Survey

Abstract— With the increasing application of gigapixel Whole-Slide Images (WSIs) in digital pathology, there has been growing interest in utilizing WSI for survival prediction. This review summarizes and analyzes the recent research on survival prediction using WSI. Firstly, the development and applications of WSI technology, as well as the significance of survival prediction in cancer management, are introduced. Then, different methods for survival prediction are reviewed, including traditional tissue feature analysis, cell-level feature analysis, and deep learning approaches. In tissue feature analysis, researchers predict survival by extracting morphological, textural, and regional features from WSIs. Cell-level feature analysis focuses on the morphological and distributional characteristics of individual cells and computes relationships between cells for more accurate predictions. Deep learning methods, utilizing techniques such as convolutional neural networks and recurrent neural networks, are able to learn complex feature representations from WSIs and achieve high-precision survival prediction. Additionally, the challenges and future directions of WSI-based survival prediction methods are discussed. Overall, utilizing WSI for survival prediction is a promising field that can provide important complementary information for clinical decision-making and personalized treatment.

Index Terms— Cancer survival analysis,Cancer survival prediction,deep learning,whole slide image

I. Introduction

Survival prediction is a direction of statistics for analyzing the duration time that is expected until the events of interest occur,such as the death of the life form of biology. With the rising importance of precision medicine, the ability to predict patient survival based on individualized characteristics becomes paramount. WSI, as a rich source of information, offers an unprecedented level of granularity by capturing the heterogeneity within tissue samples. By enabling the high-resolution digitization of entire tissue slides, WSI has not only enhanced the efficiency and accuracy of pathological assessments but has also opened new avenues for predictive modeling and survival analysis.

This paragraph of the first footnote will contain the date on which you submitted your paper for review. It will also contain support information, including sponsor and financial support acknowledgment. For example, “This work was supported in part by the U.S. Department of Commerce under Grant BS123456.”

The next few paragraphs should contain the authors’ current affiliations, including current address and e-mail. For example, F. A. Author is with the National Institute of Standards and Technology, Boulder, CO 80305 USA (e-mail:author@boulder.nist.gov).

S. B. Author, Jr., was with Rice University, Houston, TX 77005 USA. He is now with the Department of Physics, Colorado State University, Fort Collins, CO 80523 USA (e-mail: author@lamar.colostate.edu).

T. C. Author is with the Electrical Engineering Department, University of Colorado, Boulder, CO 80309 USA, on leave from the National Research Institute for Metals, Tsukuba, Japan (e-mail: author@nrim.go.jp).

This review aims to provide a comprehensive overview of the application of WSI in the context of survival prediction. We delve into the intersection of digital pathology and computational methods, exploring how WSI data can be harnessed to predict patient outcomes, especially in the field of oncology.

The introduction of WSI has given researchers and clinicians access to vast repositories of image data, offering new opportunities for feature extraction, machine learning, and deep learning techniques to model and predict patient survival. The ability to analyze tissue morphology, cellular structures, and patterns within whole slide images has the potential to uncover valuable insights into disease progression and prognosis.

In this review, we will examine the key methodologies, challenges, and recent advancements in survival prediction using WSI data. We will also discuss the implications of such predictions on personalized medicine, treatment strategies, and patient care.Additionally, we will explore various aspects of survival prediction using WSI, including the development of novel image features, the utilization of deep learning architectures for image analysis, and the challenges associated with large-scale data management. We will also consider the ethical and regulatory considerations in implementing such predictive models in clinical practice.

As the fields of digital pathology and computational biology continue to evolve, the integration of WSI into survival analysis represents a promising frontier for improving patient outcomes and advancing our understanding of disease dynamics. This review will serve as a guide for researchers and practitioners interested in leveraging WSI for survival prediction in diverse clinical and research settings.

The complexity of certain data modalities with high dimensionality poses a challenge for physicians in manually interpreting multi-modal biomedical data to decide on treatment and assess prognosis [1]. The capability to extract ”sub-visual” image features from digital pathology slide images—features provides an opportunity for more robust quantitative modeling of disease appearance and potentially improved predictions of disease aggressiveness and patient outcomes [2].

In recent years, several computer algorithms for hematoxylin and eosin (H&E) stained pathology image analysis have been developed to aid pathologists in objective clinical diagnosis and prognosis. The Cancer Genome Atlas (TCGA) are the most common datasets when discussing the topics of survival prediction.

the observation of one patient is either a survival time

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

In recent years, the integration of Whole Slide Imaging (WSI) into the realm of survival prediction has emerged as a promising frontier in the fields of digital pathology, oncology, and computational biology. WSI, a revolutionary technology, has transformed the way we analyze tissue samples, providing a comprehensive digital representation of entire pathology slides.

Traditionally, pathological assessments have been reliant on manual inspection and subjective interpretation of glass slides under a microscope. While this approach has served as the gold standard for disease diagnosis and prognosis, it is inherently limited by issues of inter-observer variability, labor intensiveness, and the inability to harness the full potential of the vast data embedded in tissue structures.

The advent of WSI has paved the way for a paradigm shift in tissue analysis. Through high-resolution digitization, WSI generates massive datasets of tissue samples, allowing for a more detailed examination of cellular structures, tissue morphology, and spatial relationships. These digital representations of tissue hold a wealth of information that extends far beyond what can be appreciated by the human eye.

One of the most compelling applications of WSI lies in the realm of survival prediction, particularly in the context of cancer. Cancer remains a leading cause of mortality worldwide, and the ability to predict patient outcomes accurately is paramount for optimizing treatment strategies and improving patient care. With the integration of WSI, researchers and clinicians can extract rich image features and employ advanced machine learning and deep learning techniques to model and predict patient survival.

Survival prediction using WSI offers the potential to uncover subtle patterns, biomarkers, and prognostic factors that may have gone unnoticed in traditional pathology assessments. The granularity and comprehensiveness of WSI data open doors to new avenues of research and personalized medicine, enabling clinicians to tailor treatments to individual patient profiles.

In this review, we delve into the intricate interplay between digital pathology and computational methods, exploring how WSI is being harnessed to predict patient survival. We will examine the methodologies, challenges, ethical considerations, and the implications of these predictions on clinical practice.

As the field of WSI continues to evolve, the fusion of digital pathology and survival prediction holds great promise for advancing our understanding of disease dynamics and ultimately enhancing patient outcomes.

As we delve deeper into this evolving landscape, it is crucial to appreciate the monumental shift brought about by WSI. Its capacity to create digital archives of pathology slides has not only expedited the diagnostic process but has also catapulted computational pathology to the forefront. This dynamic shift, driven by advancements in digital imaging and artificial intelligence, has redefined the scope of pathology by offering a more profound understanding of disease, bolstering diagnostic accuracy, and heralding a new era of prognostic modeling.

WSI's role in survival prediction is particularly significant within the domain of cancer research. It has given rise to the emergence of predictive models that harness the extensive image data contained within tissue samples. By analyzing and extracting pertinent information from these digitized slides, researchers can identify morphological and structural biomarkers, which, when combined with clinical data, offer valuable insights into patient outcomes.

Moreover, the integration of WSI data provides the foundation for the development of prognostic models capable of guiding treatment decisions and improving the overall quality of patient care. The implications extend beyond individualized medicine to population-level studies, ultimately influencing healthcare policy, resource allocation, and public health strategies.

While the potential is immense, it is not without challenges. The management and analysis of vast WSI datasets demand advanced computational infrastructure, robust machine learning algorithms, and a harmonious partnership between pathologists and data scientists. Ethical considerations surrounding patient data privacy, model interpretability, and regulatory compliance also merit attention.

