Weakly Supervised Deep Ordinal Cox Model for Survival Prediction From Whole-Slide Pathological Images

Wei Shao, Tongxin Wang, Zhi Huang, Zhi Han, *Member, IEEE*, Jie Zhang, and Kun Huang, *Senior Member, IEEE*

***Abstract*— Whole-Slide Histopathology Image (WSI) is generally considered the gold standard for cancer diagnosis and prognosis. Given the large inter-operator variation among pathologists, there is an imperative need to develop machine learning models based on WSIs for consistently predicting patient prognosis. The existing WSI-based prediction methods do not utilize the ordinal ranking loss to train the prognosis model, and thus cannot model the strong ordinal information among different patients in an efficient way. Another challenge is that a WSI is of large size (e.g., 100,000-by-100,000 pixels) with heterogeneous patterns but often only annotated with a single WSI-level label, which further complicates the training process. To address these challenges, we consider the ordinal characteristic of the survival process by adding a ranking-based regularization term on the Cox model and propose a weakly supervised deep ordinal Cox model (BDOCOX) for survival prediction from WSIs. Here, we generate amounts of bags from WSIs, and each bag is comprised of the image patches representing the heterogeneous patterns of WSIs, which is assumed to match the WSI-level labels for training the proposed model. The effectiveness of the proposed method is well validated by theoretical analysis as well as the prognosis and patient stratification results on three cancer datasets from The Cancer Genome Atlas (TCGA).**

***Index Terms*— Histopathological images, weakly supervised learning, survival analysis, ordinal cox model.**

I. Introduction

CANCER research has witnessed significant progresses during the past few decades. One of the major goals for cancer precision medicine is to identify biomarkers that can predict patients' outcome including survival time, which in turn allows the clinicians to make early decision on treatment [1], [2]. So far, a large number of biomarkers have been identified and applied to the prognosis of cancers [3], [4], including the histopathological image, genetic mutation, gene expression signatures, and protein markers. Of all these biomarkers, the Whole Slide Images (WSIs) of histopathological sections can reflect the underlying molecular processes and disease progression in cancers, and thus are generally considered the gold standard for cancer prognosis [5], [6]. Aided by the advances in digitized microscopic imaging and computational pathology, a number of image analysis algorithms have been designed for the prognosis of various types of cancers based on WSIs, including lung cancer [7], breast cancer [8], [9] and low-grade glioma [10], to name a few.

Generally, the main challenge for survival analysis from the WSIs is that a WSI is of large size (*e.g.,* 100,000-by-100,000 pixels), and traditional survival models can not directly apply to such high resolution images due to the high computational cost. In face of the above challenges, there are two main lines of approaches to predict patient outcome from WSIs including patch-based and WSI-based methods. The patch-based methods need experts to select amounts of cancer patches that can reflect the progression of the cancer for model construction. Specifically, with features derived from the annotated patches, Cox proportional hazard model [11] is the most popular survival prediction model. Based on the Cox model, several regularization methods including LASSO-COX [12] and En-Cox [13] have also been proposed in the literatures. In addition, since Deep Convolutional Neural Networks (CNN) has emerged as a powerful tool to solve many problems of computer vision, the studies in [14], [15] demonstrated end-to-end prediction models by combing CNN with traditional Cox model. Other works included Cheng *et al.* [16] have designed an auto-encoder model to extract cell patterns in WSIs for survival prediction.

Since the patch-based methods require the labour-sensitive patch-level annotation, some studies designed learning algorithms that only relies on WSI-level annotation for patient outcome prediction. Specifically, Zhu *et al.* [17] proposed a WSI-based survival analysis framework, which learns survival-related patterns from randomly selected clusters of

Manuscript received June 6, 2021; accepted July 7, 2021. Date of publication July 15, 2021; date of current version November 30, 2021. This work was supported by the Indiana University Precision Health Initiative. *(Wei Shao and Tongxin Wang contributed equally to this work.) (Corresponding authors: Zhi Han; Jie Zhang; Kun Huang.)*

Wei Shao, Zhi Han, Jie Zhang, and Kun Huang are with the School of Medicine, Indiana University, Indianapolis, IN 46202 USA (e-mail: zhihan@iu.edu; jizhan@iu.edu; kunhuang@iu.edu).

Tongxin Wang is with the Department of Computer Science, Indiana University Bloomington, Bloomington, IN 47405 USA.

Zhi Huang is with the Department of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907 USA.

Digital Object Identifier 10.1109/TMI.2021.3097319

1558-254X © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: SUN YAT-SEN UNIVERSITY. Downloaded on July 26,2022 at 09:43:40 UTC from IEEE Xplore. Restrictions apply.

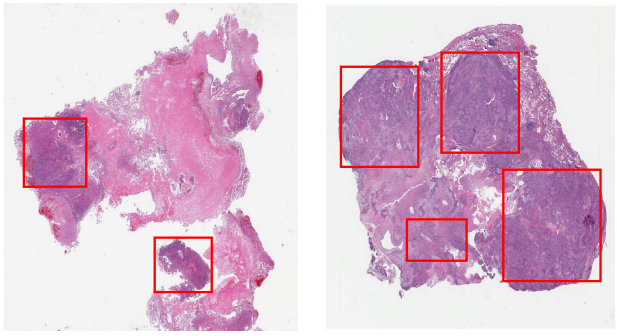

Fig. 1. WSIs of tumor samples from two lung cancer patients showing heterogeneous patterns of cell and tissue structures. The ordinal relationship between the survival time of different patients, *i.e.,* the survival time of patient A is longer than that of patient B, could be reflected by the visual appearance of WSIs, where patient B has larger area of tumor tissue regions (framed in red) than that in patient A, which suggests that the tumor in patient B progresses more aggressively.

patches. Li and Yao [18] modeled WSI as graph and then developed a graph CNN to learn global topological representations of WSI for survival prediction.

Although much progress has been achieved, the existing WSI-based prediction methods do not utilize the ordinal ranking loss to train the prognosis model, and thus cannot model the strong ordinal information among different patients in an efficient way, *e.g.,* the survival time of patient A is shorter than that of patient B. As a matter of fact, this ordinal pattern can also be reflected in WSIs. Fig. 1 presents an example, where patient B has higher risk (shorter survival time) than patient A according to the pathology reports in The Cancer Genome Atlas (TCGA). This ordinal relationship can also be observed via the visual appearance of WSIs, where patient B has larger area of tumor regions (framed in red) than that in patient A, indicating that the tumor in patient B progresses faster. Due to the lack of patch-level annotation, the existing WSI-based methods usually assumed that the survival information associated with each randomly selected patch is the same as its corresponding WSI. However, as shown in Fig. 1, WSIs usually have large heterogeneous patterns (*i.e.,* tumor region and non-tumor region) and thus the patch-level label cannot always match the WSI-level label. Accordingly, instead of only using single type of patches, we generate bags that can represent the heterogeneous patterns of WSIs for the following prediction task.

