

WSISA: Making Survival Prediction from Whole Slide Histopathological Images

Xinliang Zhu, Jiawen Yao, Feiyun Zhu, and Junzhou Huang*
Univeristy of Texas at Arlington
Tencent AI Lab

Abstract

Image-based precision medicine techniques can be used to better treat cancer patients. However, the gigapixel resolution of Whole Slide Histopathological Images (WSIs) makes traditional survival models computationally impossible. These models usually adopt manually labeled discriminative patches from region of interests (ROIs) and are unable to directly learn discriminative patches from WSIs. We argue that only a small set of patches cannot fully represent the patients' survival status due to the heterogeneity of tumor. Another challenge is that survival prediction usually comes with insufficient training patient samples. In this paper, we propose an effective Whole Slide Histopathological Images Survival Analysis framework (WSISA) to overcome above challenges. To exploit survival-discriminative patterns from WSIs, we first extract hundreds of patches from each WSI by adaptive sampling and then group these images into different clusters. Then we propose to train an aggregation model to make patient-level predictions based on cluster-level Deep Convolutional Survival (DeepConvSurv) prediction results. Different from existing state-of-the-arts image-based survival models which extract features using some patches from small regions of WSIs, the proposed framework can efficiently exploit and utilize all discriminative patterns in WSIs to predict patients' survival status. To the best of our knowledge, this has not been shown before. We apply our method to the survival predictions of glioma and non-small-cell lung cancer using three datasets. Results demonstrate the proposed framework can significantly improve the prediction performance compared with the existing state-of-the-arts survival methods.

1. Introduction

Recently, image-based precision medicine has become a very active field in healthcare research. Those image data include pathological images, CT images, MRIs, etc. cap-

tured from patients. The long-term goal in this research is to be able to improve the treatment quality of an individual based on his or her image data. For instance, how a new patient will survive in the context of known patients' image data. The employment of survival prediction would allow clinicians to make early decisions on treatments which is very crucial for patients' healthcare.

One of the most challenging problems in image-based precision medicine research is that a pathological image is of high resolution (*e.g.* measuring 1 million by 1 million pixels), compared to one regular image from ImageNet that usually measure less than 1000 pixels by 1000 pixels [18]. Currently, many approaches have been proposed to incorporate with patients' imaging data for survival analysis and present more accurate predictions [25, 27, 28, 31, 33]. Due to the computation issue, the existing methods cannot directly learn discriminative patterns from WSIs and predict patients' survival status using hand-crafted features extracted from manually labeled small discriminative patches. Apparently, the process to manually select representative patches is very labor-intensive. Moreover, one patient usually has multiple WSIs from different parts of the tissue. Because solid tumors may have a mixture of tissue architectures and structures, WSIs from the same patient actually are very heterogeneous. Therefore, only a small set of patches from manual annotations cannot completely and properly reflect the patient's tumor morphology and has a very high risk of losing survival-discriminative patterns. Fig. 1 presents an example, where both patient 1 and 2 have stage IV tumor but patient 2 has worse clinical outcome than patient 1 according to the clinical trial report. It shows that even in the same tumor stage, patients with different survival status may have very distinct visual appearances. It requires a framework that can consider large heterogeneity patterns from WSIs.

In classifying cancer subtype with WSIs, one pioneering work [8] was proposed to use a patch-level convolutional neural network (CNN) and train a decision fusion model as a two-level model for tumor classification. However, different from tumor classification, survival prediction is more challenging because the following reasons. First, survival

*Corresponding author: Dr. Junzhou Huang. Email: jzhuang@uta.edu. This work was partially supported by NSF IIS-1423056, CMMI-1434401, CNS-1405985 and the NSF CAREER grant IIS-1553687.

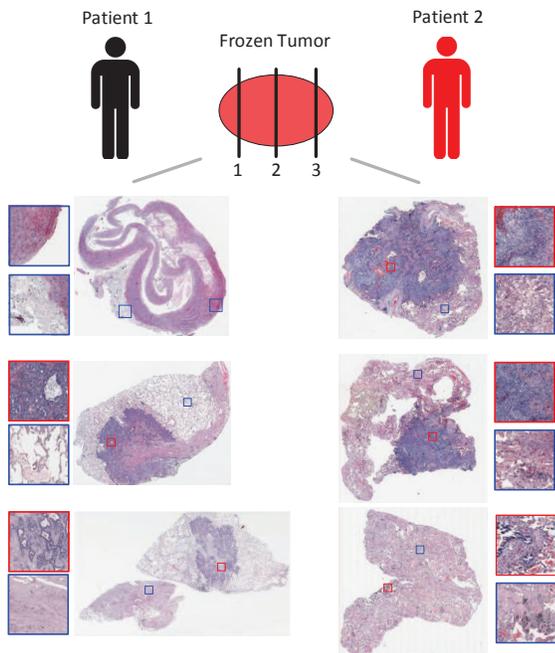


Figure 1. Gigapixel Whole Slide Histopathological Images of two lung cancer patients (best viewed in color). WSIs can present tumor in high resolution details. In this case, patches shown in red are discriminate since they show typical visual features of lung tumor. Patches in blue are non-discriminative because they only contain visual features from lower grade tumors or non-tumor tissue regions. Patient 2 has worse clinical outcome than patient 1.

