Explainable Survival Analysis with Convolution-involved Vision Transformer

**Yifan Shen[1]*, Li liu[2]*, Zhihao Tang[1], Zongyi Chen[1], Guixiang Ma[3], Jiyan Dong[2],
Xi Zhang[1]†, Lin Yang[2]†, Qingfeng Zheng[2]**

[1]Key Laboratory of Trustworthy Distributed Computing and Service (MoE),
Beijing University of Posts and Telecommunications, China
[2]National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital,
Chinese Academy of Medical Sciences and Peking Union Medical College
[3]University of Illinois at Chicago
{shenyifan, innerone, zongyichen, zhangx}@bupt.edu.cn, guixiang.ma@intel.com,
yanglin@cicams.ac.cn, {djy0823lucky, 15732027828, qfzhengpku}@163.com

Abstract

Image-based survival prediction models can facilitate doctors in diagnosing and treating cancer patients. With the advance of digital pathology technologies, the big whole slide images (WSIs) provide increased resolution and more details for diagnosis. However, the gigabyte-size or even terabyte-size WSIs would make most models computationally infeasible. To this end, instead of using the complete WSIs, most of the existing models only use a pre-selected subset of key patches or patch clusters as input, which might discard some important morphology information. In this work, we propose a novel survival analysis model to fully utilize the complete WSI information. We show that the use of a Vision Transformer (ViT) backbone, together with convolution operations involved in it, is an effective approach to improve the prediction performance. Additionally, we present a post-hoc explainable method to identify the most salient patches and distinct morphology features, making the model more faithful and the results easier to comprehend by human users. Evaluations on two large cancer datasets show that our proposed model is more effective and has better interpretability for survival prediction. We would make the code publicly available upon acceptance.

Introduction

Automatic histopathological image analysis can be used to improve the diagnosis process by reducing the workloads of pathologists and the chance of diagnosis mistakes. Histopathological image-based survival analysis aims to predict the expected duration of time until patients' death. Based on the histological details in high resolution, it would be of great benefit to make early decisions and provide better treatments for cancer patients.

Recently, many methods have been proposed for cancer diagnosis and survival prediction using pathological slides. Traditional survival analysis models (Tibshirani 1997; Bair, Tibshirani, and Golub 2004; Shedden et al. 2008) use statistical data and omics data from patients to predict the patient's health, which requires the heavy burden of feature engineering. With the advance of deep learning models, convolutional neural networks (CNNs) and their extensions have been widely used to capture the rich information of the whole slide images (WSIs) to improve prediction accuracy. However, one unique challenge for WSI-based analysis is that a pathological image is of high resolution (usually from one billion to one trillion pixels), making most deep learning models not applicable. To address this issue, these models (Zhu, Yao, and Huang 2017; Yao et al. 2017) usually pre-select a subset of key patches from the region of interest (ROI) instead of using all patches as the input of the model. But the manual selection of regions is laborious and commonly dependent on the expertise of pathologists to thoroughly examine the whole slide. To reduce such overhead, some methods (Zhu et al. 2017; Tang et al. 2019; Li et al. 2018; Yao et al. 2020) have adopted sampling strategy to generate candidate patches not limited to ROI. However, only a small set of patches may not completely capture the tumor morphology and suffers from a high risk of discarding informative patterns due to the complexity and heterogeneity of tumors. It is of prominent importance to design a framework that can consider all the patches from WSIs to improve the prediction accuracy and minimize the risk of losing important information.

Another challenge for image-based survival prediction is the requirement of the model interpretability, which is crucial to make the model faithful and the results get accepted in the diagnostics. Explainable survival analysis is very challenging due to several reasons: 1) Different from conventional image classification tasks such as tumor classification, survival analysis is formulated as a regression problem that lacks tissue-level or patch-level labels as ground-truth information for visual explanations. 2) Most of the existing methods (Li et al. 2018; Yao et al. 2020) provide pixel-wise explanations and salient regions, which ignores the domain-specific biological features and cannot be easily

*These authors contributed equally.

†Corresponding author

understood by pathologists. Although a recent study (Jaume et al. 2021b) has tried to address this issue by proposing a set of quantitative metrics based on domain-specific concepts, it is limited to its graph explainability framework based on a biological entity graph, which is not suitable for the widely adopted patch-based survival analysis models. Table 1 shows the comparison of various models in the field of survival analysis in terms of whether to manually label, select and gather patches, and interpretability.

In this paper, we aim to propose a survival prediction model named Explainable Survival Analysis using Convolution-involved Vision Transformer (ESAT) that can fully utilize the WSI information without pre-selecting ROI or patches. Enlightened by the remarkable success of Vision Transformer (Dosovitskiy et al. 2020) in a variety of computer vision tasks, we propose a ViT-based model, which naturally coincides with a typical patch-based deep learning processing paradigm. Our model first splits WSIs into discrete non-overlapping patches which are treated as tokens and then feeds the patches into transformer layers with position encoding to model their global relations for regression. To accommodate the large WSIs, we introduce convolution layers and approximately linear attention layers instead of standard self-attention mechanisms to reduce the computation complexity. Additionally, we propose a post-hoc explainable method to identify both the salient patches and the quantitative morphology indicators as explanations, which are more comprehensive and easier to understand by the pathologist. We evaluate our method on two cancer datasets and the results show that our model can outperform the state-of-the-art baselines and has better explainability.

