Medical Image Analysis 65 (2020) 101789

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks

Jiawen Yao[a], Xinliang Zhu[a], Jitendra Jonnagaddala[b], Nicholas Hawkins[c], Junzhou Huang[a,*]

[a] Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX, USA
[b] School of Public Health and Community Medicine, University of New South Wales, Sydney, NSW, Australia
[c] School of Medical Sciences, UNSW Sydney, Sydney, NSW, Australia

https://paperswithcode.com/paper/whole-slide-images-based-cancer-survival#code

ARTICLE INFO

Article history:
Received 15 July 2019
Revised 29 June 2020
Accepted 16 July 2020
Available online 19 July 2020

Keywords:
Survival prediction
Multiple instance learning
Deep learning
Whole slide images

ABSTRACT

Traditional image-based survival prediction models rely on discriminative patch labeling which make those methods not scalable to extend to large datasets. Recent studies have shown Multiple Instance Learning (MIL) framework is useful for histopathological images when no annotations are available in classification task. Different to the current image-based survival models that limit to key patches or clusters derived from Whole Slide Images (WSIs), we propose Deep Attention Multiple Instance Survival Learning (DeepAttnMISL) by introducing both siamese MI-FCN and attention-based MIL pooling to efficiently learn imaging features from the WSI and then aggregate WSI-level information to patient-level. Attention-based aggregation is more flexible and adaptive than aggregation techniques in recent survival models. We evaluated our methods on two large cancer whole slide images datasets and our results suggest that the proposed approach is more effective and suitable for large datasets and has better interpretability in locating important patterns and features that contribute to accurate cancer survival predictions. The proposed framework can also be used to assess individual patient's risk and thus assisting in delivering personalized medicine.

© 2020 Elsevier B.V. All rights reserved.

1. Introduction

Survival analysis aims to analyze the expected duration of time until events happen. It tries to find the answer of questions like: how does the proportion of a population survive past a certain time (e.g. 5 years)? What rate will they die or fail? It is a very important clinical application and many efforts have been made to search for biomarkers from omics data that are significantly related to patient death (Shedden et al., 2008; Tibshirani et al., 1997; Bair and Tibshirani, 2004; Bair et al., 2006; Park and Hastie, 2007). Recent technological innovations are enabling scientists to capture big whole slide images (WSIs) at increasing speed and resolution for diagnosis. The learning model is required to correctly predict the survival risk of each patient from his/her tumor tissue whole slide images. The more precise is risk assessment for a cancer patient, the better the patient can be treated. Compare with genomics data, pathological images can present tumor growth and morphology in extremely detailed, gigapixel resolution which is extremely useful for cancer study (Warth et al., 2012; Yuan et al., 2012).

The diagnosis is extremely laborious and highly dependent on expertise which requires pathologists to carefully examine the biopsies under the microscope (Bejnordi et al., 2017). To reduce the risk of misdiagnosis, pathologists have to conduct a thorough inspection of the whole slide which make the diagnosis quite cumbersome. Automatic analysis of histology has become one of the most rapidly expanding fields in medical imaging. Computer aided diagnostics in digital pathology can not only alleviate pathologists' workloads, but also help to reduce the chance of diagnosis mistakes. However, using WSIs for survival prediction is very challenging due to several reasons: (1) pathological images in real cancer dataset might be in terabytes (10^{12} pixels) level which makes most models computationally impossible. (2) the large variations of textures and biological structures from tumor heterogeneity, As the solid tumor may have a mixture of tissue architectures and structures, multiple WSIs from different parts of the patient's tissue are collected for diagnosis; (3) label on patient-level while each patient might have multiple WSIs for diagnosis. Those terabyte-size large WSIs from one patient will share the survival label which will make the problem more challenging.

* Corresponding author.
E-mail addresses: jiawen.yao@mavs.uta.edu (J. Yao), jzhuang@uta.edu (J. Huang).

https://doi.org/10.1016/j.media.2020.101789
1361-8415/© 2020 Elsevier B.V. All rights reserved.

1.1. Related work

During recent years, many methods have been proposed for survival prediction using pathological slides. They can be categorized into two categories: ROI-based and WSI-based methods.

Region of interest analysis Pathological images usually come with a very high resolution which makes most of exiting models and algorithms computationally infeasible even though the high resolution of image data greatly benefits survival analysis with more precise information. Previously due to the lack of computational power, most of the literature focused on regions of interest (ROI) patches which are selected by pathologists from WSIs (Gurcan et al., 2009).

Instead of handling original WSIs, ROI-based methods extracted hand-crafted features from ROIs for predictions (Yuan et al., 2012; Barker et al., 2016; Zhu et al., 2016b; Yao et al., 2016; Wang et al., 2014; Yu et al., 2016; Cheng et al., 2017; Yao et al., 2015). Wang et al. (Wang et al., 2014) proposed a novel framework to first segment cells in annotated patches and then perform cellular morphological properties from those cells which result in 166 imaging features. Yu et al. (Yu et al., 2016) extract 9879 quantitative image features from annotated regions of interest and results suggest that automatically derived image features can predict the prognosis of lung cancer patients and thereby contribute to precision oncology. Beyond classical cell detection, Yao et al. (Yao et al., 2016) used a deep subtype cell detection first to classify different cell subtypes and then extracted features from cellular subtype information. Cheng et al. (Cheng et al., 2017) used a deep auto-encoder to cluster cell patches into different types and then extracted topological features to characterize cell type distributions from ROIs for prediction. These methods extracted hand-crafted features based on nuclei detection and segmentation and those features were considered to represent prior knowledge of boundary, region or shape. However, hand-crafted features are limited in representation power and capability.

Recently, with the advance of deep neural networks, deep learning-based survival models are proposed for seeking more powerful deep representation (Katzman et al., 2016; Zhu et al., 2016a; Yao et al., 2017; Mobadersany et al., 2018). Katzman *et al.* first proposed a deep fully connected network (DeepSurv) to represent the nonlinear risk function (Katzman et al., 2016). They demonstrated that DeepSurv outperformed the standard linear Cox proportional hazard model. Another improvement is deep convolutional survival learning (DeepConvSurv) which is the first attempt to use pathological images in deep survival model (Zhu et al., 2016a). Later, Yao et al. (Yao et al., 2017) integrated genome modality with DeepConvSurv for survival prediction using multi-modality data. However, DeepConvSurv is designed to use pre-selected ROI patches by pathologists from WSIs for convolution operations. A small set of image tiles might not completely and properly reflect the patients' tumor morphology. Also, those methods perform average pooling to achieve patient-wise predictions from patch-based results. Such combination cannot effectively aggregate predictions from patch-level and needs further attention. Thus, it would be much helpful if we can facilitate knowledge discovery from big whole slide images.

Whole-slide image analysis With detailed and densely annotations on WSIs, nowadays a series of approaches in whole-slide image analysis have been proposed for a variety of applications including classification, detection or segmentation (Wang et al., 2016; Bejnordi et al., 2017; Kong et al., 2017; Li and Ping, 2018; Liu et al., 2017). Applying deep learning for supervised learning on computational pathology has achieved promising results. However, the applicability of these models in clinical practice remains in questions because of the wide variance of clinical samples. Extensive and time-consuming human manual annotations in clinical practice is impossible. Moreover, the success of those applications is built on integrating detailed patch contents and using labor-extensive annotations which might not be applicable for survival prediction.

To properly address the shortcomings of current models, one possible direction is to consider weakly supervised manner. Recently, researchers have developed many weakly supervised algorithms to medical images including weakly-supervised X-rays screening (Wang et al., 2019; Yan et al., 2018) and WSI classification (Hou et al., 2016; Mercan et al., 2018; Wang et al., 2018; 2019; Wang et al., 2019). WSI classification models are designed to find the most differentiated regions correspond to different tumor types. A two-step approach is usually used and the first step is a classifier at the tile level and then predicted scores for each tile within a WSI are aggregated with various strategies. However, learning survival from histology and developing prognosis model is considerably more difficult as risk is often reflected from a range of histology patterns that correspond to varying degrees of disease progression. Tumor heterogeneity plays an important role in cancer study which includes inter-tumor and intra-tumor heterogeneity (Jamal-Hanjani et al., 2015). Inter-tumor heterogeneity refers to the differences found between tumors in different patients. Intra-tumor heterogeneity refers to distinct tumor cell populations within the same tumor specimen. Most recent weakly-supervised WSI classification focused on localizing most differentiated regions correspond to tumor types across patients. Therefore, they are more likely to capture inter-tumor heterogeneity between tumors or subtype tumors. Understanding how to label one person's tumor type may not be enough to study the degree of tumor progression. The pathophysiology of tumor progression and proliferation is complex and thus a new image-based prognosis model which can integrate information from heterogeneous tissue regions is a better approach. Additionally, existing weakly-supervised WSI classification task is at slide-level, while survival prediction is at patient-level analysis (one patient might have multiple whole slide images). Devising patient-level decisions from slide-level results is not the objective of those studies. To achieve survival prediction from whole slide images without using annotations, Zhu et al. (2017) proposed a patch-based two-stage framework to predict patients' survival outcomes. Patches are extracted from the WSIs and clustered to different patterns defined as "phenotypes" according to their visual appearances in the first stage. Then WSISA (Zhu et al., 2017) adopted DeepConvSurv (Zhu et al., 2016a) to select important patch clusters and then aggregated those clusters for final prediction. Although this framework has practical merits to consider important patch clusters, it is hard to incorporate it into state-of-the-art deep learning paradigm as the whole approach has separate steps. In addition, it is not a scalable solution because the first stage will be significantly inefficient if more patches are sampled. One recent work (Tang et al., 2019) proposed CapSurv by introducing Capsule network (Sabour et al., 2017). However, CapSurv still has similar issues with WSISA as the main framework is following the WSISA pipeline. The relationship of tissue patterns on WSI is the great importance on survival analysis. Li et al. (2018) proposed a graph convolutional network (GCN) based method to consider such relationship of patches in the WSI and then learn effective representation for survival prediction. However, this method requires detailed graph structure knowledge to construct a complete graph representation for effective GCN training which is not flexible and needs prior knowledge.

1.2. Contributions

Though many works can be found on WSI analysis for segmentation, classification and detection, there were limited works on

weakly-supervised learning for survival prediction. Based on the literature review, a method that can adaptively learn patient-level representations with limited prior knowledge is needed. In this study, we propose a novel framework, referred to as Deep Attention Multiple-Instance Survival Learning (DeepAttnMISL) for whole slide images. In contrast to the standard supervised learning, multiple instance learning (MIL) considers a set of bags, each containing multiple feature vectors referred to as instances. The available label is only assigned to bag-level and labels of individual instances in the bag are not known. In MIL, not all the instances are necessarily relevant and some of them in the bag might not be relevant to certain labels. In observation, if the slide is from a low risk patient, most of its tiles might be benign and or contain low-grade tumor. In contrast, if the slide is from the high risk patient, it must be true that at least one of all of the possible tiles contains malignant tumor. This formalization of the WSI survival learning problem is an example of the general standard multiple instance assumption and thus MIL is a good to fit to solve such problem.