prediction is a regression problem where the ranking of patients' prediction values matters [20]. While in the tumor classification task, the prediction result of one patient's category is independent from others. Second, the information provided by different WSIs from one patient should be aggregated in survival analysis. However, in [8] the aggregation among WSIs from the same patient is not solved; Third, the ground truth label (survival time and censored status) for survival analysis is only given on patient. WSI-level and patch-level ground truth label is unknown. This complicates the survival prediction problem when training CNN due to data inefficiency.

Recent years, CNNs have seen many successful applications in computer vision [13, 6]. However, training a CNN model usually demands large volume of samples. In medical research area, the number of training samples is often very small. This makes training a CNN based survival model very challenging. Zhu *et al.* [31] proposed a deep convolutional survival model (DeepConvSurv) model to partly solve the problem by augmenting image patches. Multiple patches from ROIs were extracted and assigned with patients' labels. However, DeepConvSurv can only make patch-wise predictions. Thus, it is needed a novel method that can make patient-wise prediction from small

sample size datasets.

In this paper, we propose a Whole Slide Histopathological Images Survival Analysis (WSISA) framework to predict patients' clinical outcomes. Instead of extracting patches only from region of interests (ROIs), we adopt adaptive sampling strategy to generate numerous candidate patches from the WSIs by keeping the number of patches proportional to the WSI's size. Since not all candidate patches are survival-discriminative, we cluster the candidate patterns by applying K-Means based on their phenotype features. To find important patches clusters, we train several deep convolutional survival (DeepConSurv) models [31]. Key clusters are then distinguished by selecting models with good survival prediction performances. Tumor might have mixed patterns and one kind of distinct pattern cannot provide satisfactory prediction power. To further improve the performance using cluster-level survival models, we aggregate the selected clusters by both applying fully-connected neural network and boosting Cox's negative log likelihood.

Different from the state-of-the-arts images-based survival approaches which need manual annotations, the proposed framework can directly learn multiple patterns from WSIs and achieve end-to-end analysis. As we have discussed before, another challenge for survival prediction is the data inefficiency. For example, the world's largest datasets on lung cancer only contain less than 500 patients records, deep survival methods on such small scale dataset may achieve low performance. We demonstrate that the proposed framework can easily handle this problem. To evaluate the developed framework, we conduct extensive experiments on three different datasets and two types of diseases. The main contributions of this paper can be summarized as: 1) To best of our knowledge, we for the first time develop an end-to-end manner to predict survival on WSIs. 2) The proposed method can solve the small sample dataset problem on training deep convolutional survival network. 3) Extensive experiments are conducted to evaluate the effectiveness of the developed framework on two cancer data using three large dataset. 4) Compare to information derived from molecular profiling to classify tumors and help to make clinical decisions [19, 30], gigapixel whole slide histopathological images can better present tumor growth and morphology with very clear details which can greatly benefit cancer diagnosis. However, due to computations issues, state-of-the-art survival methods cannot built models directly from WSIs, and the proposed framework can bridge this gap to be fully applied in personalized medicine.

2. Background

The goal of survival analysis is to predict the time duration until an event occurs and the event of interest is the death of a cancer patient in our study. In survival analysis,

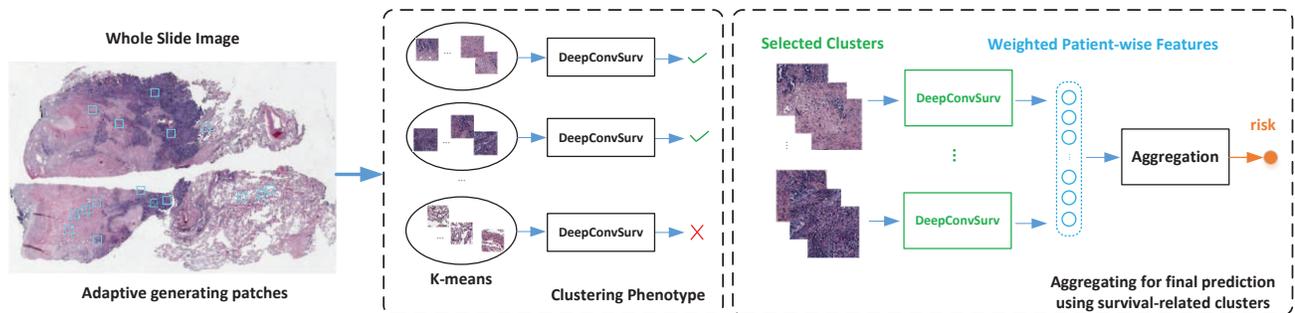


Figure 2. An overview of our WSISA framework (best viewed in color). It consists of four main stages: 1) adaptively generating patches from the WSIs; 2) clustering patch candidates according to their phenotypes; 3) selecting clusters based on patch-wise survival prediction performance; 4) aggregating the selected clusters to make final prediction.

the observation of one patient is either a survival time (O_i) or a censored time (C_i). If and only if $t_i = \min(O_i, C_i)$ can be observed during the study, the dataset is right-censored [17]. An instance in the survival data is usually represented as (x_i, t_i, δ_i) where x_i is the feature vector, t_i is the observed time, δ_i is the indicator which is 1 for a uncensored instance (death occurs during the study) and 0 for a censored instance.

