- \geq 0 \quad \forall i,
\end{aligned}
\tag{1}
$$

where the training data set is $\{(\boldsymbol{x}_i, \delta_i, y_i)\}_{i=1}^N$, N is the number of samples, $\boldsymbol{x}_i$ is the feature vector of sample i, $\delta_i \in \{0,1\}$ is the binary indicator variable that shows whether the observed survival time of sample i is censored (i.e., $\delta_i = 1$) or not (i.e., $\delta_i = 0$), and $y_i \in \mathbb{R}$ is the observed survival time of sample i (i.e., time to last follow-up if censored or time to death if uncensored). Here, $\boldsymbol{w}$ is the set of weights assigned to features, C is the non-negative regularization parameter, $\boldsymbol{\xi}^+$ and $\boldsymbol{\xi}^-$ are the sets of slack variables, D is the number of input features, ϵ is the non-negative tube width parameter, and b is the bias parameter.

The primal optimization problem in (1) has $(D + 2N + 1)$ decision variables, which makes the model computationally very costly. To integrate kernel functions into this formulation via standard kernel trick, the corresponding Lagrangian function is written as

$$
\begin{aligned}
\mathcal{L} =& \frac{1}{2}\boldsymbol{w}^\top\boldsymbol{w} + C\sum_{i=1}^{N}(\xi_i^+ + (1-\delta_i)\xi_i^-) \\
& - \sum_{i=1}^{N}\alpha_i^+(\epsilon + \xi_i^+ - y_i + \boldsymbol{w}^\top\boldsymbol{x}_i + b) \\
& - \sum_{i=1}^{N}\alpha_i^-(\epsilon + \xi_i^- - \boldsymbol{w}^\top\boldsymbol{x}_i - b + y_i) \\
& - \sum_{i=1}^{N}\beta_i^+\xi_i^+ - \sum_{i=1}^{N}\beta_i^-\xi_i^-.
\end{aligned}
$$

The derivatives of the Lagrangian function with respect to the decision variables of the primal problem are found as

$$
\begin{aligned}
\frac{\partial\mathcal{L}}{\boldsymbol{w}} &= 0 \Rightarrow \boldsymbol{w} = \sum_{i=1}^{N}(\alpha_i^+ - \alpha_i^-)\boldsymbol{x}_i \\
\frac{\partial\mathcal{L}}{\partial b} &= 0 \Rightarrow \sum_{i=1}^{N}(\alpha_i^+ - \alpha_i^-) = 0 \\
\frac{\partial\mathcal{L}}{\partial\xi_i^+} &= 0 \Rightarrow C = \alpha_i^+ + \beta_i^+ \qquad \forall i \\
\frac{\partial\mathcal{L}}{\partial\xi_i^-} &= 0 \Rightarrow C(1-\delta_i) = \alpha_i^- + \beta_i^- \qquad \forall i.
\end{aligned}
$$

Using the Lagrangian function and these derivatives, the corresponding dual optimization problem is written as

$$\begin{aligned}
\text{min.} \quad & -\sum_{i=1}^{N} y_i(\alpha_i^+ - \alpha_i^-) + \epsilon \sum_{i=1}^{N}(\alpha_i^+ + \alpha_i^-) \\
& + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i^+ - \alpha_i^-)(\alpha_j^+ - \alpha_j^-)\boldsymbol{x}_i^\top \boldsymbol{x}_j \\
\text{w.r.t.} \quad & \boldsymbol{\alpha}^+ \in \mathbb{R}^N, \ \ \boldsymbol{\alpha}^- \in \mathbb{R}^N \\
\text{s.t.} \quad & \sum_{i}^{N}(\alpha_i^+ - \alpha_i^-) = 0 \\
& C \geq \alpha_i^+ \geq 0 \quad \forall i \\
& C(1-\delta_i) \geq \alpha_i^- \geq 0 \quad \forall i.
\end{aligned} \tag{2}$$

The dual optimization problem in (2) has $2N$ decision variables instead of $(D + 2N + 1)$, which significantly reduces the computational complexity. By replacing the term $\boldsymbol{x}_i^\top \boldsymbol{x}_j$ with a kernel function $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, kernel functions can be integrated into the model.

Several recent studies showed that different cancer types have similar or same underlying biological mechanisms (The Cancer Genome Atlas Research Network et al., 2013; Choi et al., 2014; Damrauer et al., 2014; Hoadley et al., 2014; Lawrence et al., 2014; Yang et al., 2014; Khirade et al., 2015; Pappa et al., 2015; Wan et al., 2015; Anaya et al., 2016), which supports the joint modeling of multiple diseases. That is why there are existing multitask machine learning models to model multiple patient cohorts conjointly (Li et al., 2016; Wang et al., 2017). However, these methods use genomic features directly, and they are not able to extract relative importance of pathways/gene sets.

3. Our Proposed Multitask MKL Algorithm for Survival Analysis

We extended survival SVM algorithm towards multitask learning and MKL, which is named as Path2MSurv (Figure 1a). By doing so, we will be able to model multiple cohorts simultaneously and to extract survival-related pathways/gene sets to identify shared biological mechanisms among these cohorts.