Based on the above considerations, we propose a weakly-supervised bag-level deep ordinal cox model (BDOCOX) for the survival prediction from WSIs without the patch-level annotation. Our main contribution can be summarized into the following three folds:

1) We propose a new method to generate the bags from the whole-slide pathological images, where each bag can represent the heterogeneous patterns of WSIs.
2) We propose a novel ranking loss in the deep Cox proportional hazard model, which can model the strong ordinal information among different patients in an efficient way.
3) We not only conduct experiments on different cancer cohorts to demonstrate its effectiveness for prognosis and patient stratification, but also presented theoretical analysis to give an upper bound of the generalization error on the proposed BDOCOX model.

II. Related Work

The goal of survival analysis is to predict the time duration until a certain event occurs [19], [20] and the event of interest in our study is the death of a cancer patient. In survival analysis, each cancer patient can be divided into two categories *i.e.,* censored and non-censored patient. The censored patients means that the death events of them were not observed during the follow-up period, and thus their genuine survival times are longer than the recorded data, while the non-censored patients indicate that their recorded survival times are the exact time from initial diagnosis to death. We use a triplet (x_i, t_i, δ_i) to represent each observation in survival analysis, where x_i is the feature vector for the $i-$th sample, t_i is the observed time, and δ_i is the censoring indicator. Here, $\delta_i = 1$ indicates an non-censored instance and $\delta_i = 0$ is for censored instance.

By defining the survival function $Pr(O \geq t|x)$ as the probability that a patient will survive after time t, the hazard function which can assess the instantaneous rate of death is defined as follows:

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

where $x = (x^1, x^2, \ldots, x^v)$ corresponds to the extracted feature of sample x with dimensionality v. In the hazards modeling methods, Cox proportional hazard model [11] is among the most popular one. It is built based on the hypothesis that the hazard ratio between two instances is time-independent, and is defined as

$$h(t|x) = h_0(t) exp(\alpha^T x). \tag{2}$$

Here, $h_0(t)$ corresponds to the baseline hazard and $f(x) = \alpha^T x$ is being called as risk function. It is worth noting that $\alpha = (\alpha^1, \alpha^2, \ldots \alpha^v)$ is the vector of regression parameters that can be estimated by minimizing its corresponding negative log partial likelihood function [21].

$$l(\alpha) = \min \sum_{i=1}^{N} \delta_i \left(\alpha^T x_i - \log \sum_{j \in R(t_i)} \exp(\alpha^T x_j) \right) \tag{3}$$

where N denotes the number of patients, $R(t_i)$ is the risk set at time t_i, which represents the set of comparable patients that are still under risk before time t_i. In addition, δ_i is the censoring indicator, where $\delta_i = 1$ indicates an non-censored instance and $\delta_i = 0$ is for censored instance.

In addition, since deep learning has been extremely successful in the field of computer vision, Jared *et al.* [22] was the first one using deep model to capture the nonlinear relationship between the clinical feature and the risk function in Cox model, while Zhu *et al.* [14] have developed an end-to-end DeepConvSurv model for patient outcome prediction from pathological images. However, these methods ignored the fact that survival analysis is a regression problem, where the ranking of patients' survival risks matters. Accordingly, we design a deep ordinal cox model that can reflect the ordinal nature of survival time for prognosis prediction in this study.

Authorized licensed use limited to: SUN YAT-SEN UNIVERSITY. Downloaded on July 26,2022 at 09:43:40 UTC from IEEE Xplore. Restrictions apply.

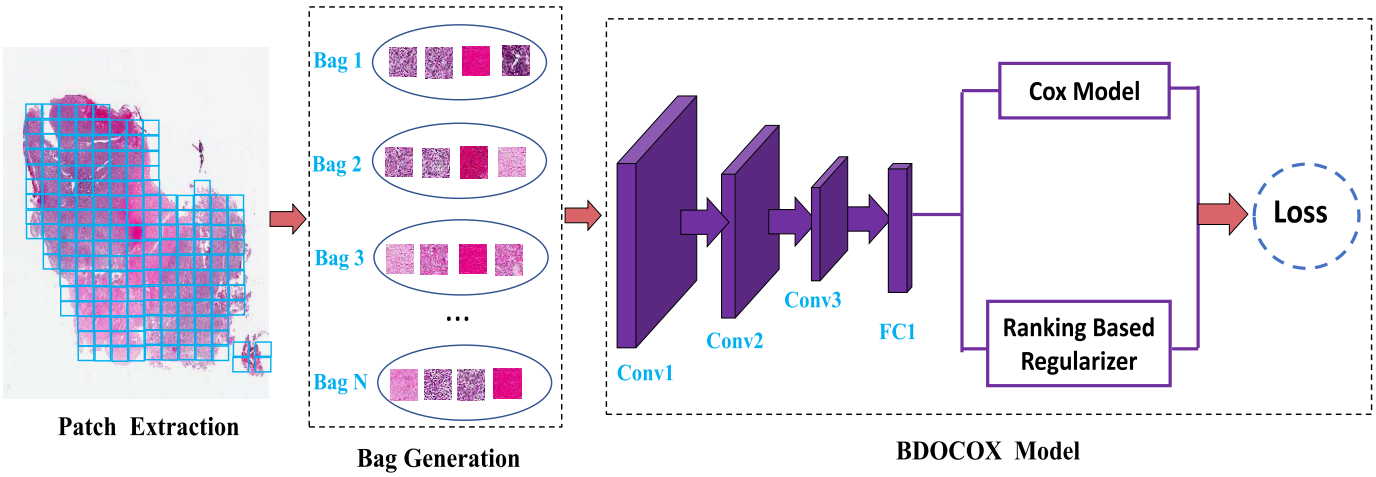

Fig. 2. A schematic diagram of our proposed method. Our method consists of three major steps. Firstly, we extract the tissue patches from the WSIs. Secondly, we generate bags from the WSIs that can represent the heterogeneous patterns of WSIs. Thirdly, we apply our proposed BDOCOX model on the generated bags for cancer prognosis prediction.