Our preliminary work that only using deep multiple instance learning can help achieve better prognosis performance was published in MICCAI 2019 (Yao et al., 2019). Compared to the previous work, we offered new contributions in following aspects. We introduced attention mechanism into deep multiple instance survival learning. The proposed DeepAttnMISL not only uses the siamese MI-FCN network to learn features from different phenotype clusters, but also largely improves performance with Attention-based MIL pooling layer to perform a trainable weighted aggregation. More importantly, the proposed framework can effectively highlight the prognosis-related clusters and has better interpretability as well as performance than our preliminary work (Yao et al., 2019). The contributions can be summarized as follows.

- Phenotype clusters provide morphology-specific representation, the proposed DeepAttnMISL first extracts phenotype-level information through a Siamese MIL-based network from patch-level features. The attention mechanism is then used to aggregate these phenotype features into patient-level information with a trainable weighted average where weights can be fully parameterized by neural networks. Such attention-based aggregation is much flexible than fixed pooling operators in recent work (Yao et al., 2019; Zhu et al., 2017; Tang et al., 2019).
- With the advantage of MIL and attention mechanism, the proposed model has a good interpretability to find important patterns of patients. Those identified important regions and patches are more likely to be associated with prognosis and overall the proposed model can achieve better patient-level predictions and improve prediction performance than our previous work (Yao et al., 2019).
- To evaluate the performance of the proposed DeepAttnMISL model, two large WSI datasets on lung and colorectal cancer are used and extensive experimental results verify the effectiveness.

Our method can efficiently exploit and utilize all discriminative patterns in whole slide pathological images to perform accurate patients' survival predictions. Additionally, we present results representing a patient's treatment group to illustrate how to view the proposed model as a treatment recommender system. Results validate that the proposed model can accurately model the risk functions of the population and thus guide treatment decisions for improving patient lifespan.

2. Methodology

Considering a set of N patients, $\{X_i\}, i = 1 \ldots N$, each patient has the follow-up label (t_i, δ_i) indicating the overall survival. The observation time t_i is either a survival time or a censored time for each patient. δ_i is the indicator which is 1 for an uncensored instance (death occurs during the study) and 0 for a censored instance. Survival model predicts a value of a target variable O for a given patient. As we discussed above, patient X_i will have multiple WSIs and our goal is to predict the corresponding target o_i from those imaging data. As we don't have pixel-level annotations but only know patient-level information, this weakly-supervised learning can be solved by Multiple Instance Learning (MIL).

In the case of MIL problem, patient X is a bag of instances, $X = \{x_1, \ldots, x_C\}$ and the number of instances C could vary for different bags. Furthermore, we assume that individual true labels exist for the instances within a bag, i.e., $y_1, \ldots, y_C$ but those values remain unknown during training. One very important assumption is that neither ordering nor dependency of instances within a bag and a MIL model must be permutation-invariant. Instances within the bag can be defined as sampling patches from WSIs and several studies (Campanella et al., 2019; Wulczyn et al., 2020) developed MIL-based deep learning approaches for automated cancer diagnosis and prognosis. In our case, we introduce phenotype cluster as the instance of the bag instead of individual patch. Cancer histology contains rich phenotypic information that reflects underlying molecular processes and disease progression. Phenotype of the pathological slides is a combination of tissue's various observable characteristics. This provides a convenient visual representation of disease aggressiveness. Recent studies have shown phenotypic information could be useful for prediction of prognosis (Zhu et al., 2017; Mobadersany et al., 2018). The purpose of the proposed framework is to predict patient outcomes from whole slides images. The study involves partitioning the original slides into a number of phenotype patterns. Each phenotype describes a type of histology pattern and includes a number of smaller patches or tiles.

instanc
es=pat
ches

2.1. DeepAttnMISL

Fig. 1 shows the overview of the proposed Deep Attention Multiple Instance Survival Learning (DeepAttnMISL). In Multiple Instance Learning, each data sample is a bag of instances and the bag can be seen as one patient in our problem. Each patient X_i may contain multiple whole slides and it is not practical to use whole slides as instances due to the extreme large size. We choose phenotypes instead of raw sampling patches as instances within the bag because it will considerably reduce the complexity of the problem as the number of heterogeneous patches is actually very huge. By using phenotype patterns which are constructed by clustering, we can build the model for different types of tissues to extract morphology-specific features. To learn patient-level information from phenotype clusters, we design a Multiple Instance Fully Convolutional Network (MI-FCN) running inside our deep learning architecture with weights being shared among them as in the siamese architecture. To detect important phenotypes associated with patients' clinical outcomes, attention-based MIL pooling layer is used to aggregate phenotype-level representation. The output is the hazard risk to represent how well for the patient behaves in the population of certain type of diseases.

2.1.1. Sampling and clustering

At the first step, we extract patches from all WSIs belong to the same patient and then cluster them into different phenotypes. To capture detailed information of the images, those patches are extracted from 20X (0.5 microns per pixel) objective magnifications and then fixed to 500 × 500 × 3 size. In one whole slide image, usually about 50% of areas are background and it is easy to select regions to contain tissues rather than background or irregular regions according to pixel values. Even we only extract tissue patches and ignore background regions, it can still get tens of

利用聚类，将所有patches重构，构成新的batch，并且，每个簇是一个bag，多个bag构成一个batch，每个bag里面有多个示例，每个示例是4096-dim的embedding feature

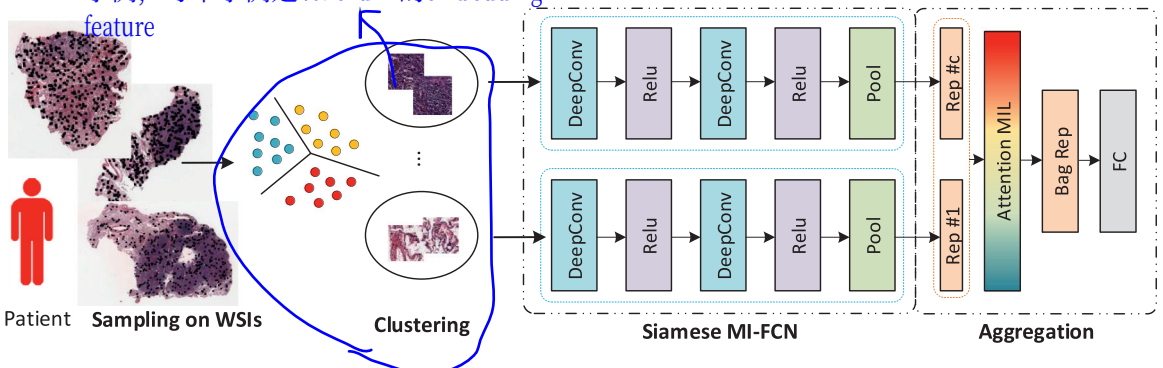
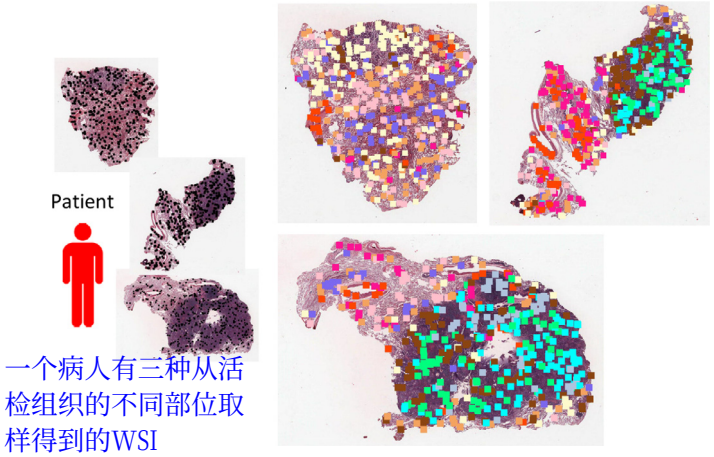


Fig. 1. An overview of the proposed DeepAttnMISL model.



一个病人有三种从活检组织的不同部位取样得到的WSI

Fig. 2. Phenotype patterns visualization after clustering on three WSIs belong to the same patient.

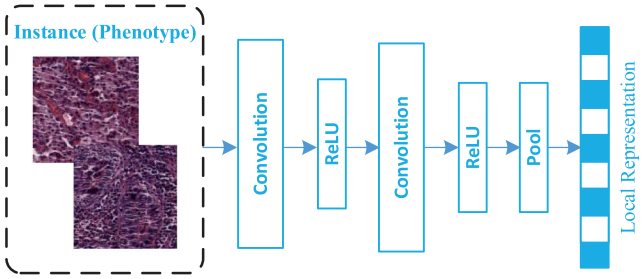


Fig. 3. The network architecture in each MI-FCN.

thousands of patches per WSI which will result in a huge number of images from the whole dataset. Different from recent segmentation and detection task in whole slide image analysis, our task is for patient-level decision aggregated from patch-level results. As pointed out in Hou et al. (2015), training patch-based CNNs for weakly supervised learning is very time costly (several weeks) and we propose to use features from pre-trained models instead of using CNNs to learn features from the scratch. We use the pre-trained model (e.g. VGG) from ImageNet (Simonyan and Zisserman, 2014) to extract features for each image patch which have more representation power than smaller size (50 × 50) thumbnail images to represent their phenotypes (Zhu et al., 2017). Then we adopt K-means clustering to cluster patches based on their deep learning features. Notice that one patient might have multiple WSIs and we actually perform clustering on patient-level instead of the whole database. Fig. 2 shows one patient's example and this patient has three WSIs that were sampled from different locations of the biopsy tissue. The corresponding phenotype clustering are shown in the right and each color means one type of phenotype clusters. In this example, we chose to cluster 10 phenotype patterns. The results show the effectiveness of this strategy as we can see similar patches are grouped into the same cluster. This could demonstrate that features from pre-trained model are capable of identifying patterns of whole slide images and we would expect them to be distinctive and informative for later survival learning task.

