Computer Methods and Programs in Biomedicine 241 (2023) 107733

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

Surformer: An interpretable pattern-perceptive survival transformer for cancer survival prediction from histopathology whole slide images

Zhikang Wang [a,b,1], Qian Gao [a,1], Xiaoping Yi [a], Xinyu Zhang [c], Yiwen Zhang [d], Daokun Zhang [c], Pietro Liò [e], Chris Bain [c], Richard Bassed [f], Shanshan Li [d], Yuming Guo [d], Seiya Imoto [g], Jianhua Yao [h,*], Roger J. Daly [b,**], Jiangning Song [b,**]

[a] Xiangya Hospital, Central South University, Changsha, Hunan, PR China
[b] Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Melbourne, Australia
[c] Faculty of Information Technology, Monash University, Melbourne, Australia
[d] Climate, Air Quality Research Unit, School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia
[e] Department of Computer Science and Technology, The University of Cambridge, Cambridge, United Kingdom
[f] Victorian Institute of Forensic Medicine, Melbourne, Australia
[g] Human Genome Center, The Institute of Medical Science, The University of Tokyo, Japan
[h] Tencent AI Lab, Tencent, Shenzhen, PR China

ARTICLE INFO

Keywords:
Survival analysis
Multiple instance learning
Whole slide image
Deep learning interpretation

ABSTRACT

Background and Objective: High-resolution histopathology whole slide images (WSIs) contain abundant valuable information for cancer prognosis. However, most computational pathology methods for survival prediction have weak interpretability and cannot explain the decision-making processes reasonably. To address this issue, we propose a highly interpretable neural network termed pattern-perceptive survival transformer (Surformer) for cancer survival prediction from WSIs.

Methods: Notably, Surformer can quantify specific histological patterns through bag-level labels without any patch/cell-level auxiliary information. Specifically, the proposed ratio-reserved cross-attention module (RRCA) generates global and local features with the learnable prototypes (p_{global}, p_{locals}) as detectors and quantifies the patches correlative to each p_{local} in the form of ratio factors (*rfs*). Afterward, multi-head self&cross-attention modules proceed with the computation for feature enhancement against noise. Eventually, the designed disentangling loss function guides multiple local features to focus on distinct patterns, thereby assisting rfs from RRCA in achieving more explicit histological feature quantification.

Results: Extensive experiments on five TCGA datasets illustrate that Surformer outperforms existing state-of-the-art methods. In addition, we highlight its interpretation by visualizing *rfs* distribution across high-risk and low-risk cohorts and retrieving and analyzing critical histological patterns contributing to the survival prediction.

Conclusions: Surformer is expected to be exploited as a useful tool for performing histopathology image data-driven analysis and gaining new insights for interpreting the associations between such images and patient survival states.

1. Introduction

Survival analysis of time-to-event data has been widely applied in varied domains, like biology, engineering, economics, and medicine [1–4]. Particularly in the area of medicine, it plays a critical role in our understanding of the effect of specific patient features with respect to survival status. The precise prognosis of malignant cancers can guide treatment and clinical management options, thereby having prominent commercial and clinical significance.

In the current clinical paradigm, survival analysis is realized based on the visual inspection of pathological alterations/features in cell morphology, invasiveness, or inflammation/infiltration in histopathology

* Corresponding author at: Tencent AI Lab, Shenzhen 518054, PR China.
** Corresponding authors at: 19 Innovation Walk, Monash University Clayton Campus, VIC 3800, Australia.
E-mail addresses: Jianhuayao@tencent.com (J. Yao), Roger.Daly@monash.edu (R.J. Daly), Jiangning.Song@monash.edu (J. Song).
[1] Cofirst authors.

https://doi.org/10.1016/j.cmpb.2023.107733
Received 21 April 2023; Received in revised form 12 June 2023; Accepted 23 July 2023
Available online 28 July 2023

slides [5–7]. For instance, Goff et al. [6] proposed that the spatial proportion of tumor-infiltrating lymphocytes (TILs) in breast cancer might serve as an important prognostic indicator. Campbell et al. [7] found that high-grade ductal carcinoma in situ (HG-DCIS) has more TILs compared to non-high-grade DCIS (nHG-DCIS). Nevertheless, cancer patient survival prediction is a challenging task for clinicians and pathologists due to its subjective nature. Clinicians and pathologists with distinct knowledge and experience can have different interpretations on the same histopathology image. Besides, inspecting gigapixel histopathology images is laborious and time-consuming, significantly intensifying the workload and pressure on pathologists. Therefore, it is urgent to study targeted and automated survival prediction algorithms with significant performance.

In recent years, the combination of whole slide imaging and deep learning techniques has advanced the development of automated whole slide image (WSI)-based cancer diagnosis/subtyping algorithms [8–10], some even achieving performances on par with expert pathologists. However, current WSI-based survival analysis algorithms still suffer from limited performance and interpretability. The performance and interpretability gap between the two tasks results from their different objectives: cancer diagnosis/subtyping algorithms merely need to detect critical instances from the whole slides while survival analysis requires to integrate instance-level and global-level features in the tumor and surrounding tissues for assessing the patient's risk of mortality. As a result, most existing multiple instance learning (MIL) methods [11–13] that follow the standard MIL assumption (if a bag contains at least one positive instance, it is labeled positive, else negative) cannot effectively interpret the correlations between the global and local features, making them unsuitable for the survival analysis task.

1.1. Related work

Typically, WSIs come with a gigapixel resolution, making it infeasible for existing hardware to execute the computation in an end-to-end function. As such, computationally efficient patch-based methods dominate this area. These techniques can be categorized into two groups: *ROI (regions of interest)-based* and *WSI-based methods*.

ROI-based methods: Traditional methods generally select several discriminative patches from manually annotated Regions of Interest (ROI) and extract features for prediction. At the early stage, ROI-based methods utilized hand-crafted features of the ROI for survival prediction [14–18]. Barker et al. [14] utilized a coarse-to-fine analysis of the localized characteristics in pathology by firstly extracting spatially localized features and secondly analyzing a single representative tile from each group for brain tumor detection. Cheng et al. [15] proposed a novel bioimage informatics pipeline for automatically characterizing the topological organization of different cell patterns in the tumor microenvironment. Notably, the proposed features provided new insights into the topological organizations for cancers and could also combine genomic data to develop new biomarkers. Yu et al. [17] extracted 9,879 quantitative image features and utilized regularized machine-learning methods to select the top features and distinguish shorter-term and longer-term survivors. With the technology advancing, deep learning-based models with better representation learning capability have outperformed traditional hand-crafted-based methods. Mobadersany et al. [19] proposed a computational approach to learn patient outcomes, which took advantage of the power of both adaptive machine learning algorithms and traditional survival models. The proposed survival convolutional neural networks (SCNNs) are able to integrate both histopathology images and genomic biomarkers.

WSI-based methods: With the advances of deep learning technologies and publish of large datasets, WSI-based methods dominate the state-of-the-art performance owing to their strong representation capability, ranging from characterizing prominent morphological phenotypes to predicting human eye invisible gene mutations.

Limited by the huge pixel size of whole slide images (WSIs) and to date hardware computational capability, most methods approach WSIs through weakly-supervised multiple instance learning (MIL). Specifically, MIL comprises two steps: (1) extracting small patches from the WSIs as independent instances, (2) generating the bag representation through the bag of instances by pooling or various aggregation methodologies, and conducting the final prediction. Yao et al. [13] proposed the deep attention multiple instance survival learning (DeepAttnMISL) by leveraging the siamese Multiple Instance Fully Convolutional Network (MI-FCN) and an attention-based MIL pooling to construct WSI features. Meanwhile, K-means clustering was adopted based on deep-transferred features to reduce computational costs. To enable more precise prediction, Wang et al. [20] proposed a deep learning framework leveraging hierarchical graph-based representations to explore multi-scale topological structures of WSIs comprehensively. In addition, Chen et al. [21] proposed the patch-based graph convolutional network (Patch-GCN), which is also a spatially resolved graph-based algorithm. Patch-GCN hierarchically aggregates instance-level histology features to model local and global topological structures in the tumor microenvironment. Huang et al. [22] introduced the Transformer to adaptatively aggregate patch-level features according to the spatial information and correlation between patches for survival analysis. Aiming to capture intratumoral heterogeneity during the survival prediction, Carmichael et al. [23] developed a novel variance pooling architecture that enables a MIL model to incorporate intratumoral heterogeneity into its prediction.