The survival function $S(t|x) = Pr(O \geq t|x)$ is used to identify the probability of being still alive at time t where $x = (x_1, \dots, x_p)^T$ is the covariates of dimension p . The hazard function is defined as

$$h(t|x) = \lim_{\Delta t \rightarrow 0} \frac{Pr(t \leq O \leq t + \Delta t | O \geq t; x)}{\Delta t}, \quad (1)$$

which assesses the instantaneous rate of death at time t . In the modeling methods, Cox proportional hazard model is among the most popular one. It models the hazards as $h(t|x) = h_0(t) \exp(\beta^T x)$ where $\beta = (\beta_1, \dots, \beta_p)^T$ is a vector of regression parameters, and $h_0(t)$ is the baseline hazard. $f(x) = \beta^T x$ is also being called as risk function. To estimate the parameters β , we can minimize the negative log partial likelihood, which is

$$l(\beta) = - \sum_{i=1}^n \delta_i \left(\beta^T x_i - \log \sum_{j \in R(t_i)} \exp(\beta^T x_j) \right). \quad (2)$$

where $R(t_i)$ is the risk set at time t_i , which is the set of all individuals who are still under study before time t_i . Clinicians and researchers at first applied the Cox model to test for significant risk factors (e.g. gender, age) affecting survival. Then they focused on molecular profiling data to predict more and accurate survival outcomes. In order to handle with high-dimensional molecular data, feature selection methods have been adapted to Cox regression setting for censored survival data [23, 2, 1, 24, 16, 3]. Besides Cox model, two recent work MTLA [14] and Deep-

Surv [12] are proposed to model more complex relationships between covariates and survival outcomes. MTLA transforms the original survival analysis into a multi-task learning problem instead of defining the hazard function by decomposing the regression component into related classification tasks. However, the number of tasks corresponds to the maximum follow-up time of all the instances. In fact, recent cancer datasets (e.g. TCGA) are collecting patient electronic health records (EHR) with a very long follow-up time. Therefore, the multi-task learning will encounter computation issues if the follow-up time is large as it needs to learn a shared representation across all tasks at different time intervals. Katzman *et al.* proposed a deep fully connected network (DeepSurv) to represent the nonlinear risk function [12]. They demonstrated that DeepSurv outperformed the standard linear Cox proportional hazard model but the architecture of DeepSurv is simple and shallow to handle complex patients' imaging data.

As pointed out in [29], tumor microenvironment is a complex milieu that includes not only the cancer cells but also the stromal cells and immune cells. All this "extra" genomic information may muddle results and therefore make molecular analysis a challenging task for cancer prognosis. Recent work [25, 27, 28] have attempted to use imaging data for survival analysis and achieve better predictions for breast and lung cancer patients. However, as we discussed before, state-of-the-arts image-based survival predictions focused on small patches and cannot exploit whole slide histopathological images (WSIs) directly and efficiently.

3. Methodology

Existing methods for survival analysis are focused on local information extracted from patches. Few work has been proposed to process the WSIs. We are for the first time developing a new effective framework for making survival prediction from WSIs. We will show that the proposed framework can catch the general information represented by

the WSIs.

3.1. The Framework

Fig.2 illustrates the pipeline of the proposed framework. It consists of four main stages: 1) adaptively generating patches from the WSIs; 2) clustering patch candidates according to their phenotypes; 3) selecting clusters based on patch-wise survival prediction performance; 4) aggregating the selected clusters to make final prediction.

3.1.1 Sampling from WSIs

The goal in this stage is to generate candidate patches from WSIs. Unlike extracting patches only from annotated ROIs, we argue that the heterogeneous patterns and their proportions in each WSI also count. We assume that those candidate patches randomly sampled from patients WSIs can catch the main patterns and their proportions. And candidate patches from different WSIs of one patient together reflect the patient's survival risk. We set a fixed area sampling ratio to sample the candidate patches (patch size of 512 by 512, 0.5 microns per pixel), making sure a fixed proportion of pixels are sampled from each WSI. This step and the assumptions are the basis of following steps.