By clustering different patches from all WSIs of the patient into several distinguished phenotype groups, we will have different phenotype groups with various prediction powers on this patient's clinical outcome. The proposed DeepAttnMISL takes phenotypes as multiple inputs and consider their connections for predicting survival outcomes. 考虑他们之间的联系来预测未来的结果

2.1.2. Siamese MI-FCN

After clustering, the patient is a set of phenotype clusters and we design a siamese Multiple Instance Fully Convolutional Networks (MI-FCN) to learn features from those patterns, similar to the work in Yao et al. (2019). Most existing well-known pre-trained models were trained based on single-instance bases, and the labels are associated with each image which is not the case of our problem. We embed multiple sub-networks running inside our deep learning architecture with weights being shared among them as in the siamese architecture. Each sub-network is based on fully convolutional neural networks (FCN) that can learn informative representation for individual phenotype of the patient.

The architecture of each Multiple Instance Fully Convolutional Networks (MI-FCN) is shown in Fig. 3. The combination of multiple layers of fully convolutional layers and non-linear activation functions has proven to be a powerful non-linear feature mapping in multiple instance problem (Yang et al., 2017). The reason to use the fully convolutional networks (FCN) without including any fully connected layers is that FCN is more flexible and can handle any spatial resolution, which is needed for the considered problem since the number of patch samples in each phenotype varies. For each phenotype, the input is a set of features from m_i patches, can be organized as $1 \times m_i \times d$ (d is the feature dimension or channel). The network consists of several layer-pairs of 1×1 conv layer and ReLU layer (we show 2 pairs in Fig. 3). The global pooling layer (e.g. average pooling) will be added at the end. For jth phenotype, its representation is denoted as $\mathbf{r}_j$. The network receives one kind of phenotypes (tensor) as input and it can focus on local information and generate representation for the phenotype. Since the number of patches in each phenotype varies, the fully convolutional network is more flexible to handle this scenario.

每个簇里面的patches数量不一致，每个簇/bag的size为[1, m_{i},d]

C个示例数

Table 1
The architecture of DeepAttnMISL.

Layer	Input	Output size
MI-FCN i	$1 \times m_i \times d$	64 ($\mathbf{r}_i$)第i个表型
Attention MIL pooling	64 × C	64
Fully-Con.	64	32
Fully-Con.	32	1 ($\mathbf{o_i}$)

在attention之前有一个stack，将所有簇的embedding连起来

2.1.3. Aggregation via attention-based MIL pooling layer

Local representations from MI-FCN encode information of the corresponding phenotype clusters and how to aggregate them into patient-level representation is one necessary step. Let $R = \{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_C\}$ be one patient with C phenotype local representations and the goal is to get patient-level representation $\mathbf{z}$. The very straightforward choice is to use maximum or the mean operator, but drawbacks are very clear that they are pre-defined and non-trainable which might not be flexible and adjustable to the specific task. Previous work (Zhu et al., 2017) used weighted average of features from clusters to get the patient feature but they performed such patient-level aggregation in a separate stage and the whole approach cannot be trained end-to-end from instance-level to patient-level. A better way to integrate phenotype-level information is to leverage an attention mechanism that considers the importance of each phenotype. In this paper, we propose to use the attention-based MIL pooling (Ilse et al., 2018) for aggregation which is flexible and adaptive. By using such pooling operator, the patient-level representation can be calculated as

$$\mathbf{z} = \sum_{k=1}^{C} a_k \mathbf{r}_k, \tag{1}$$

where

$$a_k = \frac{\exp\{\mathbf{w}^\top \tanh(\mathbf{V}\mathbf{r}_k^\top)\}}{\sum_{j=1}^{C} \exp\{\mathbf{w}^\top \tanh(\mathbf{V}\mathbf{r}_j^\top)\}}. \tag{2}$$

In the weight a_k calculation, $\mathbf{w} \in \mathbb{R}^{L\times 1}$ and $\mathbf{V} \in \mathbb{R}^{L\times M}$ are trainable parameters. Tangent $\tanh(.)$ element-wise non-linearity is introduced both negative and positive values for proper gradient flow. The attention-based MIL pooling allows to assign different weights to phenotype clusters within one patient and hence the final patient-level representation could be highly informative for survival prediction. In other words, it should be able to locate key clusters and provide potential ROIs. Different from traditional attention mechanism that all instances are sequentially dependent (Lin et al., 2017; Raffel and Ellis, 2015), multiple instance learning assumes all instances are independent. As phenotype in our problem is more natural to be independent to each other, attention mechanism used in MIL pooling will be beneficial to achieve good results.

2.1.4. Loss function

After attention-based MIL pooling, we will generate the patient-level aggregation from all local representations. For ith patient sample passing through the proposed model, the output of this patient's hazard risk is denoted as $\mathbf{o}_i$. Table 1 presents architecture details of the proposed DeepAttnMISL. Input of our model is the set of patients' phenotype features, organized as $[(1 \times m_1 \times d), (1 \times m_2 \times d), \ldots, (1 \times m_c \times d)]$ where C is the number of phenotypes and m_i means the number of patches in ith phenotype.

如果num_clusters=C=10，那么就是聚十类，第c类的instances=patches数是m_{c}

Denote the label of the ith patient as (t_i, δ_i) where t_i is the observed time, We assume that censoring data ($\delta = 0$, death not observed) is non-informative in that, given $\mathbf{x}_i$, the event and censoring time for the jth patient are independent. Let $t_1 < t_2 < \ldots < t_N$ denote the ordered event times. The risk set $R(t_i)$ is the set of all individuals who are still under study. For example, the patient j in risk set has the survival time is equal or larger than t_i ($t_j \geq t_i$). Conditioned upon the existence of a unique event at some particular time t the probability that the death event occurs in the patient i is

$$L_i = \frac{\exp(\mathbf{o}_i)}{\Sigma_{j\in R(t_i)} \exp(\mathbf{o}_j)}, \tag{3}$$

Assuming the patients' events were statistically independent, the joint probability of all death events conditioned upon the existence of events at those times is the partial likelihood:

$$L = \prod_{i:\delta_i=1} \frac{\exp(\mathbf{o}_i)}{\Sigma_{j\in R(t_i)} \exp(\mathbf{o}_j)}, \tag{4}$$

The corresponding log partial likelihood is

$$\begin{aligned} l = \log(L) &= \sum_{i:\delta_i=1} \left(\mathbf{o}_i - \log \sum_{j:R(t_i)} \exp(\mathbf{o}_j) \right) \\ &= \sum_i \delta_i \left(\mathbf{o}_i - \log \sum_{j:R(t_i)} \exp(\mathbf{o}_j) \right), \end{aligned} \tag{5}$$

The function can be maximized over the network parameters to produce maximum partial likelihood estimates. It is equivalent to minimize the negative log partial likelihood. We then use the negative log partial likelihood as the loss function in our model as shown in below

$$L(\mathbf{o}_i) = \sum_i \delta_i \left(-\mathbf{o}_i + \log \sum_{j:t_j>=t_i} \exp(\mathbf{o}_j) \right). \tag{6}$$

In a simplified view, the loss function contributes to overall concordance by penalizing any discordance in any values of higher risk patients if they are greater than lower those of lower risk. Different with other deep models used the same loss function (Katzman et al., 2016; Zhu et al., 2016a; 2017Huang and Zhan, 2019), the proposed model can better fit realistic patients' whole slide imaging data and learn complex interactions using deep multiple instance representation that cover both holistic and local information. Since patient's risk is correlated with phenotypes from WSIs, the proposed framework can efficiently exploit phenotypes by deep multiple instance learning and attention mechanism for clinical outcome prediction at patient-level.

3. Experiments

3.1. Dataset description

To validate the performance of the proposed DeepAttnMISL, we used two very large datasets on lung and colorectal cancers with high-resolution WSIs. They are the National Lung Screening Trial (NLST) (Team et al., 2011) and the Molecular and Cellular Oncology (MCO) study (Ward and Hawkins, 2015; Jonnagaddala et al., 2016). 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). The MCO study is a collection of imaging, specimen, clinical and genetic data from over 1500 Australian individuals who underwent curative resection for colorectal cancer from 1994 to 2010. Clinical and pathological data were collected on all those cases, including follow-up data. The WSIs collection in MCO study consists of more than 1500 WSIs representing at least one typical section from each tumour case, stained with Hematoxylin and eosin, and scanned using a 40x objective. We have different experiment comparison settings on two datasets because we only have annotations that locate tumor regions in NLST. Both datasets are good for WSI-based models

Table 2
The numbers of WSIs, patients, patches, and the average number of patches per WSI extracted in each dataset.

Dataset	NLST	MCO_130 K	MCO_1M
#patients	387	1146	1146
#WSIs	1177	1614	1,614
#patches	275,244	132,910	915,324
#patches/WSI	234	82	567

as those models without requiring ROI labelling but more extensive experiments with ROI-based comparisons can only be made on NLST dataset.

The numbers of WSIs and patients in each dataset are shown in Table 2. State-of-the-art WSI models (Zhu et al., 2017; Tang et al., 2019) need to control the scale of data as they will have significant computational issues on the very large number of patches. They sampled hundreds of patches per WSI and collected around 20 K–200 K patches in total. One advantage of the proposed model is the computational efficiency because it uses MIL with attention to aggregate 1D deep features from pre-trained models instead of training patch-based CNNs which is very time costly (Hou et al., 2015). For the purpose of training baseline WSI survival model, we first extract in total of 130 K and 275 K patches for MCO and NLST, respectively. We then sample more patches on MCO dataset and collect 915 K patches and each WSI will have more than 500 patches. MCO study has more than 1000 patients which is much larger than data used in recent work (Zhu et al., 2017; Tang et al., 2019).

3.2. Implementation details

For training, we use Adam optimization with weight decay 5×10^{-4}. The learning rate is set to 10^{-4} and the training monitors the loss on validation dataset and it will early stop if the loss goes increased much. To evaluate the performances in survival prediction, we take the concordance index (C-index) and area under curve (AUC) as our evaluation metrics (Heagerty and Zheng, 2005). The C-index quantifies the ranking quality of rankings and is calculated as follows

$$c = \frac{1}{n} \sum_{i \in \{1 \ldots N | \delta_i = 1\}} \sum_{t_j > t_i} I[f_i > f_j] \tag{7}$$

where n is the number of comparable pairs and $I[.]$ is the indicator function. $t.$ is the actual time observation. $f.$ denotes the corresponding risk. The value of C-index ranges from 0 to 1. The larger the value is, the better the model predicts.