1.2. Contributions

This study proposes an innovative MIL neural network termed pattern-perceptive survival transformer (Surformer) for WSI-based survival analysis. Briefly, Surformer comprises three components: a ratio-reserved cross-attention module (RRCA), a multi-head self-attention module (MHSA) [24], and a multi-head cross-attention module (MHCA). RRCA simultaneously detects global features and multiple pattern-specific local features through a learnable global prototype p_{global} and multiple local prototypes p_{locals} and quantifies the patches correlative to each p_{locals} in the form of ratio factors (*rfs*). The ratio information is subsequentially embedded into the feature space for representation enhancement. As a quantification index, *rfs* differentially analyze the high-risk and low-risk patients and indirectly interpret the model prediction. The MHSA and MHCA establish and disentangle the contextual connections between features, aiming to further optimize the learnable prototypes and improve the network's ability against noise. Weight sharing between different modules is adopted for stabilizing the network training. Eventually, the proposed disentangling loss $\mathcal{L}_{dis}$ constrains local features to focus on distinct patterns, thereby assisting rfs from RRCA in achieving more explicit quantification. The proposed RRCA and $\mathcal{L}_{dis}$ jointly contribute to the advanced interpretability of the Surformer.

To validate the performance of Surformer, we conduct benchmarking experiments on five TCGA tumor datasets, including Bladder Urothelial Carcinoma (BLCA), Breast Invasive Carcinoma (BRCA), Glioblastoma & Lower Grade Glioma (GBMLGG), Lung Adenocarcinoma (LUAD), and Uterine Corpus Endometrial Carcinoma (UCEC). The experimental results demonstrate that the proposed Surformer can accurately model the risk function of the population, standing out from state-of-the-art methods. Besides, the proposed Surformer can quantify specific histological patterns and explicitly interpret the statistical correlations between patterns and overall survival. In summary, both the state-of-the-art survival prediction performance and great interpretability of Surformer are significant in precision medicine.

The contribution of this paper can be summarized as follows:

(1) The proposed Surformer achieves significant performance improvement compared with current state-of-the-art methods on five benchmarking datasets;

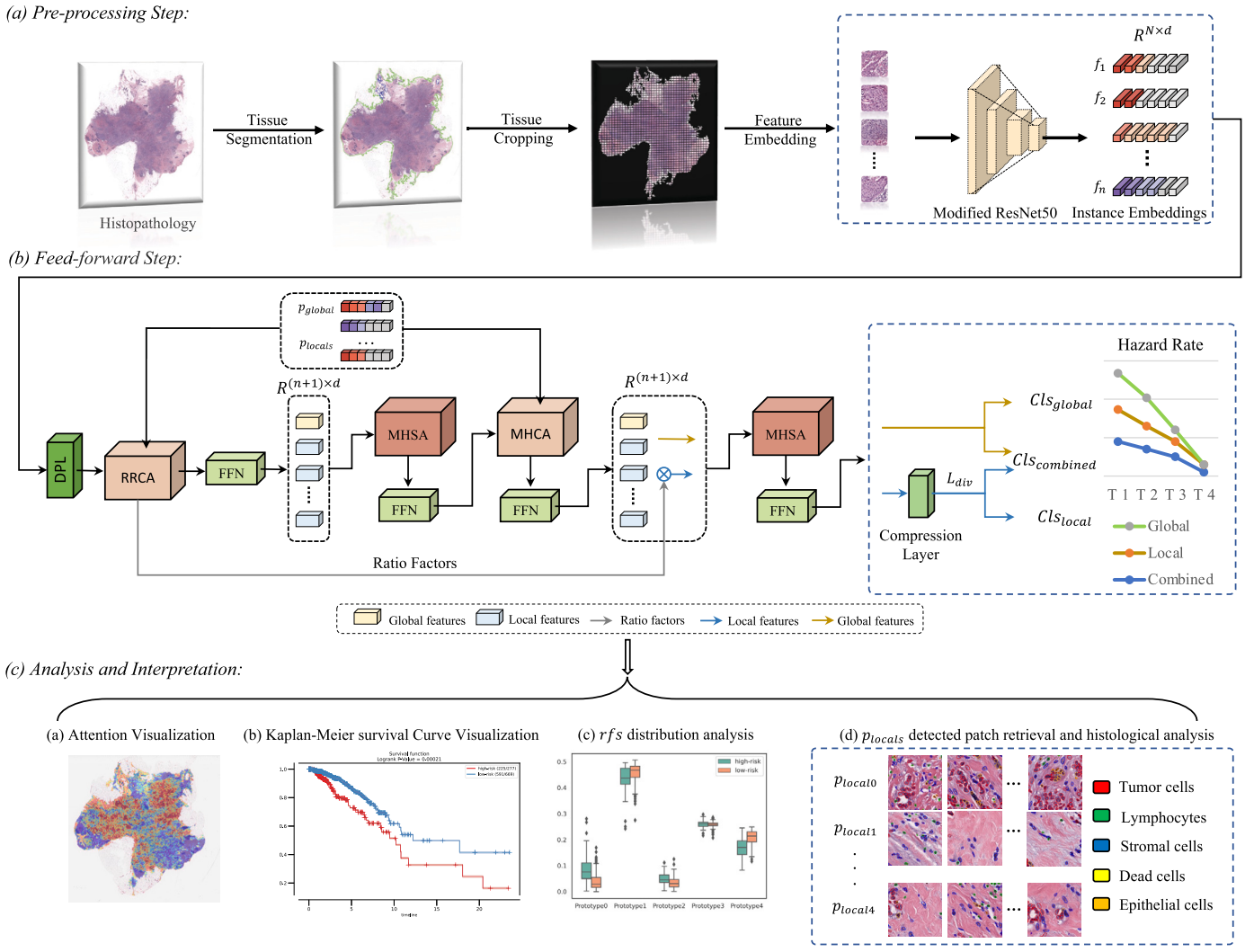

Fig. 1. Overview of the proposed pattern-perceptive survival transformer, termed Surformer. It consists of two steps: (1) the pre-processing step for converting the gigapixel WSI into computable feature vectors; (2) the feed-forward step for predicting the overall survival of patients. Here, we denote the ratio factors, local features, and global features with gray, yellow, and green lines, respectively. We apply the disentangling loss on the multiple local features to guide them in having distinct pattern attention. Eventually, both local features, and global features, and their combination will be utilized for hazard rate analysis through independent layers. Meanwhile, we interpret and analyze the proposed Surformer in terms of attention visualization (a), Kaplan-Meier survival curve visualization (b), rfs analysis (c), and p_{locals} detected patch retrieval and histological analysis.

(2) The novel ratio-reserved cross-attention module and the disentangling loss cooperate to achieve the quantification of the critical histological patterns in the form of ratio factors rfs;

(3) We interpret the model by statistically analyzing the rfs across high-risk and low-risk cohorts and retrieving and analyzing the critical histological patterns (detected by each p_{local} of RRCA) that contribute to the overall survival.

2. Method

Given a patient i, he/she is annotated with labels t_i and δ_i and is paired with at least one WSI. Here, t_i indicates the overall survival, whereas δ_i represents the censorship (i.e., uncensored or censored). Our objective is to train a deep neural network with a strong capability in overall survival analysis through the provided WSIs and labels. The overall framework is shown in Fig. 1. In the following section, we will introduce the pre-processing step, the proposed RRCA, the feed-forward step, and loss functions in sequence.

2.1. Pre-processing step

Firstly, we use an automated segmentation algorithm to distinguish each WSI's foreground (tissue region) and background, and then crop the tissue region into patches with a fixed size. Secondly, an ImageNet pre-trained model is leveraged to embed each patch from the original high-dimensional image space into the low-dimensional feature space. Following [21], we construct the pre-trained model by the first Convolution Block and the first three Residual Blocks of the ResNet50 model [25]. Accordingly, each patch is embedded into a 1024-dimensional feature vector, with the bag of instances represented as $F = \{f_1, f_2, \ldots, f_N\} \in \mathbb{R}^{N\times 1024}$, in which the first and second dimensions are termed as *instance dimension* and *feature dimension*, respectively. After the feature embedding, training and inference can be operated in the low-dimensional feature space rather than the original high-dimensional image space, which significantly reduces the subsequent computational costs.