The training data sets defined over multiple cohorts are given as $\{\{(\boldsymbol{x}_{ti}, \delta_{ti}, y_{ti})\}_{i=1}^{N_t}\}_{t=1}^{T}$, where T denotes the number of tasks (i.e., cohorts), N_t represents the total number of samples for task t, $\boldsymbol{x}_{ti}$ is the feature vector of sample i of task t, δ_{ti} is the binary indicator variable that shows whether the observed survival time of sample i of task t is censored (i.e., $\delta_{ti} = 1$) or not (i.e., $\delta_{ti} = 0$), and $y_{ti} \in \mathbb{R}$ is the observed survival time of sample i of task t. The primal optimization problem of our formulation can be written as

$$\begin{aligned}
\text{min.} \quad & \sum_{t=1}^{T}\left[\frac{1}{2}\boldsymbol{w}_t^\top \boldsymbol{w}_t + C\sum_{i=1}^{N_t}(\xi_{ti}^+ + (1-\delta_{ti})\xi_{ti}^-)\right] \\
\text{w.r.t.} \quad & \boldsymbol{w}_t \in \mathbb{R}^{D_t}, \ \ \boldsymbol{\xi}_t^+ \in \mathbb{R}^{N_t}, \ \ \boldsymbol{\xi}_t^- \in \mathbb{R}^{N_t}, \ \ b_t \in \mathbb{R} \\
\text{s.t.} \quad & \epsilon + \xi_{ti}^+ \geq y_{ti} - \boldsymbol{w}_t^\top \boldsymbol{x}_{ti} - b_t \quad \forall (t,i) \\
& \epsilon + \xi_{ti}^- \geq \boldsymbol{w}_t^\top \boldsymbol{x}_{ti} + b_t - y_{ti} \quad \forall (t,i) \\
& \xi_{ti}^+ \geq 0 \quad \forall (t,i) \\
& \xi_{ti}^- \geq 0 \quad \forall (t,i),
\end{aligned} \tag{3}$$

where $\boldsymbol{w}_t$ is the vector of weights assigned to features for task t, C is the non-negative regularization parameter, $\boldsymbol{\xi}_t^+$ and $\boldsymbol{\xi}_t^-$ are the sets of slack variables for task t, D_t is the number of input features for task t, ϵ is the non-negative tube width parameter, and b_t is the bias parameter for task t.

We formulated the corresponding dual optimization problem, where we have a combined objective function over all tasks with a single set of constraints on the kernel weights.

$$\begin{aligned}
\text{min.} \quad & \sum_{t=1}^{T} J_t(\boldsymbol{\eta}) \\
\text{w.r.t.} \quad & \boldsymbol{\eta} \in \mathbb{R}^P \\
\text{s.t.} \quad & \sum_{m=1}^{P} \eta_m = 1 \\
& \eta_m \geq 0 \quad \forall m.
\end{aligned} \tag{4}$$

The inner optimization model $J_t(\boldsymbol{\eta})$ for each task is basically a single-kernel survival SVM defined as

$$\begin{aligned}
\text{min.} \quad & -\sum_{i=1}^{N_t} y_{ti}(\alpha_{ti}^+ - \alpha_{ti}^-) + \epsilon \sum_{i=1}^{N_t}(\alpha_{ti}^+ + \alpha_{ti}^-) \\
& + \frac{1}{2}\sum_{i=1}^{N_t}\sum_{j=1}^{N_t}(\alpha_{ti}^+ - \alpha_{ti}^-)(\alpha_{tj}^+ - \alpha_{tj}^-)k_{\boldsymbol{\eta}}(\boldsymbol{x}_{ti}, \boldsymbol{x}_{tj}) \\
\text{w.r.t.} \quad & \boldsymbol{\alpha}_t^+ \in \mathbb{R}^{N_t}, \ \ \boldsymbol{\alpha}_t^- \in \mathbb{R}^{N_t} \\
\text{s.t.} \quad & \sum_{i=1}^{N_t}(\alpha_{ti}^+ - \alpha_{ti}^-) = 0 \\
& C \geq \alpha_{ti}^+ \geq 0 \quad \forall i \\
& C(1-\delta_{ti}) \geq \alpha_{ti}^- \geq 0 \quad \forall i,
\end{aligned} \tag{5}$$

where $k_{\boldsymbol{\eta}}(\boldsymbol{x}_{ti}, \boldsymbol{x}_{tj})$ corresponds to $\sum_{m=1}^{P} \eta_m k_m(\boldsymbol{x}_{ti}, \boldsymbol{x}_{tj})$. We are guaranteed to obtain a sparse set of kernel weights in Path2MSurv since $\boldsymbol{\eta}$ lies on a simplex, i.e., $\boldsymbol{\eta} \in \mathbb{R}^P$, $\sum_{m=1}^{P} \eta_m = 1$, and $\eta_m \geq 0$.

It is not possible to find the global optimal solution of the overall optimization problem in (4) since it is not jointly convex with respect to decision variables $\boldsymbol{\eta}$ and