3.1.2 Clustering on Phenotypes

The candidate patches from the first stage are heterogeneous. Some of them may be extracted from tumor sections, some of others may be generated from normal tissue section and others may contain both. A simple but effective way to distinguish them is by clustering based on their phenotypes. Due to the computation concerns, we generate much smaller size (50×50) thumbnail images to represent their phenotypes. The features are from concatenating row pixels of the thumbnail images, thus there're 2,500 dimensions of features. Even that, the features are relatively high for clustering. We employ PCA to reduce the dimension to 50 prior to the K-means clustering process. This step will cluster different patches into several distinguished phenotype groups.

3.1.3 Selecting Clusters

As pointed above, different candidate clusters contain distinguished pattern patches. Those patches may have various prediction power on patients' survival. To distinguish their predicting power and select the candidate clusters, we train separate deep convolutional survival model (DeepConvSurv) on each cluster. The DeepConvSurv models are trained patch-wise. That is, we assign each patch with the survival label from that patient. Then we train the model with all the patches in the cluster. We select the clusters with predicting accuracy a little bit better than random guess as

Table 1. The architecture of DeepConvSurv

Layer	Filter size, stride, number
Conv (ReLU)	7×7 , 3, 32
Max-pooling	2×2
Conv (ReLU)	5×5 , 2, 32
Conv (ReLU)	3×3 , 2, 32
Max-pooling	2×2
FC	32

the base learners in the aggregation step. The architecture of DeepConvSurv is shown in Table 1. The difference between DeepConvSurv and traditional deep models in classification or regression relies at the loss function. The loss function of DeepConvSurv is replaced with Cox model's loss function (2). The corresponding models of the selected clusters become features generator in the aggregation stage.

3.1.4 Aggregation

Aggregation is the key stage of the framework that outputs patient-wise survival prediction values. This stage can be further divided into two main sub-steps: Generate weighted features and Aggregation.

Generating Weighted Features This substep distinguishes our framework with traditional patch-based survival prediction methods to a large extent. We take the proportions of various heterogeneous patterns into consideration. And we solve the problem of various numbers and sizes of WSIs among different patients by unifying the contributions from those WSIs. Here, we show how various patches' contributions from one patient are unified. From the first stage, patches are extracted based on a fixed area sampling ratio. When keep the patch size fixed, the number of patches extracted are proportional to the WSI's size for one WSI. The patient's total number of patches is the sum of all the patches extracted from his/her WSIs. To estimate the weight of separate pattern, we need to count the number of patches in each cluster. Suppose there're total n_i patches extracted from patient i . For each selected cluster, the patient has n_{ij} patches in cluster j . Then the contribution of cluster j to the patients survival prediction can be calculated as:

$$w_{ij} = \frac{n_{ij}}{n_i}, \quad i \in \{1, \dots, N\}, j \in \{1, \dots, J\}, \quad (3)$$

where N is the number of patients, J is the number of selected clusters and w_{ij} is the weight for cluster j in patient i . Since each patient may have different numbers of patches in a selected cluster, the features for that patient in the selected

cluster are calculated as

$$\mathbf{x}_{ij} = w_{ij} \sum_{k=1}^{k=K} \mathbf{x}_{ijk} / K, \quad (4)$$

where \mathbf{x}_{ij} is the output features in cluster j for patient i . It can be either the prediction risks from each cluster or the output of FC layer in the DeepConvSurv. K is the number of patches for patient i in cluster j . By randomly sampling and setting a large enough sampling ratio, the weights for survival correlated patches can be estimated well for a patient.

Aggregation After generating patient-wise weighted features, the last step is aggregating those features to make a final survival prediction. As pointed above, separate patches lack ability of representing patients’ holistic information. It is needed to integrate them to better predict patients’ survival. In this problem, from extensive experiments on three different cancer datasets, the simple Cox model with Lasso [23] can predict survivals very well based on the weighted features. The reasons are: 1) because the sample size is relatively small, simple model will not easily be overfitted; 2) if the features are highly related to the survival labels, simple model will work well. In WSISA, the prediction model can also be easily changed to other state-of-the-arts models, such as random survival forests [9]. The algorithm for WSISA is shown in algorithm 1. It shows the general procedure of WSISA and will not include the details like splitting the training, validation and testing sets.

3.2. Solving the Small Sample Data Problem

The proposed WSISA solves the problem of small sample data in survival prediction with WSIs by splitting the task into two parts: the estimation of patch-wise risks and the aggregation of patch-wise risks into patients’ risks. In training the separate patch-wise DeepConvSurv models, we have relatively large samples. The input of DeepConvSurv is $512 \times 512 \times 3$. After getting the features from selected clusters, the aggregation task becomes simple. Thus the small sample data problem in this task is solved appropriately.

4. Experiments

In this section, we will first describe the datasets used in our experiments and then demonstrate the performances of different methods.