III. Method

Fig. 2 shows a schematic diagram of our proposed method, which consists of the following *three* major steps including patch extraction, bag generation, and the design of the proposed BDOCOX model.

A. Extract Patches From WSIs

The goal of this step is to generate patches from WSIs and we divide each whole-slide image into 512×512 patches without overlap. To capture detailed information of the images, those patches are extracted from $20\times$ (0.5 microns per pixel) objective magnifications. Because the generated patches may contain background regions that do not contain enough cells for further investigation, we select the patches whose image density are larger than 0.7. Here, the image density is calculated as the percentage of non-white (at least one of the red, green, and blue values was below 200 in the 24-bit RGB colour space) pixels in that patch. After the patch extraction step, we can derive around 1000 patches for each WSI. Given that such amount of patches can capture the heterogeneous patterns of WSI, we can also efficiently train our model with lower computational burden.

B. Bag Generation

Since tumors usually have a mixture of different structural and textural properties, the labels of the derived patches are not necessarily consistent with their corresponding WSI. In order to use the WSI-level information for model construction, we form amounts of bags, where each one of them is consisted of patches with heterogeneous patterns and thus can be assumed to share the same survival information with WSI.

Specifically, given the n patches sampled from a WSI, we first resize each of them to the size of 224×224 and then feed them to the pre-trained Resnet128 [23], by which we derive a 4096-dimensional feature vector from its last fully connected layer. Next, based on the derived features, we apply K-means clustering algorithm basing on Euclidean distance to aggregate the patches into $d = 5$ distinct phenotype groups. In order to capture the major patterns of each WSI, we generate amounts of bags, where each bag B_i is compromised of D_i patches $B_i = (p_i^1, p_i^2, \ldots, p_i^{D_i})$ that are randomly sampled from $d = 5$ clusters. Here, the number of the randomly selected patches *i.e.,* k_j $(j = 1, 2 \ldots d)$ in each cluster are proportional to the cluster capacity, and thus can be calculated as follows:

$$k_j = \left[Z \times \frac{m_j}{\sum_{j=1}^{d} m_j} \right], \tag{4}$$

where m_j $(j = 1, 2, 3, 4, 5)$ corresponds to the number of patches in each selected cluster, and Z controls the bag patch number and we set it as 10 in this paper. In addition, $[A]$ indicates the nearest integer to A. It is worth noting that, our method can generate numerous bags that share the WSI-level label, which makes it possible to use the deep learning model to learn the survival-related patterns from WSI.

C. Weakly Supervised Deep Ordinal Cox Model

Compared to the shallow hand-crafted features, deep learning models can learn the representation of images at high levels. Obviously, we could directly apply CNN to achieve survival prediction from pathological images by replacing the linear part $\alpha^T x$ of the Cox model (*i.e.,* shown in Eq.(3)) with non-linear deep fully connected network. Here, we design a neural network which is compromised of 3 convolutional layers, 2 max-pooling layers, and 1 fully-connected layer, for model training. We adopt batch normalization (BN) right after each convolution and before activation, the detailed structure is shown in Table I.

Specifically, suppose we have generated N bags from the WSIs of different patients, where each bag B_i is consisted of D_i patches $B_i = (p_i^1, p_i^2, \ldots, p_i^{D_i})$ with the survival time t_i, the deep CNN model extracts the layer-wise representation

Authorized licensed use limited to: SUN YAT-SEN UNIVERSITY. Downloaded on July 26,2022 at 09:43:40 UTC from IEEE Xplore. Restrictions apply.

TABLE I
THE ARCHITECTURE OF THE PROPOSED NEURAL NETWORK

Layer	Filter size	Stride	Number
Conv1	7×7	3	32
Batch Normalization			
Relu			
Max-pooling	2×2		
Conv2	5×5	2	32
Batch Normalization			
Relu			
Max-pooling	2×2		
Conv3	3×3	2	32
Batch Normalization			
Relu			
Max-pooling	2×2		
FC	16		

for each input patch p_i^j. We denote the output of its last fully connected layer (*i.e.,* FC1 in Fig. 2) as $\sigma(p_i^j) \in R^m$, and assume that the representation of a bag (*i.e.,* B_i) is obtained by averaging the representation of the patches it covers, *i.e.,* $\sigma(B_i) = \frac{1}{D_i}\sum_{j=1}^{D_i}\sigma(p_i^j)$. Then, the negative log partial likelihood function of the Cox model (shown in Eq.(3)) can be reformulated as:

$$\min_{\alpha}\sum_{i=1}^{N}\delta_i\left(\alpha^T\sigma(B_i) - \log\sum_{j\in R(t_i)}\exp(\alpha^T\sigma(B_j))\right) \quad (5)$$

Although the above deep Cox model could be directly used for survival prediction from WSIs, it fails to utilize the ordinal ranking loss *i.e.,* the survival time for patient A is longer than that for patient B, to train the prognosis model. Since the generated bags from a patient share the same WSI-level survival information, we design a ranking loss function to capture the ordinal survival information among different bags. It is noteworthy that we cannot explicitly derive the survival function in Cox model (shown in Eq.(2)), since the baseline hazard function $h_0(t)$ is unknown. However, the ordinal relationship of two bags (*i.e.,* B_i, B_j) can be easily derived by calculating the ratio r_{ij} between them as following:

$$\begin{aligned} r_{ij} &= \frac{h(t|B_i)}{h(t|B_j)} = \frac{h_0(t)\exp(\alpha^T\sigma(B_i))}{h_0(t)\exp(\alpha^T\sigma(B_j))} \\ &= \exp(\alpha^T(\sigma(B_i) - \sigma(B_j))) \end{aligned} \quad (6)$$

Obviously, the survival time for bag B_i should be longer than that for bag B_j if $r_{ij} \leq 1$ and vice versa. By utilizing the above ordinal relationship indicated by Cox model, we define the ranking loss function below:

$$RankLoss = \sum_{i=1}^{N}\sum_{j\in R(t_i)} I(B_i, B_j)\max(0, 1 - r_{ij}), \quad (7)$$

where $I(B_i, B_j) = 1$ if B_i and B_j are a comparable ranking pair, otherwise, $I(B_i, B_j) = 0$. Here, B_i and B_j are *comparable* if B_i belongs to a non-censored patient, whose recorded survival time t_i is shorter than that of a patient bag B_j (*i.e.,* t_j). It is worth noting that B_j may also belong to a censored patient since its genuine survival time is longer than the recorded time t_j.