3.3. MCO results

3.3.1. Settings and parameters

To see effects from phenotype patterns, we tested different cluster numbers changing from 6 to 12. We split the data into 80% training and 20% testing. 10% of training data will be used as validation data for achieving early stop training. We would like to note that the number of phenotype clusters is the maximum number that allows each patient sample can have. The proposed model is flexible to handle patients with fewer patterns (e.g. smaller biopsy tissue). We implement this by setting the corresponding weight d_k to zero if there are no patches in this cluster. To evaluate the use of different pooling ways, we built two baselines by replacing attention MIL pooling layer in DeepAttnMISL with commonly used Max and Mean pooling layer, and we indicate them as "DeepMIL + Max/Mean" below. Table 3 presents results of each model. We first notice DeepAttnMISL can achieve best results in all cases which demonstrate attention MIL pooling is more flexible and better than fixed pooling operators. Second, when the phenotype is set to large values, results get worse which show more clusters actually cannot guarantee prediction benefits.

Table 3
Performances with different number of phenotypes.

Model	c = 6	c = 8	c = 10	c = 12
DeepAttnMISL	**0.652**	**0.648**	**0.624**	**0.607**
DeepMIL + Max	0.594	0.606	0.606	0.540
DeepMIL + Mean	0.604	0.578	0.604	0.601

Table 4
Results under different network configurations on testing data. The cluster number is set to 6.

c = 6	1 layer	2 layers	3 layers
Global Average Pooling	**0.652**	0.634	0.644
Global Max Pooling	0.615	0.640	0.623

Table 5
Validation of Siamese on MCO-130K dataset.

c = 6	fold 1	fold 2	fold 3	fold 4	fold 5
No Siamese	0.564	0.538	0.515	0.527	0.564
w Siamese	0.652	0.579	0.609	0.564	0.573

Table 6
Results with different feature extractor and clustering on MCO-130K dataset.

c = 6	fold 1	fold 2	fold 3	fold 4	fold 5
InceptionV3+k	0.670	0.540	0.630	0.603	0.549
InceptionV3+sp	0.630	0.565	0.611	0.554	0.607

The basic MI-FCN network of our DeepAttnMISL consists of one convlolutional layer, one ReLU layer, one pooling layer. We study the effects of different number of convolution and ReLu layer-pairs and report results in Table 4. For the 1 layer, we used 64 filters in the convlolutional layer. We used {2048, 64} number of filters in 2 layers and {2048, 1024, 64} for 3 layers setting, respectively. From the table, we decide to choose one convolutional-ReLU layer pair with Global Average Pooling in MI-FCN network.

To validate the effectiveness of Siamese, we then remove the Siamese network and only use attention pooling layer on input features. In this case, no phenotype clusters are considered. This scenario will be the direct application of attention aggregation without using phenotype clusters (Ilse et al., 2018). 5 fold cross-validation is performed with the cluster number 6 on MCO-130K. Results can be found in Table 5. We can see the overall performance is not good as the DeepAttnMISL which means the importance of Siamese network. This validates the effectiveness of phenotype clusters in Siamese network. The final c-index across 5 folds is 0.542 ± 0.022 for model without Siamese and 0.595 ± 0.036 for model with Siamese, respectively. Results suggest the usefulness of phenotype patterns and the Siamese architecture.

To validate effects of different components, we add more evaluations by changing encoder/clustering part, and results can be found in the Table 6. The more advanced InceptionV3 (Szegedy et al., 2016) model is tested and we also introduce spectral clustering as the alternative method for Kmeans. All other settings and architectures are kept the same. Details about each fold can be seen in the Table 6. For model with InceptionV3 and Kmeans clustering, C-index result is 0.598 ± 0.054 on 5-fold cross validation. When changing Kmeans clustering to spectral clustering, the performance is 0.593 ± 0.032. Compared with the model using VGG-16 and Kmeans clustering (0.595 ± 0.036),

Table 7
Results of different attention mechanisms on MCO-130K dataset.

	Gated-Attention	Attention
c = 6	0.596 (0.029)	0.595 (0.036)
c = 8	0.586 (0.043)	0.599 (0.049)
c = 10	0.561 (0.048)	0.585 (0.036)
c = 12	0.579 (0.031)	0.591 (0.026)

Table 8
C-index values of the proposed model and WSISA with different settings.

Method	Settings	c = 6	c = 8	c = 10	c = 12
DeepAttnMISL	130 K	0.595	0.599	0.585	0.591
	1 M	**0.606**	**0.600**	**0.603**	**0.599**
DeepMISL	130 K	0.557	0.547	0.587	0.543
	1 M	0.569	0.575	0.573	0.567
W-MTLSA	130 K	0.558	0.567	0.524	0.547
W-LassoCox	130 K	0.552	0.546	0.503	0.523
W-EnCox	130 K	0.552	0.545	0.504	0.522

Table 9
AUC values of the proposed model and WSISA with different settings.

Method	Settings	c = 6	c = 8	c = 10	c = 12
DeepAttnMISL	130 K	0.623	**0.640**	**0.636**	0.622
	1 M	**0.644**	0.638	0.633	**0.637**
DeepMISL	130 K	0.564	0.552	0.590	0.547
	1 M	0.570	0.587	0.579	0.576
W-MTLSA	130 K	0.560	0.560	0.531	0.555
W-LassoCox	130 K	0.531	0.541	0.495	0.495
W-EnCox	130 K	0.532	0.544	0.497	0.496

performances from different variants of models are quite similar. Therefore, we decide to use VGG-16 and Kmeans clustering for comparisons.

We also try with the more advanced gating mechanism (Ilse et al., 2018; Dauphin et al., 2017) together with *tanh*(.) nonlinearity in eq (2). Results on MCO-130K are reported in Table 7. We can find gated-attention and plain attention mechanism behave similarly in different phenotype cluster settings but the plain attention is slightly better.

3.3.2. Comparisons

WSISA (Zhu et al., 2017) is one representative WSI-based survival learning but it only extracts features from WSIs and needs a separate survival learning to get final predictions. We choose three top survival models according to settings in WSISA (Zhu et al., 2017), they are Lasso-Cox (Tibshirani et al., 1997), En-Cox (Yang and Zou, 2012) and MTLSA (Li et al., 2016). As WSISA has the computational issue when there are too many patches in the whole dataset and thus the scale of 100K-200K patches is acceptable for experiments. We have a another collection of patches with around 1 million patches to see effects from the patch scale but only perform our model on this scale because training with WSISA is not endurable.

Our preliminary work DeepMISL (Yao et al., 2019) has shown the effectiveness of using both global and local representation from Multiple Instance Learning can benefit survival prediction. However, the model still treats phenotype clusters equally and cannot recognize clusters that contribute more on patients' survival. We perform 5 fold cross-validation and report the average values of C-index and AUC on Table 8 and 9, respectively. From both tables, one can see that the proposed method achieves best results than models using WSISA features in all cluster number settings on MCO-130k. Improvements can be related to the following differences. First, clustering is performed on patient-wise while recent WSI-based approaches (Zhu et al., 2017; Tang et al., 2019) need to cluster on all patches from patients of the database. Because WSISA (Zhu et al., 2017) needs independent DeepConvSurv to select important clusters and it has to divide the whole dataset into different types by clustering on all patches. DeepMISL (Yao et al., 2019) can combine both local and bag representation with MIL but it is still unable to treat phenotype clusters differently which will limit its use on larger datasets. With the advantage of MIL and attention mechanism, the proposed DeepAttnMISL can easily find important instances (clusters) within the bag are more likely to achieve better patient-level predictions. There is no need to perform clustering on the whole dataset. A trainable and adaptive attention-based MIL pooling in DeepAttnMISL can adjust to a task and data which could help succeed in calculating the better patient representation. Increases with 1%-3% are observed when we use more patches from MCO-1M data and this reminds us more patches can benefit predictions but actually cannot offer significant improvements. This demonstrates the robustness of the proposed DeepAttnMISL that is not rely on the number of sampling patches.

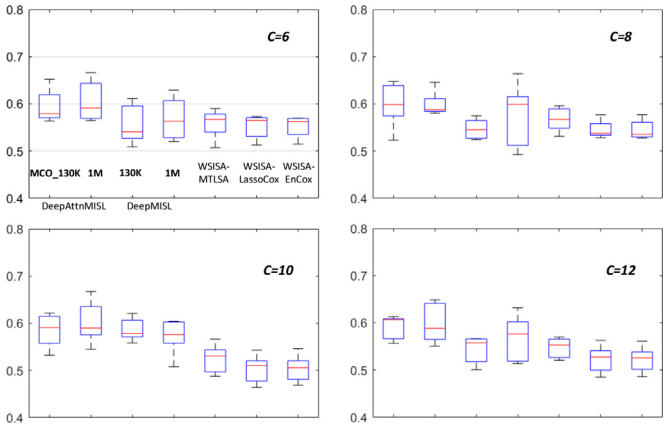


Fig. 4. Boxplots of C-index values with different numbers of phenotype patterns.

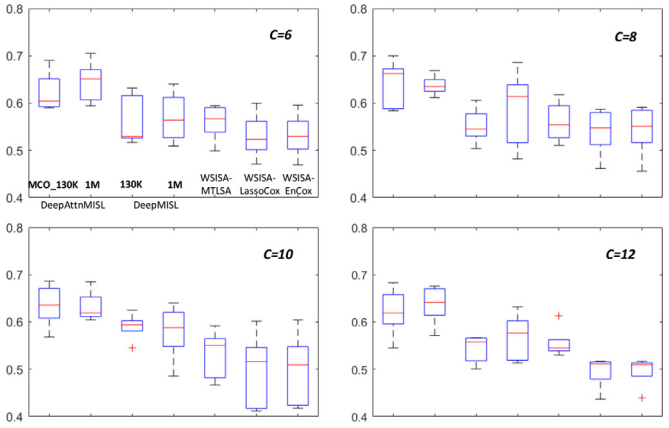


Fig. 5. Boxplots of AUC values with different numbers of phenotype patterns.

Fig. 4 and Fig. 5 present boxplots of C-index and AUC values from each model with different phenotype cluster numbers. We only show captions in the top left figure and others will also share this description. We can see that results of our method on MCO-1M and MCO-130K don't have significant differences. This shows sampling strategies will not affect final results of the proposed method in cross-validation settings. One can observe that our models consistently perform better than WSISA models across different phenotype cluster numbers.