This review embarks on a journey to explore the multifaceted landscape of survival prediction using WSI. It endeavors to survey the methodologies, showcase the milestones achieved, and address the formidable challenges that lie ahead. In doing so, it aspires to serve as a guiding compass for researchers, pathologists, clinicians, and healthcare stakeholders invested in leveraging the symbiotic relationship between digital pathology and survival prediction.

As we advance in this ever-evolving field, the fusion of digital pathology with survival prediction promises to unravel the intricacies of disease dynamics, enhance patient care, and pave the way for a new era of data-driven medicine.

II. FORMULATION OF SURVIVAL PREDICTION

A. Survival Analysis

In traditional survival modeling, it is assumed that the time durations follow an unknown distribution. Among the popular modeling methods, the Cox proportional hazard model stands out, as it focuses on modeling hazards rather than the survival function:

$$h(t|x) = h_0(t)exp(\beta^T x) \tag{1}$$

The expression involves the time variable t, covariates x of dimension p where $x = (x_1, ..., x_p)^T$, a vector of

regression parameters $\beta = (\beta_1, ..., \beta_p)^T$, and the baseline hazard $h_0(t)$. The risk function, denoted as $f(x) = \beta^T x$, is also referred to as the regression function. The estimation of regression parameters involves minimizing the negative log partial likelihood:

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

n represents the patient count, and t_i corresponds to the survival time (censored or observed) for patient i, and δ_i is an indicator variable indicating whether the survival time is censored ($\delta_i = 0$) or observed ($\delta_i = 1$). The term $R(t_i)$ refers to the risk set at time t_i, defined as the group of individuals still under study before time t_i.

Compared with traditional survival analysis, Katzman et al. [3] introduced DeepSurv, a deep fully connected network designed to capture nonlinear relationships between covariates and the risk function. This model replaces the exponential component $\beta^T x$ in the traditional Cox model's risk function $f(x)$ with a nonlinear deep fully connected network.

B. Measurement Metrics

To assess the predictive performances in survival analysis, the ordinary model utilizes the concordance index (C-Index) as the standard evaluation metric, a widely accepted measure for model assessment in survival prediction [4]. For a set of Whole Slide Images (WSIs) from $\mathcal{M}$ patients along with their respective labels, the input data is represented as:

$$\mathbb{D}_{in} = \{C_1, C_2, ..., C_M\}, \quad M \geq 2,$$

The C-Index [5] is expressed as:

$$\mathcal{C}_{index}(\mathbb{D}_{in}) = \frac{1}{\mathcal{M}} \sum_{i:\delta_i=1} \sum_{j:T_i<T_j} \mathcal{I}[(T_i, X_i) < (T_j, X_j)] \quad (3)$$

Here, $\mathcal{M}$ is the number of comparable pairs, $T[.]$ denotes the indicator function, and T is the actual observation. The C-index ranges from 0 to 1, where a larger C-index indicates better prediction performance, and vice versa. Specifically, 0 represents the worst condition, 1 is the best, and 0.5 is the value for a random guess. It is worth noting that the indicator matrix of the C-index has a unique characteristic, as illustrated in the following formulation:

$$F(X_i, X_j) = 1 - F(X_j, X_i) \quad (4)$$

The equation represents the ratio of correctly ordered pairs of subjects' survival times to all possible ranking pairs. The Cox proportional hazards model provides a relative risk assessment for a patient compared with others, making the CI a suitable evaluation index. The Kaplan-Meier curve, illustrating the trend of the survival function $S(t) = \Pr(T > t)$, is employed for assessing the model's effectiveness. The survival function adheres to the subsequent relationship at the k-th time point (for uncensored samples):

$$S(t_k) = S(t_{k-1})(1 - \frac{e_k}{r_k}) \quad (5)$$

Here, $S(t_{k-1})$ represents survival probability at the time point t_{k-1}, e_k is the count of events occurring between t_{k-1} and t_k, and r_k is the number of patients with observed time exceeding t_k.

To evaluate the survival model intuitively, the entire dataset is stratified into high-risk and low-risk groups using the some risk score quantiles of the predicted relative risk as a threshold. Model performance is assessed by comparing the Kaplan-Meier curves of these groups, with greater separation indicating better performance. The significance of the difference is determined by the P value obtained through the log-rank test, where $P < 0.05$ signifies statistical significance.

Survival prediction, a regression problem, involves distinguishing high-hazard and low-hazard patients by adopting a threshold, transforming it into a binary classification task. Positive samples include uncensored patients with survival time under the threshold, while negative samples comprise patients with survival/censored time exceeding the threshold. Additionally, the model's performance is evaluated by plotting the Receiver Operating Characteristic (ROC) curve and calculating the Area Under the ROC Curve (AUC).

C. Graph Construction

Furthermore, it has been widely recognized that the topological properties of instances on pathological images are crucial in medical tasks.Graph is widely employed to represent topological structures.

Generally ,we let $G = (V, E)$ denote a graph with nodes V and edges E. We define $X \in \mathbb{R}^{N \times F}$ as a feature matrix of N nodes.A in V with F-dimensional features, and $A \in \mathbb{R}^{N \times N}$ as the djacency matrix that holds the graph topology. Given the graph adjacency matrix A, the graph node embedding X is updated by the message passing function defined in **GCN**. We consider a multi-layer Graph Convolutional Network (GCN) with the following layer-wise propagation rule [6]:

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^{(l)} W^{(l)}\right). \quad (6)$$

Here, $\tilde{A} = A + I_N$ is the adjacency matrix of the undirected graph g with added self-connections. I_N is the identity matrix, $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$ and $W^{(l)}$ is a layer-specific trainable weight matrix. $\sigma(\cdot)$ denotes an activation function, such as the $\mathrm{ReLU}(\cdot) = \max(0, \cdot)$.$H^{(l)} \in \mathbb{R}^{N \times D}$ is the matrix of activations in the l^{th} layer; $H^{(0)} = X$. The existing methods for constructing graph on WSI are mainly based on patches and cells.

Graph Construction based on patches : Given a set of sampled patch images $\mathbf{P} = \{\mathbf{P_i}\}$ from WSI, we have to dump those patches from the margin areas which contains few cells, therefore, the cardinality $\|\mathcal{P}\|$ differs by WSI.

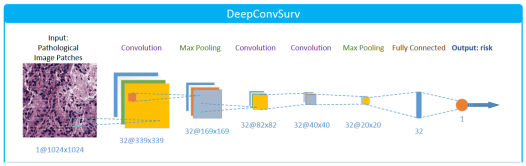

Fig. 1: the framework of DeepConvSurv(image from [9])

Consequently, the graphs we construct for WSIs are of different sizes. Given patches as vertices, vertex features are generated by the different feature extractors

Graph Construction based on cells :For this method, accurate nuclear segmentation is crucial.For already segmented cell nuclei, Pathomic Fusion [7] use the K-Nearest Neighbors (KNN) algorithm and hypothesize that adjacent cells will have the most significant cell-cell interactions and limit the adjacency matrix to K nearest neighbours. Using KNN, adjacency matrix A is defined as:

if $j \in$KNN(i) and $D(i,j) < d$ otherwise 0

III. Methodology

A. Region of Interest

There exist many methods developed for predicting survival through the information provided by the whole slide images(WSIs). Rather than utilizing the overall patches from the gigapixel pathology images, the traditional models usually pre-select a subset of critical patches from the region of interest (ROI) as the input data. Apart from the features extracted from ROIs with the deep neural networks(DNNs), some morphological features of the image patches can be extracted and accessed by the image analysis software named CellProfiler [8], commonly used for cell phenotypes measurements at that time. It's the features including cell shape, size, the distribution of pixel intensity in the cells and nuclei and texture of cells and nuclei that the quantitative analysis tool can extract.