By combining Eq.(5) with Eq.(7), the objective function of the proposed BDOCOX model in the output layer can finally be formulated as:

$$\min_{\alpha} l(\alpha) = \min_{\alpha}\sum_{i=1}^{N}\delta_i J(\sigma(B_i)), \quad (8)$$

where

$$\begin{aligned} J(\sigma(B_i)) = &\left(\alpha^T\sigma(B_i) - \log\sum_{j\in R(t_i)}\exp(\alpha^T\sigma(B_j))\right) \\ &+ \lambda\sum_{j\in R(t_i)} I(B_i, B_j)\max(0, 1 - r_{ij}) \end{aligned} \quad (9)$$

with λ being a regularization parameter. The parameter α can be learnt by using the stochastic gradient decent method.

D. Theoretical Results

In this part, we provide a risk bound analysis for the proposed BDOCOX model.

Theorem 1: Let H represents a family of loss function associated to J (shown in Eq.(9)). Then, for any $\delta > 0$ over $\sigma(B) = [\sigma(B_1), \sigma(B_2), \ldots, \sigma(B_N)] \in R^{m\times N}$, each of the following holds for any $J \in H$ with probability at least $1-\delta$.

$$E[J(\sigma(B))] \leq \frac{1}{N}\sum_{i=1}^{N} J(\sigma(B_i)) + R_S(J) + \sqrt{\frac{2\log\frac{1}{\delta}}{N}} \quad (10)$$

where $E[J(\sigma(B))] = E_{B\sim T^N}\sum_{i=1}^{N} J(\sigma(B_i))$ and B are sampled from the distribution T^N with N samples. $R_S(J)$ represents the Rademacher complexity of J with respect to $\sigma(B)$.

Proof Sketch: We use Theorem 10.2 and Theorem 10.3 described in [24] to prove the risk bound. Here, using McDiarmid inequality and Rademacher complexity is the core idea to measure the function class H.

Remark 1: As shown in Theorem 1, the generalization bound of the BDOCOX model is largely determined by the Rademacher complexity of J. Here, we suppose that $\sigma(B_i)$ and α are L_2 norm-bounded in the BDOCOX model. Then, we can derive the upper bound of the Rademacher complexity of J by the following Theorem 2.

Theorem 2: For $\sigma(B) = [\sigma(B_1), \sigma(B_2), \ldots, \sigma(B_N)] \in R^{m\times N}$. Suppose that $\sigma(B_i) \leq M_B$ and $\alpha \leq M_\alpha$ are L_2 norm-bounded in the BDOCOX model. Let L be the Lipschitz constant for the function $f(z) = max(0, 1-exp(z))$, then the Rademacher complexity of J satisfies the following inequality:

$$\begin{aligned} R_S(J) \leq (3L\sqrt{\frac{N}{|\mathrm{R}(t_i)|}} + 2)M_\alpha M_B + \sqrt{\frac{8\log(2/\delta)}{|\mathrm{R}(t_i)|}} \\ - \log|\mathrm{R}(t_i)| \end{aligned}$$

where $|\mathrm{R}(t_i)|$ indicates the cardinality of the risk set $R(t_i)$

Authorized licensed use limited to: SUN YAT-SEN UNIVERSITY. Downloaded on July 26,2022 at 09:43:40 UTC from IEEE Xplore. Restrictions apply.

Proof: According to the Theorem 1 in [25], the Rademacher complexity of the second term in Eq.(9) J can be formulated as:

$$E(\sum_{j\in R(t_i)} I(B_i, B_j)\max(0, 1-r_{ij})) \le M_\alpha M_B\sqrt{\frac{N}{|\mathrm{R}(t_i)|}} + \sqrt{\frac{8\log(2/\delta)}{|\mathrm{R}(t_i)|}}$$

For the first term in Eq.(9), we have:

$$\begin{aligned}&\alpha^T\sigma(B_i) - \log\sum_{j\in R(t_i)}\exp(\alpha^T\sigma(B_j))\\ &\le \alpha^T\sigma(B_i) - \log|\mathrm{R(t_i)}| - \log\exp(\sum_{j\in R(t_i)}\frac{\alpha^T\sigma(B_i)}{|\mathrm{R(t_i)}|}))\\ &\le 2M_\alpha M_B - \log|\mathrm{R(t_i)}|\end{aligned}$$

Accordingly, we can get Theorem 2 by combining the above two parts together. Then, by combining Theorem 1, Theorem 2, we get the following Theorem 3.

Theorem 3: Let H represent a class of mappings (shown in Eq.(9)), where $\|\alpha\|_2 \le \mathrm{M}_\alpha$ and $\max_i \|\sigma(B_i)\|_2 \le M_B$. Then, for any $\delta \le 0$, each of the following holds for all $J \in H$ with the probability $1-\delta$.

$$E[J(\sigma(B))] \le \frac{1}{N}\sum_{i=1}^{N} J(\sigma(B_i)) + (3L\sqrt{\frac{N}{|\mathrm{R}(t_i)|}} + 2)M_\alpha M_B + \sqrt{\frac{8\log(2/\delta)}{|\mathrm{R}(t_i)|}} - \log|\mathrm{R}(t_i)| + \sqrt{\frac{2\log\frac{1}{\delta}}{N}}$$

Remark 2: As shown in Theorem 3. On one hand, the generalization error of the proposed BDOCOX model is reduced by the increased sample size in the risk set $|\mathrm{R}(t_i)|$. Actually, $|\mathrm{R}(t_i)|$ is correlated with the number of comparable pairs for bag B_i, and we can get more comparable pairs if the ratio of non-censored patients is higher in the training cohort. Accordingly, the proposed BDOCOX will perform better if more non-censored patients are involved for model training. On the other hand, Theorem 3 also indicates that the error bound will get worse when the L2-norm bound of $\sigma(B_i)$ increases, which reminds us of applying the batch normalization trick to avoid the risk of model explosion that will increase the scale of $\sigma(B_i)$.

IV. Experimental Results

A. Dataset

The Cancer Genome Atlas (TCGA) has generated multi-modal genome, epigenome, transcriptome, and imaging data for thousands of tumor samples across more than 30 types of cancers [26]. In this study, we conduct experiments on three early-stage(*i.e.*, stage I and stage II) cancer datasets *i.e.,* Kidney Renal Clear Cell Carcinoma (KIRC), Liver Hepatocellular Carcinoma(LIHC) and Lung Squamous Cell Carcinoma (LUSC) with high resolution whole slide pathological images in TCGA. The details of the cohort information are listed in Table II, where the ratio of non-censored patients in LUSC dataset is higher than the other two cancer cohorts. In this study, all the patients' (including both censoring and non-censoring patients) survival information is derived from TCGA, and the survival time for the censoring patients in TCGA refers to their last follow-up time.