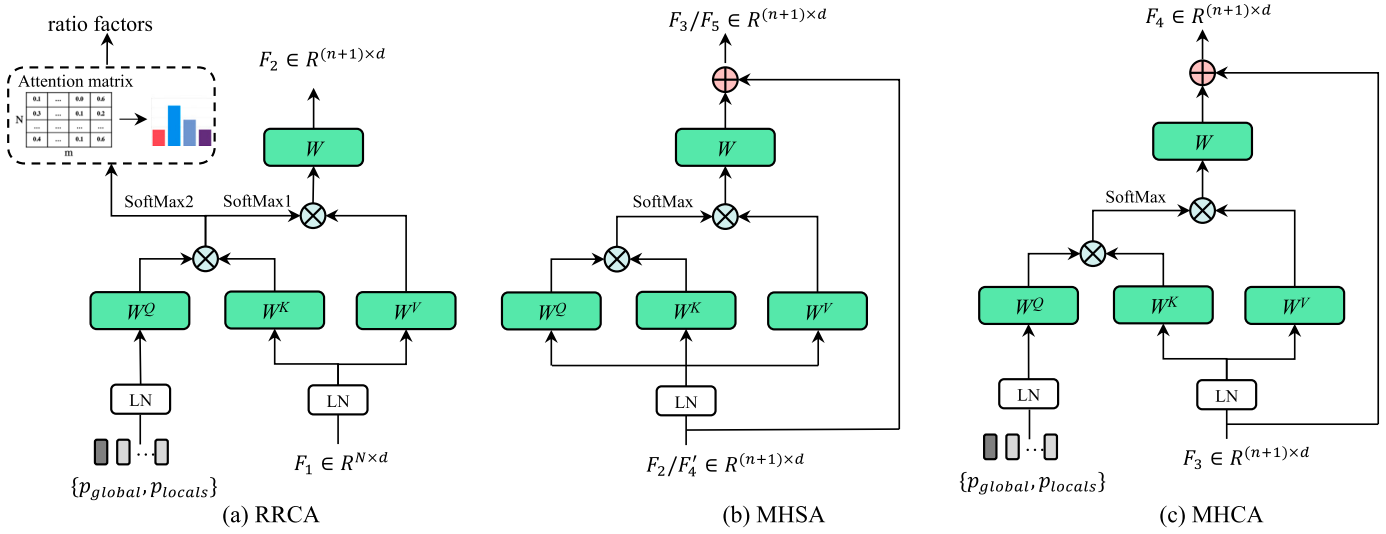

Fig. 2. Illustration of the ratio-reserved cross-attention module (RRCA), multi-head self-attention module (MHSA), and multi-head cross-attention module (MHCA).

2.2. Ratio-reserved cross-attention module

In Fig. 2, we illustrate the ratio-reserved cross-attention module (RRCA), multi-head self-attention module (MHSA) and multi-head cross-attention module (MHCA). In this subsection, we first detail the proposed RRCA module and then summarize the differences between the three modules from both functional and architectural perspectives.

RRCA is the most important component of Surformer. As shown in Fig. 2 (a), RRCA has a query, gallery and key. Specifically, query Q arises from the p_{global} and p_{locals}, and key K and value V arise from the transformed features F_1. $p_{global} \in \mathbb{R}^{1\times d}$ and $p_{locals} \in \mathbb{R}^{n\times d}$ are both predefined learnable parameters and will be optimized through the loss functions in the training phase. Formally,

$$\begin{aligned} Q &= LN([p_{global}, p_{locals}])W^Q + b^Q, \\ K &= LN(f')W^K + b^K, \\ V &= LN(f')W^V + b^V, \end{aligned} \tag{1}$$

where $W \in \mathbb{R}^{d\times d'}$ indicates a linear projection operation and b is the bias. We keep the multi-head strategy of MHSA, partitioning the features into m segments along the *feature dimension* and getting $Q \in \mathbb{R}^{(n+1)\times m\times(d'/m)}$ and $K, V \in \mathbb{R}^{N\times m\times(d'/m)}$. Next, we conduct the cross-product multiplication operation to get the attention matrix $mt \in \mathbb{R}^{m\times(n+1)\times N}$ and apply $softmax$ operation along the third dimension to get normalized mt'. For simplicity, the number of heads m is supposed as 1 and the process can be formulated as follows:

$$mt'_{i,j} = \frac{exp(mt_{i,j})}{\sum_{j=1}^{N} exp(mt_{i,j})}, \quad mt_{i,j} = \frac{Q_i K_j}{\sqrt{d_k}}, \tag{2}$$

where $\sqrt{d_k}$ is a scaling factor and, i and j present the indices of *query* and *key*, respectively. Each element of the attention matrix indirectly indicates the similarity between *query* and *key*. The final aggregated features are generated as follows:

$$F_2 = (mt' \times V)W + b. \tag{3}$$

In the above computation process, both p_{global} and p_{locals} function as detectors to gather correlative features from WSIs. We denote the features detected by p_{global} and p_{locals} as global features $F_{global2}$ and local features F_{local2}, respectively. Here, the p_{global} is similar to the classification token of Bert [26] and targets to aggregate all valuable features for survival prediction. Regarding p_{locals}, the disentangling loss (2.4) is employed on the detected features. By optimizing the indirect constraints, each p_{local} is encouraged to exhibit distinct histological pattern attention, thereby achieving pattern-specific local feature aggregation from the whole image.

The *rfs* are also derived from the mt. Essentially, rfs are the quantification of the patches detected by each p_{local}. Specifically, we first slice the attention matrix that is related to the p_{locals} and get $mt^{locals} \in \mathbb{R}^{m\times n\times N}$. Then, $average$ and $softmax$ operations are conducted along the first and second dimensions for generating the final affinity scores between each instance and p_{locals}. The softmax function can be formulated as follows:

$$mt''_{i,j} = \frac{exp(mt^{locals}_{i,j})}{\sum_{i=1}^{n} exp(mt^{locals}_{i,j})}. \tag{4}$$

Afterwards, we adopt the *maximum* operation to get the highest affinity score for each instance. To relieve the adverse influence of the noise, a pre-defined threshold is utilized to filter out the instances that are not correlated to any p_{locals}. Generally, the threshold is $1/n+0.1$. At last, we quantify the remaining instances corresponding to each p_{local} and the final ratio factors (rfs) can be expressed as:

$$rfs = \{\frac{num_i}{\sum_{j=1}^{n} num_j}; i = 1, ..., n\}, \tag{5}$$

where i and j are the indices and $num_{i/j}$ indicates the number of detected patches correlated i/j_{th} p_{local}.

Although RRCA, MHCA, and MHSA look similar, they essentially have big differences. From the functional perspective, the three modules have distinct characteristics. As the key module in Surformer, RRCA has three functions: (1) detect and aggregate both global and local critical features related to survival analysis from the transformed feature vectors F_1 with the assistance of the p_{global} and p_{locals}, (2) reduce the instance dimension from N to $n+1$ ($N \gg n$), making the subsequent computation efficiently, (3) generate the rfs that represent the spatial ratio of specific histological features related to p_{locals}. Likewise, MHCA takes the aggregated features, p_{locals} and p_{global} as input, aiming to further distill critical features and optimize the learnable prototypes (p_{locals} and p_{global}). MHSA aims to establish long-range dependencies between the generated features, thereby mutually enhancing their representativeness. From the architectural perspective, the three modules are highly similar. Both RRCA and MHCA are derived from the MHSA. The main differences between RRCA and MHSA lie in the rfs generation process and the residual operation between the input and output features. In the term of input, RRCA takes p_{locals}, p_{global}, and transformed features F_1, while MHSA only takes the aggregated global and local features. Regarding MHCA, it takes aggregated features as input and has an additional residual operation between input and output fea-

tures. Meanwhile, MHCA does not generate the rfs. As a result, the commonality between RRCA and MHCA ensures that their parameters can be shared during computation.

Algorithm 1: Overview of feed-forward step

Input: one bag of instances $F = \{f_1, ..., f_N\} \in \mathbb{R}^{N\times 1024}$, one global prototype $p_{global} \in \mathbb{R}^{1\times d}$ and multiple local prototypes $p_{locals} \in \mathbb{R}^{n\times d}$
Output: The patient hazard rate $P \in \mathbb{R}^4$
1: DPL for feature transformation:
$F_1 = \text{DPL}(F)$;
2: RRCA and FFN for feature detection and aggregation:
$F_2, rfs = \text{FFN}(\text{RRCA}(F_1, p_{global}, p_{locals}))$,
$F_2 = \{F_{global2} \in \mathbb{R}^{1\times d}, F_{local2} \in \mathbb{R}^{n\times d}\}$;
3: MHSA and FFN for feature engineering:
$F_3 = \text{FFN}(\text{MHSA}(F_2))$;
4: MHCA and FFN for feature engineering:
$F_4 = \text{FFN}(\text{MHCA}(F_3, p_{global}, p_{locals}))$;
5: Spatial ratio factors embedding:
$F'_{local4} = F_{local4} \times (\alpha + rfs)$, $F'_4 = \{F'_{local4}, F_{global4}\}$;
6: MHSA and FFN for feature engineering:
$F_5 = \text{FFN}(\text{MHSA}(F'_4))$;
7: Compression Layer (CompL) for local feature dimension reduction:
$F'_{local5} = CompL(F_{local5})$;
8: Hazard rate prediction based on F'_{local5}, $F_{global5}$ and their combination.