Taking whole slide images with various size as inputs, in order to do end-to-end survival prediction, DeepConvSurv proposed by [9] randomly chooses the patches among the ROIs annotated by the professional pathologists. The deep convolutional neural network of DeepConvSurv is shown in fig.1. And the experiments showed it could extract more abstract information different from the hand-crafted features generated by the state-of-the-art analysis tool CellProfiler mentioned above. The primary distinction between the model and the traditional Cox model lies in the exponential component. The loss function for the model is:

$$l(w,b) = -\sum_{i:R_i=1} h_{w,b}^{last}(x_i) + log \sum_{j:t_j>=t_i} exp(h_{w,b}^{last}(x_j)), \tag{7}$$

In the equation, $h_{w,b}^{last}(x)$ represents the output of DeepConvSurv, specifically the output of the last layer, where w and b denote the parameters of the network, and x_i and x_j refer to the i^{th} and j^{th} inputs. To simplify, we will omit the parameters of each layer and express $h_{w,b}^{last}(x)$ as $h^{last}(x)$. Therefore, the loss function's gradient is changed to:

$$\delta^{last}(x_i) = \frac{\partial l(w,b)}{\partial h^{last}(x_i)} = -1 + \frac{exp(h^{last}(x_i))}{\sum_{j:t_j>=t_i} exp(h^{last}(x_j))}, \tag{8}$$

Except this, [10], [11], [12], [13], [14] have shown that random sampling of patches within the tissues in WSIs still makes sense in stratifying phenotypic information which can be improved. With the help of online tool, Mobadersany et al. [15] manually selected the ROIs without tissue-processing artifacts containing overstaining or understaining areas from alternates, which have viable tumor features. Faced the difficulty of intratumoral heterogeneity and few availability of labeled data, to obtain better and more robust effects, the model uses the data augmentation techniques and sorted median risks to get prediction results. And the model eventually gets more accurate outcomes. Moreover, the paper provides some publicly accessible datasets with ROIs. With the assistance of the assembled diagnostic slides offered by [15] with ROIs, characteristics of tumor microenvironment could be got in the built graphs in [7]. Because semantic segmentation is executed in ROIs to recognize and localize relevant cells acting as set of nodes in the spatial graph for abstract graph representations. Not only dose the model in [16] use the public cancer survival dataset TCGA, but it also adopts a core sample set from UT MD Anderson Cancer Center during the holistic procedure. And it takes advantage of the annotations of ROIs to locate the possible tumor regions in pathological images for subsequent steps. Some methods adopt sampling strategy to generate candidate patches not limited to ROIs. Since revealing ROIs requires specialized prior knowledge and expensive labor costs, Wang et al. proposed an automatic model aimed at finding ROIs. The proposed model in [17] can identify tumor regions as ROIs in hematoxylin and eosin (H&E) stained pathology images using predicted likelihood of image patches, each patch tagged as the highest probability category. In this way, some tumor area-related features can be extracted as the descriptors of above ROIs including area, perimeter, convex area, filled area, major axis length, minor axis length and so on. For the purpose of training the prognostic model indicating that the risk group defined by tumor shape features is an independent prognostic factor, the features are used in the training process. The model in [18] has a automated pipeline excluding the background white space among the identified tissue, then overlooking the sparse cellularity regions and randomly sampling the potential patches from the foreground area.

In a short, as the methods previously implemented, ROIs routinely ask artificial marking and rigorous reviews to produce passable survival prediction and focusing on ROI ignores the impact of other regions. Additionally speaking, the ROI-based methods discussed above require pathologists to hand-annotate ROIs, a tedious task.

B. Feature Extraction

Previous methods often extract image features from patches of whole slide images(WSIs) using the pre-trained model based on the numerous natural images datasets ImageNet theoretically being able to withdraw the low-dimensional features such as edge and texture.However,they have ignored the enormous difference between the WSIs and the natural images. Recently, some new methods have been proposed to suitably get features from the WSIs to overcome the significant shortcomings. Without the knowledge of each patch-level labels, the self-supervised learning methods can autonomously impart the outstanding feature extraction ability to the model.

The method named SimCLR in [19], one of the self-supervised learning(SSL) methods, using the contrastive learning has excellent feature extraction capacity even comparable to the supervised learning model thanks to the collected 57 digital histopathology datasets with none labels. Aside from SimCLR as a strategy used in [20] and [21], KimiaNet from [22] has also been one of the most welcome pre-trained models used for survival analysis exploiting a variable, multi-organ open image repository lick TCGA, which has been employed directly to extract the feature embeddings of image patches in [23] [21].

In the research of [24], Liu et at. introduced the EOCSA framework, which focuses on the survival analysis of EOC by utilizing deep learning techniques to process WSIs. They developed a prediction model called DCAS using CBAM [25], which demonstrated excellent feature extraction capabilities and strong predictive performance. The DCAS model employs a cluster-selection strategy to efficiently eliminate redundant information. Unlike other deep survival prediction models, they integrated spatial and channel attention modules in the DCAS model to capture tumor-related information. Moreover, their weight calculation method enabled the generation of more distinctive patient-level features. However, there are still some limitations in the EOCSA framework. Firstly, the cluster selection process, while effective in choosing discriminative patches, can be computationally expensive. The efficiency of the EOCSA framework may decrease when handling an extremely large number of WSIs. Secondly, not all selected patches are guaranteed to be highly discriminative, which could potentially impact the accuracy of survival analysis if non-discriminative patches are included.

In the study of [26], two deep-learning algorithms were developed using whole-slide digitized histological slides (whole-slide imaging; WSI) to predict the survival of patients with hepatocellular carcinoma (HCC) who underwent surgical resection. The study utilized both a discovery set (Henri Mondor Hospital) for algorithm development and an independent validation set (The Cancer Genome Atlas [TCGA]) for validation. Whole-slide images were divided into smaller squares ("tiles"), and features were extracted using a pretrained convolutional neural network. The first algorithm, "SCHMOWDER," incorporates an attention mechanism for tumor areas annotated by a pathologist, while the second, "CHOWDER," does not require human expertise. In the discovery set, both SCHMOWDER and CHOWDER achieved high c-indices for survival prediction surpassing a composite score based on baseline variables. These results were further confirmed in the TCGA dataset, where both models outperformed the baseline variables. Pathological analysis revealed that specific tumor characteristics, such as vascular spaces, macrotrabecular architectural patterns, and the absence of immune infiltration, were strongly associated with poor survival. Overall, this study demonstrated the potential of artificial intelligence in refining HCC prognosis prediction and emphasized the significance of collaboration between pathologists and machine learning for deep-learning algorithm construction that benefited from expert knowledge and contributed to a better biological understanding of their predictions.