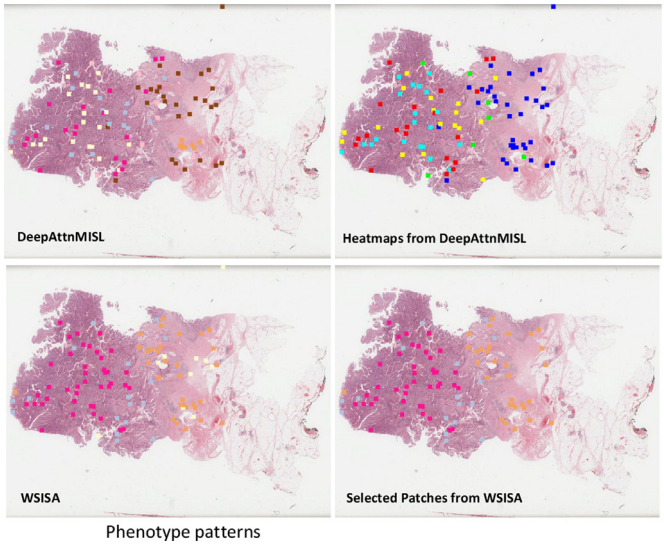


Fig. 6. Comparison of phenotype patterns distribution in the first column. The second column shows heatmap and selected patches from the proposed model and WSISA on MCO-130K, respectively.

Fig. 6 visualizes clustered phenotype patterns and selected patches from DeepAttnMISL and WSISA on MCO-130K when cluster number is set as 6. The first row shows results from DeepAttnMISL while the second one presents results from WSISA. In MCO-130K, around 100 patches per WSI are sampled and it clearly can see that clustering based on VGG-16 features is capable of identifying patches from different layers of WSI and grouping similar patches into the same category. The most important advantage of DeepAttnMISL is its good interpretability and we create a heatmap by showing the corresponding attention weight of each phenotype cluster. We rescaled the attention weights using $a_k^{'} = (a_k - \min(a))/(\max(a) - \min(a))$. Red color indicates the highest attention weight while blue means the lowest values. From the obtained heatmap, we can see the proposed approach can identify higher risk regions properly because most of patches with high attention weights are from tumor regions. When we look at selected patches from WSISA, we can observe that many patches from non-tumor regions are also selected. That is because WSISA selects clusters based on patches from the whole database and thus it cannot guarantee reliable selection on the specific patient due to the heterogeneity across patients.

More clear visualizations can be found in Fig. 7 on MCO-1M set and more patches (about 1000) are sampled per WSI. The first column shows phenotype patterns from the proposed model with different numbers. The second column shows the corresponding heatmaps. Attention mechanism in DeepAttnMISL allows to easily interpret the provided decision in terms of instance-level labels. From heatmaps, we can see results from $c = 6$ and $c = 8$ look better as most patches from cancerous regions are given by high attention weights.

To better visually validate the effect of attention mechanism, we collect and examine the attention weights as well as their corresponding patch images on MCO-1M data in Fig. 8. The bottom shows randomly selected patches from each phenotype and the frame colors of patches correspond to pattern colors in Fig. 8(b). We use threshold as 0.8 to only show patterns with higher attention weights in Fig. 8d. Each color represents each phenotype pattern of the whole slide image and we can see the proposed model has higher interest on patches more related to tumor regions. Relative low attention weights are given to normal tissue regions. More surprisingly, the model can also give low attentions on background

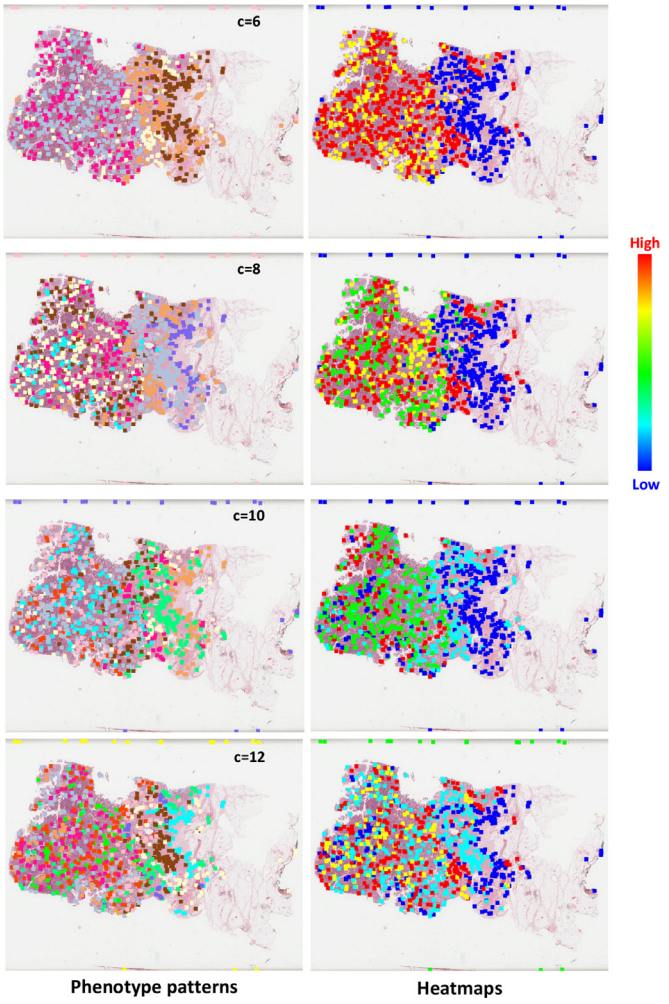


Fig. 7. Phenotype patterns clustering visualizations and the corresponding heatmaps from the proposed model on MCO-1M.

regions as they don't provide any information and are noisy images. Fig. 9 shows another example. From the figure, we can see most patches from tumor regions are found and our model can successfully assign higher attention weight for such pattern. For patches with relatively less complex structures and textures, our model can identify them as not very important regions by giving lower attention weights.

3.4. Lung cancer dataset results

3.4.1. Baseline models

As we have annotations on NLST dataset, we can conduct more extensive experiments with ROI-based survival models. Following the recent framework (Yu et al., 2016), we extracted 10 dense image patches from ROIs and calculated hand-crafted features using CellProfiler (Carpenter et al., 2006) which serves as a state-of-the-art medical image feature extracting and quantitative analysis tool. A total of 1795 quantitative features were obtained from each image tile. Then we averaged those features across different patches for each patient. 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 can summarize the comparison methods into five categories as follows:

- **Cox models**: The Cox proportional hazards model is the most commonly used semi-parametric model in survival

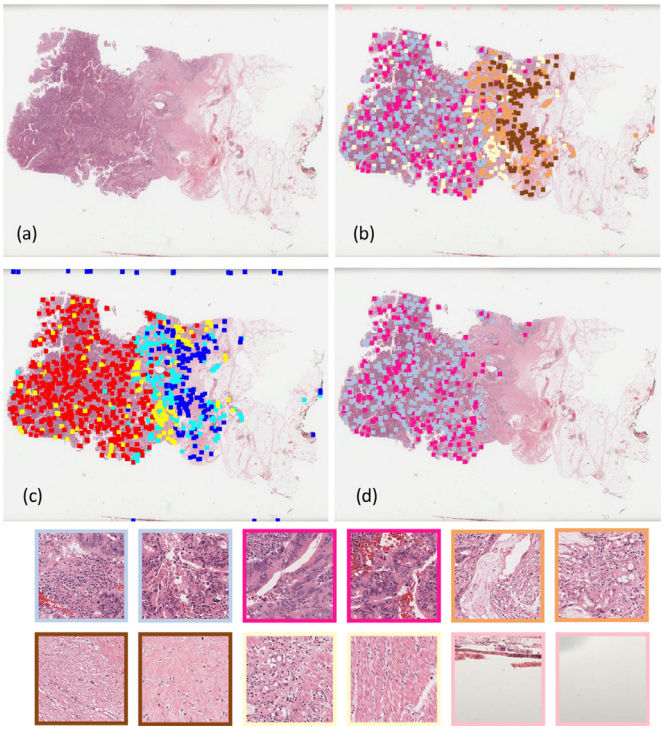


Fig. 8. (a) Original WSI, (b) Phenotype patterns distribution, (c) Heatmaps from our model, (d) Selected patches with highest attention weights. The bottom shows representative patches from each phenotype.

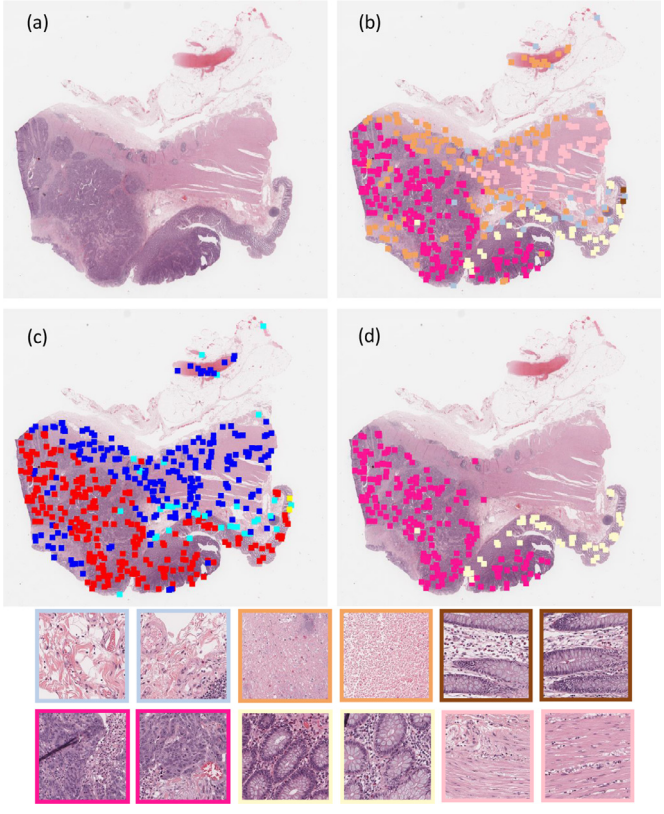


Fig. 9. (a) Original WSI, (b) Phenotype patterns distribution, (c) Heatmaps from our model, (d) Selected patches with highest attention weights. The bottom shows representative patches from each phenotype.

聚类数选多少好

Table 10
Performances with different number of phenotypes.

No.	6	8	10	12
CI	0.673	0.769	0.775	0.742

analysis. Two regularized Cox models l_1-norm (LASSO-Cox) (Tibshirani et al., 1997) and boosting cox model (Cox-boost) (Binder and Schumacher, 2008) are compared in experiments.

- **Parametric censored regression models**: PCR models formulates the joint probability of the uncensored and censored instances as a product of death density function and survival functions, respectively (Lee and Wang, 2003). We choose Weibull, Logistic distribution to approximate the survival data.
- **MTLSA**: Multi-Task Learning model for Survival Analysis (MTLSA) (Li et al., 2016) reformulates the survival model into a multi-task learning problem.
- **WSISA**: WSISA can learn effective features from WSIs (Zhu et al., 2017). We train LassoCox and MTLSA using WSISA learned features as they are top models based on their report. To investigate performance from pre-trained network, a ResNet34 (He et al., 2016) model is used and then fine-tuned as the backbone network in WSISA.
- **DeepMISL**: Deep Multiple Survival Learning combined both local and global representation to predict outcomes (Yao et al., 2019).