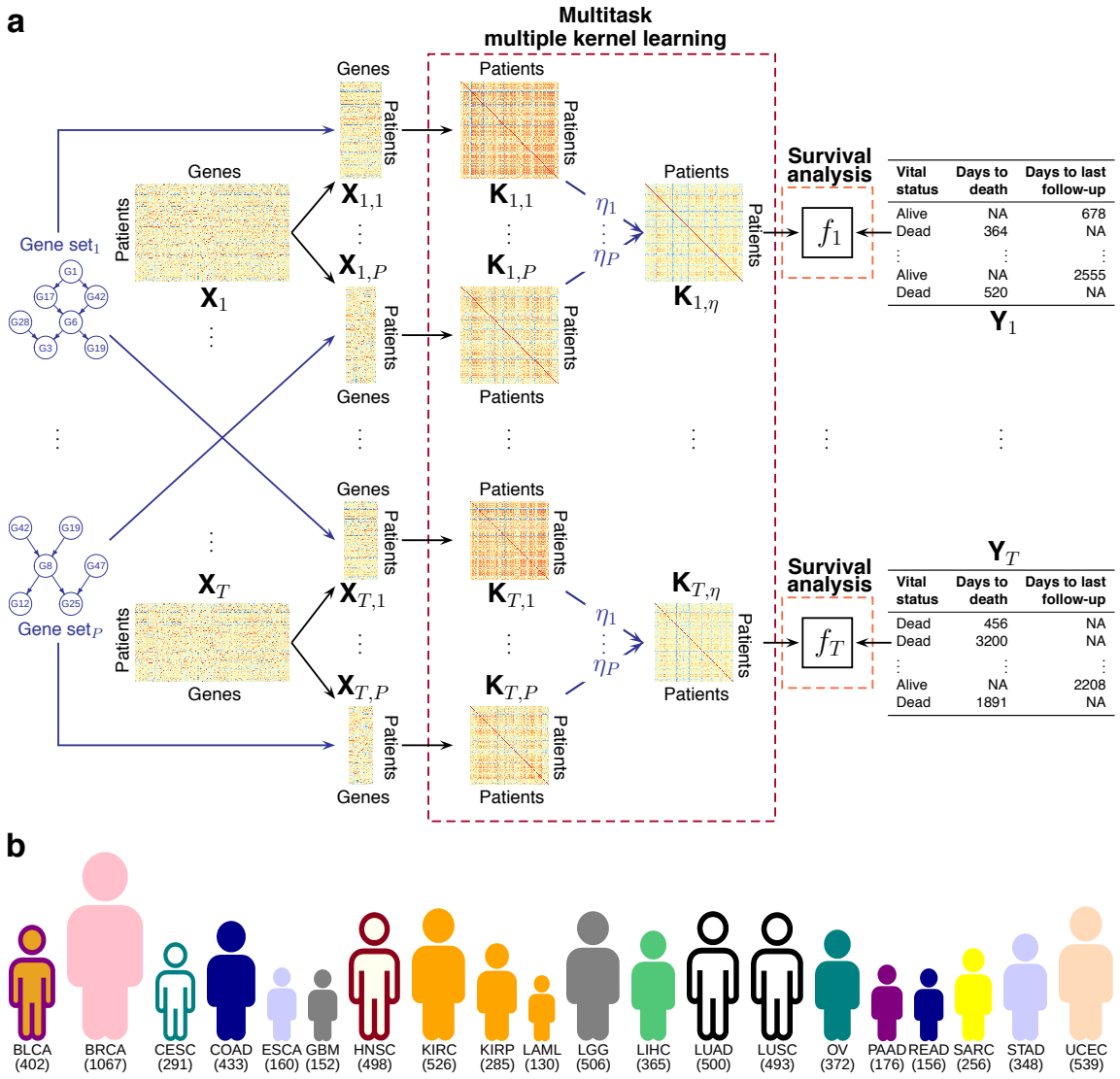

Figure 1. Overview of the proposed Path2MSurv algorithm together with the summary of data sets used in our computational experiments. (**a**) Path2MSurv algorithm takes gene expression profiles of patients from each cohort, i.e., $\{\mathbf{X}_t\}_{t=1}^T$, a pathway/gene set collection with P pathways/gene sets, and clinical information including vital status, days to death, and days to last follow-up, i.e., $\{\mathbf{Y}_t\}_{t=1}^T$, as its inputs. It then calculates kernel matrices, i.e., $\{\mathbf{K}_{t,p}\}_{p=1}^P$, on data matrix slices, i.e., $\{\mathbf{X}_{t,p}\}_{t=1,p=1}^{T,P}$, obtained by mapping pathways/gene sets on gene expression profiles. The weighted sums of these kernel matrices, i.e., $\{\mathbf{K}_{t,\eta}\}_{t=1}^T$, are used to predict survival times of cancer patients using the prediction functions, i.e., $\{f\}_{t=1}^T$. (**b**) Data sets used in our computational experiments and their corresponding numbers of patients after filtering steps.

$\{(\boldsymbol{\alpha}_t^+, \boldsymbol{\alpha}_t^-)\}_{t=1}^T$. Instead, we formulated an alternating optimization approach to the overall optimization problem by following the idea proposed by Xu et al. (2010). Kernel weights are initialized to uniform values, i.e., $\eta_m^{(s)} = 1/P$, at the first iteration, i.e., $s = 0$. In each iteration, using kernel weights $\boldsymbol{\eta}^{(s)}$, we solve the inner optimization problem in (5) for each task to obtain its corresponding support vector coefficients $\{\boldsymbol{\alpha}_t^{+(s)}, \boldsymbol{\alpha}_t^{-(s)}\}$. Kernel weights are then updated for the next iteration $(s+1)$ using the support vector coefficients of all tasks in the following update equation:

$$\eta_m^{(s+1)} = \frac{\sum_{t=1}^{T} \eta_m^{(s)} \sqrt{\sum_{i=1}^{N_t} \sum_{j=1}^{N_t} \alpha_{ti}^{(s)} \alpha_{tj}^{(s)} k_m(\boldsymbol{x}_{ti}, \boldsymbol{x}_{tj})}}{\sum_{t=1}^{T} \sum_{o=1}^{P} \eta_o^{(s)} \sqrt{\sum_{i=1}^{N_t} \sum_{j=1}^{N_t} \alpha_{ti}^{(s)} \alpha_{tj}^{(s)} k_o(\boldsymbol{x}_{ti}, \boldsymbol{x}_{tj})}} \quad \forall m,$$

where $\alpha_{ti}^{(s)} = (\alpha_{ti}^{+(s)} - \alpha_{ti}^{-(s)})$. The convergence of this alternating optimization approach is guaranteed since we monotonically decrease the objective function value of (4).

4. Experiments

We performed an extensive set of computational experiments on several cancer data sets, where we compared the predictive performance of our proposed method Path2MSurv against survival RF (Ishwaran et al., 2008), survival SVM (Shivaswamy et al., 2007; Khan & Zubek, 2008), and single-task variant of our algorithm (i.e., Path2Surv) trained on each data set separately (Dereli et al., in press).