Related Work

In this section, we briefly review relevant works on survival prediction, explainability in digital pathology, and Vision Transformer.

Survival prediction with digital pathology Various studies have shown that tumor morphology and growth displayed in the pathological images is useful in cancer diagnosis (Warth et al. 2012; Yuan et al. 2012). Due to the large image size of WSIs, most approaches only selected a set of patches from regions of interest (ROI) to represent the whole WSIs. Among them, early studies commonly extracted hand-crafted features from ROI (Yuan et al. 2012; Wang et al. 2014; Yao et al. 2015; Barker et al. 2016; Cheng et al. 2018). Specifically, they used nuclei detection and segmentation to extract features that rely on human annotations. More recent studies applied deep learning-based models to improve the representation of patches and reduce human efforts. For example, CNNs were first used by some work (Zhu, Yao, and Huang 2017; Yao et al. 2017) to extract the feature from pre-selected ROI. WSISA (Zhu et al. 2017) adopted the way to select important patches and sampled more patches randomly. Then it aggregated them to different clusters and selected important clusters to get a patient-level result. This framework utilized more information from WSIs and reduced the chance of diagnosis mistakes. CapSurv (Tang et al. 2019) followed WSISA by introducing the capsule network. Graph convolutional networks (GCNs) are used in the work (Li et al. 2018) to consider the similarity relationship of patches features and learn a WSI-level representation. Another work (Yao et al. 2020) used the siamese multiple instance learning network to learn features from clusters and attention-based pooling layer to aggregate the clusters. Different from the existing models which use selective patches, the model we proposed considers all the patches on WSIs and takes advantage of the long-distance dependence between patches to predict the overall survival risk of the patient.

Explainability in digital pathology Although the deep learning-based diagnostic models have achieved remarkable performance, their lack of interpretability limits their application in practical scenarios. To address this issue, a set of explainability methods have been proposed, including feature-attribution and attention-based technique (Hägele et al. 2020; Lu et al. 2021). However, most of the existing methods are designed for cancer classification models, aiming to indicate the difference between cancer cells (or regions) and normal cells (or regions), and can be validated by ground-truth annotations or expert knowledge. On the contrary, for the survival analysis problem, almost all instances contain cancer cells and regions, thus we need to show distinct patterns between cancer cells or regions, which is more difficult. Moreover, they commonly provide explainability results in the form of heatmaps (Zhu, Yao, and Huang 2017), graph nodes (You et al. 2020), and path clusters (Yao et al. 2020). These forms cannot be intuitively understood by human experts such as pathologists due to the ignorance of biological entities like the nucleus. Although a recent study (Jaume et al. 2021b) has proposed to use quantitative metrics involving domain-specific concepts, it is limited to its graph-based model with sparsely distributed cells. To this end, we make slight changes to their approach to make it applicable to more types of cancers such as small cell lung cancer and the more commonly used patch-based models.

Vision Transformer Vision Transformer (ViT) (Dosovitskiy et al. 2020) is the first to prove that a pure Transformer architecture can attain state-of-the-art performance in Computer Vision. Specifically, ViT decomposes each image into a series of flattened patches and then applies multiple standard Transformer layers to model these tokens. This idea coincides with the processing of handing WSIs and satisfies the needs of survival analysis to consider the impact of different patches on the overall WSI. Despite the emergence of attention-based models in pathological diagnosis, there is still no survival analysis model that uses a transformer to capture long-distance connections between all patches. The application of ViT needs to deal with the problem caused by WSIs' oversized pixels that leads to excessive computational overhead. One idea is to use a small number of large patches to reduce the cost of calculating the relationship between patches. We tried this method, and the reasons for its poor performance will be discussed in the "Results and Discussion" section. Inspired by the works which combine CNNs and Transformers to model both local and global dependencies (Wu et al. 2021; Yuan et al. 2021), we incorporate the

Table 1: Prior survival analysis methods *v.s.* the proposed method

Models	Need annotations	Approach of patch selection	Approach of patch aggregation	Explanation
NHIA (Warth et al. 2012)	Yes	All	No aggregation	Statistical features
BOEH (Cheng et al. 2018)	Yes	ROI	Cluster	Statistical features
DeepConvSurv (Zhu, Yao, and Huang 2017)	No	ROI	Fully connected layer	Statistical features
RankDeepSurv (Jing et al. 2019)	No	ROI	Fully connected layer	Statistical features
WSISA (Zhu et al. 2017)	No	Random sample	Cluster and Fully connected layer	No
CapSurv (Tang et al. 2019)	No	Random sample	Cluster and capsule network	No
DeepGraphSurv (Li et al. 2018)	No	Random sample	Graph convolutional neural network	Heat map
DeepAttnMISL (Yao et al. 2020)	No	Random sample	Siamese MI-FCN with attention	Heat map
ESAT (ours)	No	All	Vision transformer	Important patches, quantitative indicators

convolutional layers in ViT to aggregate the adjacent small patches to reduce the calculations. Besides, we use an approximately linear attention layer (Xiong et al. 2021) to replace the original attention layer to reduce the time complexity of the calculations. In contrast to the concurrent works, this work attempts to use convolutional layers to reduce the number of patches in the input phase of ViT.