2.3. Feed-forward step

In this subsection, we introduce the feed-forward computation of the proposed Surformer. The pseudocode is illustrated in Algorithm 1. We first use a deep projection layer (DPL) [10] to perform a non-linear transformation on the bag of features F and obtain the transformed features $F_1 = \{f_{1\_1}, f_{1\_2}, \dots, f_{1\_N}\} \in \mathbb{R}^{N\times d}$. Particularly, the DPL is constructed by two fully connected layers with intermediate *ReLU* and *LayerNorm* functions. Then, RRCA operates on the pre-defined learnable prototypes (global prototypes $p_{global} \in \mathbb{R}^{1\times d}$ and local ones $p_{locals} \in \mathbb{R}^{n\times d}$) and transformed bag features F_1. RRCA adaptatively aggregates critical features over the whole bag of instances (output both global features $f_{global} \in \mathbb{R}^{1\times d}$ and local features $f_{locals} \in \mathbb{R}^{n\times d}$) and generates spatial ratio factors rfs corresponding to different histological features. Subsequently, the MHSA and MHCA components explore long-range dependencies between the aggregated features in an encoder-decoder pipeline, which can better optimize the learnable prototypes (p_{global} and p_{locals}) and empower the model's robustness against the noise. Afterward, we explicitly encode the rfs into the local features after MHCA with the soft multiplication operation, which can be formulated as follows:

$$F'_{local4} = F_{local4} \times (\alpha + rfs) \tag{6}$$

where α is a predefined hyperparameter. By adding the hyperparameter α, we can weaken the ratio fluctuations caused by sample specificity, thereby improving the training stability. Eventually, we utilize the MHSA to establish the long-range dependencies between the *rfs*-encoded local features and global ones. Each attention module is followed by one independent Pre-LN feed-forward network (FFN) [27] for feature recalibration.

Although the proposed RRCA successfully compresses the feature dimension from $\mathbb{R}^{N\times d}$ to $\mathbb{R}^{n\times d}$ ($n \ll N$), the dimension of flattened local features grows significantly as n increasing, which leads to the computational overhead for optimizing the classification layers. To combat this issue, we apply a compression layer on the local features to reduce the channel dimensions before the final prediction. Then, the proposed disentangling loss $\mathcal{L}_{dis}$ is imposed on the compressed local features F'_{local} to indirectly constrain p_{locals} with distinct histological attention. Module reuse is also adopted to facilitate training and increase the stability of models. As such, our two MHSA modules shared the same parameters. Not only that, RRCA and MHCA also share parameters for their common architecture. Three classification layers are utilized to predict the overall survival on top of the global features $F_{global5}$, local features F'_{local5}, and their combination.

2.4. Loss function

Disentangling Loss: In RRCA, we utilize the cross-attention mechanism to calculate the affinities between instances and p_{locals} and quantify the spatial ratio information of histological patterns. Although p_{locals} are initially randomly initialized, the end-to-end training cannot guarantee p_{locals} with distinct visual attention, thus the n local features will have overlapping characteristics and the affinity between each instance and p_{locals} will be indistinct. Considering that our intention is quantifying each category of histological patches in the form of rfs, we propose a disentangling loss $\mathcal{L}_{dis}$ to disentangle the local features by applying constraints in the feature space. Instead of directly applying this loss function onto the features after the final FFN, we insert a compression layer, which comprises a linear layer for dimension reduction, a $LayerNorm$ for regularization, and a $ReLU$ activation function for non-linearization. Meanwhile, the compression layer, which acts as a buffer between the strong feature constraints and the reused feature extraction modules, can improve the training stability. Specifically, $\mathcal{L}_{dis}$ can be formulated as follows:

$$\mathcal{L}_{dis} = \frac{2}{N(N-1)} \sum_{i}^{N} \sum_{j,j\neq i}^{N} \frac{< F'_{local5,i}, F'_{local5,j} >}{||F'_{local5,i}||_2, ||F'_{local5,j}||_2}. \tag{7}$$

Essentially, it aims to increase the mutual distances among $F_{locals'}$ and indirectly pass the constraints to the detectors p_{locals}.

Cross-entropy-based Cox proportional loss [28]: The survival prediction datasets have both censored and uncensored data. To achieve the prediction, we first convert the continuous overall survival time into four non-overlapping bins: $[t_0, t_1)$, $[t_1, t_2)$, $[t_2, t_3)$, $[t_3, t_4)$, where $t_0 = 0$, $t_4 = \infty$, and t_1, t_2, t_3 are the quartiles of overall survival for uncensored patients. For patient j with t_j, we get his/her discretised class label y_j by referring to the above bins. Therefore, the final loss function can be formulated as:

$$\begin{aligned} \mathcal{L}_{cox} = &- c_j \cdot log(f_{surv}(y_j, F'_j)) \\ &- (1 - c_j) \cdot log(f_{surv}(y_j - 1, F'_j)) \\ &- (1 - c_j) \cdot log(f_{hazard}(y_j, F'_j)), \end{aligned} \tag{8}$$

where F'_j indicates the bag-level representation of j_{th} patient and f_{hazard} represents the prediction of hazard rates. In terms of f_{surv}, it can be formulated as:

$$f_{surv}(y_j, F'_j) = \prod_{i=0}^{y_j} (1 - f_{hazard}(i, F'_j)). \tag{9}$$

Total Loss: In total, there are four loss functions in the training process. One $\mathcal{L}_{dis}$ enforces constraints in the feature space. Three cross-entropy-based Cox proportional loss functions $\mathcal{L}_{cox}$ [28] optimize the hazard rates based on the global features, local features, and their combination.

Eventually, the total loss function for the whole model can be formulated as:

$$\begin{aligned} \mathcal{L} = &\mathcal{L}_{cox}(F_{global5}, c, y) + \mathcal{L}_{cox}(F'_{local5}, c, y) \\ &+ \mathcal{L}_{cox}([F_{global5}, F'_{local5}], c, y) + \mathcal{L}_{dis}, \end{aligned} \tag{10}$$

where c indicates the censoring of the data, y is the label, and $[\cdot]$ represents the concatenation operation.

2.5. Implementation details

In the *pre-processing step*, we crop each WSI into a series of 256 × 256 non-overlapping patches and discard the background patches (saturation < 15). The modified ImageNet pre-trained ResNet50 model (one

Table 1
Data details of the BLCA, BRCA, GBMLGG, LUAD and UCEC datasets. CS, US, and AP represent censored samples, uncensored samples and average patches, respectively.

Number of	Samples	CS	US	AP
BLCA	436	236	200	15014
BRCA	1022	889	133	9760
GBMLGG	1041	700	341	7495
LUAD	515	314	201	10973
UCEC	538	460	78	16142

Convolutional Block and three *Residual Blocks*) embeds original patches into a collection of feature vectors. In the *feed-forward process*, we train our network in an end-to-end fashion through the Adam optimizer for 20 epochs. The learning rate and weight decay are initialized as 2e-4 and 1e-5, respectively. We train the model with a batch size of 1 and 32 steps for gradient accumulation. All the experiments are conducted on one NVIDIA GeForce RTX 3090 Graphic Card.

In the experiments, each dataset is randomly split into training and testing sets with a ratio of 0.8 and 0.2. We introduce the five-fold cross-validation for verifying the feasibility, stability, and effectiveness of the proposed algorithm. Furthermore, we compare our proposed Surformer against several state-of-the-art weakly-supervised deep learning approaches. For a fair comparison, all these methods follow the same splitting and training strategy. We utilize the cross-validated concordance index (c-index) to measure the predictive performance of each model. Surformer is available for academic purposes at https://github.com/ZacharyWang-007/Surformer.