4.1. Dataset Description

We focus on lung and brain cancer in our study and used three public cancer survival datasets with high resolution whole slide pathological images, The National Lung Screening Trial (NLST) [21] and The Cancer Genome Atlas

Algorithm 1 WSISA Algorithm

Input: WSIs, time t , status δ , sample ratio r and patch size p

- 1: /*The 1st stage: **Sampling patches***/
- 2: **for all** WSIs **do**
- 3: $num_patches = \frac{WSIsize \times r}{p}$
- 4: **end for**
- 5: /*The 2nd stage: **Clustering***/
- 6: **for all** patches in training set **do**
- 7: $features = createThumbnail(patch, size)$
- 8: $pca = PCA(train_features)$
- 9: $clusters = kMeans(pca, numclusters)$
- 10: **end for**
- 11: /*The 3rd stage: **Selecting clusters***/
- 12: **for all** c in clusters **do**
- 13: $model = trainDeepConvSurv(c_features, t, \delta)$
- 14: $valid_accuracy = evaluate(valid_c_features)$
- 15: **end for**
- 16: $selected_clusters = selectCluster(valid_accuracy)$
- 17: /*The 4th stage: **Aggregating***/
- 18: $patient_features = weightedFeatures(sc_features)$
- 19: $aggre_model = trainAggre(patient_features, t, \delta)$

Output: The patients’ risk

Table 2. The numbers of WSIs, patients, patches originally extracted and filtered in each dataset. Because some patches are extracted from background parts (they are mostly white) in WSIs, we filter out the valid patches which are non-white.

Dataset	NLST	TCGA-LUSC	TCGA-GBM
#patients	404	121	126
#WSIs	1104	485	255
#patches	67834	70738	60623
#valid patches	41303	24387	27551

(TCGA). TCGA project [11] can provide large-scale molecular profiling data and pathological images for each patient. NLST is a very large lung cancer dataset collected by the National Cancer Institute’s Division of Cancer Prevention (DCP) and Division of Cancer Treatment and Diagnosis (DCTD).

We conducted experiments on two cancer subtypes of brain and lung cancer in TCGA: glioblastoma multiforme (GBM) and lung squamous cell carcinoma (LUSC). We adopted a core sample set from UT MD Anderson Cancer Center [30] in which each sample has information for the overall survival time, pathological images and molecular data related to gene expression. The numbers of WSIs and patients in each dataset are shown in Table.2.

4.2. Comparison methods and Evaluation Metric

To analyze pathological images in comparison survival models, we have annotations that locate the tumor regions in whole slide images (WSIs) with the help of pathologists. We calculated hand-crafted features using CellProfiler [4] which serves as a state-of-the-art medical image feature extracting and quantitative analysis tool. Motivated by recent work [28, 32], a total of 1,795 quantitative features were calculated from each image tile. These types of image features include cell shape, size, texture of the cells and nuclei, as well as the distribution of pixel intensity in the cells and nuclei.

We compare our framework with seven popular state-of-the-art survival models. They are classified into five categories:

- **Regularized Cox models:** The Cox proportional hazards model [5] is the most commonly used semi-parametric model in survival analysis. The l_1 -norm (LASSO-Cox) [23] and elastic-net penalized Cox (EN-Cox) models [26] are used in this paper.
- **Parametric censored regression models:** This type of survival models formulates the joint probability of the uncensored and censored instances as a product of death density function and survival functions, respectively. The likelihood function can be defined by combining these two components [10]. We choose Weibull, Logistic distribution to approximate the survival data.
- **Random survival forests:** Random survival forests (RSF) improves the survival prediction performance by ensembling base learning trees [9].
- **Boosting concordance index (BoostCI):** It is an approach where the concordance index metric is modified to an equivalent smoothed criterion using the sigmoid function [15].
- **Multi-task learning models:** We compared our method with recent "Multi-Task Learning model for Survival Analysis" (MTLSA) [14] which reformulates the survival model into a multi-task learning problem.

WSISA is the proposed framework that can make survival prediction from patients' all WSIs. Since the choices of survival models in aggregation stage are multiple, we try all the above state-of-the-art methods to fully compare our framework with the traditional ways.

To evaluate the performances in survival prediction, we take the concordance index (C-index) as our evaluation metric [7]. The C-index quantifies the ranking quality of rank-

ings and is calculated as follows:

$$c = \frac{1}{n} \sum_{i \in \{1 \dots N | \delta_i = 1\}} \sum_{s_j > s_i} I[X_i \hat{\beta} > X_j \hat{\beta}] \quad (5)$$

where n is the number of comparable pairs, $I[\cdot]$ is the indicator function and s_i is the actual observation. The value of C-index ranges from 0 to 1. The larger CI value means the better prediction performance of the model and *vice versa*. 0 is the worst condition, 1 is the best and 0.5 is the value as a random guess.

4.3. Implementation details

The source codes of MTLISA are downloaded from the authors' website¹. All other methods in our comparisons were implemented in R. LASSO-Cox and EN-Cox are built using the *cocktail* function from the *fastcox* package [26]. RSF is from the *randomForestSRC* package [9]. The implementation of BoostCI can be found in the supplementary materials of [15]. The parametric censored regression models are from the *survival* package [22].

We set 80% of the patients as training set and the left 20% as testing set. From the training set, we split 25% of them as the validation set. All the sets are split stratified by the ratio of censored data.