Methodology

In this section, we start by introducing the problem definition of survival analysis and the overview of the framework. Then we describe the main components of the framework in detail, including the convolution-involved Vision Transformer, the survival risk loss, and the explainable survival analysis.

Preliminaries

Considering a set of N patients, P_i, $i = 1, \dots, N$, each patient has a binary label (t_i, δ_i) and a set of $\{W_j\}_{j=1}^{k} \in P_i$, where the observation time t_i is either a survival time or a censored time, and the censorship status δ_i is the indicator which is 1 for an uncensored instance (death occurs during the study) and 0 for a censored instance. W is the WSIs set of the patient. Survival models aim to predict the hazard risk to present how well the patient behaves. Fig. 1 shows the overview of the proposed ESAT. The WSIs of patients are fed into the convolution-involved Vision Transformer module to extract features, which are then used to calculate the survival risk loss. Next, WSIs features are trained with the patient-level ground-truth labels. Additionally, a post-hoc explainable module is proposed to provide important patches and quantitative indices to make the prediction results trustworthy. Different from existing models, ESAT discards the step of randomly or manually selecting patches from WSIs but considers all the patches. It also adopts a model-agnostic interpretable module that only uses the input and output of the model to provide useful explanations.

Convolution-involved Vision Transformer

The role of this module is to extract effective features from WSIs. Its architecture is shown in the bottom right of Fig 1. Due to different sizes and shapes of WSIs, we first resize them into a uniform square size and then split each WSI into $n \times n$ patches. Existing models (Zhu et al. 2017; Yao et al. 2020) have found that using pre-trained CNNs can facilitate the feature extraction process. Therefore, we use Resnet34 (He et al. 2016) instead of the original random parameters to get initial d-dimensional patch embeddings. Then follow the standard procedure in Vision Transformers, we organize the total n^2 tokens with dimension size d into $X \in \mathbf{R}^{n^2 \times d}$, which is then projected with three matrices $W^Q \in \mathbf{R}^{d \times d_q}$, $W^K \in \mathbf{R}^{d \times d_k}$ and $W^V \in \mathbf{R}^{d \times d_v}$ to extract feature representations Q, K, and V, representing query, key and value in the attention mechanism. We adopt multi-head attention to model the tokens, that is,

$$Q = XW^Q, \quad K = XW^K, \quad V = XW^V \tag{1}$$

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{2}$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{3}$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O(4)$$

The time complexity of the multi-head attention is $O(n^2d^2 + n^4d)$. Since the feature dimension d is independent of the input data, it is necessary to reduce the number of patches n^2 to reduce the complexity due to the excessive size of WSIs. An intuitive way is to divide WSIs into patches with larger sizes. But experimental results show that the performance drops significantly (see Table. 3). We will discuss it in “Results and Discussion” section. Therefore we propose to retain the small patch size and use convolutional layers to aggregate adjacent small patches into a bigger one to reduce the number of patches. Specifically, one filter in the convolutional layer is used to aggregate one dimension of the d-dimensional features of the adjacent patches. We merge such d filters to obtain the next-layer feature representations as shown in the bottom right of Fig. 1. Next, following the steps of ViT, we flatten the patch embedding and add position embedding. Meanwhile, to further reduce the complexity, we adopt the Nystrom-based linear transformers (Xiong et al. 2021) to replace the standard self-attention transformers. The Nystrom method is adopted to approximate the softmax matrix in self-attention by sampling a subset of columns and rows. Consequently, the time complexity in this module can be reduced to $O(nd^2 + n^2d)$ as n is small.

Survival Risk Loss

In this step, we use WSIs embeddings produced by the convolution-involved Vision Transformer to generate a