TABLE II
Patient Demographics for Different Cancer Datasets

Data Sets	# No.Patient	# No.Censored	# Total.Size
KIRC	278	232	275 GB
LIHC	218	165	252 GB
LUSC	345	207	324 GB

B. Experimental Settings

For parameter settings, we generate 20 bags for each WSI and the parameter λ in Eq.(9) is tuned from $\{10^0, 10^1, 10^2\}$. We empirically set the mini-batch size and learning rate as 10 and 1e-3, respectively. During the experiments, we randomly partition the images in each cohort into 5 folds, and the subjects within one fold are used as testing data at each time, while all the other subjects in the remaining 4 folds are used for model training. For the training dataset, we split 20% samples in it as the validation dataset to tune model parameters.

We evaluate the performance of the proposed BDOCOX and other comparing methods using the metric *Concordance Index* (CI) and AUC [27]. Since each WSI contains 20 bags, we calculate $\sigma(W_i)$ by averaging the representation of the bags it contains. The value of CI and AUC ranges from 0 to 1, and the larger CI value means the better prediction performance and vice versa.

C. Bag Generated From WSI

Fig. 3 shows an example of a bag generated from a WSI. We frame the patches within the bag that are derived from different clusters with diverse colors. From Fig. 3 we can see that the patches framed in red correspond to the cluster of tumor regions. Since the overall majority of the WSI shown in Fig. 3 are tumor tissues, 5 of 10 patches in this sample bag are tumor patches. Next, the patches framed in green color come from the cluster that contains the immune cells while the ones framed in blue and green are extracted from stromal regions. Extensive studies have verified that the progression of cancer is supported by the abnormal development of stroma [28] and growth of the immune cells [29]. These results show that the generated bag can reflect both the heterogeneous patterns of WSI and the aggressiveness of the cancer, and thus can share its corresponding WSI-level label.

D. Survival Prediction

We compare our proposed model BDOCOX with the following baseline methods using the metrics CI and AUC, the results are shown in Table III.

Authorized licensed use limited to: SUN YAT-SEN UNIVERSITY. Downloaded on July 26,2022 at 09:43:40 UTC from IEEE Xplore. Restrictions apply.

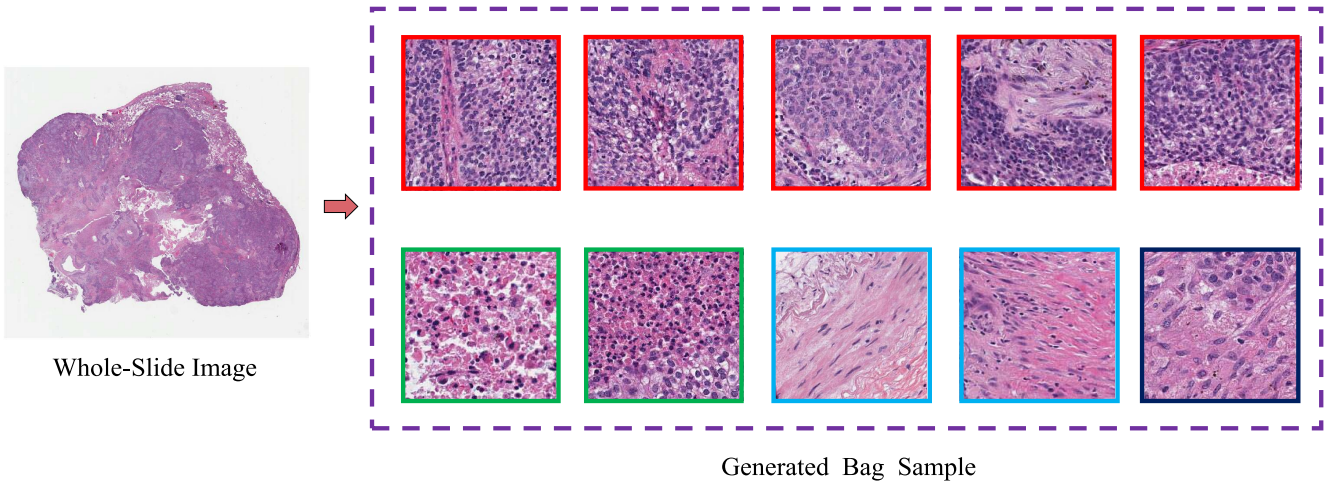

Fig. 3. Example of a bag generated from the WSI, the patches derived from different clusters are framed with different colors, where the red frames correspond to the tumor patches, green frames indicate the immune cells and the patches with blue and black frames are related to the stromal regions in the WSI.

TABLE III
PERFORMANCE COMPARISON AMONG DIFFERENT SURVIVAL PREDICTION METHODS BY THE MEASUREMENTS OF CONCORDANCE INDEX

	KIRC		LIHC		LUSC	
	CI	AUC	CI	AUC	CI	AUC
LASSO-COX	0.541 ± 0.054	0.584 ± 0.047	0.534 ± 0.048	0.544 ± 0.031	0.527 ± 0.043	0.551 ± 0.063
En-Cox	0.530 ± 0.081	0.554 ± 0.067	0.5371 ± 0.061	0.552 ± 0.055	0.552 ± 0.085	0.589 ± 0.077
RSF	0.538 ± 0.078	0.562 ± 0.091	0.542 ± 0.089	0.571 ± 0.079	0.561 ± 0.067	0.593 ± 0.055
DeepConvSurv	0.578 ± 0.056	0.597 ± 0.066	0.562 ± 0.070	0.587 ± 0.072	0.604 ± 0.032	0.629 ± 0.042
DeepAttnMISL	0.647 ± 0.041	0.674 ± 0.039	0.671 ± 0.034	0.702 ± 0.034	0.654 ± 0.041	0.677 ± 0.051
DeepGraphSurv	0.652 ± 0.073	0.682 ± 0.083	0.654 ± 0.048	0.694 ± 0.045	0.687 ± 0.082	0.717 ± 0.071
DOCOX	0.605 ± 0.067	0.629 ± 0.071	0.622 ± 0.043	0.657 ± 0.061	0.633 ± 0.039	0.651 ± 0.035
BDCOX	0.667 ± 0.052	0.704 ± 0.065	0.678 ± 0.038	0.699 ± 0.041	0.696 ± 0.041	0.719 ± 0.038
BDOCOX	$\mathbf{0.699 \pm 0.027}$	$\mathbf{0.714 \pm 0.031}$	$\mathbf{0.701 \pm 0.041}$	$\mathbf{0.727 \pm 0.039}$	$\mathbf{0.726 \pm 0.022}$	$\mathbf{0.751 \pm 0.034}$