4.4. Results and Discussions

4.4.1 Sampling Patches

We extract patches of size 512×512 from WSIs. To capture detailed information of the images, those patches are extracted from 20X (0.5 microns per pixel) objective magnifications. This step generates numerous heterogeneous patches from one patient. Among them some are survival related and some are not. Even some of them are the background patches (mostly white). Table 2 gives the statistics on the extracted patches from three datasets. In TCGA-LUSC and TCGA-GBM datasets, over 54% of the original patches are background patches. In NLST dataset, there're about 38% are background patches. The background patches are easily filtered out according to the variance of pixel values.

4.4.2 Clustering and Selecting Clusters

During the experiments, we cluster patches into 10 groups in each dataset. Part of the clustering results can be found in Fig. 3. From Fig. 3, we can see the heterogeneity among each cluster. Their patterns are different. For the cluster selection, we set the threshold of patch-wise prediction C-index as 0.5. That is if the model predicts better than random guess, it is selected. The patches surrounded by red

¹<https://github.com/yanlirock/MTLSA>

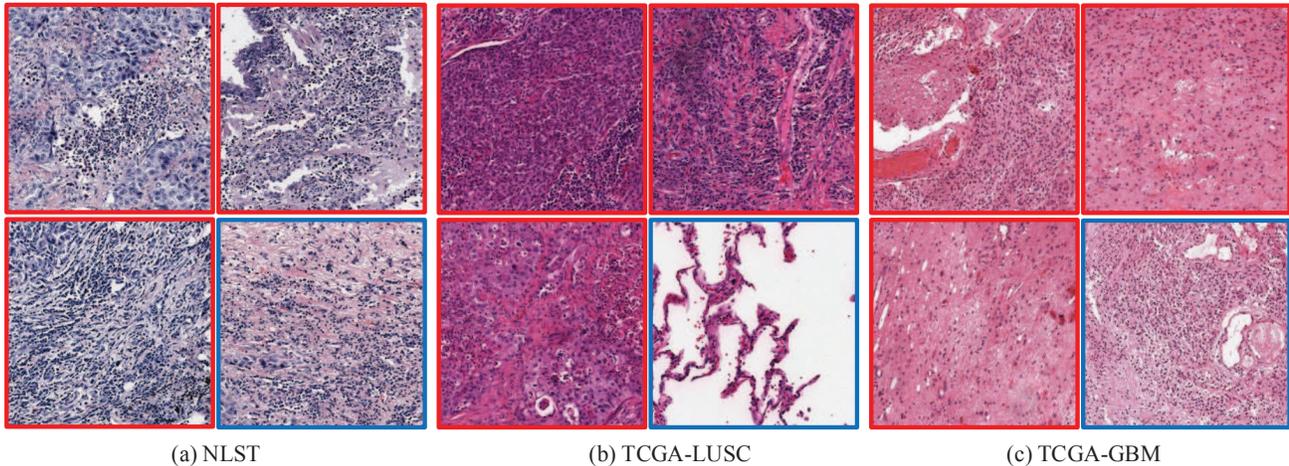


Figure 3. Some sample patches in different clusters from three datasets. Patches in each dataset are from the same patient. Sample patches from selected clusters are shown in red and patches in blue mean they are from non-survival related clusters.

lines are the effective patterns selected and by the blue lines are ineffective patterns. The results also show that phenotype based cluster method is effective to distinguish survival related patterns.

4.4.3 Predicting Survival

As pointed out before, the features used to train aggregation model in WSISA can either be the output risk from each DeepConvSurv or the FC layer (the layer before output layer) values in each DeepConvSurv. In this paper, we choose FC layer values for NLST and output risks for both TCGA-LUSC and TCGA-GBM. It is based on the sample sizes of the datasets. There are two benefits to aggregate the risks: 1) if the sample sizes in datasets with histopathological images are much smaller, high dimension features with small training sample will not fit a good prediction model. The dimension of output risks equals to the number of selected clusters, which is much smaller than the training data size; 2) the output weighted risks of one cluster can partly estimate the patients' survival. Thus it can be used as features in aggregation stage due to the high level survival information it may contain. However, due to NLST has relatively larger number of samples compared to TCGA datasets, we feed the features from FC layer to train the aggregation model.

Table 3 presents the C-index values by various survival regression methods on three datasets. The C-index value serves as a standard evaluation metric in survival analysis [20]. It shows the prediction power of different survival models. From Table 3, two groups of experiments are conducted on two different types of features (ROI-based features and WSISA holistic features). In ROI-based experiments, the patients' features are extracted from ROI patches. If one patient has more than one ROI in his/her

WSIs, we sample one patch from each ROI and then average the features extracted from them. The dimension of those features is 1,795. WSISA provides a way to represent the holistic information of one patient. The features in WSISA are concatenated from selected clusters by weighting them first.