3.4.2. Results

We reported results from a few possible numbers of phenotypes, such as {6, 8, 10, 12} on the testing dataset. From the Table 10, we can see models using fewer clusters are unable to achieve good results. The reason might be patches of lung cancer patients are very heterogeneous and it is relative difficult to learn survival-related representations from fewer phenotypes. Results suggest the number of 10 achieves slightly better predictions which is consistent with findings in WSISA (Zhu et al., 2017). Thus, we decide to choose to cluster 10 phenotypes in our model. Other parameters are kept the same with settings in MCO experiments.

Table 11 shows C-index and AUC values by various survival regression methods on 5-fold cross validation. It shows the prediction power of the proposed method compared with different survival models. One can see that the proposed method achieves both highest C-index and AUC values which present the best prediction performance among all methods. From the table, baseline models using hand-crafted features perform not well due to following reasons: (1) the limitation of local information provided by the patches extracted from the ROI using hand-crafted features; (2) the non-effective aggregation way to represent the heterogeneity of tumor and patient from patch-based results. Instead of using a small set of patches and human-designed features, the proposed method can effectively learn complex deep bag representation from phenotype patterns to predict patient survival outcomes.

WSISA achieves better results than baseline models which shows the good representative ability of features from WSISA. However, WSISA needs a separate stage to train several DeepConvSurv models independently and will discard some phenotypes in the final stage, the performance actually depends on how well to select important clusters and WSISA still has the chance to lose in selecting survival-related clusters for a good final survival prediction. To investigate results from pre-trained model, we replaced the original 2DCNN network of WSISA by using a pre-trained ResNet34 (He et al., 2016). The whole model will be fine-tuned following the same process. It is clear to see improved C-index which can demonstrate that fine-tuned models can bring

Table 11
Performance comparison of the proposed methods and other existing related methods using C-index values on NLST dataset.

Type	Method	C-index	AUC
Deep Learning	DeepAttnMISL	**0.6963** (0.0660)	**0.7143** (0.0541)
	DeepMISL (Yao et al., 2019)	0.6476 (0.0698)	0.6693 (0.0866)
	Finetuned-WSISA-LassoCox (Zhu et al., 2017)	0.6123 (0.0216)	0.6427 (0.0575)
	Finetuned-WSISA-MTLSA (Zhu et al., 2017)	0.6428 (0.0259)	0.6963 (0.0668)
	WSISA-LassoCox (Zhu et al., 2017)	0.5996 (0.0750)	0.5957 (0.0674)
	WSISA-MTLSA (Zhu et al., 2017)	0.6305 (0.0575)	0.6479 (0.0936)
Cox-based	Lasso-Cox (Tibshirani et al., 1997)	0.4842 (0.0508)	0.4903 (0.1011)
	Cox-boost (Binder and Schumacher, 2008)	0.5474 (0.0370)	0.5271 (0.0386)
Parametric models	Logistic (Kalbfleisch and Prentice, 2011)	0.4998 (0.0881)	0.5013 (0.1146)
	Weibull (Kalbfleisch and Prentice, 2011)	0.5577 (0.0395)	0.5618 (0.0976)
Multi-task based	MTLSA (Li et al., 2016)	0.5053 (0.0509)	0.5362 (0.0416)
Ranking based	BoostCI (Mayr and Schmid, 2014)	0.5595 (0.0610)	0.5487 (0.0532)

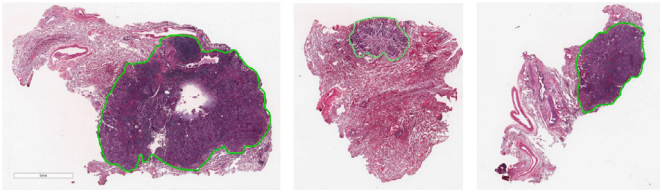


Fig. 10. WSI Annotations of one example patient.

Table 12
Results of single and ensemble models.

	fold 1	fold 2	fold 3	fold 4	fold 5
MCO single	0.666	0.571	0.591	0.565	0.636
MCO ensemble	0.665	0.564	0.585	0.549	0.637
NLST single	0.750	0.775	0.613	0.663	0.680
NLST ensemble	0.766	0.797	0.555	0.656	0.699

benefits but our DeepAttnMISL is still better than finetuned-WSISA by a large margin. When introducing MIL into survival learning, DeepMISL and the proposed model can improve predictions from WSISA by C-index metric. Instead of selecting phenotypes, DeepMISL and the proposed model are designed to consider all possible patterns. Performance is further improved when we use the more flexible attention mechanism to learn informative and discriminate patterns. This architecture makes the proposed method can better learn heterogeneous information encoded in WSIs which will make it more practical and have better intepretability than DeepMISL in real applications.

We pick one patient as the example to show visualization results. Fig. 10 presents this patient's all WSIs and the corresponding tumor region annotations. Figs. 11 and 12 show results from the proposed model and WSISA, respectively. In Fig. 11, the first row shows attention weights heatmaps and the second row shows phenotype pattern distributions on original WSIs. The bottom presents randomly selected patches with higher attention weights (patches with red colors in heatmaps). It is clear to see that most patches from tumor regions are highlighted with high attentions while patches from normal tissues are treated with lower attentions. Compared with results from WSISA shown in Fig. 12, we can see that WSISA will miss many tumor patches and select many normal patches as discriminative patterns. Patches from cancerous regions can be grouped in similar clusters but not all of them will be selected by WSISA as the selection is performed via DeepConvSurv on all patches of the database. Selected phenotypes are more likely discriminative for the whole database with all patients and they are not well interpreted for the specific patient.

Given the trained survival models, we can use the estimated testing risk scores to classify patients into low or high-risk group for personalized treatments. Two groups are classified by the median of predicted risk scores. We evaluate if those models can correctly classify death patients (uncensored data) into two groups since uncensored data is more informative. Patients with longer survival time should be classified into low risk group and vice versa. If the model cannot correctly distinguish high and low risk death patients, two average death times should be very close. We plot Kaplan-Meier survival curves on one testing fold in Fig. 13. From the figure, one can see that the proposed model can more successfully group testing death patients into two groups than other methods in all datasets. The log rank test is conducted to test the difference of two curves. It is shown that the proposed method can achieve the most significant log rank test outcome (p-value $= 4.527 \times 10^{-3}$) while some of others do not reach statistical significances. Kaplan-Meier curves suggest that the proposed comprehensive prediction model can offer personalized risk scores which can better group individuals into two groups. The proposed model has a significant impact on population survival times. It can be used as a recommendation system for offering personalized treatments by determining the relationship between a patient's whole slide pathological images and his or her risk of an event (death). p value of log-rank test less than 0.05 is considered significant and $p < 0.1$ is marginal significant. For five testing folds, our model can achieve four significant results and one marginal significant result with $p = 0.09$, DeepMISL achieves two significant results and one marginal significant with $p = 0.07$, Finetuned-WSISA-MTLSA achieves one significant result and three marginal significant results. Logistic model achieves only one significant result ($p = 0.049$), shown in Fig.13. BoostCI has one fold significant result ($p = 0.0123$). All other baseline models cannot have significant results on all testing folds. Overall, deep learning models perform well to achieve significant results than models with handcrafted features. The proposed model achieves the most number of significant results which could validate that the predictor from our model is an important prognostic factor which could be used for a good patient risk stratification.

3.5. Ensemble models

We investigated if ensemble models could benefit final results. During each fold, we train five models and then average prediction score on the corresponding testing fold. The maximum cluster number is set to 6 and 10, respectively. Table 12 shows C-index values using single and ensemble models on MCO-1M and NLST dataset. The average c-index across five folds is 0.606 for MCO single and 0.600 for MCO ensemble, respectively. On NLST dataset, the averaged c-index of single model is 0.696 and the ensemble model is 0.695. From the table, it can be seen that ensemble models cannot provide additional power for predictions.

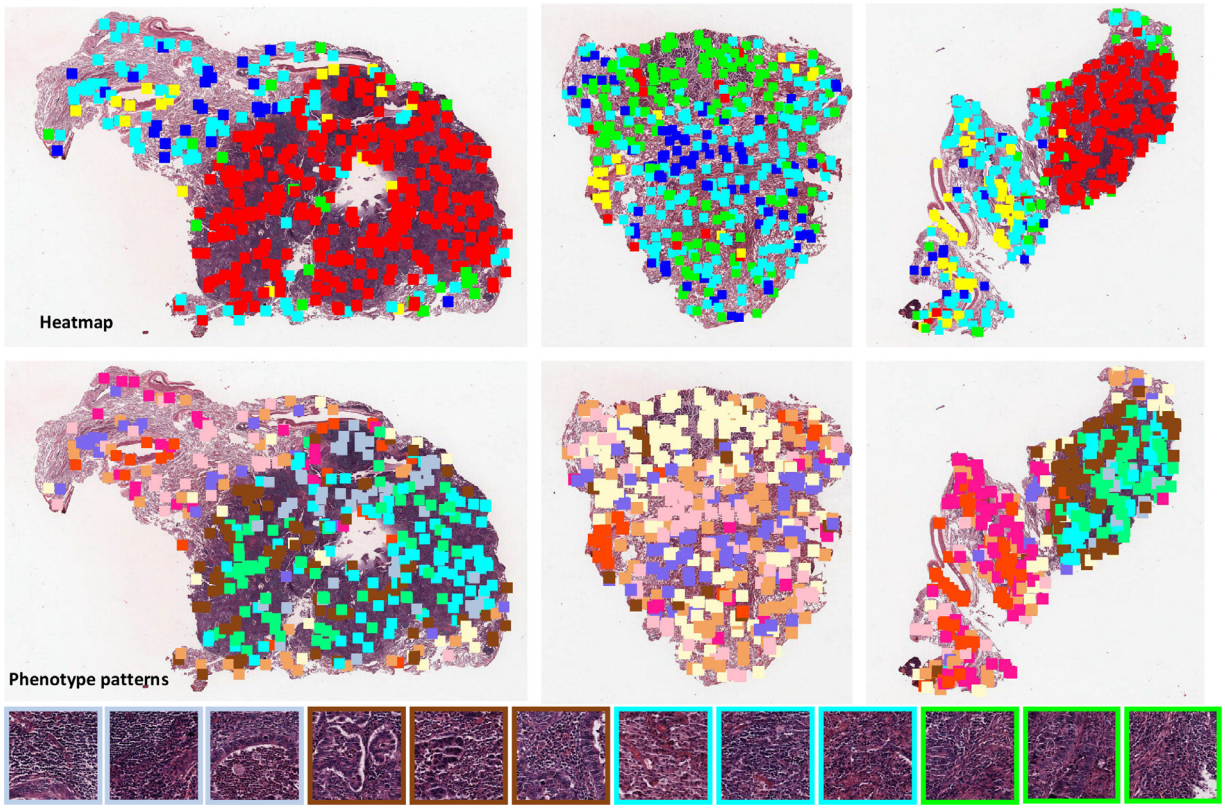


Fig. 11. Phenotype pattern distribution and the corresponding heatmaps from the proposed model on three WSIs of the same patient. The bottom shows patches from phenotypes with high attention values.