3. Results

3.1. Dataset description

In this paper, we conduct experiments across five different tumor types from The Cancer Genome Atlas (TCGA). The datasets are selected under rigorous criteria, including dataset size and balanced distribution of uncensored-to-censored patients. All the WSIs are processed at 20× magnification. The details of each dataset are shown in Table 1. Specifically, there contains Bladder Urothelial Carcinoma (BLCA) (n = 437), Breast Invasive Carcinoma (BRCA) (n = 1,022), Glioblastoma & Lower Grade Glioma (GBMLGG) (n = 1,011), Lung Adenocarcinoma (LUAD) (n = 515), and Uterine Corpus Endometrial Carcinoma (UCEC) (n = 538). After the segmentation and cropping operations, WSIs in BLCA, BRCA, GBMLGG, LUAD and UCEC have an average of 15,014, 9,760, 7,495, 10,973, and 16,142 patches, respectively. Meanwhile, we also introduce the number of censored samples (CS) and uncensored samples (US) of each dataset.

3.2. Ablation study

Analysis of the number of local prototypes. Different kinds of tumor are generally caused by different genetic aberrations, resulting in diversification in histological phenotypes. In Surformer, we utilize RRCA and p_{locals} to detect diversified histological patterns relative to the overall survival. In this process, the number of p_{locals} determines the varieties of patterns in consideration. Here, we specifically analyze the influence of the number of p_{locals} for different tumors. The experiment results are shown in Table 2. For BLCA and BRCA datasets, Surformer achieves the best performance with five p_{locals}, reaching 0.571 and 0.668, respectively. For GBMLGG, LUAD and UCEC datasets, Surformer performs best with four p_{locals}, reaching 0.861, 0.651, and 0.686, respectively. In terms of overall performance, models with four p_{locals} achieve the best results. Notably, the performance of Surformer on most tumor types drops dramatically when the numbers of p_{locals} are two or six. This suggests a small number of p_{locals} cannot cover the wide range of histological patterns but superfluous p_{locals} may inversely introduce too much noise. Generally, Surformer performs best with different numbers of p_{local} on different tumor datasets, which indirectly verifies the inter-tumor heterogeneity.

Table 2
Analysis of the number of local prototypes on TCGA datasets in terms of c-index.

Number	BLCA	BRCA	GBMLGG
2	0.556 ± 0.047	0.621 ± 0.071	0.849 ± 0.026
3	0.553 ± 0.045	0.646 ± 0.064	0.852 ± 0.028
4	0.552 ± 0.047	0.644 ± 0.058	**0.861 ± 0.018**
5	**0.571 ± 0.032**	**0.668 ± 0.063**	0.855 ± 0.020
6	0.566 ± 0.046	0.640 ± 0.073	0.853 ± 0.021
Number	LUAD	UCEC	Overall
2	0.635 ± 0.067	0.620 ± 0.121	0.652
3	0.637 ± 0.059	0.637 ± 0.025	0.665
4	**0.651 ± 0.042**	**0.686 ± 0.014**	**0.679**
5	0.615 ± 0.076	0.645 ± 0.027	0.671
6	0.626 ± 0.067	0.610 ± 0.092	0.659

Table 3
Ablation study on the TCGA-LUAD and TCGA-BRCA datasets. Specifically, we evaluate the significance of the proposed global prototype p_{global}, local prototypes p_{locals}, ratio factors *rfs* and $\mathcal{L}_{dis}$ in terms of c-index.

		BRCA	LUAD
0	p_{local}	0.619 ± 0.093	0.605 ± 0.076
1	$p_{local} + rfs$	0.632 ± 0.074	0.615 ± 0.065
2	$p_{global} + p_{local}$	0.636 ± 0.083	0.629 ± 0.050
3	$p_{global} + p_{local} + rfs$	0.645 ± 0.071	0.647 ± 0.064
4	$p_{global} + p_{local} + \mathcal{L}_{dis}$	0.641 ± 0.063	0.632 ± 0.059
5	Surformer	0.644 ± 0.058	0.651 ± 0.042

Ablation study of each proposed module. In this subsection, we verify the effectiveness of the RRCA and $\mathcal{L}_{dis}$ through experiments on the BRCA and LUAD datasets. Specifically, RRCA is analyzed in the form of global prototype p_{global}, local prototypes p_{locals}, and ratio factors rfs. All the experimental results are shown in Table 3. Here, we set the number of p_{locals} as four for a fair comparison. $Model_0$ only utilizes the p_{locals} to detect and aggregate features all over the instances. Compared with $Model_0$, $Model_1$ additionally embeds the rfs into the feature space. $Model_2$ and $Model_3$ add p_{global} on top of $Model_0$ and $Model_1$. The outstanding performance of $Model_0$ indicates that our cross-attention operation using p_{locals} for feature detection is effective in aggregating critical features related to overall survival. By comparing with $Model_0$, $Model_1$ achieves 1.3% and 1.0% improvements on BRCA and LUAD, respectively, which can conclude that embedding rfs in the feature space is helpful in improving the feature representation of WSIs. By comparing $Model_0$ and $Model_2$, $Model_2$ achieves 1.7% and 2.4% improvements on BRCA and LUAD, respectively, which proves that p_{global} can be complementary to p_{locals} for the feature engineering. Meanwhile, it also verifies that combining global and local features together can significantly improve network performance. In $Model_4$, we add $\mathcal{L}_{dis}$ on top of the $Model_2$. Although the improvements are limited compared with other components, it helps to regulate the feature distribution and indirectly optimize p_{locals}, thus greatly improving the algorithm's interpretability. $Model_5$, which integrates all the components, achieves the best performance. In summary, we can conclude that each component can work well independently and cooperatively.

3.3. Comparison with state-of-the-art methods

In Table 4, we compare our approach with other weakly-supervised learning methods for WSI-based overall survival prediction. As we can see, Surformer significantly outperforms all previous techniques on the other four tumor types, with the exception of the BLCA dataset. GBMLGG is known for its intertumoral and intratumoral heterogeneity,

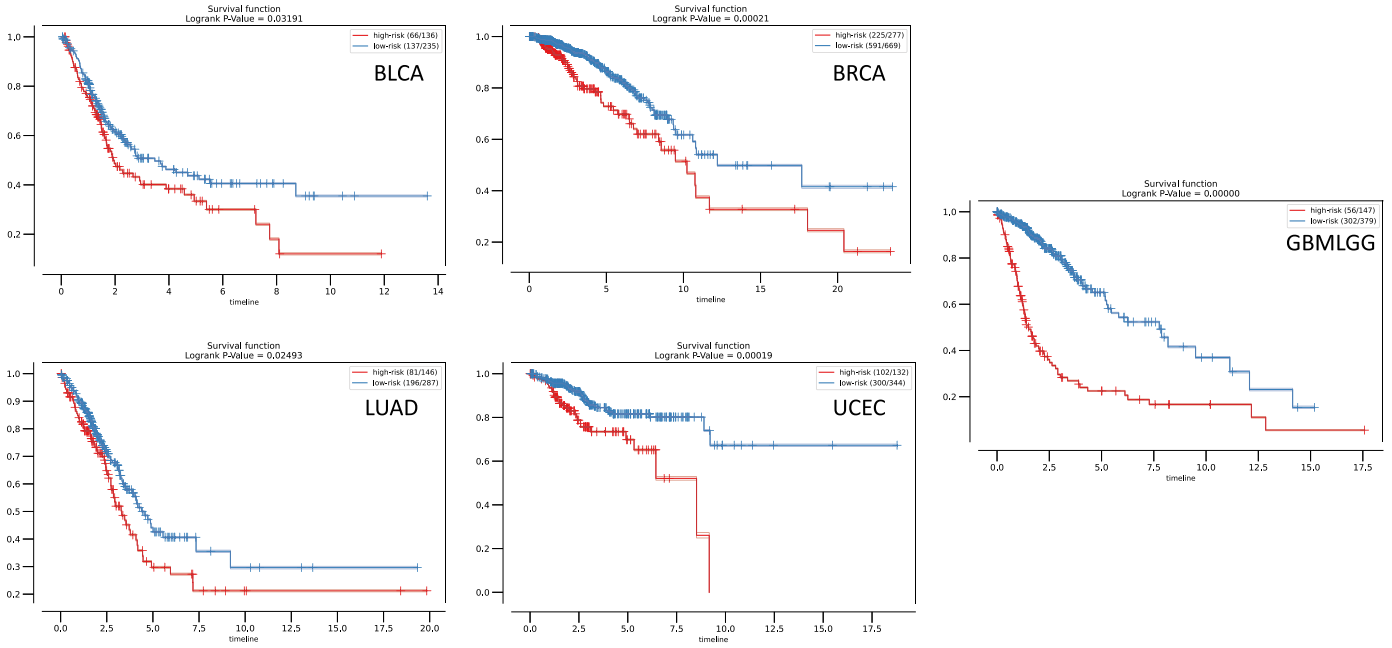

Fig. 3. Kaplan-Meier survival curves of our proposed Surformer across five cancer types. High-risk and low-risk patients are represented by red and blue lines, respectively. The x-axis shows the time in months and the y-axis presents the probability of survival. Log-rank test is used to test for statistical significance in survival distributions between low-risk and high-risk patients (*P-Value* < 0.05).