4.1. Data Sets

We used gene expression profiles and clinical annotation data of cancer patients provided by The Cancer Genome Atlas (TCGA) at the Genomics Data Commons (GDC) data portal (`https://portal.gdc.cancer.gov`). To integrate prior biological knowledge about cancer-specific pathways/gene sets into our model, we used one pathway and one gene set collection.

4.1.1. TCGA Data Sets

TCGA data sets include genomic characterizations and clinical information of more than 10,000 cancer patients for 33 different cancer types. We used gene expression profiles and survival characteristics (i.e., days to death for dead patients and days to last follow-up for alive patients) of patients. We downloaded 9,911 `HTSeq-FPKM` and 10,949 `Clinical Supplement` files to obtain gene expression profiles and clinical annotation data, respectively. We did not include metastatic tumors in this study since their underlying mechanisms might be significantly different than primary tumors. We included the patients who have both gene expression profile and survival information available in our analyses. Patients with `vital_status` as `Dead` (`Alive`) and `days_to_death` (`days_to_last_followup`) as non-positive or `NA` were also discarded. We only included cohorts with at least 20 patients having `vital_status` as `Dead` and at least 100 patients in total. After these filtering steps, we obtained 20 TCGA data sets including 7,655 patients in total (Figure 1b). The following 20 cancer types were included in our experiments: bladder urothelial carcinoma (`BLCA`), breast invasive carcinoma (`BRCA`), cervical squamous cell carcinoma and endocervical adenocarcinoma (`CESC`), colon adenocarcinoma (`COAD`), esophageal carcinoma (`ESCA`), glioblastoma multiforme (`GBM`), head and neck squamous cell carcinoma (`HNSC`), kidney renal clear cell carcinoma (`KIRC`), kidney renal papillary cell carcinoma(`KIRP`), acute myeloid leukemia (`LAML`), brain lower grade glioma (`LGG`), liver hepatocellular carcinoma (`LIHC`), lung adenocarcinoma (`LUAD`), lung squamous cell carcinoma (`LUSC`), ovarian serous cystadenocarcinoma (`OV`), pancreatic adenocarcinoma (`PAAD`), rectum adenocarcinoma (`READ`), sarcoma (`SARC`), stomach adenocarcinoma (`STAD`), and uterine corpus endometrial carcinoma (`UCEC`).

4.1.2. Pathway/Gene Set Databases

We used two pathway/gene set databases in addition to gene expression profiles of primary tumors to understand which biological mechanisms are predictive of overall survival times of cancer patients. We extracted gene sets in `Hallmark` collection (Liberzon et al., 2015) and pathways in `Pathway Interaction Database` (`PID`) collection (Schaefer et al., 2009) from the Molecular Signatures Database (MSigDB) (`http://software.broadinstitute.org/gsea`). These collections consist of group of genes that play joint roles in metabolism, gene regulation, and signaling in cells. `Hallmark` is a computationally constructed gene set collection including 50 gene sets with sizes between 32–200. It summarizes and represents specific well-defined biological states or processes displaying coherent expression of gene sets. `PID` is a manually curated and peer-reviewed pathway collection including 196 human signaling and regulatory pathways with sizes between 10–137.

4.2. Experimental Settings

We divided each cohort into training and test partitions by randomly picking 80% of samples as the training set and using the remaining 20% as the test set. We tried to keep the ratio between the number of patients having `vital_status` as `Dead` and the number of patients having `vital_status` as `Alive` for training and test sets as close as possible. We repeated this procedure 100 times for each cohort to get more robust performance values.

We first $\log_2$-transformed the gene expression profiles. For each data set, we then normalized the training set to zero mean and unit standard deviation, whereas we used the mean and the standard deviation of the original training data to normalize the test set. In each replication, we used 4-fold inner cross-validation on the training set to pick the hyper-parameters of Path2MSurv (i.e., regularization parameter C) and baseline algorithms (i.e., number of trees to grow, `ntree`, for survival RF; regularization parameter C for survival SVM and Path2Surv). We chose `ntree` from the set $\{500, 1000, \ldots, 2500\}$ and C from the set $\{10^{-4}, 10^{-3}, \ldots, 10^{+5}\}$.

For survival RF, we used `randomForestSRC` R package version 2.5.1 (Ishwaran & Kogalur, 2017). We implemented survival SVM, Path2Surv, and Path2MSurv in R using CPLEX version 12.7.1 (IBM, 2017) to solve quadratic optimization problems. Our implementations are publicly available at `https://github.com/mehmetgonen/path2msurv`. We used the Gaussian kernel function in kernel-based algorithms, namely, survival SVM, Path2Surv, and Path2MSurv:

$$k_{\mathcal{G}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top (\boldsymbol{x}_i - \boldsymbol{x}_j)/(2\sigma^2)\right),$$

where the kernel width parameter, i.e., σ, was set to the average pairwise Euclidean distance between training data points. We chose the Gaussian kernel function since it is more likely to better capture highly non-linear dependency between gene expression profiles and overall survival times. We calculated kernel matrices on subset of gene expression profiles which includes only corresponding genes of each pathway/gene set. We set the tube width parameter, i.e., ϵ, in kernel-based algorithms to zero. For Path2Surv and Path2MSurv, the convergence is usually observed in tens of iterations. That is why we picked the number of iterations as 200 to guarantee the convergence.