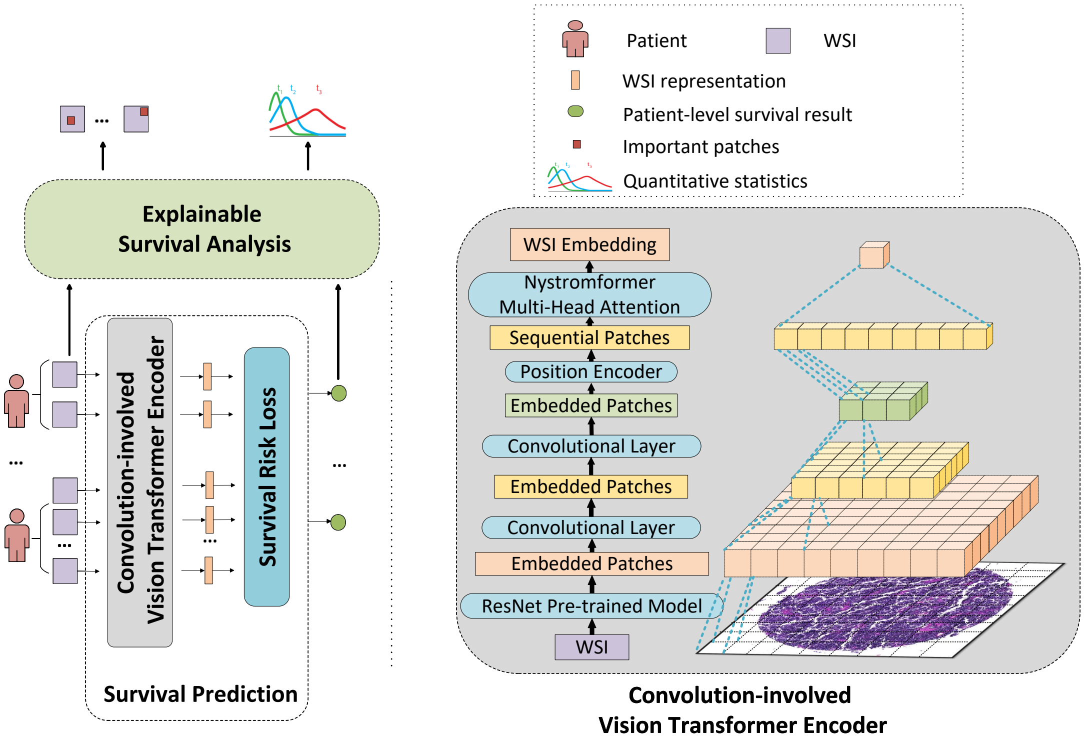

Figure 1: **Left**: An overall framework consisting of a survival prediction module and an explainable survival module. **Top right**: Legend of the framework. **Bottom right**: Details of the convolution-involved Vision Transformer and its corresponding three-dimensional display. The blue component on the left corresponds to the blue dashed line on the right, and the remaining data layers correspond to colors.

patient-wise hazard risk which measures the probability of the expected development of cancer. For the i-th patient, the output of this step is denoted as O_i, and its patient-level label is (t_i, δ_i). As censoring data ($\delta_i = 0$) means patients were alive during the study and their survival time does not represent the true survival time, we assume that the censoring data is non-informative and matches the WSIs embeddings of the uncensored data from the same patient with the consistent patient-level label. Let $t_1 < t_2 < \cdots < t_N$ denote a sequence of ordered event time and $R(t_i)$ denotes the risk set of patients who live longer than the i-th patient. In other words, it means the survival time of the patients in the risk set is equal or larger than the i-th patient ($t_j \geq t_i$). Following the work (Katzman et al. 2016; Zhu et al. 2017), we define the conditional probability upon the existence of the death event that occurs at some particular time t for the i-th patient is:

$$\mathbf{L_i} = \frac{exp(O_i)}{\sum_{j \in R(t_i)} exp(O_j)} \tag{5}$$

Assuming the patient's events are independent, we can get the joint probability of all conditioned probability as the partial likelihood:

$$\mathbf{L} = \prod_{i=1} \frac{exp(O_i)}{\sum_{j \in R(t_i)} exp(O_j)} \tag{6}$$

To maximize the partial likelihood, we take the logarithm on both sides of the equation. It is equivalent to minimizing the negative log partial likelihood as follows:

$$\mathbf{L(O_i)} = - \sum_{i=1}^{n} \delta_i \left(O_i - \log \sum_{j \in R(t_i)} exp(O_j) \right) \tag{7}$$

We use the negative log partial likelihood as loss function which is the same as (Zhu et al. 2017; Yao et al. 2020). For a set of model predictions, it can contribute to the consistency of the risk set and penalize predictions that are not in the correct order. For a patient with multiple WSIs, the same number of survival risk predictions can be obtained with our method. As the WSI with a worse predictive result is more likely to reflect the actual state for cancer patients, we choose the prediction with the highest survival risk as our final result.

Explainable Survival Analysis

To provide interpretable explanations for predictions, we propose a post-hoc explainable method for survival analysis. Different from most existing explaining methods for survival analysis, our method is model-agnostic and can be applied to a variety of patch-based survival models. Specifically, it takes the trained model and its predictions as input and returns the explanations in the form of patch-level salient areas together with human-intelligible cell-level quantitative indices.