Table 4
Performance comparison with state-of-the-art methods on TCGA datasets in terms of c-index.

Methods	BLCA	BRCA	GBMLGG
MIL (Deep Set) ([11])	0.500 ± 0.000	0.500 ± 0.000	0.498 ± 0.014
Attention MIL ([12])	0.536 ± 0.038	0.564 ± 0.050	0.787 ± 0.028
DeepAttnMISL ([13])	0.504 ± 0.042	0.524 ± 0.043	0.734 ± 0.029
DeepGraphConv ([29])	0.499 ± 0.057	0.574 ± 0.044	0.816 ± 0.025
Patch-GCN ([21])	0.560 ± 0.034	0.580 ± 0.025	0.824 ± 0.024
Patch-GCN+VarPool ([23])	**0.573 ± 0.027**	0.587 ± 0.021	0.832 ± 0.016
Ours	0.571 ± 0.032	**0.668 ± 0.063**	**0.861 ± 0.018**
Methods	LUAD	UCEC	Overall
MIL (Deep Set) ([11])	0.496 ± 0.008	0.500 ± 0.000	0.499
Attention MIL ([12])	0.559 ± 0.060	0.625 ± 0.057	0.614
DeepAttnMISL ([13])	0.548 ± 0.050	0.597 ± 0.059	0.581
DeepGraphConv ([29])	0.552 ± 0.058	0.659 ± 0.056	0.620
Patch-GCN ([21])	0.585 ± 0.012	0.629 ± 0.052	0.636
Patch-GCN+VarPool ([23])	0.577 ± 0.021	0.641 ± 0.043	0.642
Ours	**0.651 ± 0.042**	**0.686 ± 0.014**	**0.687**

on which we achieve a c-index of 83.2%, surpassing other methods by at least 3.9%. The outstanding improvements demonstrate Surformer's superiorities in feature aggregation and generalization. VarPool [23] is known for incorporating intratumoral heterogeneity into its prediction. They achieve this by computing the variance of learned projections of instances. According to the experimental results, VarPool only slightly surpasses us on the BLCA dataset by 0.2%, and Surformer outperforms it on the other four datasets by a large margin. Therefore, we can conclude that optimizing detection prototypes (*e.g.,* p_{global} and p_{locals}) with the whole training data is a more advanced strategy for tackling the heterogeneity challenge. Patch-GCN ([21]) is a novel context-aware method using the graph neural network. Edges in the graph help to build instance connections; however, many connections may not contribute to the final survival prediction (*e.g.,* instance communications between benign cells). In Surformer, we adopt the learnable prototypes to collect critical features and generate their ratio information. Critical features and ratio information cooperate to make the final prediction in the following computation. In conclusion, our proposed Surformer is a more sophisticated method for WSI-based survival prediction.

4. Discussion

4.1. Kaplan–Meier curve analysis

The Kaplan–Meier estimator, also known as the product limit estimator, is a non-parametric statistic used to estimate the survival function from lifetime data. In medical research, it is widely used to measure the fraction of patients living for a certain amount of time after treatment. In Fig. 3, we utilize Kaplan-Meier curves to visualize the quality of patient stratification between predicted low-risk and high-risk patient populations on BLCA, BRCA, LUAD, UCEC, and GBMLGG datasets. Specifically, we classify patients into high-risk and low-risk cohorts according to the survival probability in the middle of the uncensored patients' timeline. To increase the reliability of the curves, we collect and present all the testing data across the five-fold cross-validation. It is obvious that Surformer can distinguish high-risk (red) and low-risk (blue) patients with only the paired WSI. At the same time, we use the log-rank test to measure the statistical difference between the two cohorts. The *P-Values* on BLCA, BRCA, LUAD, UCEC, and GBMLGG datasets are 0.03191, 0.00021, 0.02493, 0.00019, and 0.00000, respectively. All of the *P-Values* are below 0.05, demonstrating the tremendous statistical significance of the observed difference derived by our Surformer.

4.2. Differential analysis of rfs

This section conducts the differential analysis of *rfs* across high-risk and low-risk patients in five cancer types. As introduced in Section 2.2, *rfs* are the quantification of the patches detected by each p_{local}. Therefore, for easy understanding, we simply replace the index of *rfs* with corresponding p_{local} in Fig. 4. Meanwhile, deep learning models are typically denounced by their generalization ability and robustness. We specifically visualize the ratio distribution on both training and testing data. As we see, Surformer achieves a high-level distribution consistency on the two sets, thereby validating its outstanding capacity for generalization and robustness.

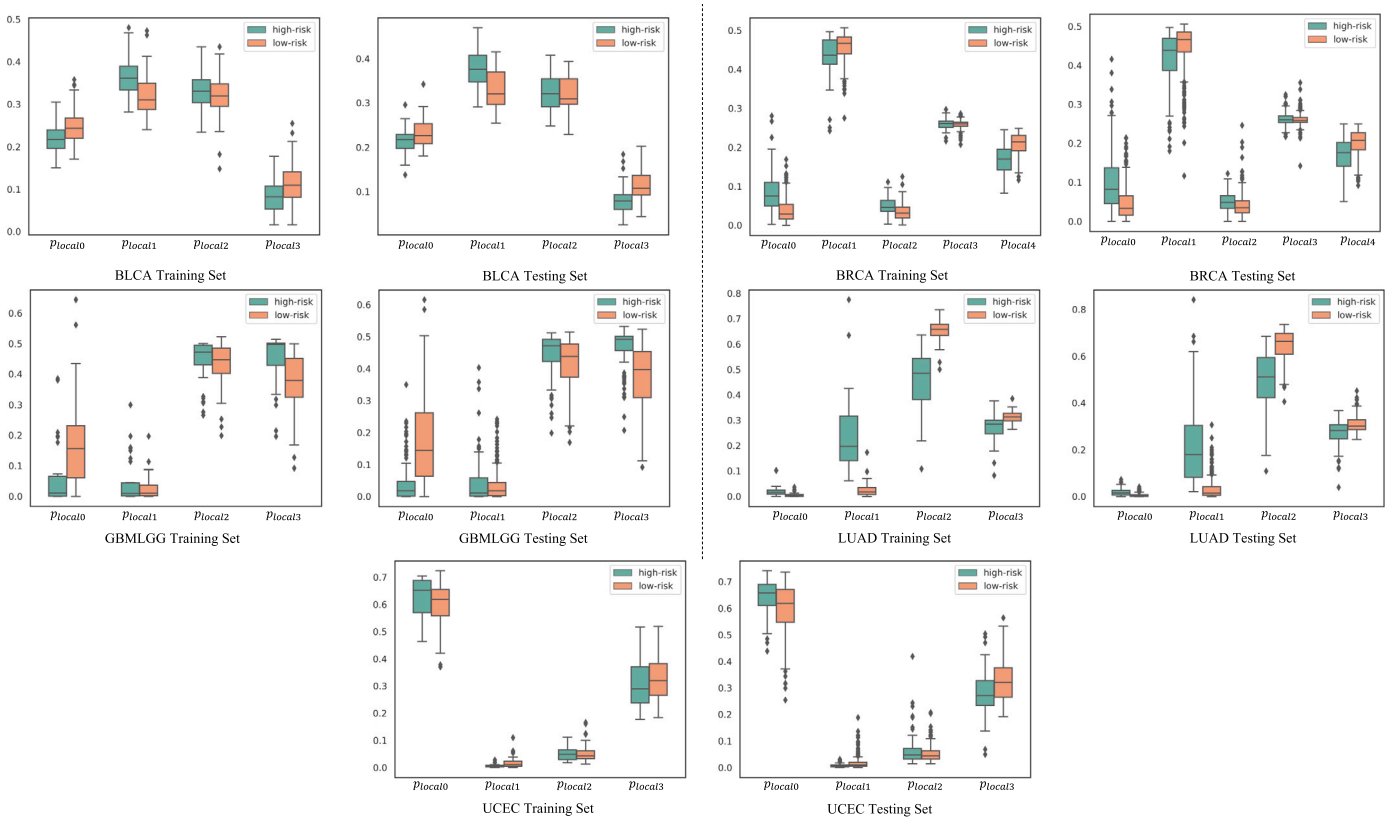

Fig. 4. Illustration of the RRCA generated rfs on five different tumor types. Box plots are utilized to present the distributions on both training and testing sets. High-risk and low-risk patients are colored green and orange, respectively.