4.3. Performance Measure

We used the concordance index (C-index) to compare the predictive performances of baseline algorithms and our proposed Path2MSurv algorithm. C-index gives the ratio between the number of concordant pairs and the number of all comparable pairs. Comparable pairs consist of two `Dead` patients, or one `Dead` patient and one `Alive` patient with an observed survival time longer than the `Dead` patient. A comparable pair is called concordant if the predicted survival times can be ordered in the same way with the observed survival times. Higher C-index indicates better predictive performance. C-index can be formulated as

$$\frac{\sum_{i=1}^{N}\sum_{j\neq i}\Delta_{ij}1((y_i-y_j)(\hat{y}_i-\hat{y}_j)>0)}{\sum_{i=1}^{N}\sum_{j\neq i}\Delta_{ij}},$$

where $\hat{y}_i$ is the predicted survival time of patient i and

$$\Delta_{ij}=\begin{cases}1, & (\delta_i=0,\delta_j=0) \text{ or } (\delta_i=0,\delta_j=1,y_i<y_j),\\ 0, & \text{otherwise.}\end{cases}$$

4.4. Experimental Results

We compared four different machine learning algorithms, namely, survival RF (denoted as `RF`), survival SVM (denoted as `SVM`), single-task version of our algorithm Path2Surv (denoted as `MKL`), and our multitask MKL algorithm Path2MSurv (denoted as `MTMKL`) on 20 TCGA data sets (Figure 1b). We added `[H]` or `[P]` to the algorithm name when `Hallmark` or `PID` pathway/gene set collection was used in the corresponding algorithm.

`RF`, `SVM`, `MKL[P]`, `MTMKL[P]`, `MKL[H]`, and `MTMKL[H]` are compared in terms of their predictive performance on 20 TCGA data sets in Figure 2, where they predicted overall survival times of cancer patients from their gene expression profiles at the diagnosis time. For each cancer type, we reported C-index values over 100 replications in the corresponding box-and-whisker plots and used two-tailed paired t-tests to see whether there is a significant predictive performance difference between the algorithm pairs. `SVM`, `MKL`, and `MTMKL` were compared against `RF`. `MTMKL` was also compared against `MKL` to see the added benefit of multitask learning.

Figure 2 indicates that Path2MSurv with `PID` pathways (i.e., `MTMKL[P]`) and Path2MSurv with `Hallmark` gene sets (i.e., `MTMKL[H]`) outperformed `RF` on 13 out of 20 data sets. On the other hand, `RF` outperformed `MTMKL[P]` on `COAD` and `READ` data sets, while it outperformed `MTMKL[H]` only on `READ` data set. When compared against `RF`, most successful predictive performance for overall survival times of patients was obtained using `MTMKL`, especially on `CESC`, `GBM`, `HNSC`, `LUAD`, `LUSC`, `PAAD`, and `UCEC` data sets by improving the C-index values more than 4%. Single-task algorithms `MKL[P]` and `MKL[H]` outperformed `RF` on 10 and 13 out of 20 data sets, respectively. However, both algorithms were outperformed by `RF` on `COAD`, `LAML`, and `READ` data sets.

We observed that `MTMKL[P]` outperformed `MKL[P]` on 15 out of 20 data sets, especially by improving prediction performances on `BLCA`, `CESC`, `GBM`, `OV`, `PAAD`, `SARC`, and `STAD` data sets more than 2%. When we considered the results obtained using `Hallmark` gene sets, `MTMKL[H]` outperformed `MKL[H]` on 14 out of 20 data sets, especially by improving prediction performances on `BLCA`, `LAML`, and `UCEC` data sets more than 2%. `MKL[P]` and `MKL[H]` outperformed `MTMKL[P]` and `MTMKL[H]` simultaneously only on `BRCA` and `LUSC` data sets. These results clearly showed the benefit of multitask learning over single-task learning (i.e., modeling each cohort separately).

In our experiments, we observed that `RF` could not make satisfactory predictions on `GBM` and `LUSC`, where the corresponding median C-index values were below 0.5. `MKL[P]`, `MTMKL[P]`, and `MKL[H]` obtained median C-index values higher than 0.5 on all data sets. Only `MTMKL[H]` gave a median C-index value lower than 0.5 on `READ` data set. These results clearly showed that MKL-based algorithms `MKL` and `MTMKL` are better in capturing highly non-linear dependency between gene expression profiles and overall survival times.

Our method Path2MSurv used a shared set of kernel weights to identify informative pathways/gene sets during training for included cohorts. A pathway/gene set was considered as included into the final model if the corresponding kernel weight was greater than 0.01. For each pathway/gene set, we counted the number of replications in which that pathway/gene set was included into the final model. Figure 3 and Figure 4 show the selection frequencies of `Hallmark` gene sets and top 50 `PID` pathways used by Path2MSurv algorithm over 100 replications, respectively.