Formally, we use $f(\mathbf{x}^s;\boldsymbol{\theta})$ to represent the trained survival model. As the parameter $\boldsymbol{\theta}$ is fixed, it can thus be omitted. Given the WSI as input, we set $\mathbf{M} \in [0,1]^{n\times n}$ as corresponding masks with the same size as the input. We aim to learn $\mathbf{M}$ in a self-supervised learning (SSL) manner to minimize the loss of prediction results with the original input $\mathbf{x}$ and with the selected input $\mathbf{M}\otimes\mathbf{x}$, that is,

$$\min_{\mathbf{M}} \quad \ell\left(f(\mathbf{x}), f(\mathbf{M}\otimes\mathbf{x})\right) + \gamma\|a(\mathbf{M})\|. \tag{8}$$

where $\gamma\|a(\mathbf{M})\|$ is used to sparse the weight of M. With the learned mask $\mathbf{M}$, we can extract the salient areas of $\mathbf{x}$ to maximally preserve information in $\mathbf{x}$. A larger value in $\mathbf{M}$ indicates that the corresponding area provides more contribution to the prediction. In this study, we can thus obtain the most important patches and least important patches by ranking the element values in $\mathbf{M}$.

Practically, merely providing important patches is not easy to understand by pathologists. Different from classification problems with very distinct positive or negative instances for comparisons, our problem can be viewed as a regression problem and the important patches for patients with different survival risks may not be easily differentiated for visualizations. To address this issue, we adopt domain-specific quantitative metrics that can facilitate the pathologist in spotting the differences on the selected important patches. We will illustrate the metrics and results in detail in the “Experiments” and “Results and Discussion” section.

Experiments

Datasets

We use two datasets to evaluate the performance of our model. One is a public National Lung Screening Trial (Team 2011) (NLST) dataset collected by the National Cancer Institute's Division of Cancer Prevention (DCP) and Division of Cancer Treatment and Diagnosis (DCTD), which can be downloaded from Internet via application. The other is collected by the Chinese Academy of Medical Science (CHCAMS), and has been approved by the Ethics Committee and Institutional Review Boards of Cancer Hospital. The number of WSIs and patients, as well as the types of cancer in each dataset are shown in Table 2.

Table 2: The number of WSIs, patients, and types of cancer

Datasets	Patients	WSIs	Types of cancer
NLST	449	1041	adenocarcinoma, squamous cell carcinoma
CHCAMS	343	686	small cell lung cancer

Baseline models

We reproduce several popular survival models as follows:

Cox model The Cox proportional hazard model is one of the most commonly used semi-parametric model in survival analysis. We used l_1-norm (LASSO-Cox) (Tibshirani 1997) model as baseline.

Logistic regression model It formulated the joint probability of the uncensored and censored instances as a product of death density function and survival function by logistic distribution (Lee and Wang 2003).

RankDeepSurv It learned the patient-level survival prediction based on ranking information on survival data function (Jing et al. 2019).

WSISA WSISA (Zhu et al. 2017) sampled a set of patches from WSIs and clustered them into different categories to predict the risk.

DeepAttnMISL The DeepAttnMISL (Yao et al. 2020) followed the way of clustering the patterns and used an attention-based pooling layer to consider all the patterns.

ViT Vision Transformer (ViT) (Dosovitskiy et al. 2020) decomposes WSIs into a series of flattened big patches and then applies standard Transformer layers to model these tokens to get the WSIs' embeddings.

Implementation details

As cox model and logistic regression model depend on hand-crafted features, we use Histocartography (Jaume et al. 2021a) which is a state-of-the-art medical image feature extracting tool to obtain the hand-crafted features. The extracted features include size, shape, pixel intensity distribution, texture of the objects, as well as the relation between neighboring objects. The source codes of WSISA and DeepAttnMISL are obtained from the websites of the authors. All other methods are built using the functions from the lifelines package, which is a survival analysis library available on Github [1]. For data preprocessing in ViT, we split WSIs into flatten patches with a size of 512×512 to meet the constraints of computing resources. All the experiments run on NVIDIA V100 GPU.

We split the NLST dataset and CHCAMS dataset into training, validation, and testing set with a split ratio of 8:1:1. For training, the parameters are optimized using the Adam algorithm, where the learning rate is initialized at 0.01. We set the dimension of the hidden feature vector as 256. The batch size is set to 64. The training process is iterated upon 1000 epochs. The patch size is set to 16×16.

Evaluation Metric

Survival Prediction To evaluate the performance in survival prediction, following previous studies (Zhu et al. 2017; Yao et al. 2020; Li et al. 2018), we take the concordance index (C-index) as the evaluation metric. It refers to the proportion of pairs whose predicted results are consistent with actual results among all patient pairs. The formal definition of C-index is

$$\mathbf{c} \;=\; \frac{1}{n}\sum_{i\in\{1\ldots N|\delta_i=1\}}\;\sum_{t_j>t_i} I[f_i > f_j], \tag{9}$$

where n is the number of comparable pairs, $I[.]$ is the indicator function, t_i is the actual observation time of patient i

[1] https://github.com/CamDavidsonPilon/lifelines

and f_i means the corresponding risk of patient i. δ_i is the indicator which is 1 when death occurs during the study. The value of C-index ranges from 0 to 1. A larger value indicates the better prediction performance and vice versa. 0.5 is the value as a random guess.