- *Regularized Cox Models:* Follow the study in [30], which extracts 150-dimensional morphology features from WSIs, and then use l_1-norm LASSO-Cox [12] and EN-Cox [13] models for survival prediction.
- *Random Survival Forests(RSF) [31]:* First extract the 150-dimensional morphological feature introduced in [30], and then apply random forest based survival prediction model introduced in [31] for clinical outcome prediction.
- *DeepConvSurv [14]:* Extract patches from the ROI (region of interest) of WSIs, and then feed them into CNN for survival prediction.
- *DeepAttnMISL [27]:* An attention-based algorithm for survival analysis from WSIs.
- *DeepGraphSurv [27]:* A graph based algorithm for survival analysis from WSIs.
- *DOCOX:* A variant of BDOCOX, which assumes that the survival information of each randomly selected patch matches the WSI-level survival information, we select 20 patches for each WSI for model training.
- *BDCOX:* A variant of BDOCOX, in comparison with the proposed BDOCOX model, BDCOX does not take the ordinal ranking loss (*i.e.,* the second term in Eq. (9)) for model training.

As can be seen from Table III, firstly, the prognosis power of the regularized Cox models (*i.e.,* LASSO-Cox and En-Cox) is inferior than the other deep model based algorithms. This is because the deep model can better represent WSIs than the hand-crafted low-level features. Secondly, the proposed BDOCOX and its variant BDCOX can achieve higher CI and AUC values than the comparing methods. These results clearly demonstrate the advantage of generating bags that can represent the heterogeneous patterns of WSIs for model training can improve the final prognosis prediction performance. Moreover, we find that the BDOCOX model can provide the mean CI values of 0.699, 0.701 and 0.726 and AUC values of 0.714, 0.727 and 0.751 for KIRC, LIHC and LUSC datasets, respectively, which is better than the comparing BDCOX algorithm for prognostic prediction. These results shows the advantage of taking the ordinal relationship among the survival time of different patients into consideration for prognosis prediction. Finally, we also observe that the CI and AUC values for the proposed BDOCOX method on LUSC dataset

Authorized licensed use limited to: SUN YAT-SEN UNIVERSITY. Downloaded on July 26,2022 at 09:43:40 UTC from IEEE Xplore. Restrictions apply.

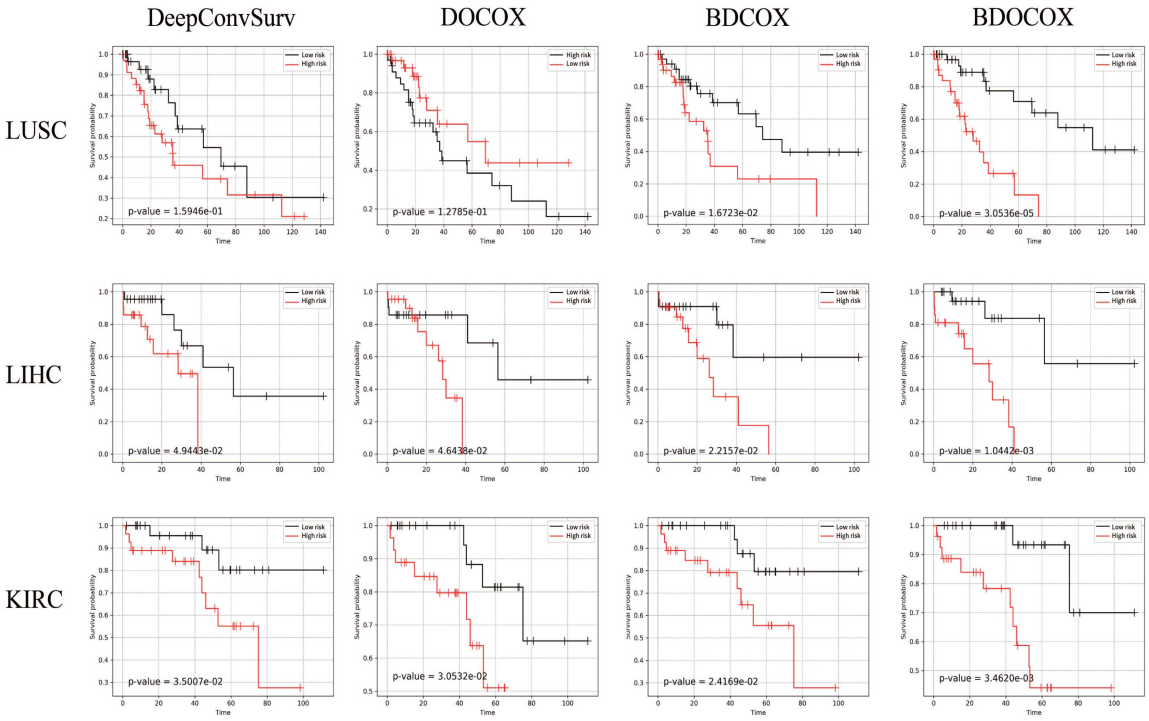

Fig. 4. The survival curves by applying different methods on LUSC,LIHC and KIRC datasets. Here, the better patient stratification performance comes with the smaller p-values by the log-rank test.

TABLE IV
THE EFFECTS OF THE CLUSTER NUMBER d FOR THE PROGNOSIS PREDICTION ON DIFFERENT CANCER COHORT

	$d=2$	$d=3$	$d=5$	$d=6$	$d=7$
KIRC	0.611	0.674	0.699	0.707	0.689
LIHC	0.637	0.654	0.701	0.695	0.697
LUSC	0.626	0.677	0.726	0.734	0.729

is significantly better than those on LIHC and KIRC datasets. This is consistent with our theoretical results (i.e., Theorem 3) that the lower ratio of censored patients in LUSC cohort(shown in Table II) will come with lower generalization error for the proposed BDOCOX model.