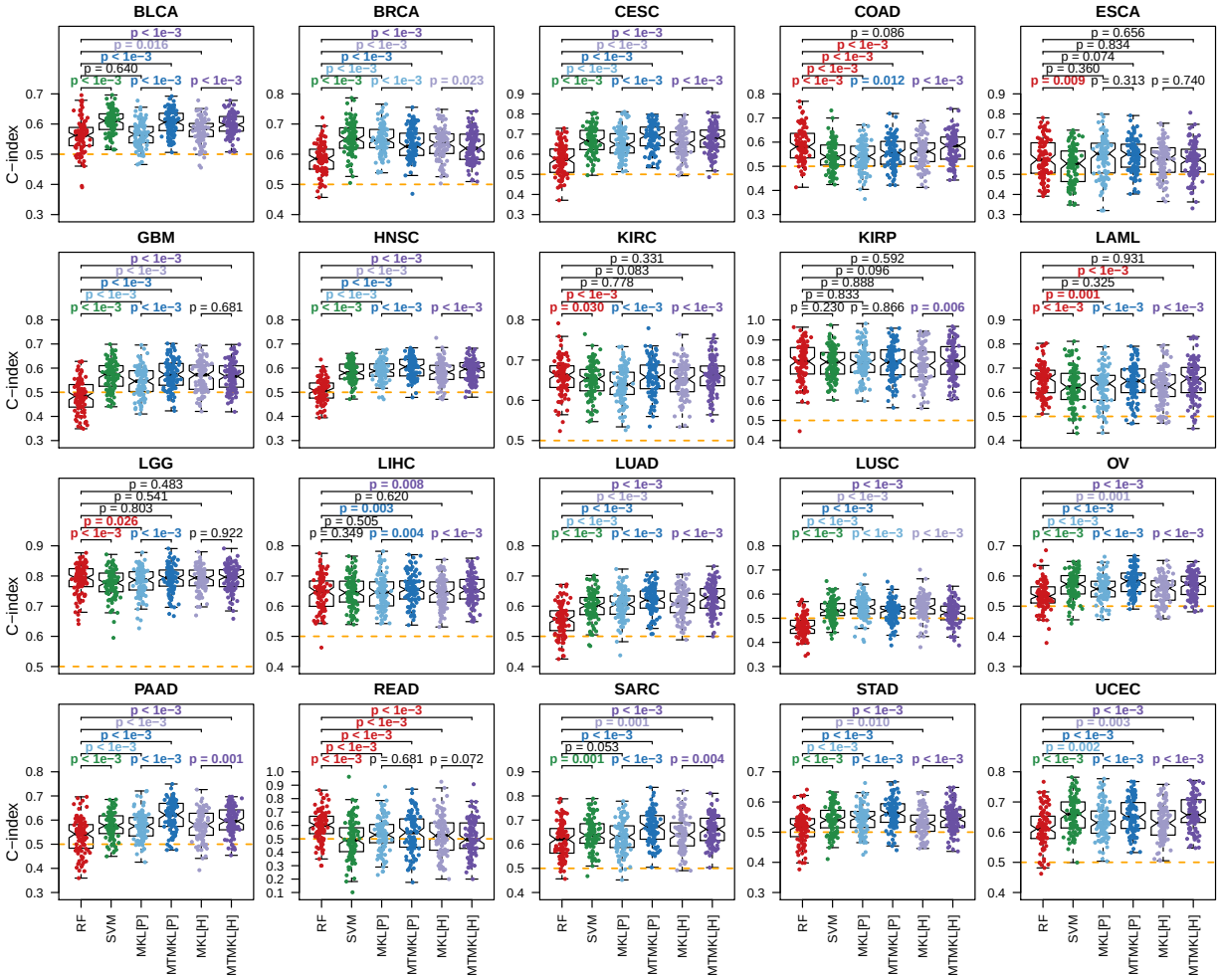

Figure 2. Predictive performances of survival RF (`RF`) algorithm, survival SVM (`SVM`) algorithm, single-task MKL algorithm Path2Surv with `PID` pathway collection (`MKL[P]`) and with `Hallmark` gene set collection (`MKL[H]`), multitask MKL algorithm Path2MSurv with `PID` pathway collection (`MTMKL[P]`) and with `Hallmark` gene set collection (`MTMKL[H]`) on 20 cancer data sets. Each box-and-whisker plot shows C-index values over 100 replications. Two-tailed paired t-tests are used to see whether there are significant differences between pairs of algorithms. For P-value results, **red**: RF is better; **green**: `SVM` is better; **light blue**: `MKL[P]` is better; **dark blue**: `MTMKL[P]` is better; **light magenta**: `MKL[H]` is better; **dark magenta**: `MTMKL[H]` is better; black: no difference. **Orange**: baseline performance level where C-index = 0.5.

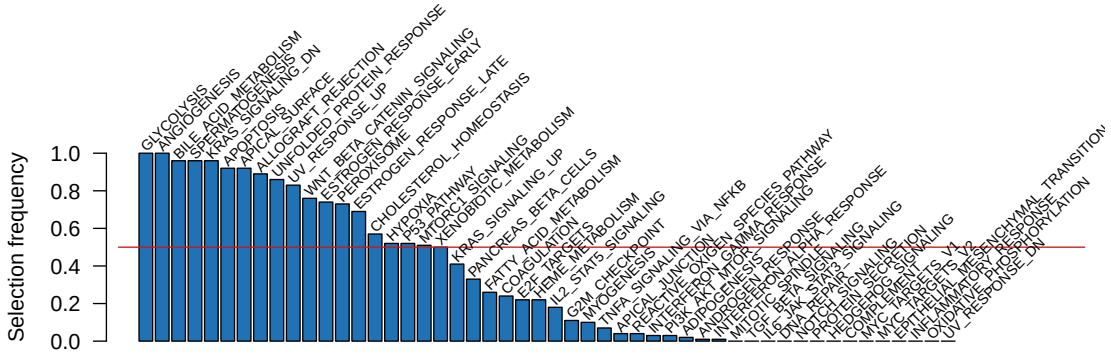

Figure 3. Selection frequencies of 50 gene sets in the `Hallmark` collection over 100 replications by Path2MSurv algorithm. The red line shows where the selection frequency is 50%.

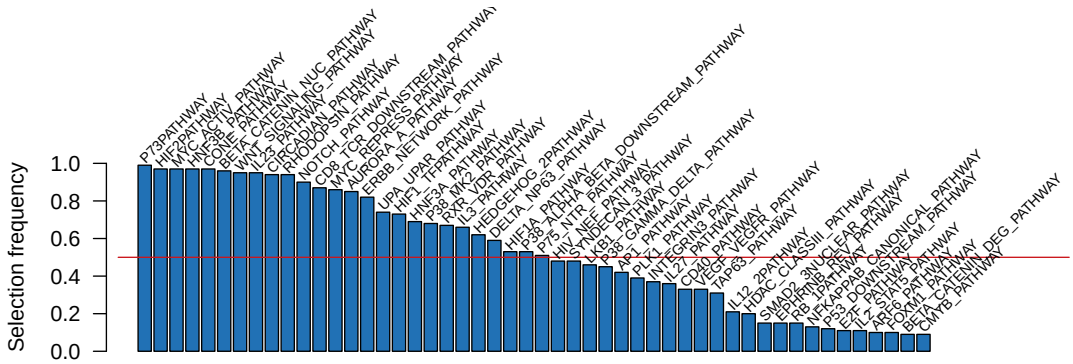

Figure 4. Selection frequencies of top 50 out of 196 pathways in the PID collection over 100 replications by Path2MSurv algorithm. The red line shows where the selection frequency is 50%.