Explainable Survival Analysis To evaluate the explainability in our model, we adopt the similar metrics as those proposed in a recent study (Jaume et al. 2021b). They designed a set of quantitative metrics based on the statistics of the class separability using pathologically measurable concepts (like nuclear shape) to characterize graph-structural models. Since it is devised for graph model, to make it applicable to our framework, we choose nucleus on important patches instead of nucleus on graphs. We also notice that they use a set of features which are suitable for some specific type of cancer whose WSIs have sparse distributed cells, but are not fit for other types of cancers whose cells are closely distributed or even overlapped in WSIs. To address this issue, we use pathologically measurable features (like nuclear area) as our basic unit in order to get the suitable information to distinguish more types of cancer.

Based on the aforementioned motivations, we propose our explainability evaluation method as follows. Firstly, we pick out important patches for every WSI with the approach proposed in the "Post-hoc Explanations" Section, and extract the nucleus from the patches with pre-trained Hover-net (Graham et al. 2019). Then we divide the nucleus into two categories according to the types of patients they belong to. Note that the patients are categorized into two types according to their survival time lengths, with the median value as the division point (more details are provided in the next section). For both categories of nucleus, we extract pathologically measurable features using toolkits provided in (Jaume et al. 2021a), and calculate the probability distribution for every feature. Given the probability distributions, we convert them into the probability density functions for every feature in both types. Then we can compute the separability score based on the optimal transport as the Wasserstein distance (Panaretos and Zemel 2018) between the feature density functions of two types. Finally, the evaluation metrics for explainability can be calculated as follows:

$$\begin{aligned} s_{\max} &= \max_{f \in \mathcal{F}} S_f \\ s_{\text{avg}} &= \frac{1}{|F|} \sum_{f \in \mathcal{F}} S_f \end{aligned} \tag{10}$$

where F is the set of pathologically measurable features, and S_f denotes the separability score of feature f. $s_{\max}$ and s_{avg} represent the utmost and expected separability between nuclear features of different categories respectively.

Results and Discussion

Prediction Survival

Table 3 shows the performance of various survival models on two datasets in terms of C-index values. The results show that our ESAT outperforms all other competitors on both datasets. Compared to the state-of-the-art baseline DeepAttnMISL, our model improves C-index by a large margin (more than 10%).

LASSO-Cox and logistic models are traditional methods that use ROI and hand-crafted features which depend on the experience of pathologists in selecting pathological areas and extracting features. The predicted C-index values of these two models are close to the reported results in previous studies (Zhu et al. 2017; Yao et al. 2020).

The performance of RankDeepSurv, DeepAttnMISL, and WSISA outperform LASSO-Cox and logistic, demonstrating the superior ability of deep learning models in learning effective image representations for survival analysis. However, they still perform worse than our ESAT. One possible reason is that ESAT considers all the patches rather than sampled areas. Another reason is that our model can better capture distinct image patterns with the ability of convolution-involved vision transformer.

Although ViT also uses all the patches, ESAT still outperforms it by a large margin. That's because the vanilla ViT has a very large computing complexity and has to reduce the number of patches. Thus, they use a small number of large patches as input, and a larger patch may bring more noises that are difficult to filter, leading to poor performance. On the contrary, in ESAT, with the help of the convolutional layer, we can accommodate a large number of small patches, and the patches with a large portion of noises can be easily discarded by the attention mechanism of ViT, thus yielding superior performance.

Table 3: Performance comparison of different models using C-index values on two datasets

Methods	**NLST**	**CHCAMS**
LASSO-Cox (Tibshirani 1997)	0.517	0.474
Logistic (Lee and Wang 2003)	0.514	0.500
ViT(Dosovitskiy et al. 2020)	0.563	0.536
RankDeepSurv (Jing et al. 2019)	0.541	0.541
WSISA (Zhu et al. 2017)	0.662	0.631
DeepAttnMISL (Yao et al. 2020)	0.630	0.628
ESAT (ours)	**0.730**	**0.707**

Ablation Study

To validate the effectiveness of different components in ESAT, we use the leave-one-out scheme to evaluate the impact of two modules: ResNet pre-trained model and the survival risk loss. Note that, the effectiveness of the convolution layer in our model has been validated in the previous section when comparing ESAT with ViT. Here we develop two variants: (1) **ESAT-pre** which uses random parameters instead of the ResNet pre-trained model; (2) **ESAT-SRL** which uses BCEloss instead of the survival risk loss.

Figure 2 compares the C-index values of these variations in two datasets. We can observe that the pre-trained

ResNet and the survival risk loss modules are both beneficial to our task. The performance of ESAT-SRL is still better than the state-of-the-art baselines. This demonstrates that despite using BCEloss, using all the patches with our proposed Convolution-involved Vision Transformer can still yield superior performance.

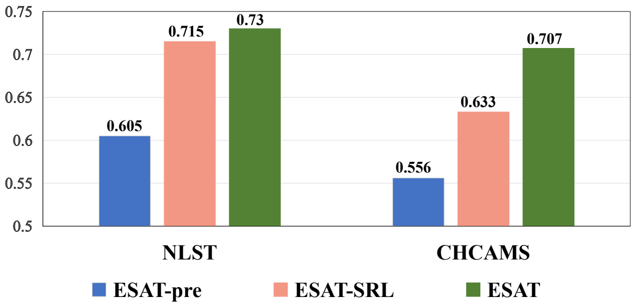

Figure 2: The C-index performance of ESAT and its variants on the NLST and CHCAMS datasets.