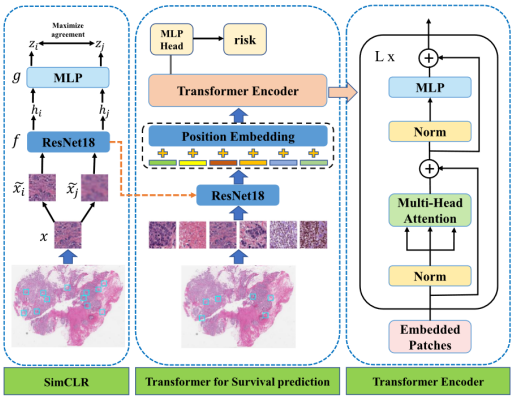

Fig. 2: the framework of SeTranSurv(image from [27])

SeTranSurv, the proposed model in [27](and shown in fig.2) applies SimCLR to train the initial feature extraction model ResNet18 [28] to get the specialized model for downstream task survival prediction. During the training the model applies the contrastive loss [29] to enhance feature extraction ability. Firstly augmentation module in SeTranSurv using two methods to transform a random image example to a position pair, then the model uses the fine-tuned feature extractor raised above to get the features. Considering other images as negative examples ensures that the views of different slide images are far apart in the high-dimensional space and the views of the same image are closer in the process of training. Additionally, position encodings capturing spatial information and self-attention modules learning correlation between patches are put into the training of the model above to obtain slide-level features and therefore the patient-level features.

The workflow of SimCLR is depicted on the left side of fig.2. This consistency is maximized using the contrastive loss [29] in the potential space. The framework encompasses four major components. A data augmentation module that randomly transforms any given data instance into two correlated views. The image, denoted as x undergoes two different data augmentation methods to

produce $\widetilde{x}_i$ and $\widetilde{x}_j$, forming a position pair. The goal of the contrastive prediction task is to identify $\widetilde{x}_j$ in $\{\widetilde{x}_k\}_{k\neq i}$ for a given $\widetilde{x}_i$. A neural network-based encoder f that extracts representation vectors from augmented data instances. A small MLP g that maps representations to the space where the contrastive loss is applied. The MLP is employed to obtain Eq.9

$$z_i = g(h_i) = W^{(2)}ReLU(W^{(1)}h_i) \tag{9}$$

A contrastive loss function defined for a contrastive prediction task. For a given batch size N, the set $\{\widetilde{x}_k\}, k \in \{0...N\}$ comprises a positive pair of examples $\widetilde{x}_i$ and $\widetilde{x}_j$.

As illustrated on the right side of fig.2, the Transformer Encoder [30] comprises multiple encoding blocks, each with constant widths. A learnable embedding is prepended to the sequence of embedded patches (z_0^0) like BERT [31], and its state at the output of the Transformer encoder (z_L^0) serves as the WSI representation y. The Transformer encoder consists of alternating layers of multi-headed self-attention (MSA) and MLP blocks (Equations 10 and 11).

$$t_l = MSA(LN(z_{l-1})) + z_{l-1}, \quad l = 1...L \tag{10}$$

$$z_l = MLP(LN(t_l)) + t_l, \quad l = 1...L \tag{11}$$

$$L(\boldsymbol{R}) = \sum_{i\in\{i:S_i=1\}} (-R_i + log \sum_{j\in\{j:T_j\leq T_i\}} exp(R_j)) \tag{12}$$

Layer normalization (LN) is applied before every block, and residual connections follow every block. The MLP comprises two layers with a GELU non-linearity. The Transformer Encoder is composed of six encoding blocks, each with four heads, and the hidden size of the MLP is 128. The resultant $y = LN(z_L^0)$ embodies the amalgamation of high-level semantic features by the Transformer Encoder. Passing through an MLP Head module denoted as $R = W^{(2)}ReLU(W^{(1)}y)$, it directly produces predicted risks. The loss function corresponds to the negative Cox log partial likelihood (Eq.12) designed for censored survival data, where S_i and T_i represent the censoring status and survival time of the i-th patient, respectively.

In the model of [32], particularly, a single filter within the convolutional layer is utilized for aggregating one dimension of the multi-dimensional characteristics from neighboring patches. The combination of these d filters results in the subsequent layer's feature representations. Subsequently, following the ViT's procedures, Shen et al. flattened the patch embedding and introduced position embedding. In addition, for the purpose of further simplification, they opted for Nystrom-based linear transformers [33] to replace the conventional self-attention transformers. The Nystrom approach is applied to approximate the softmax matrix in self-attention by randomly selecting a subset of columns and rows. As a result, the time complexity within this module can be reduced. Moreover, in order to offer understandable rationales for prediction outcomes, they introduced a post-hoc explainability technique for survival analysis.

The approach in [34] fuses Red, Green, Blue (RGB) color image data with Nucleus, Tumor, Lymphocyte (NTL) data, which includes nuclear segmentation maps, tumor region segmentation maps, and tumor infiltrating lymphocyte (TIL) maps generated through deep learning techniques. Essentially, Liu et al. treated the input data as composite images with six channels, incorporating the RGB channels and the NTL channels. They adapted a recent state-of-the-art deep learning survival analysis method developed by [35] to function with the RGB and NTL channels as well as multi-resolution data. The morphological characteristics of nuclei and cells exhibit variations dependent on cancer subtype and stage, making them useful for predictive models. Likewise, the spatial properties and distributions of tumor regions serve as important indicators of cancer progression and treatment response. In addition to augmenting Whole Slide Images (WSIs) with the TIL, tumor, and nuclear segmentation channels, they trained deep learning models with different resolutions of input data to capture both local and global perspectives on the distribution of TILs, tumor regions, and nuclei in tissue.

Without using convnet-based methods, the model in [36] also considers the self-supervised learning (SSL) methods mainly for making full use of plentiful color information designed for WSI patches or pixels without hand-actuated labels, totally colorization and cross-channel as the pretext tasks. The colorization model is trained to predict corresponding color channels based on the lightness channel and the letter, on the other hand, is trained to get lightness channel using given color channels data, after which the visualized results indicating highlighted overall structure of nuclei and tissue.

The model DSCA introduced in [37], inspired by modern feature pyramid networks [38], comprises two essential elements: the dual-stream and cross-attention. The dual-stream module independently processes low- and high-resolution patches, enabling efficient learning of hierarchical WSI representations. In the high-resolution stream, a square pooling layer is introduced to drastically reduce the number of high-resolution patches, thereby minimizing computational costs during network training. This square pooling operation is implemented using a cross-attention mechanism, where high-resolution patches are pooled under the guidance of a global low-resolution patch. This strategy effectively addresses the potential semantic gap between features at different resolutions, promoting seamless fusion of dual-stream features. The proposed DSCA offers several key contributions. Firstly, it introduces a distinctive dual-stream network with cross-attention that fully harnesses image pyramids to enhance the visual representation of WSIs. Secondly, the square pooling layer significantly reduces computational costs, serving as a critical component for computational efficiency. Lastly, the cross-attention-based pooling method effectively handles the semantic gap between high- and

low-resolution features, facilitating improved fusion of dual-stream features and enhancing hierarchical representation for superior prediction performance.

In the paper [39], Diao et al. introduce a fully automated dual-branch global fusion pipeline at the cellular level to forecast survival based on Whole-Slide Images (WSI). In contrast to prognosis methods that focus on nuclei, their pipeline effectively leverages global contextual information, encompassing various cell types and their locations. Initially, they generated an embedded WSI map by segmenting and classifying all cell nuclei in the WSI, thereby preserving their spatial information. Subsequently, they calculated features separately for the overall WSI and the relationships between different regions within the WSI based on the embedded maps. The former features emphasize global cell type and distribution, while the latter features focus on contextual relationships among different or within the same cell types. Finally, these two types of global features are fused to evaluate their prognostic implications. The proposed method demonstrated higher robustness compared to individual global features. Moreover, this comprehensive framework holds potential for predicting survival and treatment responses in other cancer malignancies.