Figure 3 shows that 19 out of 50 Hallmark gene sets were selected as informative in at least 50 replications. The most informative gene sets were GLYCOLYSIS and ANGIOGENESIS with 100% selection frequencies. These two gene sets are known to be key mechanisms cancer cells benefit from. KRAS_SIGNALING_DN, SPERMATOGENESIS, APOPTOSIS, APICAL_SURFACE and BILE_ACID_METABOLISM were selected in more than 90 replications. Figure 4 indicates that 26 out of 196 PID pathways were selected as informative in at least 50 replications. The most informative pathways were P73PATHWAY, BETA_CATENIN_NUC_PATHWAY, HIF2PATHWAY, CONE_PATHWAY, HNF3B_PATHWAY, MYC_ACTIV_PATHWAY, WNT_SIGNALING_PATHWAY, and IL23_PATHWAY, which were selected in almost all replications and were known to be key biological mechanisms in cancer.

We also observed that multitask MKL algorithms (i.e., MTMKL[P] and MTMKL[H]) used slightly more pathways/gene sets than MKL algorithms (i.e., MKL[P] and MKL[H]), hence the increased predictive performance of MTMKL can be attributed to this. MTMKL were modeling multiple patient cohorts conjointly, so it needed more pathways/gene sets than MKL to capture underlying survival mechanisms of all cohorts simultaneously. Even with this increased number of pathways/gene sets, MTMKL used significantly fewer gene expression features than RF and SVM.

5. Conclusions

Identification of biologically important mechanisms for predicting disease-related phenotypes (e.g, overall survival time) is quite important to better understand the formation and progression characteristics of the diseases. In this study, we extended survival SVM algorithm towards MKL (Gönen & Alpaydın, 2011) and multitask learning (Caruana, 1997), which is known to improve the predictive performance of machine learning algorithms when modeling related tasks.

To test our proposed Path2MSurv algorithm (Figure 1a), we used gene expression profiles of patients from 20 different cancer types provided by TCGA consortium (Figure 1b). We used two cancer-specific pathway/gene set databases, namely, Hallmark gene set collection (Liberzon et al., 2015) and PID pathway collection (Schaefer et al., 2009), to identify key biological mechanisms for survival.

We reported predictive performance of our Path2MSurv algorithm and compared its performance against survival RF (Ishwaran et al., 2008), survival SVM (Shivaswamy et al., 2007; Khan & Zubek, 2008), and single-task variant of our algorithm (Figure 2). Path2MSurv algorithm obtained the best predictive performance on most of the data sets using significantly fewer gene expression features than survival RF and survival SVM algorithms.

We envision extending our work towards task clustering in the future. In this study, we trained a shared MKL model on all data sets, which makes sense if all of the tasks are related. If we have disease groups with different underlying biological mechanisms, forcing all tasks to use the same pathways/gene sets for prediction might not be meaningful. We will extend Path2MSurv algorithm to conjointly perform the following three steps: (i) clustering of data sets, (ii) learning shared kernel weights for each cluster, and (iii) learning a survival analysis model for each data set.

Acknowledgments

This work was supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK) under Grant EEEAG 117E181. Onur Dereli was supported by the Ph.D. scholarship (2211) from TÜBİTAK. Mehmet Gönen was supported by the Turkish Academy of Sciences (TÜBA-GEBİP; The Young Scientist Award Program) and the Science Academy of Turkey (BAGEP; The Young Scientist Award Program). Computational experiments were performed on the OHSU Exacloud high performance computing cluster.

References

Anaya, J., Reon, B., Chen, W.-M., Bekiranov, S., and Dutta, A. A pan-cancer analysis of prognostic genes. *PeerJ*, 3: e1499, 2016.

Bakker, B., Heskes, T., Neijt, J., and Kappen, B. Improving Cox survival analysis with a neural-Bayesian approach. *Stat. Med.*, 23:2989–3012, 2004.

Breiman, L. Random forests. *Mach. Learn.*, 45:5–32, 2001.

Caruana, R. Multitask learning. *Mach. Learn.*, 28:41–75, 1997.

Choi, W., Porten, S., Kim, S., Willis, D., Plimack, E., et al. Identification of distinct basal and luminal subtypes of muscle-invasive bladder cancer with different sensitivities to frontline chemotherapy. *Cancer Cell*, 25:152–165, 2014.

Cortes, C. and Vapnik, V. Support-vector networks. *Mach. Learn.*, 20:273–297, 1995.

Costello, J. C., Heiser, L. M., Georgii, E., Gönen, M., Menden, M. P., et al. A community effort to assess and improve drug sensitivity prediction algorithms. *Nat. Biotechnol.*, 32:1202–1212, 2014.

Cox, D. R. Regression models and life-tables. *J. R. Stat. Soc. Ser. B-Stat. Methodol.*, 34:187–220, 1972.

Cox, D. R. and Oakes, D. *Analysis of Survival Data*. Chapman and Hall, London, 1984.

Damrauer, J., Hoadley, K., Chism, D., Fan, C., Tiganelli, C., et al. Intrinsic subtypes of high-grade bladder cancer reflect the hallmarks of breast cancer biology. *Proc. Natl. Acad. Sci. U. S. A.*, 111:3110–3115, 2014.