Explainable Result

We perform the explainability evaluation from two perspectives. From the model-level perspective, we compare ESAT with other explainable models in terms of separability scores defined in Eq.(10), aiming to show that our model can successfully identify the important patches that can be used to distinguish patients with different survival times. From the feature-level perspective, we aim to provide a set of most important features according to their separability scores.

Model-level explainablity comparison To compare the explainability of models using separability scores, we divide survival times into two categories by median: (1) 0-3 years, (2) 4+ years. For our model, the top 5 important patches with size 512 × 512 pixels are extracted with our post-hoc explainable module. We compare our model with DeepAttnMISL and a RANDOM explainer with the random patch selection strategy for each WSI for comparison. For DeepAttnMISL, it aggregates patches into different clusters and uses the attention mechanism to select the clusters with the largest weight as explanations. Based on the important patches obtained by these models and the associated nucleus, we extract 24 nuclear morphology features as (Jaume et al. 2021b), and obtain the avg and max separability scores according to Eq. (10). The results on CHCAMS are shown in Table 4. Similar results on the NLST dataset can be found in the Appendix.

It can be observed that ESAT achieves the best maximum and average separability scores. Both our model and DeepAttnMISL outperform RANDOM which conveys that the attention mechanism in deep learning models are effective in selecting important patches that can be used to distinguish different categories of patients.

Features ranked by separability scores After obtain the important patches for different patients, we aim to investigate their fine-grained differences. Specifically, we output the most important nuclear features that can be used to distinguish the two types of patients. Specifically, we rank the 24 nuclear features by their separability scores and output the top 5 and bottom 5 in Fig.3. It can be observed that the nuclear roundness is the most distinctive feature between the two types, followed by solidity and glcmASM. In contrast, the nuclear area is one of the least distinctive features. Compared to the important patches, these features can provide fine-grained and quantitative characteristics that can be more easily comprehended by the pathologists.

Table 4: Explainability comparison of the proposed models and other methods using two separability metrics on CHCAMS dataset.

Methods	s_{avg}	$s_{\max}$
RANDOM	0.0879	0.1146
DeepAttnMISL	0.0939	0.1315
ESAT	**0.1262**	**0.2271**

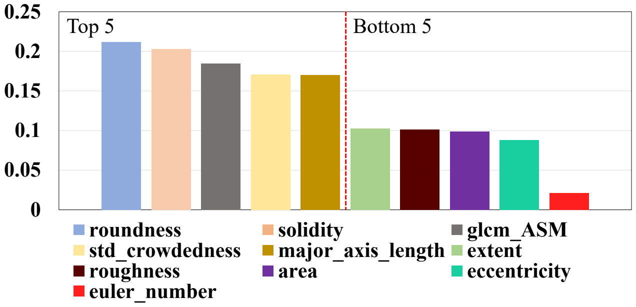

Figure 3: The separability scores of the top 5 and bottom 5 features, which are presented on the left and right parts, respectively.

Conclusion

In this paper, we propose an explainable survival analysis framework with a convolution-involved vision transformer, which can learn effective and explainable survival patterns from the whole slide histopathological images. Compared to existing image-based survival models, our proposal can learn from the complete image information and thus achieve better performance. Moreover, the interpretable part of our framework can provide both important patches and informative quantitative indicators as explanations, which are easy to comprehend and can facilitate the doctors in cancer diagnosis. The explanation module is post-hoc, model-agnostic and can be easily deployed in other image-based survival models. Evaluations on two cancer datasets demonstrate that our model can outperform the state-of-the-art baselines and provide explanations with better separability ability. In future, it would be interesting to introduce the transfer learning approach to make our model work on some specific type of cancer that lacks sufficient labeled data.

Acknowledgements

This work was supported by the Natural Science Foundation of China (No.61976026, No.61902394) and 111 Project (B18008). The authors would like to thank the National Cancer Institute for access to NCI's data collected by the National Lung Screening Trial. The statements contained herein are solely those of the authors and do not represent or imply concurrence or endorsement by NCI.

References

Bair, E.; Tibshirani, R.; and Golub, T. 2004. Semi-supervised methods to predict patient survival from gene expression data. *PLoS biology*, 2(4): e108.

Barker, J.; Hoogi, A.; Depeursinge, A.; and Rubin, D. L. 2016. Automated classification of brain tumor type in whole-slide digital pathology images using local representative tiles. *Medical image analysis*, 30: 60–71.

Cheng, J.; Mo, X.; Wang, X.; Parwani, A.; Feng, Q.; and Huang, K. 2018. Identification of topological features in renal tumor microenvironment associated with patient survival. *Bioinformatics*, 34(6): 1024–1030.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Graham, S.; Vu, Q. D.; Raza, S. E. A.; Azam, A.; Tsang, Y. W.; Kwak, J. T.; and Rajpoot, N. 2019. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical Image Analysis*, 101563.