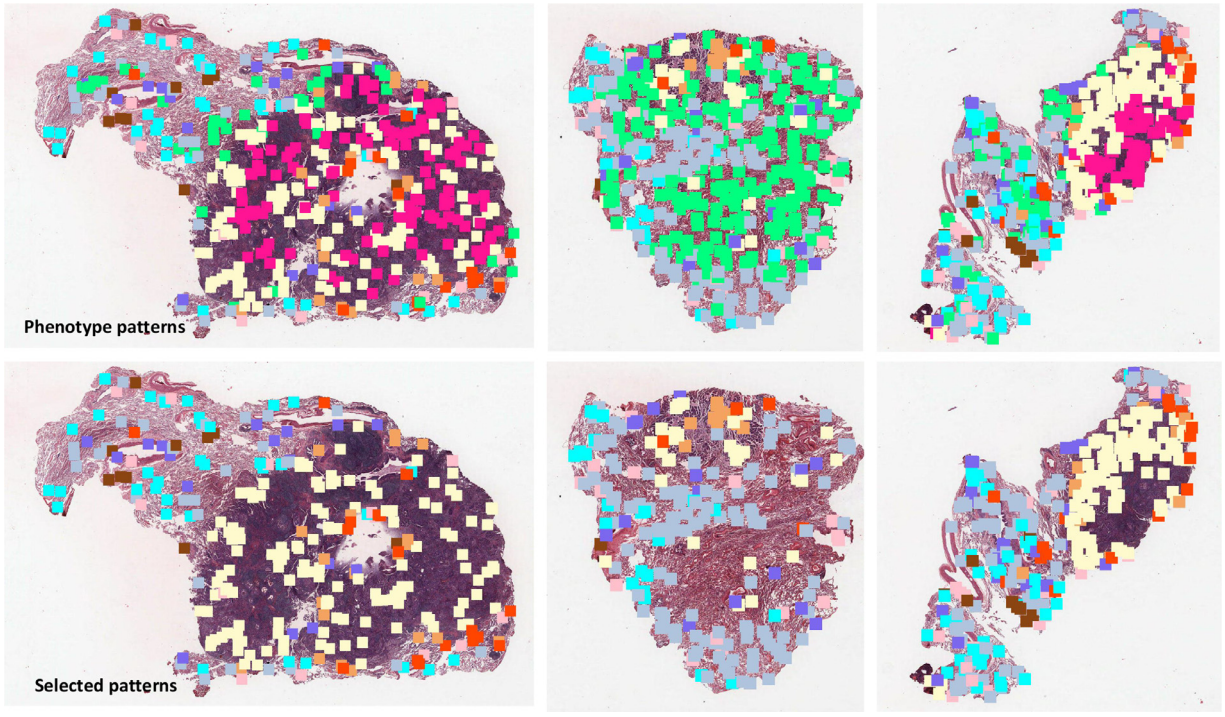


Fig. 12. Phenotype pattern distribution and selected patterns from WSISA. Missing tumor patches can be observed from selected patterns by WSISA.

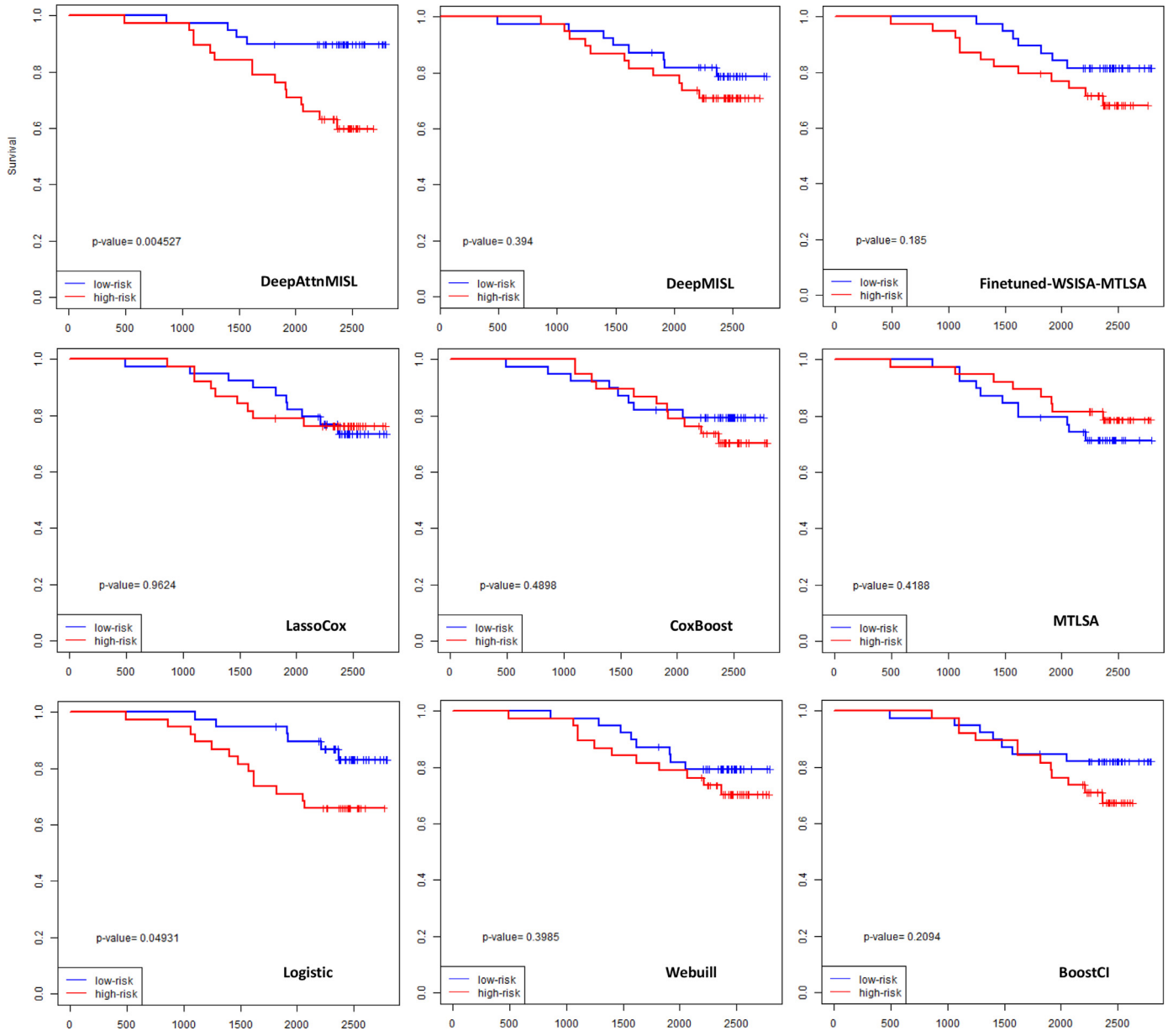


Fig. 13. Kaplan-Meier survival curves of different models for one testing fold. High risk (great than median) groups are plotted as green lines, and low risk (less than or equal to median) groups are plotted as red lines. The x axis shows the time in days and y axis presents the probability of overall survival. Log rank *p* value is shown on each figure. "+" means the censored patient. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4. Conclusion

In this paper, we proposed a deep multiple instance model to directly learn survival patterns from gigapixel images without annotations which make it more easily applicable in large scale cancer dataset. Compared to existing image-based survival models, the developed framework can handle various numbers and sizes whole slide images among different patients. It can learn holistic information of the patient using bag representations and achieve much better performance compared to the ROI patch based methods. Moreover, the flexible and interpretable attention-based MIL pooling can overcome drawbacks from fixed aggregation techniques in state-of-the-art survival learning models. We showed that our approach provides an interpretation of the clinical outcome prediction by presenting reasonable ROIs which is very important in such practical application. Additionally, We illustrated the proposed method can provide personalized treatment for patients and can be used by doctors to guide their treatment decisions for improving patient lifespan. With future research and development, the proposed approach has the potential to be applied in other tumor types.

Declaration of Competing Interest

No conflict of interest.

CRediT authorship contribution statement

Jiawen Yao: Conceptualization, Methodology, Formal analysis, Software, Visualization, Writing - original draft. **Xinliang Zhu:** Methodology, Data curation, Writing - review & editing. **Jitendra Jonnagaddala:** Funding acquisition, Writing - review & editing. **Nicholas Hawkins:** Funding acquisition, Supervision, Writing - re-

view & editing. **Junzhou Huang:** Conceptualization, Supervision, Funding acquisition, Data curation, Writing - review & editing.

Acknowledgments

This work was partially supported by U.S. National Science Foundation IIS-1718853, the CAREER grant IIS-1553687 and Cancer Prevention and Research Institute of Texas (CPRIT) award (RP190107).

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 of the authors and do not represent or imply concurrence or endorsement by NCI.

References

Bair, E., Hastie, T., Paul, D., Tibshirani, R., 2006. Prediction by supervised principal components. J. Am. Stat. Assoc. 101 (473).

Bair, E., Tibshirani, R., 2004. Semi-supervised methods to predict patient survival from gene expression data. PLoS Biol. 2 (4), E108.

Barker, J., Hoogi, A., Depeursinge, A., Rubin, D.L., 2016. Automated classification of brain tumor type in whole-slide digital pathology images using local representative tiles. Med. Image Anal. 30, 60–71.

Bejnordi, B.E., Veta, M., Van Diest, P.J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J.A., Hermsen, M., Manson, Q.F., Balkenhol, M., et al., 2017. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 318 (22), 2199–2210.

Binder, H., Schumacher, M., 2008. Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. BMC Bioinform. 9 (1), 1–10.

Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Silva, V.W.K., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J., 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat. Med. 25 (8), 1301–1309.

Carpenter, A.E., Jones, T.R., Lamprecht, M.R., Clarke, C., Kang, I.H., Friman, O., Guertin, D.A., Chang, J.H., Lindquist, R.A., Moffat, J., et al., 2006. Cellprofiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol. 7 (10), R100.