In this work, they proposed a hierarchical visual converter (HVTSurv) [40]for predicting patient horizontal survival, which gradually explores the local horizontal space, WSI horizontal context, and patient level hierarchical interaction in the patient level package. For each WSI, the local level interaction layer will first encode the local spatial information. Then, spatial mixing is applied to make the model perform similarity calculations on features in different local windows. Finally, all WSI level features will be connected to perform attention pooling, and we use patient level representations to predict the patient's risk of danger. It is worth noting that in the local window block, this method adds relative position offset to the self attention calculation and uses Manhattan distance to measure distance,it can make the model more sensitive to short rather than long distances [41]. In the shuffling window block, this method shuffles patient features in different local windows, and then performs window partitioning and self attention calculations.

They prposed kernel attention Transformer (KAT) [42] . Compared to ViT [30], they use cross-attention between the tokens and a set of kernels to replace the token-wise self-attention to achieve information transmission. The information flowing in KAT is achieved by cross-attention operation between the kernels and patches, which includes the procedures of information gathering (I.G), information broadcast (I.B), and information aggregation (I.A).Through I.G flow, the local information described by the patch representations is reported to their nearby kernels for information gathering. Then, through the I.B flow, the regional information summarized by the kernels is broadcast back to the patches. Based on the bi-directional message passing flow, communication among the patch representations of the WSI can be accomplished.Finally, through I.A flow ,a classification token is used for summing up the information from all the kernels.In this information transformation process, they bind a set of anchor points defined by spatial positions in the patch to the kernel of KAT, creating a soft mask based on layered anchor points, guiding cross attention to perceive multi-scale features of WSI.The experiments have demonstrated that the kernel-based cross-attention contributes to a competitive performance for WSI classification and meanwhile significantly reduces the computational complexity of ViT in both the training and the inference stages.

Inspired by the potent multi-head attention mechanism initially introduced in the transformer [43] model, the researchers innovatively adapted this mechanism to function efficiently in the realm of multiple instance learning. Specifically, the transformer model's core function is to learn context-sensitive representations for every element within an input sequence. However, this necessitates attention between every pair of elements, leading to a quadratic increase in computational cost with longer input sequences. In the case of the prognostic information extraction task, the sole focus is on the global representation of the input sequence. This allowed the elimination of computationally intensive pairwise attentions among the input sequence, retaining only the relevant attentions between global queries and local keys. This modification effectively ensures that computational costs scale linearly with input length. Building upon this innovation, a multi-head attention framework for cancer survival prediction, named MHAttnSurv, was introduced, designed to work with whole slide images (WSIs) without the need for region-of-interest (ROI) annotations. In the MHAttnSurv framework, a backbone ResNet model is employed to extract features from randomly sampled WSI patches. These features are subsequently projected into values and keys, which are segmented into several chunks along with a learnable query vector. Within each chunk, parallel attention processes work concurrently to pinpoint discriminative regions, and the results from each attention map are concatenated to facilitate survival prediction.

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 selfattention 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 pglobal and multiple local prototypes plocals and quantifies the patches correlative to each plocals in the form of ratio factors (rfs). The ratio information is subsequentially embedded into the feature space for representation enhancement.

They proposed a survival prediction method [44] based on tissue pathology and tissue region features extracted from WSIs. Use the DeepConvSurv model to extract histopathological features from plaques of actual tissue types (tumors, lymphocytes, stroma, and mucus), and

extract tissue region features from the tissue map of WSI by using image processing techniques to locate and quantify tissue regions (tumors, cells, and stroma). Firstly, this method uses the DeepConvSurv model to extract the histopathological features of tumors, lymphocytes, stroma, and mucus. Secondly, by evaluating the area and proportion of tumors, lymphocytes, and stroma, tissue area features are retrieved from the tissue map. Thirdly, we use the extracted histopathological and tissue regional features to predict patient risk using six survival models.

Inspired by GAN [45] and other applications other than image generation [], the author proposes a new framework for survival analysis of gigabit pixel entire slide images, called Adversarial Multiple Instance Learning (AdvMIL). This framework no longer relies on the classical paradigm of event time modeling; On the contrary, it is based on event modeling for adversarial time and integrates multi instance learning required for WSI representation learning. The AdvMIL framework is based on event modeling for adversarial time and combined with MIL for WSI representation learning. The paradigm combining GAN and MIL is implemented through an RLIP (Region Level Instance Projection), which can be used to implement discriminators for WSI survival analysis through fusion networks.

Unlike prior studies that pick and choose specific regions of interest (ROIs) from whole-slide images (WSIs) for survival analysis, new method can utilize the complete tissue and tumor micro-environment without the need for detailed annotations within WSIs. Additionally, it can seamlessly incorporate multiple diagnostic slides of various dimensions from a single patient sample for both training and inference in a unified framework.

C. Multimodal

One limitation for some existing survival models is that they initially focus on one modality and cannot sufficiently handle multi-modalities data. Actually, multi-modalities information could provide complementary and auxiliary information for tumor diagnosis. For instance, molecular data and whole slide images share relevant representations to describe the same event in tumor growth and symptoms which are very critical for tumor diagnosis. Therefore,it is essential and necessary to combine and integrate multi-modal data such as pathological images, genotypic information and clinical data for explaining and understanding cancerous heterogeneity and complex symptoms for customized treatments and healings, consequently boosting the survival predictions. Although hematoxylin and eosin (H&E)-stained slides are enough to build a comprehensive diagnosis, other modal data can provide a deeper description of the tumor. For example, genomic profiles being comprising of ten thousand dimensional sequences can provide a molecular characterization of the tumor. Additionally,during the past several decades, multiple clinicians have made clinical cancer survival prediction on the basis of clinical variates and experience, therefore the clinical data is also an important source modality for multi-modal survival prediction. However, multi-modal survival prediction faces an important challenge due to the huge data heterogeneity gap between WSIs and other modalities, and many proposed approaches use simple multi-modal fusion mechanisms for feature incorporation, which give up mining important multi-modal relationships. Unlike needle-in-a-haystack challenges, survival outcome prediction involves the modeling of diverse visual elements within the tumor micro-environment. These elements are not distinguishable by traditional Multiple Instance Learning (MIL) methods. For instance, they include the co-occurrence of tumor cells and lymphocyte infiltrates, which is linked to a positive prognosis. This necessitates the modeling of interactions over medium to long distances among instances within Whole Slide Images (WSI). While it is commonly tackled as a weakly-supervised problem utilizing solely gigapixel Whole Slide Images (WSIs), the conventional perspective on survival outcome prediction views it as a multi-modal learning task. Due to the variant costs and trauma degree of multi-omics examination, the scenarios of missing modalities are very common.

The model in [46] gets the features from the pathological images extracted using Cellprofiler [8] from the image tiles in ROIs,namely geometry, texture, holistic features.In addition,the model uses the preprocessed genetic data and one feature selection operation SPACE [47] to select representative features to integrate data for the principal component regression model for survival prediction. The experiments focusing on ADC lung cancer revealed that the results could be better than only using data of genes or images, which demonstrated that the genetic data could actually enhance the prediction performance of the survival analysis model to some extent.

Differently, the experiments in [16] were conducted on other two cancer types: glioblastoma multiforme (GBM) and lung squamous cell carcinoma (LUSC) using pathological images and molecular data including protein data, Copy number variation(CNV) data and so on. But similarly, to fuse the data from different modality for better results, the model firstly learn deep representations from two kinds of data using separate Convolutional Neural Networks (CNNs).Next, the representations passing through the sub-network in the model are connected to get the new representation, which serves as the input of the correlational layer. Certainly, the deep correlation layer is used to decrease the discrepancy by maximizing the correlation. After the common layer, the output acts as the input of the survival prediction layer using the negative log partial likelihood as survival loss function. Compared with the models handling the linear condition, the new model DeepCorrSurv can learn complex correlation using deep neural networks by using the unsupervised method to learn the interactions and the survival loss network to fine-tune the model, eventually getting better results in the comparison experiments about lung and brain cancer. The results showed that the common representation after

maximizing step could bring better performance to the survival prediction measured by the metric named concordance index values or c-index values.