Hägele, M.; Seegerer, P.; Lapuschkin, S.; Bockmayr, M.; Samek, W.; Klauschen, F.; Müller, K.-R.; and Binder, A. 2020. Resolving challenges in deep learning-based analyses of histopathological images using explanation methods. *Scientific reports*, 10(1): 1–12.

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

Jaume, G.; Pati, P.; Anklin, V.; Foncubierta, A.; and Gabrani, M. 2021a. HistoCartography: A Toolkit for Graph Analytics in Digital Pathology. *arXiv preprint arXiv:2107.10073*.

Jaume, G.; Pati, P.; Bozorgtabar, B.; Foncubierta, A.; Anniciello, A. M.; Feroce, F.; Rau, T.; Thiran, J.-P.; Gabrani, M.; and Goksel, O. 2021b. Quantifying explainers of graph neural networks in computational pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8106–8116.

Jing, B.; Zhang, T.; Wang, Z.; Jin, Y.; Liu, K.; Qiu, W.; Ke, L.; Sun, Y.; He, C.; Hou, D.; et al. 2019. A deep survival analysis method based on ranking. *Artificial intelligence in medicine*, 98: 1–9.

Katzman, J. L.; Shaham, U.; Cloninger, A.; Bates, J.; Jiang, T.; and Kluger, Y. 2016. Deep survival: A deep cox proportional hazards network. *stat*, 1050(2): 1–10.

Lee, E. T.; and Wang, J. 2003. *Statistical methods for survival data analysis*, volume 476. John Wiley & Sons.

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

Lu, M. Y.; Williamson, D. F.; Chen, T. Y.; Chen, R. J.; Barbieri, M.; and Mahmood, F. 2021. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6): 555–570.

Panaretos, V. M.; and Zemel, Y. 2018. Statistical Aspects of Wasserstein Distances. *arXiv: Methodology*.

Shedden, K.; Taylor, J. M.; Enkemann, S. A.; Tsao, M. S.; Yeatman, T. J.; Gerald, W. L.; Eschrich, S.; Jurisica, I.; Venkatraman, S. E.; Meyerson, M.; et al. 2008. Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study: Director's Challenge Consortium for the molecular classification of lung adenocarcinoma. *Nature medicine*, 14(8): 822.

Tang, B.; Li, A.; Li, B.; and Wang, M. 2019. CapSurv: capsule network for survival analysis with whole slide pathological images. *IEEE Access*, 7: 26022–26030.

Team, N. L. S. T. R. 2011. The national lung screening trial: overview and study design. *Radiology*, 258(1): 243–253.

Tibshirani, R. 1997. The lasso method for variable selection in the Cox model. *Statistics in medicine*, 16(4): 385–395.

Wang, H.; Xing, F.; Su, H.; Stromberg, A.; and Yang, L. 2014. Novel image markers for non-small cell lung cancer classification and survival prediction. *BMC bioinformatics*, 15(1): 1–12.

Warth, A.; Muley, T.; Meister, M.; Stenzinger, A.; Thomas, M.; Schirmacher, P.; Schnabel, P. A.; Budczies, J.; Hoffmann, H.; and 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. *Journal of clinical oncology*, 30(13): 1438–1446.

Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; and Zhang, L. 2021. Cvt: Introducing convolutions to vision transformers. *arXiv preprint arXiv:2103.15808*.

Xiong, Y.; Zeng, Z.; Chakraborty, R.; Tan, M.; Fung, G.; Li, Y.; and Singh, V. 2021. Nystr\" omformer: A Nystr\" om-Based Algorithm for Approximating Self-Attention. *arXiv preprint arXiv:2102.03902*.

Yao, J.; Ganti, D.; Luo, X.; Xiao, G.; Xie, Y.; Yan, S.; and Huang, J. 2015. Computer-assisted diagnosis of lung cancer using quantitative topology features. In *International Workshop on Machine Learning in Medical Imaging*, 288–295. Springer.

Yao, J.; Zhu, X.; Jonnagaddala, J.; Hawkins, N.; and Huang, J. 2020. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical Image Analysis*, 65: 101789.

Yao, J.; Zhu, X.; Zhu, F.; and Huang, J. 2017. Deep correlational learning for survival prediction from multi-modality data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 406–414. Springer.

You, Y.; Chen, T.; Sui, Y.; Chen, T.; Wang, Z.; and Shen, Y. 2020. Graph contrastive learning with augmentations. *NeurIPS*.

Yuan, L.; Chen, Y.; Wang, T.; Yu, W.; Shi, Y.; Jiang, Z.; Tay, F. E.; Feng, J.; and Yan, S. 2021. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *arXiv preprint arXiv:2101.11986*.

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.; and Huang, J. 2017. Deep convolutional neural network for survival analysis with pathological images. In *IEEE International Conference on Bioinformatics Biomedicine*.

Zhu, X.; Yao, J.; Zhu, F.; and Huang, J. 2017. Wsisa: Making survival prediction from whole slide histopathological images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7234–7242.