For the BLCA dataset, high-risk patients have more patterns related to p_{local1} and fewer patterns related to p_{local0} and p_{local3}. In terms of p_{local2}, although correlated patches are also taken for final survival prediction, there are no apparent differences between high-risk and low-risk cohorts. For the BRCA dataset, high-risk patients have more patterns related to p_{local0} and p_{local2} and fewer patterns related to p_{local1} and p_{local4}. In terms of p_{local3}, there is no significant difference between high-risk and low-risk patients. For the GBMLGG dataset, a big difference only occurs between p_{local0} and p_{local3}. High-risk patients have more patterns related to p_{local3} and fewer patterns related to p_{local0}. For the LUAD dataset, noticeable differences exist on p_{local1}, p_{local2} and p_{local3}. High-risk patients have more patterns related to p_{local0}. For the UCEC dataset, there is no significant difference between high-risk and low-risk patients. High-risk patients have a little bit more patterns related to p_{local0} and a little bit fewer patterns related to p_{local3}.

Through the visualization and analysis of rfs in Fig. 4, we can conclude that p_{locals} plays a critical role in model interpretability in the form of rfs. The apparent distribution results are more convincing than the direct prediction from the end-to-end models. Furthermore, in subsequent Subsection 4.3, we will continue to analyze the p_{locals} by visualizing the attentive patches and deciphering their cell types.

4.3. Patch retrieval and analysis of p_{global} and p_{locals}

For sets of WSIs belonging to different patient cohorts, we perform p_{global} interpretation by visualizing attention regions on the original WSIs and p_{locals} interpretation by conducting histological analysis of the patches detected by each p_{local}. Here, we take two WSIs from BRCA and LUAD testing datasets as examples.

p_{global} interpretation: In Fig. 5 (a,d), we present the original WSIs after the automated segmentation operation. The tissue regions and inside blanks are circled with green and blue lines, respectively. In Fig. 5 (b,e), we visualize the attention heatmaps of Surformer. Typically, tumor cells appear darker in the H&E-stained WSIs. Apparently, the final predictions are not merely based on the tumor cells, but also the interactions with surrounding benign cells. These intuitionistic visualizations can help to improve the pathologists' efficiency in recognizing critical regions in survival analysis.

p_{locals} interpretation of BRCA: In accordance with Table 4 and Fig. 4, Surformer functions best with five p_{locals}, and patches correlated to p_{local0}, p_{local1} and p_{local4} are critical for stratifying high-risk and low-risk patients. In Fig. 5 (c), we retrieve the top five patches corresponding to each p_{local} and segment and categorize the cells within the patches through trained HoverNet [30]. It is easy to find that p_{local0} and p_{local4} are related to vast, and few immune cell infiltrates in tumor cells, respectively. p_{local1} correlates to patches with a small number of immune cells. By correlating with the rfs distribution in Fig. 4, Surformer figures out that vast immune cell infiltrates are one of the most important signatures for breast cancer patients; high-risk patients tend to have more vast immune cell infiltrates. Besides, a modest number of immune cell infiltrates and immune cell counts have a negligible impact on the ultimate survival prognosis. Our findings about the immune cell infiltrates can also be supported by some clinical research [6,31], demonstrating our potency in BRCA survival analysis.

p_{locals} interpretation of LUAD: In terms of the LUAD dataset, according to Table 4 and Fig. 4, Surformer performs best with four p_{locals} and patches correlated to p_{local1} and p_{local2} are critical for stratifying high-risk and low-risk patients. Here, we carry out cell segmentation and classification on the obtained top patches, which are identical to the previous process. Particularly, p_{local1} and p_{local2} correspond to tumor cells and immune cells, respectively. Patients with a high portion of immune cells and a low portion of tumor cells tend to have better survival conditions. In terms of immune cell infiltrations, Surformer does not detect big differences between high-risk and low-risk cohorts.

In conclusion, the model prediction is relying on the comprehensive analysis of the key histological features. Meanwhile, Surformer can

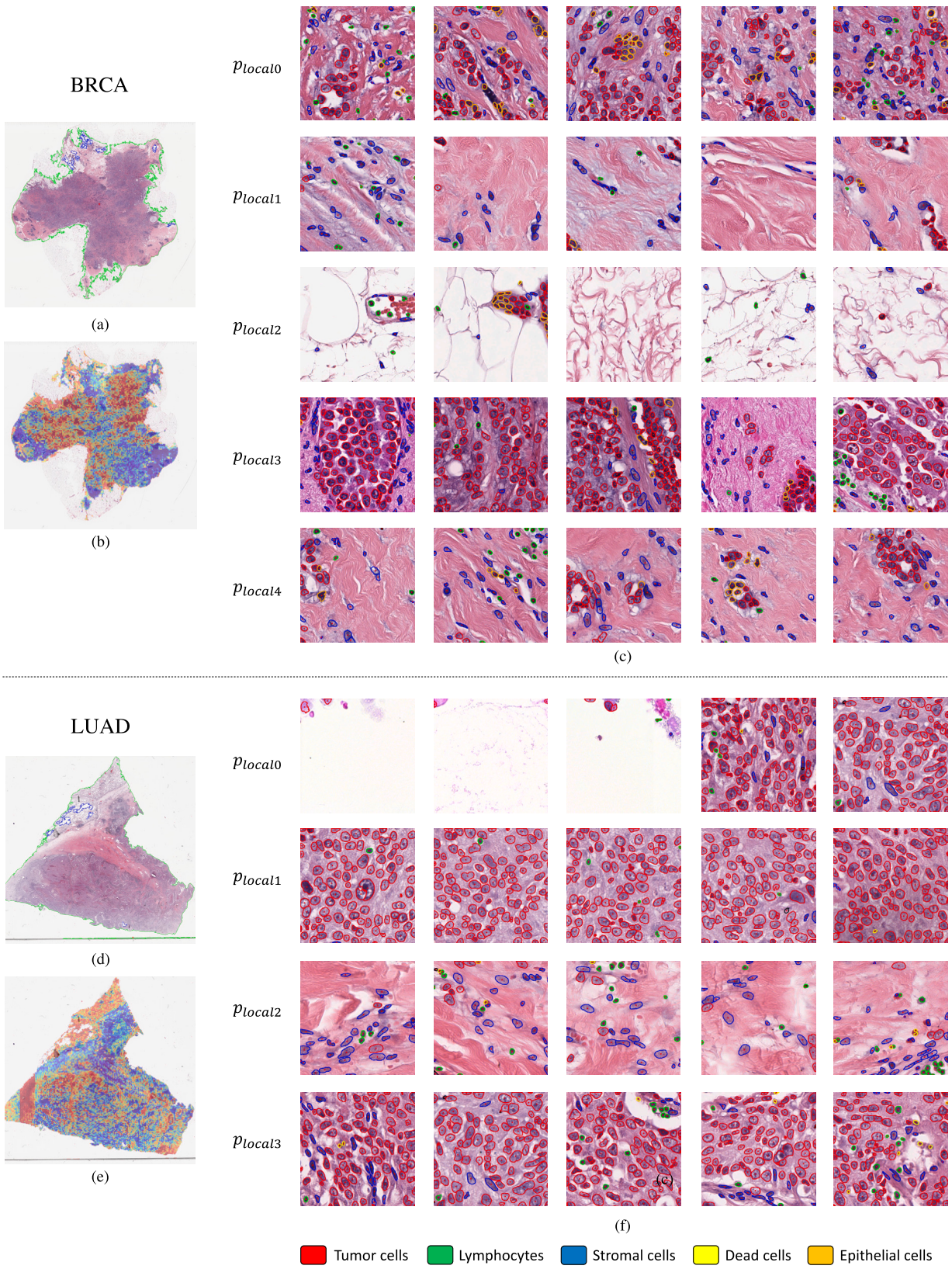

Fig. 5. Global and local interpretation of WSIs. Two WSIs from BRCA and LUAD datasets are presented: (a,d) are the original segmented images and (b,e) present the attentive patches onto the images. (c,f) present the patches detected by each p_{local}. Cells in patches are segmented and classified by the HoverNet [30].

provide visualization and statistical enrichment analysis of informative histological patterns derived from whole slide images, which can be of particular interest and assistance to pathologists with regard to survival analysis and associated variables.