E. Patient Stratification

In the field of survival analysis, another important task is to stratify cancer patients into subgroups with different predicted outcomes, by which we can intervene with personalized treatment plan for the patients [32]. We follow the method described in [16], which used the median risk score (*i.e.,* $\sigma^T(W_i)$) in the training set as a threshold to stratify patients in the testing set into low-risk and high-risk groups, and then test if these two groups have significantly different survival times using log-rank test [33]. Here, better prognosis prediction performance comes with the smaller p-value by the log-rank test, we only show the fold with the best stratification performance for different methods in Fig. 4.

As can be seen from Fig. 4, the proposed BDOCOX achieves superior stratification performance (consistently lowest p-values) than the comparing method (DOCOX, BDCOX and DeepConvSurv) on KIRC, LIHC, and LUSC datasets. This is because the BDOCOX model considers both the ordinal characteristics in survival analysis and the heterogeneous patterns in the WSIs, and thus its prognostic power is effectively improved.

F. The Effects of the Cluster Number for the Prognosis Prediction

In the proposed BDOCOX model, we apply clustering algorithm to aggregate the patches into $d = 5$ distinct phenotype groups to generate bags. In this section, we investigate the effect of d in the BDOCOX model. Specifically, we tune the cluster number d from $\{2, 3, 5, 6, 7\}$ and record their mean Concordance Index for KIRC, LIHC and LUSC datasets, respectively, with the results are shown in Table.IV.

From Table. IV, we can clearly see that the BDOCOX model yields inferior prognosis prediction performance with $d = 2, 3$, one possible reason lies in that different types of image patches (*i.e.,* tumor patches, stromal patches, and immune cell patches) cannot be involved in each generated bag if the selected cluster number is relatively small ($d = 2, 3$). As the selected number of clusters (*i.e.,* $d = 5, 6, 7$) increases, each bag is generally consisted of patches with the

Authorized licensed use limited to: SUN YAT-SEN UNIVERSITY. Downloaded on July 26,2022 at 09:43:40 UTC from IEEE Xplore. Restrictions apply.

TABLE V
THE EFFECTS OF THE BAG NUMBER FOR THE PROGNOSIS PREDICTION ON DIFFERENT CANCER COHORTS

Bag Number	10	20	30	40
KIRC	0.664	0.699	0.704	0.706
LIHC	0.675	0.701	0.696	0.711
LUSC	0.669	0.726	0.714	0.721

TABLE VI
THE EFFECTS OF THE BAG PATCH NUMBER FOR THE PROGNOSIS PREDICTION ON DIFFERENT CANCER COHORTS

	$Z = 9$	$Z = 10$	$Z = 11$	$Z = 12$
KIRC	0.702	0.699	0.714	0.685
LIHC	0.687	0.701	0.691	0.713
LUSC	0.698	0.726	0.723	0.705

heterogeneous patterns, which is helpful for the downstream prognosis prediction task.

G. The Effects of the Bag Number for the Prognosis Prediction

In the proposed BDOCOX model, we generate 20 bags for each WSI and use these bags for model training. In this section, we investigate if the prognosis results will be changed with the variation of the bag number. Specifically, we tune the number of generated bags for each WSI over the range of $\{10, 20, 30, 40\}$ and record their corresponding average Concordance Index (CI) for different cancer cohorts in Table V. As can be seen from Table V, the CI values will be decreased when the number of selected bags for each WSI is small(i.e., 10). One possible reason is that the proposed deep neural network cannot learn the survival-related patterns from such limited number of bags. On the other hand, the CI values will be stable if the number of generated bags for each slide is greater than or equal to 20. These results indicate that our results will not be affected by the bag number if enough bags have been used for model training.

H. The Effects of Bag Patch Number for the Prognosis Prediction

As introduced in Section III.B, we set the parameter Z that controls the bag patch number as 10. In this section, we investigate if the bag patch number will affect the final prognosis results. Specifically, we tune Z from 9 to 12 with intervals of 1 and record their corresponding CI values for KIRC, LIHC, and LUSC datasets in Table VI. As can be seen from Table VI, the CI values are generally stable with different bag patch numbers for all of these three cancer cohorts, which indicates that our proposed method is insensitive to the parameter of bag patch number.

V. DISCUSSION

Histopathology images have been recognized as the gold standard for the diagnosis and prognosis of human cancers. Many studies have been focusing on using machine learning algorithms to predict patients' survival. In this study, we present a weakly supervised deep ordinal model for survival prediction basing on histopathological images. By considering the progression of tumor is not only affected by unlimited growth of the tumor cells, but also supported, stimulated, and nurtured by the microenvironment around it, we generate the bags that can represent the heterogeneous patterns of WSI for model training. Our method provides a way of predicting the patients' survival from WSI only with WSI-level survival information. The experimental results also demonstrate that our proposed method could be with stronger prognostic power than other comparing methods on three different cancer cohorts.

Our study has several limitations. Firstly, the complexity of the new algorithm ($O(N^2)$) is relatively high and further investigation on how to efficiently train our model with a larger batch size is needed. Secondly, we only calculate the ranking loss basing on the Cox model. In the future, we plan to design the ranking loss on other survival analysis methods (i.e., random survival forest) to further validate its effectiveness. Thirdly, the ranking loss applied in this study focuses on predicting prognosis in patients with histopathological images, and how to apply it to predict survival with multi-modality data is very worth discussing. Finally, in the proposed deep neural network, we use the averaging pooling strategy to aggregate the patch-level representations, and how to use the attention mechanism that can consider the importance of each patch for aggregation is also an important research direction in the future.

VI. CONCLUSION

In this study, we present a novel WSI-based survival prediction analysis framework BDOCOX, in which we take patients ordinal survival information and cancer tissue heterogeneity into consideration. Compared with existing WSI-based survival prediction models, our method generate amounts of bags that can represent the heterogeneous patterns of WSIs for model training. We demonstrate the effectiveness of the proposed method with superior experimental results on three

Authorized licensed use limited to: SUN YAT-SEN UNIVERSITY. Downloaded on July 26,2022 at 09:43:40 UTC from IEEE Xplore. Restrictions apply.

different cancer datasets. BDOCOX is a general WSI-based learning framework, which can be used to predict the clinical outcome and the response to specific treatment for different cancer types. In addition, we will explore more ways based on the BDOCOX model for the integration of WSI with genomic data to gain new insights regarding the relationships between genomic data and WSIs as well as improved survival prediction.

REFERENCES

[1] J. Liu *et al.*, "An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics," *Cell*, vol. 173, no. 2, pp. 400–416, 2018.