In the paper [15], the model named GSCNN aims for fusing genomic data and second modal data from The Cancer Genome Atlas (TCGA) Lower-Grade Glioma (LGG) and Glioblastoma (GBM) projects. If the second modal data are added to the network during the whole training process, the median c index will improve more than simply integrating the second type data into the fully connected layers. Then the model proves that the Molecular subtype is significant in the multi-variable regression model.

Again, for genomic information, Sun et al. in [48] considered building distinct models to make survival predictions according to the five major molecular subtypes of the breast cancer. The genomic data chosen in the model consists of gene expression, copy number alteration (CNA), gene methylation and protein expression. The main idea in the paper is how to merge different types of data as one type, and in this way using multiple kernel learning (MKL) is a choice. To integrate the genotypic information and image data, GPMKL model was proposed using 5 independent Gaussian kernels and having a integrating step. The baseline algorithms used that time for comparison includes LASSSO-Cox [49], elastic-net penalized Cox (EN-Cox) [50], Parametric censored regression models (PCRM) [51], Random survival forests (RSF) [52], Boosting concordance index (BoostCI) [53], Supervised principal components regression (superPC) [54]. Additionally, two independent models named GMKL and PMKL was constructed only adopting genomic data or pathological images data and four single dimensional models using four types of genomic data were also built. Subsequently, the results showed that the gene expression and protein information play relatively more important role than others and CNA makes a little contribution to holistic prediction accuracy.

In the model proposed in [55] named Multimodal-Prognosis, clinical, genomic and WSI images data are processed by FC layers, deep highway networks [56], the SqueezeNet architecture [57] separately. The architecture except SqueezeNet is demonstrated in fig.3. The microRNA data is also passed through deep highway networks discussed. One notable problem about the microRNA and clinical data is the missing data. However, the function of the highway networks dose not stand out without comparison with other methods in the paper. Then, the similarity loss(Eq.13-Eq.15) is used to train the model to recognize the patient-distinguishing patterns and correspond the data from different modality to generate associated representations.

$$sim_0(x,y) = \sum_{i,j \in \text{modalities}} \frac{\hat{h}_{0,i}(x_i) \cdot \hat{h}_{0,j}(y_j)}{|\hat{h}_{\theta,i}(x_i)||\hat{h}_{\theta,j}(y_j)|} \tag{13}$$

$$L_\theta(x,y) = \max(0, M - sim_0(x,y) + sim_0(x,x)) \tag{14}$$

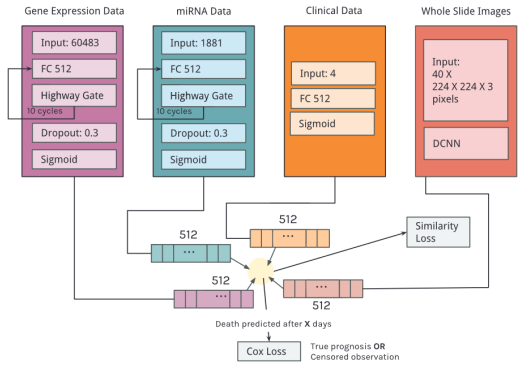

Fig. 3: the framework of multimodalprognosis(image from [55])

$$l_{sim}(\theta) = \sum_{x,y} L_\theta(x,y) \tag{15}$$

The model aggregates the similarity loss for each pair of modalities that are available. In the above expression, x_i represents the data for modality i, and $h_{\theta i}$ denotes the predictive model for modality i. It's important to note that the parameter M regulates the 'tightness' of the clustering. A higher value of M allows feature vectors for a given patient to be relatively different, as long as they remain similar to a certain extent. Conversely, a lower value of M compels feature vectors for a patient to be much closer together, which is generally more desirable but may lead to mode collapse. This loss is calculated for every pair of patients in a batch. Consequently, the unsupervised model could discern crucial, patient-distinguishing patterns in both genomic and image data. Thus, the final loss function is composed of cox loss and similarity loss in the unsupervised model. The multi-modal dropout was also invented that is dropping whole feature vectors of each modality and accordingly increase the weights of other modalities to build robust representations. The solution above was then validated in the experiments including visualizing the encodings of the pancancer patient cohort and calculating the C-index values and the results also demonstrated the essence of the adopted modalities.

One more variation usually ignored is the ordinal relationship among the survival time of different patients, which is assumed independent among patients. The model proposed in [58] named OSCCA intends to take advantage of the information. For extracting features, the model uses the methods in [59] to generate segmented nucleus and get the specific features. As to gene expression data, the co-expression network analysis algorithms are utilized to derive eigen-gene features. Considering the correlation between imaging and eigen-gene features, sparse canonical correlation analysis(i.e. SCCA) is used to choose features. Among the datasets, most patients are censored, denoting that their actual living time is longer than recorded data, while the uncensored patients have the real survival time. To make full use of the censored state, an ordinal sparse canonical correlation analysis (OSCCA) method

is proposed. In the newly proposed method, the equation estimates the uncensored information, while linear inequalities restrains the ordered relationship between censored and uncensored data.

After the model above, Shao et al. developed a new model named OMMFS in [60] using CNV data and level-3 DNA Methylation (DME) data additionally. The experiments showed that the log-rank test [61] had better stratification performance than univariate Cox regression method. Furthermore, the model implements the second feature selection function based on the Generalized Sparse Canonical Correlation Analysis (GSCCA) framework [62] to get the inherent relationship among different modalities. Likewise the modality in OSCCA, the model under discussing also notices the survival information of patients. The validation experiments demonstrated that the new model even had superior stratification of early-stage KIRC patients. To make comparison, the SGSCCA model was proposed with the same objective function without ordered survival information. However, the feature pre-selection strategy is based on the median value, which is too arbitrary to consider the accurate relevance between features and cancer type. Also, the image features relies on the regions of interest annotated by pathologists.

This model in [63] has a two-stage feature extraction process used in a deep learning model for pathology-specific layers. In the first stage, a pre-trained convolution neural network (CNN) is used to identify survival-discriminative features in patches of pathological images. These features are obtained through dilated convolution layers and max-pooling. In the second stage, global survival-discriminative features for a whole slide image (WSI) are generated by aggregating feature scores from multiple patches. A two-stage pooling approach, including 3-norm pooling, is used to rank and aggregate the most important features and patches. The resulting vector of aggregated survival-discriminative features represents a WSI for a patient, contributing to the integrative deep learning model. The genome and demography-specific layers in this model are adapted from the Cox-PASNet [64], which is a pathway-based sparse deep neural network. The genome-specific layers consist of a gene layer, a pathway layer, and two hidden layers (H1 and H2). The gene layer serves as an input layer for gene expression data, with each node representing a gene. The pathway layer incorporates prior biological knowledge from databases like KEGG for biological interpretation. Connections between the gene layer and pathway layer are established based on biological pathway databases, with pathway nodes representing specific biological pathways. The two hidden layers capture nonlinear and hierarchical relationships between the pathways. Clinical patient data are integrated into the demography-specific layer and combined with genomic features from gene expressions and aggregated survival-discriminative features from pathological images in the final hidden layer of the integrative model. To address overfitting in deep learning models with high-dimensional, low-sample-size data, the training technique from Cox-PASNet is applied. Instead of training the entire network, small networks are randomly selected, and sparse coding is used to create sparse connections for model interpretability. Training continues until convergence, with validation data used to monitor errors and prevent overfitting through early stopping. PAGE-Net statistically outperformed Cox-EN with histopathological images only and Cox-PASNet with genomic data only.