Dereli, O., Oğuz, C., and Gönen, M. Path2Surv: Pathway/gene set-based survival analysis using multiple kernel learning. *Bioinformatics*, in press.

Evers, L. and Messow, C. M. Sparse kernel methods for high-dimensional survival data. *Bioinformatics*, 24:1632–1638, 2008.

Gönen, M. and Alpaydın, E. Multiple kernel learning algorithms. *J. Mach. Learn. Res.*, 12:2211–2268, 2011.

Gönen, M., Weir, B. A., Cowley, G. S., Vazquez, F., Guan, Y. F., et al. A community challenge for inferring genetic predictors of gene essentialities through analysis of a functional screen of cancer cell lines. *Cell Syst.*, 5:485–497, 2017.

Hoadley, K. A., Yau, C., Wolf, D. M., Cherniack, A. D., Tamborero, D., et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*, 158:929–944, 2014.

IBM. *ILOG CPLEX Interactive Optimizer. Version 12.7.1.0*, 2017.

Ishwaran, H. and Kogalur, U. B. *randomForestSRC: Random Forests for Survival, Regression, and Classification (RF-SRC) R package version 2.5.1*, 2017.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. Random survival forests. *Ann. Appl. Stat.*, 2:841–860, 2008.

Khan, F. M. and Zubek, V. B. Support vector regression for censored data (SVRc): A novel tool for survival analysis. In *Proc. 8th IEEE ICDM*, 2008.

Khirade, M. F., Lal, G., and Bapat, S. A. Derivation of a fifteen gene prognostic panel for six cancers. *Sci. Rep.*, 5: 13248, 2015.

Kiaee, F., Sheikhzadeh, H., and Mahabadi, S. E. Relevance vector machine for survival analysis. *IEEE Trans. Neural Netw. Learn. Syst.*, 27:648–660, 2016.

Lawrence, M., Stojanov, P., Mermel, C., Robinson, J., Garraway, L., et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, 505:495–501, 2014.

Li, Y., Wang, J., Ye, J., and Reddy, C. K. A multi-task learning formulation for survival analysis. In *Proc. 22nd ACM KDD*, 2016.

Liberzon, A., Birger, C., Thorvaldsdottir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.*, 1:417–425, 2015.

Mogensen, U. B. and Gerds, T. A. A random forest approach for competing risks based on pseudo-values. *Stat. Med.*, 32:3102–3114, 2013.

Pang, H., Datta, D., and Zhao, H. Pathway analysis using random forests with bivariate node-split for survival outcomes. *Bioinformatics*, 26:250–258, 2010.

Pang, H., Hauser, M., and Minvielle, S. Pathway-based identification of SNPs predictive of survival. *Eur. J. Hum. Genet.*, 19:704–709, 2011.

Pang, H., George, S. L., Hui, K., and Tong, T. Gene selection using iterative feature elimination random forests for survival outcomes. *IEEE-ACM Trans. Comput. Biol. Bioinform.*, 9:1422–1431, 2012.

Pappa, K. I., Polyzos, A., Jacob-Hirsch, J., Amariglio, N., Vlachos, G. D., et al. Profiling of discrete gynecological cancers reveals novel transcriptional modules and common features shared by other cancer types and embryonic stem cells. *PLoS One*, 10:e0142229, 2015.

Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., et al. PID: The Pathway Interaction Database. *Nucleic Acids Res.*, 37:D674–D679, 2009.

Schölkopf, B. and Smola, A. J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, 2002.

Shivaswamy, P. K., Chu, W., and Jansche, M. A support vector approach to censored targets. In *Proc. 7th IEEE ICDM*, 2007.

The Cancer Genome Atlas Research Network, Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genet.*, 45:1113–1120, 2013.

Van Belle, V., Pelckmans, K., Van Huffel, S., and Suykens, J. A. Support vector methods for survival analysis: A comparison between ranking and regression approaches. *Artif. Intell. Med.*, 53:107–118, 2011a.

Van Belle, V., Pelckmans, K., Van Huffel, S., and Suykens, J. A. Improved performance on high-dimensional survival data by application of Survival-SVM. *Bioinformatics*, 27: 87–94, 2011b.

Wan, Q., Dingerdissen, H., Fan, Y., Gulzar, N., Pan, Y., et al. BioXpress: An integrated RNA-seq-derived gene expression database for pan-cancer analysis. *Database*, 2015.

Wang, L., Li, Y., Zhou, J., Zhu, D., and Ye, J. Multi-task survival analysis. In *Proc. 17th IEEE ICDM*, 2017.

Wang, Y., Chen, T., and Zeng, D. Support vector hazards machine: A counting process framework for learning risk scores for censored outcomes. *J. Mach. Learn. Res.*, 17: 1–37, 2016.

Xu, Z., Jin, R., Yang, H., King, I., and Lyu, M. Simple and efficient multiple kernel learning by group Lasso. In *Proc. 27th ICML*, 2010.

Yang, Y., Han, L., Yuan, Y., Li, J., Hei, N., et al. Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nat. Commun.*, 5:3231, 2014.

Yousefi, S., Amrollahi, F., Amgad, M., Dong, C., Lewis, J. E., et al. Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. *Sci. Rep.*, 7:11707, 2017.

Yuan, Y., Van Allen, E. M., Omberg, L., Wagle, N., Amin-Mansour, A., et al. Assessing the clinical utility of cancer genomic and proteomic data across tumor types. *Nat. Biotechnol.*, 32:644–652, 2014.

Zhang, X., Li, Y., Akinyemiju, T., Ojesina, A. I., Buckhaults, P., et al. Pathway-structured predictive model for cancer survival prediction: A two-stage approach. *Genetics*, 205: 89–100, 2017.