[2] W. Fridman, L. Zitvogel, C. Fridman, and G. Kroemer, "The immune contexture in cancer prognosis and treatment," *Nature Rev. Clin. Oncol.*, vol. 14, no. 12, pp. 717–718, 2017.

[3] W. Shao *et al.*, "Integrative analysis of pathological images and multi-dimensional genomic data for early-stage cancer prognosis," *IEEE Trans. Med. Imag.*, vol. 39, no. 1, pp. 99–110, Jan. 2020.

[4] Z. Huang *et al.*, "SALMON: Survival analysis learning with multi-omics neural networks on breast cancer," *Frontiers Genet.*, vol. 10, pp. 166–176, Mar. 2019.

[5] J. Xu *et al.*, "Stacked sparse autoencoder (SSAE) for nuclei detection on breast cancer histopathology images," *IEEE Trans. Med. Imag.*, vol. 35, no. 1, pp. 119–130, Jul. 2015.

[6] H. Tu, Y. Liu, D. Turchinovich, M. Marjanovic, and E. Chaney, "Stain-free histopathology by programmable supercontinuum pulses," *Nature Photon.*, vol. 10, no. 8, pp. 534–535, 2016.

[7] K.-H. Yu *et al.*, "Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features," *Nature Commun.*, vol. 7, no. 1, pp. 1–12, Nov. 2016.

[8] M. Veta, J. P. W. Pluim, P. J. van Diest, and M. A. Viergever, "Breast cancer histopathology image analysis: A review," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 5, pp. 1400–1411, May 2014.

[9] P. Mobadersany *et al.*, "Predicting cancer outcomes from histology and genomics using convolutional networks," *Proc. Nat. Acad. Sci. USA*, vol. 115, no. 13, pp. E2970–E2979, Mar. 2018.

[10] S. Ryall, U. Tabori, and C. Hawkins, "A comprehensive review of paediatric low-grade diffuse glioma: Pathology, molecular genetics and treatment," *Brain Tumor Pathol.*, vol. 34, no. 2, pp. 51–61, Apr. 2017.

[11] D. Lin and Z. Ying, "Checking the COX model with cumulative sums of martingale-based residuals," *Biometrika*, vol. 80, no. 3, pp. 385–395, 1993.

[12] R. Tibshirani, "The lasso method for variable selection in the COX model," *Statist. Med.*, vol. 16, no. 4, pp. 385–395, Feb. 1997.

[13] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for Cox's proportional hazards model via coordinate descent," *J. Stat. Softw.*, vol. 39, no. 5, pp. 1–10, 2011.

[14] X. Zhu, J. Yao, and J. Huang, "Deep convolutional neural network for survival analysis with pathological images," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2016, pp. 544–547.

[15] P. Mobadersanya, S. Yousefi, M. Amgad, D. A. Gutman, and S. Barnholtz-Sloan, "Predicting cancer outcomes from histology and genomics using convolutional networks," *Proc. Nat. Acad. Sci. USA*, vol. 20, no. 4, pp. 11–21, 2018.

[16] J. Cheng, X. Mo, X. Wang, A. Parwani, Q. Feng, and K. Huang, "Identification of topological features in renal tumor microenvironment associated with patient survival," *Bioinformatics*, vol. 34, no. 6, pp. 1024–1030, Mar. 2018.

[17] X. Zhu, J. Yao, F. Zhu, and J. Huang, "WSISA: Making survival prediction from whole slide histopathological images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7234–7242.

[18] R. Li and J. Yao, "Graph CNN for survival analysis on whole slide pathological images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2018, pp. 174–182.

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

[20] A. Viganò, M. Dorgan, J. Buckingham, E. Bruera, and M. E. Suarez-Almazor, "Survival prediction in terminal cancer patients: A systematic review of the medical literature," *Palliative Med.*, vol. 14, no. 5, pp. 363–374, Jul. 2000.

[21] J. P. Sy and J. M. G. Taylor, "Estimation in a Cox proportional hazards cure model," *Biometrics*, vol. 56, no. 1, pp. 227–236, Mar. 2000.

[22] J. Katzman, U. Shaham, J. Bates, A. Cloninger, T. Jiang, and Y. Kluger, "DeepSurv: Personalized treatment recommender system using a cox proportional hazards deep neural network," 2016, *arXiv:1606.00931*. [Online]. Available: http://arxiv.org/abs/1606.00931

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[24] M. Mehryar and A. Rostamizadeh, *Foundations of Machine Learning* (Adaptive Computation and Machine Learning series). Cambridge, MA, USA: MIT Press, 2012.

[25] O. Chapelle and M. Wu, "Gradient descent optimization of smoothed information retrieval metrics," *Inf. Retr.*, vol. 13, no. 3, pp. 216–235, Jun. 2010.

[26] T. Katarzyna, P. Czerwińska, and M. Wiznerowicz, "The cancer genome atlas (TCGA): An immeasurable source of knowledge," *Contemp. Oncol.*, vol. 19, no. 1, pp. 68–74, 2015.

[27] J. Yao, X. Zhu, J. Jonnagaddala, N. Hawkins, and J. Huang, "Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks," *Med. Image Anal.*, vol. 65, Oct. 2020, Art. no. 101789.

[28] C. Zeltz, I. Primac, P. Erusappan, J. Alam, A. Noel, and D. Gullberg, "Cancer-associated fibroblasts in desmoplastic tumors: Emerging role of integrins," *Seminars Cancer Biol.*, vol. 62, pp. 166–181, May 2020.

[29] D. Visser and A. Eichten, "Paradoxical roles of the immune system during cancer development," *Nature Rev. Cancer*, vol. 6, no. 1, p. 24, 2006.

[30] W. Shao, J. Cheng, L. Sun, D. Zhang, and K. Huang, "Ordinal multi-modal feature selection for survival analysis of early-stage renal cancer," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2018, pp. 648–656.

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

[32] G. Gandaglia *et al.*, "Impact of the site of metastases on survival in patients with metastatic prostate cancer," *Eur. Urol.*, vol. 68, no. 2, pp. 325–334, Aug. 2015.

[33] J. Xie and C. Liu, "Adjusted Kaplan–Meier estimator and log-rank test with inverse probability of treatment weighting for survival data," *Statist. Med.*, vol. 24, no. 20, pp. 3089–3110, 2005.

Authorized licensed use limited to: SUN YAT-SEN UNIVERSITY. Downloaded on July 26,2022 at 09:43:40 UTC from IEEE Xplore. Restrictions apply.