In the model proposed in [7], the cell graphs from histology images are supposed to get the cell-to-cell interactions and cell neighborhood structure. To build cell graphs, the first step is generating accurate nuclei segmentation, in which a conditional generative adversarial network (cGAN) is used to learn the appropriate loss function for semantic segmentation. The edge set and adjacency matrix of the graph are constructed using the K-Nearest Neighbors (KNN) algorithm from segmented cell nuclei. In addition to manually computed statistics, an unsupervised technique called Contrastive Predictive Coding (CPC) is employed to extract 1024-dimensional features from tissue regions centered around each cell. Graph Convolution Networks (GCNs) learn abstract feature representations for each node by aggregating feature vectors from their neighborhood through message passing. In scenarios with a high-dimensional feature space and limited training samples, traditional feed-forward neural networks are susceptible to over-fitting. To address the challenge and apply more robust regularization techniques when training feed-forward networks on high-dimensional, low-sample-size genomics data, the model adopts normalization layers inspired by Self-Normalizing Networks introduced by Klambaeur et al [65]. Moreover, the Kronecker Product is used to construct a multi-modal representation. The feature vectors of histology images, cell graphs, and genomic features undergo matrix outer product operations to create a multi-modal tensor. This tensor captures important interactions among these three modalities in terms of single-modal, bimodal, and tri-modal relationships. Ultimately, a neural network is trained using fully connected layers with the multi-modal tensor as input. The central aim of this method is to fuse heterogeneous modalities with distinct structural dependencies, thereby enhancing research and analysis in cancer pathology. To mitigate the impact of noisy uni-modal features during multi-modal training, a gating-based attention mechanism [66] is introduced to control the expressive power of features within each modality. When fusing histology images, cell graphs, and genomic features, the gating mechanism helps reduce the feature space's size before performing the Kronecker Product calculation.

To integrate the multi-modality data with different weights, the model proposed in [67] uses an asymmetrical Transformer encoder. The main idea to fuse other modality data unevenly is to add the new nodes and edges into the original graphs. Different from the normal self-attention in Transformer, the noisy genomic nodes cannot impact the image features because they do not have the outgoing edges, which can only improve themselves by the

influence of the imaging features. The GPDBN framework in [68] comprises an inter-modality bilinear feature encoding module (Inter-BFEM) and two intra-modality bilinear feature encoding modules (Intra-BFEMs) to efficiently handle information interactions both across and within genomic data and pathological images. This work needs further more data to enhance prognostic performance.

Similar to [55], the model in [69] can process the missing data including tabular clinical data (herein simply referred to as "clinical"), gene expression ("mRNA"), microRNA expression(miRNA), DNA methylation(DNAm), gene copy number variation(CNV) data and WSI. If some patients loss the whole modality data, they will be excluded when training the model using uni-modal data. And the missing data will be replaced by zero matrix when training the multi-modal model. To avoid over-fitting, the model has the mechanism to select some patients randomly and replace their specific modality data with zero input. The t-SNE visualization reveals that patients with different cancer types occupy distinct clusters in a two-dimensional space, aligning well with known cancer type prognosis. The model is trained end-to-end and delivers non-proportional outputs, achieving accurate long-term predictions across various cancer entities. Clinical data is identified as the most informative unimodal data modality.

Inspired by methods in Visual Question Answering (VQA), the model in [70] introduces new approach to let histology patches attend to genes in the survival prediction. Unlike late fusion-based architectures that simply concatenate WSI-level bag representations with genomic features, the genomic-guided co-attention(GCA) layer captures multi-modal interactions, connecting histology-based visual concepts with gene embeddings, similar to the approach used in VQA. These interactions are visualized as attention heatmaps at the WSI level for each genomic embedding. Additionally, the GCA layer reduces the effective "sequence length" of WSI bags from M instance-level patch features to N gene-guided visual concepts, where N represents the effective sequence length of gene embeddings (with $M > N$). This reduction enables more advanced feature aggregation techniques using self-attention and Transformers, allowing for supervision with entire WSIs, which was previously unattainable. One limitation of the research was that they utilized a gene set that had been curated previously, and it contained genes with overlapping biological functional impact.

The PG-TFNet proposed in [71] comprises three modules: a transformer-based multi-scale pathological feature fusion module, a cross-attention transformer-based multi-modal feature fusion module, and a final Cox layer for clinical data integration. The multi-scale feature fusion module processes pathological images at different magnification levels using transformer encoders, capturing relationships between image patches. The feature vectors from various magnification levels are concatenated, and a learnable class token and fixed positional encoding vector are introduced to create a multi-scale feature sequence. The multi-scale feature sequence is then input into a stack of transformer encoders for multi-scale feature fusion. This approach enables the extraction of morphological features at various field-of-view scales, improving the model's ability to understand the relationships between image patches. The model utilizes a cross-attention transformer module to integrate pathological and genomic data. This module combines feature representations from both data types, enabling effective multi-modal data fusion for improved analysis of cancer prognosis.

To overcome the limitation of Kronecker product, HFB-Surv extended GPDBN [68] mentioned with the factorized bilinear model. The model introduces a hierarchical multi-modal fusion approach, which employs factorized bilinear models to progressively integrate information from different levels, reducing computational complexity. Genomic data processing involves data cleaning, normalization, and feature selection using randomForestSRC [72]. To process the missing values, the model uses the weighted nearest neighbors algorithm like [73]. Pathological images are quantitatively analyzed to extract relevant features.

The main contribution Richard et al. in [74] is the development of a research tool, the Pathology-Omics Research Platform for Integrative Survival Estimation (PORPOISE). PORPOISE is an interactive platform that provides prognostic markers learned by model for thousands of patients across 14 cancer types. It allows users to visualize H&E images with interpretability overlays, local explanations of molecular features, and global patterns of feature importance. They used PORPOISE to analyze high-attention morphological regions in whole slide images and confirmed that the presence of tumor-infiltrating lymphocytes correlates with favorable cancer prognosis, as identified by their model. This tool facilitates the discovery of joint image-omic biomarkers. PORPOISE consists of three network components: an attention-based multiple instance learning network (AMIL) for WSI inputs, a self-normalizing network (SNN) for molecular features, and a multimodal fusion layer (MMF) to integrate feature representations from AMIL and SNN. AMIL localizes prognostically relevant regions in WSIs without selective patch sampling and aggregates them for feature representation. SNN transforms molecular data into low-dimensional features. The model constructs a joint multimodal feature representation for histology and molecular data interactions, which is used for survival analysis. One drawback of the platform is that while PORPOISE can elucidate "what," it may not always provide an explanation for "why."

In the model introduced in [75], Xie et al. utilized a multimodal attention module to generate well-structured aggregated feature embeddings for patients and their association matrix. Using these embeddings as the foundation, they constructed a graph neural network for predicting survival outcomes. Thanks to the superior graph quality and effective node aggregation process, the GNN model achieves precise predictions in both datasets with complete and incomplete data, offering a viable solution for handling

the problem of missing data. Each patient is treated as a node in a KNN-affinity graph, enabling the learning and encoding of topological structure and neighborhood information into each patient's feature vector. Unlike previous studies that treated patches of WSIs as nodes in a graph, this work innovatively constructs a graph of patients in datasets, motivated by the idea that patients with similar WSI and gene expression data are likely to have similar survival outcomes. This approach helps enhance patient representations, especially given the noisy and missing data often present in WSIs and gene expression datasets.