From Table 3, the best performance is achieved by WSISA based methods in each dataset. For NLST and TCGA-LUSC datasets, the best results are achieved by simple Cox based models. The improvement in NLST by WSISA can even be over 15% compared to the best performance achieved by ROI-based methods. The reason why simple Cox based models work is due to the good representative ability of features from WSISA. ROI based models perform generally not well due to three reasons: 1) the limitation of local information provided by the patches extracted from the ROI; 2) the non-effective way to learn the heterogeneous features from the patches; and 3) the small sample available to train the model.

The results also show only the patches from ROIs can not provide enough information for survival prediction. And the proposed WSISA is good at estimating the patient's survival with the information provided by WSIs not just from ROIs. Thus WSISA is good at finding the survival related patterns and making a better prediction on patient's survival from small sample datasets. What's more, it does not need the annotations on histopathological images, which will make it more practical in real applications.

4.4.4 Discussion

From the extensive experiments on three datasets, the patches selected by WSISA are discriminative and the aggregation results on predicting patients' survival are better. Thus the main contribution of this paper are as follows:

Table 3. Performance comparison of the proposed methods and other existing related methods using C-index values on three datasets. The larger C-index value is better. The results highlighted with red bold show the best performance in those datasets. The results highlighted with black bold indicate the features that can generate better C-index values with specific methods.

Methods	NLST		TCGA-LUSC		TCGA-GBM	
	ROI-based	WSISA	ROI-based	WSISA	ROI-based	WSISA
LASSO-Cox [23]	0.503	0.703	0.540	0.638	0.440	0.600
En-Cox [26]	0.502	0.703	0.613	0.638	0.440	0.603
Cox-Log [10]	0.466	0.440	0.548	0.397	0.504	0.645
Cox-Weibull [10]	0.480	0.295	0.491	0.388	0.384	0.400
RSF [9]	0.485	0.595	0.347	0.578	0.560	0.518
BoostCI [15]	0.511	0.610	0.339	0.273	0.507	0.510
MTLSA [14]	0.609	0.680	0.536	0.603	0.571	0.510

- WSISA is for the first time being developed to make survival prediction based on whole slide histopathological images. What's more, it is annotation free, which will make it close to the real applications.
- We solve the problem of training deep survival models on small sample datasets as a by-product in WSISA. The framework provides large number of training samples by clustering numerous candidate patches from patients' whole slides.
- The developed WSISA can catch the holistic information of one patient regardless of the number and size of whole slide histopathological images the patient provides. This distinguishes to all the ROI based methods and make the performance of WSISA improve a lot.
- Extensive experiments on three datasets and two types of cancers are conducted to make our conclusion more concrete.

However, there are still some open problems needed to be solved in our future work. To make WSISA work well, it is needed to extract hundreds to thousands patches from one patient, which will definitely take much disk space. Thus it is valuable to explore a way to reduce the disk space consumption. The selected clusters may contain certain phenotypes that related to specific disease. The discovery of those phenotypes will be significantly meaningful.

5. Conclusion

We presented a whole slide histopathological images-based survival analysis framework (WSISA) which can directly learn discriminative and survival-related patterns from patients' gigapixel images. Compared to existing

patch-based survival models, the developed framework can handle various numbers and sizes whole slide histopathological images among different patients. It can learn holistic information of the patient and achieve much better performance compared to the ROI patch based methods. The proposed framework can also be applied to other tasks based on the whole slide histopathological images like tumor grade estimation. In the future, we will explore more ways to optimize the training process so that it scales up to the large scale histopathological datasets with other types of cancers.

Acknowledgment

The authors would like to thank the National Cancer Institute for access to NCI's data collected by the National Lung Screening Trial. The statements contained herein are solely those of the authors and do not represent or imply concurrence or endorsement by NCI. We also thank Nvidia for supporting our research with NVIDIA Tesla K40 GPUs.

References

- [1] E. Bair, T. Hastie, D. Paul, and R. Tibshirani. Prediction by supervised principal components. *Journal of the American Statistical Association*, 101(473), 2006. 3
- [2] E. Bair and R. Tibshirani. Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol*, 2(4):E108, 2004. 3
- [3] H. M. Bøvelstad, S. Nygård, H. L. Størvold, M. Aldrin, Ø. Borgan, A. Frigessi, and O. C. Lingjærde. Predicting survival from microarray data: a comparative study. *Bioinformatics*, 23(16):2080–2087, 2007. 3
- [4] A. E. Carpenter, T. R. Jones, M. R. Lamprecht, C. Clarke, I. H. Kang, O. Friman, D. A. Guertin, J. H. Chang, R. A. Lindquist, J. Moffat, et al. Cellprofiler: image analysis software for identifying and quantifying cell phenotypes. *Genome biology*, 7(10):R100, 2006. 6