Cheng, J., Mo, X., Wang, X., Parwani, A., Feng, Q., Huang, K., 2017. Identification of topological features in renal tumor microenvironment associated with patient survival. Bioinformatics 34 (6), 1024–1030.

Dauphin, Y.N., Fan, A., Auli, M., Grangier, D., 2017. Language modeling with gated convolutional networks. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, pp. 933–941.

Gurcan, M.N., Boucheron, L., Can, A., Madabhushi, A., Rajpoot, N., Yener, B., 2009. Histopathological image analysis: a review. IEEE Rev. Biomed. Eng. 2, 147.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Heagerty, P.J., Zheng, Y., 2005. Survival model predictive accuracy and ROC curves. Biometrics 61 (1), 92–105.

Hou, L., Samaras, D., Kurc, T.M., Gao, Y., Davis, J.E., Saltz, J.H., 2015. Efficient multiple instance convolutional neural networks for gigapixel resolution image classification. arXiv preprint arXiv:1504.07947, 7.

Hou, L., Samaras, D., Kurc, T.M., Gao, Y., Davis, J.E., Saltz, J.H., 2016. Patch-based convolutional neural network for whole slide tissue image classification.. In: CVPR, pp. 2424–2433.

Ilse, M., Tomczak, J.M., Welling, M., 2018. Attention-based deep multiple instance learning. arXiv preprint arXiv:1802.04712.

Huang, Zhi, Zhan, Xiaohui, et al., 2019. SALMON: Survival Analysis Learning With Multi-Omics Neural Networks on Breast Cancer. Frontiers in Genetics 10. doi:10.3389/fgene.2019.00166, In this issue.

Jamal-Hanjani, M., Quezada, S.A., Larkin, J., Swanton, C., 2015. Translational implications of tumor heterogeneity. Clin. Cancer Res. 21 (6), 1258–1266.

Jonnagaddala, J., Croucher, J.L., Jue, T.R., Meagher, N.S., Caruso, L., Ward, R., Hawkins, N.J., 2016. Integration and analysis of heterogeneous colorectal cancer data for translational research. p. 387.

Kalbfleisch, J.D., Prentice, R.L., 2011. The Statistical Analysis of Failure Time Data, 360. John Wiley & Sons.

Katzman, J., Shaham, U., Cloninger, A., Bates, J., Jiang, T., Kluger, Y., 2016. Deep survival: a deep COX proportional hazards network. arXiv preprint arXiv:1606.00931.

Kong, B., Wang, X., Li, Z., Song, Q., Zhang, S., 2017. Cancer metastasis detection via spatially structured deep network. In: International Conference on Information Processing in Medical Imaging. Springer, pp. 236–248.

Lee, E.T., Wang, J., 2003. Statistical Methods for Survival Data Analysis, 476. John Wiley & Sons.

Li, R., Yao, J., Zhu, X., Li, Y., Huang, J., 2018. Graph CNN for survival analysis on whole slide pathological images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 174–182.

Li, Y., Ping, W., 2018. Cancer metastasis detection with neural conditional random field. Medical Imaging with Deep Learning.

Li, Y., Wang, J., Ye, J., Reddy, C.K., 2016. 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).

Lin, Z., Feng, M., Santos, C.N.D., Yu, M., Xiang, B., Zhou, B., Bengio, Y., 2017. A structured self-attentive sentence embedding. arXiv preprint arXiv:1703.03130.

Liu, Y., Gadepalli, K., Norouzi, M., Dahl, G.E., Kohlberger, T., Boyko, A., Venugopalan, S., Timofeev, A., Nelson, P.Q., Corrado, G.S., et al., 2017. Detecting cancer metastases on gigapixel pathology images. arXiv preprint arXiv:1703.02442.

Mayr, A., Schmid, M., 2014. Boosting the concordance index for survival data–a unified framework to derive and evaluate biomarker combinations. PLoS One 9 (1), e84483.

Mercan, C., Aksoy, S., Mercan, E., Shapiro, L.G., Weaver, D.L., Elmore, J.G., 2018. Multi-instance multi-label learning for multi-class classification of whole slide breast histopathology images. IEEE Trans. Med. Imaging 37 (1), 316–325.

Mobadersany, P., Yousefi, S., Amgad, M., Gutman, D.A., Barnholtz-Sloan, J.S., Vega, J.E.V., Brat, D.J., Cooper, L.A., 2018. Predicting cancer outcomes from histology and genomics using convolutional networks. Proc. Natl. Acad. Sci. 115 (13), E2970–E2979.

Park, M.Y., Hastie, T., 2007. L1-regularization path algorithm for generalized linear models. J. R. Stat. Soc. 69 (4), 659–677.

Raffel, C., Ellis, D.P., 2015. Feed-forward networks with attention can solve some long-term memory problems. arXiv preprint arXiv:1512.08756.

Sabour, S., Frosst, N., Hinton, G.E., 2017. Dynamic routing between capsules. In: Advances in Neural Information Processing Systems, pp. 3856–3866.

Shedden, K., Taylor, J.M., Enkemann, S.A., Tsao, M.-S., Yeatman, T.J., Gerald, W.L., Eschrich, S., Jurisica, I., Giordano, T.J., Misek, D.E., et al., 2008. Gene expression–based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. Nat. Med. 14 (8), 822–827.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826.

Tang, B., Li, A., Li, B., Wang, M., 2019. Capsurv: Capsule network for survival analysis with whole slide pathological images. IEEE Access.

Team, N.L.S.T.R., et al., 2011. The national lung screening trial: overview and study design. Radiology.

Tibshirani, R., et al., 1997. The lasso method for variable selection in the COX model. Stat. Med. 16 (4), 385–395.

Wang, D., Khosla, A., Gargeya, R., Irshad, H., Beck, A.H., 2016. Deep learning for identifying metastatic breast cancer. arXiv preprint arXiv:1606.05718.

Wang, H., Xing, F., Su, H., Stromberg, A., Yang, L., 2014. Novel image markers for non-small cell lung cancer classification and survival prediction. BMC Bioinform. 15 (1), 310. doi:10.1186/1471-2105-15-310.

Wang, S., Zhu, Y., Yu, L., Chen, H., Lin, H., Wan, X., Fan, X., Heng, P.-A., 2019. Rmdl: recalibrated multi-instance deep learning for whole slide gastric image classification. Med. Image Anal. 58, 101549.

Wang, X., Chen, H., Gan, C., Lin, H., Dou, Q., Huang, Q., Cai, M., Heng, P.-A., 2018. Weakly supervised learning for whole slide lung cancer image classification. Med. Imaging Deep Learn..

Wang, X., Chen, H., Gan, C., Lin, H., Dou, Q., Tsougenis, E., Huang, Q., Cai, M., Heng, P., 2019. Weakly supervised deep learning for whole slide lung cancer image analysis. IEEE Trans. Cybern. 1–13. doi:10.1109/TCYB.2019.2935141.

Wang, Y., Lu, L., Cheng, C.-T., Jin, D., Harrison, A.P., Xiao, J., Liao, C.-H., Miao, S., 2019. Weakly supervised universal fracture detection in pelvic x-rays. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 459–467.

Ward, R.L., Hawkins, N.J., 2015. Molecular and cellular oncology (MCO) study tumour collection. UNSW Australia. https://doi.org/10.4225/53/5559205bea135.

Warth, A., Muley, T., Meister, M., Stenzinger, A., Thomas, M., Schirmacher, P., Schnabel, P.A., Budczies, J., Hoffmann, H., Weichert, W., 2012. The novel histologic international association for the study of lung cancer/american thoracic society/european respiratory society classification system of lung adenocarcinoma is a stage-independent predictor of survival. J. Clin. Oncol. JCO–2011.

Wulczyn, E., Steiner, D.F., Xu, Z., Sadhwani, A., Wang, H., Flament-Auvigne, I., Mermel, C.H., Chen, P.-H.C., Liu, Y., Stumpe, M.C., 2020. Deep learning-based survival prediction for multiple cancer types using histopathology images. PLoS One 15 (6), 1–18. doi:10.1371/journal.pone.0233678.

Yan, C., Yao, J., Li, R., Xu, Z., Huang, J., 2018. Weakly supervised deep learning for thoracic disease classification and localization on chest x-rays. In: Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, pp. 103–110.

Yang, H., Zhou, J.T., Cai, J., Ong, Y.S., 2017. MIML-FCN+: Multi-instance multi-label learning via fully convolutional networks with privileged information. In: CVPR, pp. 1577–1585.

Yang, Y., Zou, H., 2012. A cocktail algorithm for solving the elastic net penalized COX's regression in high dimensions. Stat. Interface 6 (2), 167–173.

Yao, J., Ganti, D., Luo, X., Xiao, G., Xie, Y., Yan, S., Huang, J., 2015. Computer-assisted diagnosis of lung cancer using quantitative topology features. In: Zhou, L., Wang, L., Wang, Q., Shi, Y. (Eds.), Machine Learning in Medical Imaging. In: Lecture Notes in Computer Science, 9352. Springer International Publishing, pp. 288–295.

Yao, J., Wang, S., Zhu, X., Huang, J., 2016. Imaging biomarker discovery for lung cancer survival prediction. In: MICCAI. Springer International Publishing, pp. 649–657.

Yao, J., Zhu, X., Huang, J., 2019. Deep multi-instance learning for survival prediction from whole slide images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 496–504.
Yao, J., Zhu, X., Zhu, F., Huang, J., 2017. Deep correlational learning for survival prediction from multi-modality data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 406–414.
Yu, K.-H., Zhang, C., Berry, G.J., Altman, R.B., Ré, C., Rubin, D.L., Snyder, M., 2016. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. Nat. Commun. 7 (12474).
Yuan, Y., Failmezger, H., Rueda, O.M., Ali, H.R., Gräf, S., Chin, S.-F., Schwarz, R.F., Curtis, C., Dunning, M.J., Bardwell, H., et al., 2012. Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. Science translational medicine 4 (157). 157ra143–157ra143.
Zhu, X., Yao, J., Huang, J., 2016. Deep convolutional neural network for survival analysis with pathological images. In: BIBM. IEEE, pp. 544–547.
Zhu, X., Yao, J., Luo, X., Xiao, G., Xie, Y., Gazdar, A., Huang, J., 2016. Lung cancer survival prediction from pathological images and genetic data - an integration study. In: ISBI, pp. 1173–1176.
Zhu, X., Yao, J., Zhu, F., Huang, J., 2017. WSISA: Making survival prediction from whole slide histopathological images. In: CVPR, pp. 7234–7242.