4.4. Conclusions

Survival analysis can provide constructive guidance to pathologists and physicians for proposing precision therapy, which has significant commercial and clinical value. In this study, we have proposed a novel neural network with high interpretability, termed pattern-perceptive transformer (Surformer), for WSI-based survival prediction. Specifically, Surformer can quantify specific histological patterns and explicitly interpret the final prediction with statistical analysis through bag-level labels. Extensive benchmarking experiments on five TCGA benchmark datasets illustrated that Surformer outperformed other existing state-of-the-art methods and highlighted a superiority in terms of both predictive performance and interpretability of deep learning models for cancer survival prediction. While Surformer exhibits impressive performance, it fails to model the spatial contextual features of WSIs, thereby ignoring the significance of the tumor microenvironment in survival analysis. Consequently, our future work will particularly improve the algorithm by jointly encoding contextual and long-range global features for better survival analysis. In conclusion, the development and availability of data-driven deep learning-based tools such as Surformer proposed in this study represent a useful step forward towards the implementation of digital pathology tools for informed clinical decision-making underpinning precision oncology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was based on the open-source public dataset The Cancer Genome Atlas (TCGA, https://www.cancer.gov/tcga). The patients involved in the database have given ethical approval.

This work was supported by Major and Seed Inter-Disciplinary Research (IDR) projects awarded by Monash University and a Grant from the International Joint Usage/Research Center, Institute of Medical Science, The University of Tokyo (K23-2074).

References

[1] J.P. Klein, M.L. Moeschberger, Survival Analysis: Techniques for Censored and Truncated Data, vol. 1230, Springer, 2003.
[2] J.G. Ibrahim, M.-H. Chen, D. Sinha, J. Ibrahim, M. Chen, Bayesian Survival Analysis, vol. 2, Springer, 2001.
[3] D.G. Kleinbaum, M. Klein, et al., Survival Analysis: A Self-Learning Text, vol. 3, Springer, 2012.
[4] N.R. Latimer, Survival analysis for economic evaluations alongside clinical trials—extrapolation with patient-level data: inconsistencies, limitations, and a practical guide, Med. Decis. Mak. 33 (6) (2013) 743–754.
[5] N. West, M. Dattani, P. McShane, G. Hutchins, J. Grabsch, W. Mueller, D. Treanor, P. Quirke, H. Grabsch, The proportion of tumour cells is an independent predictor for survival in colorectal cancer patients, Br. J. Cancer 102 (10) (2010) 1519–1523.
[6] S.L. Goff, D.N. Danforth, The role of immune cells in breast tissue and immunotherapy for the treatment of breast cancer, Clin. Breast Cancer 21 (1) (2021) e63–e73.
[7] M.J. Campbell, F. Baehner, T. O'Meara, E. Ojukwu, B. Han, R. Mukhtar, V. Tandon, M. Endicott, Z. Zhu, J. Wong, et al., Characterizing the immune microenvironment in high-risk ductal carcinoma in situ of the breast, Breast Cancer Res. Treat. 161 (1) (2017) 17–28.
[8] Y. Wang, N. Coudray, Y. Zhao, F. Li, C. Hu, Y.-Z. Zhang, S. Imoto, A. Tsirigos, G.I. Webb, R.J. Daly, et al., Heal: an automated deep learning framework for cancer histopathology image analysis, Bioinformatics 37 (22) (2021) 4291–4295.
[9] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji, et al., Transmil: transformer based correlated multiple instance learning for whole slide image classification, Adv. Neural Inf. Process. Syst. 34 (2021) 2136–2147.
[10] Z. Wang, Y. Bi, T. Pan, X. Wang, C. Bain, R. Bassed, S. Imoto, J. Yao, R.J. Daly, J. Song, Targeting tumor heterogeneity: multiplex-detection-based multiple instance learning for whole slide image classification, Bioinformatics 39 (3) (2023) btad114.
[11] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R.R. Salakhutdinov, A.J. Smola, Deep sets, Adv. Neural Inf. Process. Syst. 30 (2017).
[12] M. Ilse, J. Tomczak, M. Welling, Attention-based deep multiple instance learning, in: International Conference on Machine Learning, PMLR, 2018, pp. 2127–2136.
[13] J. Yao, X. Zhu, J. Jonnagaddala, N. Hawkins, J. Huang, Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks, Med. Image Anal. 65 (2020) 101789.
[14] J. Barker, A. Hoogi, A. Depeursinge, D.L. Rubin, Automated classification of brain tumor type in whole-slide digital pathology images using local representative tiles, Med. Image Anal. 30 (2016) 60–71.
[15] J. Cheng, X. Mo, X. Wang, A. Parwani, Q. Feng, K. Huang, Identification of topological features in renal tumor microenvironment associated with patient survival, Bioinformatics 34 (6) (2018) 1024–1030.
[16] J. Yao, S. Wang, X. Zhu, J. Huang, Imaging biomarker discovery for lung cancer survival prediction, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2016, pp. 649–657.
[17] K.-H. Yu, C. Zhang, G.J. Berry, R.B. Altman, C. Ré, D.L. Rubin, M. Snyder, Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features, Nat. Commun. 7 (1) (2016) 1–10.
[18] X. Zhu, J. Yao, X. Luo, G. Xiao, Y. Xie, A. Gazdar, J. Huang, Lung cancer survival prediction from pathological images and genetic data—an integration study, in: International Symposium on Biomedical Imaging (ISBI), IEEE, 2016, pp. 1173–1176.
[19] P. Mobadersany, S. Yousefi, M. Amgad, D.A. Gutman, J.S. Barnholtz-Sloan, J.E. Velázquez Vega, D.J. Brat, L.A. Cooper, Predicting cancer outcomes from histology and genomics using convolutional networks, Proc. Natl. Acad. Sci. 115 (13) (2018) E2970–E2979.
[20] Z. Wang, J. Li, Z. Pan, W. Li, A. Sisk, H. Ye, W. Speier, C.W. Arnold, Hierarchical graph pathomic network for progression free survival prediction, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 227–237.
[21] R.J. Chen, M.Y. Lu, M. Shaban, C. Chen, T.Y. Chen, D.F. Williamson, F. Mahmood, Whole slide images are 2d point clouds: context-aware survival prediction using patch-based graph convolutional networks, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 339–349.
[22] Z. Huang, H. Chai, R. Wang, H. Wang, Y. Yang, H. Wu, Integration of patch features through self-supervised learning and transformer for survival analysis on whole slide images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 561–570.
[23] I. Carmichael, A.H. Song, R.J. Chen, D.F. Williamson, T.Y. Chen, F. Mahmood, Incorporating intratumoral heterogeneity into weakly-supervised deep learning models via variance pooling, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2022, pp. 387–397.
[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[25] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[26] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: pre-training of deep bidirectional transformers for language understanding, arXiv preprint, arXiv:1810.04805, 2018.
[27] R. Xiong, Y. Yang, D. He, K. Zheng, S. Zheng, C. Xing, H. Zhang, Y. Lan, L. Wang, T. Liu, On layer normalization in the transformer architecture, in: International Conference on Machine Learning, PMLR, 2020, pp. 10524–10533.
[28] S.G. Zadeh, M. Schmid, Bias in cross-entropy-based training of deep survival networks, IEEE Trans. Pattern Anal. Mach. Intell. 43 (9) (2020) 3126–3137.
[29] R. Li, J. Yao, X. Zhu, Y. Li, J. Huang, Graph cnn for survival analysis on whole slide pathological images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2018, pp. 174–182.
[30] S. Graham, Q.D. Vu, S.E.A. Raza, A. Azam, Y.W. Tsang, J.T. Kwak, N. Rajpoot, Hover-net: simultaneous segmentation and classification of nuclei in multi-tissue histology images, Med. Image Anal. 58 (2019) 101563.
[31] S. Zuo, M. Wei, S. Wang, J. Dong, J. Wei, Pan-cancer analysis of immune cell infiltration identifies a prognostic immune-cell characteristic score (iccs) in lung adenocarcinoma, Front. Immunol. 11 (2020) 1218.