- [5] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 187–220, 1972. 6
- [6] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, pages 1026–1034, 2015. 2
- [7] P. J. Heagerty and Y. Zheng. Survival model predictive accuracy and roc curves. *Biometrics*, 61(1):92–105, 2005. 6
- [8] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz. Patch-based convolutional neural network for whole slide tissue image classification. In *CVPR*, pages 2424–2433, 2016. 1, 2
- [9] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer. Random survival forests. *The annals of applied statistics*, pages 841–860, 2008. 5, 6, 8
- [10] J. D. Kalbfleisch and R. L. Prentice. *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, 2011. 6, 8
- [11] C. Kandath, M. D. McLellan, F. Vandin, K. Ye, B. Niu, C. Lu, M. Xie, Q. Zhang, J. F. McMichael, M. A. Wyczalkowski, et al. Mutational landscape and significance across 12 major cancer types. *Nature*, 502(7471):333–339, 2013. 5
- [12] J. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger. Deep survival: A deep cox proportional hazards network. *arXiv preprint arXiv:1606.00931*, 2016. 3
- [13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 2
- [14] Y. Li, J. Wang, J. Ye, and C. K. Reddy. A multi-task learning formulation for survival analysis. In *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16)*, 2016. 3, 6, 8
- [15] A. Mayr and M. Schmid. Boosting the concordance index for survival data—a unified framework to derive and evaluate biomarker combinations. *PLoS one*, 9(1):e84483, 2014. 6, 8
- [16] M. Y. Park and T. Hastie. L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):659–677, 2007. 3
- [17] C. K. Reddy and Y. Li. A review of clinical prediction models. In *Healthcare Data Analytics*, pages 343–378. Chapman and Hall/CRC, 2015. 3
- [18] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 1
- [19] K. Shedden, J. M. Taylor, S. A. Enkemann, M.-S. Tsao, T. J. Yeatman, W. L. Gerald, S. Eschrich, I. Jurisica, T. J. Giordano, D. E. Misek, et al. Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nature medicine*, 14(8):822–827, 2008. 2
- [20] H. Steck, B. Krishnapuram, C. Dehing-oberije, P. Lambin, and V. C. Raykar. On ranking in survival analysis: Bounds on the concordance index. In *Advances in neural information processing systems*, pages 1209–1216, 2008. 2, 7
- [21] N. L. S. T. R. Team et al. The national lung screening trial: overview and study design. *Radiology*, 2011. 5
- [22] T. Therneau. A package for survival analysis in s. r package version 2.37-4. URL <http://CRAN.R-project.org/package=survival>. *Box*, 980032:23298–0032, 2013. 6
- [23] R. Tibshirani et al. The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395, 1997. 3, 5, 6, 8
- [24] H. C. van Houwelingen, T. Bruinsma, A. A. Hart, L. J. van't Veer, and L. F. Wessels. Cross-validated cox regression on microarray gene expression data. *Statistics in medicine*, 25(18):3201–3216, 2006. 3
- [25] H. Wang, F. Xing, H. Su, A. Stromberg, and L. Yang. Novel image markers for non-small cell lung cancer classification and survival prediction. *BMC Bioinformatics*, 15(1):310, 2014. 1, 3
- [26] Y. Yang and H. Zou. A cocktail algorithm for solving the elastic net penalized coxs regression in high dimensions. *Statistics and its Interface*, 6(2):167–173, 2012. 6, 8
- [27] J. Yao, S. Wang, X. Zhu, and J. Huang. Imaging biomarker discovery for lung cancer survival prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 649–657. Springer International Publishing, 2016. 1, 3
- [28] K.-H. Yu, C. Zhang, G. J. Berry, R. B. Altman, C. R. D. L. Rubin, and M. Snyder. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nature Communications*, 7(12474), 2016. 1, 3, 6
- [29] Y. Yuan, H. Failmezger, O. M. Rueda, H. R. Ali, S. Gräf, S.-F. Chin, R. F. Schwarz, C. Curtis, M. J. Dunning, H. Bardwell, et al. Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. *Science translational medicine*, 4(157):157ra143–157ra143, 2012. 3
- [30] Y. Yuan, E. M. Van Allen, L. Omberg, N. Wagle, A. Amin-Mansour, A. Sokolov, L. A. Byers, Y. Xu, K. R. Hess, L. Diao, et al. Assessing the clinical utility of cancer genomic and proteomic data across tumor types. *Nature biotechnology*, 32(7):644–652, 2014. 2, 5
- [31] X. Zhu, J. Yao, and J. Huang. Deep convolutional neural network for survival analysis with pathological images. In *Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on*, pages 544–547. IEEE, 2016. 1, 2
- [32] X. Zhu, J. Yao, X. Luo, G. Xiao, Y. Xie, A. Gazdar, and J. Huang. Lung cancer survival prediction from pathological images and genetic data: an integration study. In *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on*, pages 1173–1176. IEEE, 2016. 6
- [33] X. Zhu, J. Yao, G. Xiao, Y. Xie, J. Rodriguez-Canales, E. R. Parra, C. Behrens, I. I. Wistuba, and J. Huang. Imaging-genetic data mapping for clinical outcome prediction via supervised conditional gaussian graphical model. In *Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on*, pages 455–459. IEEE, 2016. 1