Apart from this, another approach to add the genotypic information is using the biological pathway databases to teach the model about the hidden biological functionality. PONET proposed in [76] uses a sparse biological pathway-informed embedding network for gene expression, additionally adopting the Multi-modal Factorized Bilinear pooling (MFB) method instead of original bilinear model to generate unimodal fusion to catch the modality-specific representations. Getting each uni-modal fusion, the model can use the output representations as the input of the bimodal and tri-modal fusion respectively utilizing the bimodal attention and tri-modal attention. Finally the model is trained through the Cox partial likelihood loss proposed by [55] used for the multi-modalities survival prediction to get the prediction results.

The model in [77] introduces multi-modal graphs as a foundation for a novel Hypergraph Convolution Network (HGCN) designed to extract prognostic-related information. The HGCN employs a node message passing mechanism for intra-modal interactions and a hyperedge mixing module for advanced modal interaction. Survival predictions were made by combining modalities using an online Multi-view Autoencoder (MAE) proposed in [78] paradigm during model inference. The research systematically explored the robustness of the multimodal survival prediction model, addressing an overlooked clinical challenge. To handle missing modalities, an online MAE method captures intrinsic dependencies and generates hyperedges. Several techniques are employed for handling missing modalities, including zero padding, multimodal factorized method (MFM) for reconstruction, autoencoders [79] [80], and a lower-bound approach without the online MAE. The experiments showed that the online MAE module played a key role in improving the prediction robustness. The study underscored the potential of masked signal modeling for self-supervised learning across different scientific domains.

This study in [81] developed a deep learning-based prognostic model, called MultiDeepCox-SC, for predicting survival outcomes in stomach cancer patients. The model integrates histopathological images, clinical data, and gene expression data to improve prognostic accuracy. The MultiDeepCox-SC model outperformed the current clinical benchmark model based on pathologic grade, stage, and clinical data, with a higher C-index of 0.744 compared to 0.660. It automatedly selects informative patches in histopathological images and identifies genetic and clinical risk factors associated with survival in stomach cancer. The risk score generated by the MultiDeepCox-SC model remained an independent predictor of survival outcome even after adjusting for confounding factors. The study also validated the model's performance on external datasets. The proposed fully automated prognostic tool based on histopathological images, clinical data, and gene expression data has the potential to improve pathologists' efficiency and accuracy and aid clinicians in selecting appropriate therapies.

D. Graph Neural Network

The extremely high gigapixel resolution of Whole Slide Images (WSI) requires researchers to divide them into smaller patches for analysis. Most of the previous work has primarily centered around aggregating images based on patches. But based on [82], these methods may also overlook the crucial context between image patches and their neighbors for accurate predictions. Graphs are mathematical models depicting connections between pairs of elements. They are particularly effective for illustrating relationships between individual patches within a Whole Slide Image based on their spatial proximity or correlation. Currently, in risk prediction based on GNN through WSI images, many researchers have also achieved state-of-the-art results.

DeepGraphSurv [83] is the first GCN-based survival prediction model that uses WSIs as input. The authors believe that intermediate patch-wise features are a suitable choice for constructing a graph. They integrate global topology features and local patch features of WSIs using spectral convolution operators, which is the core of the entire architecture. In their approach, each patch is treated as a node, and node features are generated using a pre-trained VGG-16 model on ImageNet. The initial WSI graph is built based on patch features, and the VGG-16 feature extractor is not fine-tuned for WSI patches due to the absence of patch labels. Consequently, the initial graph may not accurately represent the topology between WSI patches, which is attributed to the limited training of feature networks.To address this issue, the authors design an independent graph G and L to describe the topological relationships between specific survival-related WSI patches. This framework can simultaneously learn both local and global representations of Whole Slide Images by integrating local patch features with global topological structures through convolution. Typically, only a few local Regions of Interest (RoIs) in WSIs are relevant to survival analysis. Random sampling may not guarantee that all patches originate from RoIs. The attention mechanism is employed to selectively choose patches by learning their importance. So they introduce a parallel network with attention to adaptively learn attention on node features for selecting more important patches.

Chen [7] utilized a Graph Convolutional Network (GCN) approach to extract morphometric cell features from histology images. This article proposes a method

called Pathomic Fusion, which is a comprehensive framework for integrating histopathological and genomic features for cancer diagnosis and prognosis. In this approach, cells found within histological tissue are represented as nodes in a graph. These cells are initially identified using a deep learning-based nuclei segmentation algorithm, and connections between cells are established using the K-Nearest Neighbors (KNN) method. The features of each cell are initialized through a combination of handcrafted features and deep features obtained via contrastive predictive coding. The aggregation and combination functions employed in this approach are drawn from the GraphSAGE architecture, with the additional inclusion of node masking and hierarchical pooling strategies from SAGEPool. The primary objective of this method is to extract and learn cell morphological features from histological images.

DeepGraphSurv proposed sampling patches in a WSI as nodes, followed by constructing edges between patches via feature similarity on the embedding space and using spectral convolutions. Alternatively, "Pathomic Fusion" constructed a cell-based graph for small image ROIs followed by spectral convolutions. However, in this approach for graph construction, GCNs are unable to learn context-aware features as message passing as feature interactions between adjacent image patches are not modeled. To enhance the modeling of feature interaction between adjacent image patches during message transmission, Chen [84] introduced a context aware, spatially resolved patch based graph convolutional network. This network hierarchically aggregates instance level histological features to capture local and global topological structures in the tumor microenvironment Their method uses Graph Convolutional Network (GCN), which iteratively aggregates and combines node features of different hidden layers through message passing. The message passing function of the network is adapted from DeepGCN [85], including message construction, permutation invariant aggregation, and update functions. Unlike previous graph based methods, they create neighborhoods based on the nearest neighbors in the embedded space, and construct graphs in Euclidean space. This allows them to use spatial convolution to perform local neighborhood aggregation functions similar to convolutional neural networks (CNNs). Compared with other methods of connecting nodes through adjacent image patches, Patch-GCN has improved performance, allowing for learning coarse-grained to fine-grained topological structures in tumor microenvironment.This model is suitable for any weakly supervised learning task in computational pathology that uses slide level or patient level labels, which helps to gain a more comprehensive understanding of representation learning in the tumor microenvironment.

This paper introduces a gastric cancer survival prediction method called Gc-Splem [75], which employs multimodal learning by combining whole-slide image (WSI) data and gene expression information. The method consists of three key components: the WSI feature extractor, the modal fusion network, and the predictor based on a Graph Neural Network (GNN). Firstly, the WSI feature extractor is utilized to extract features from the WSI images. Subsequently, the modal fusion network integrates WSI features with gene expression data to generate high-quality feature embeddings. Lastly, the GNN-based predictor leverages patient associations to make survival predictions, ensuring accuracy through high-quality graph and node aggregation mechanisms, even in the presence of missing data. In a GNN based predictor, there are three key steps: first, construct a K-nearest neighbor (KNN) affinity graph to represent patient relationships, and use cosine distance to connect; Secondly, apply a double-layer graph convolutional network to process patient feature vectors and construct graphs, including element by element multiplication, linear calculation, and ReLU activation function; Thirdly, using the survival prediction layer, similar to the final fully connected layer in modal fusion networks, including the loss function and number of neurons, based on the output of graph convolutional layers for survival prediction.

Attention mechanisms have become almost a de facto standard in many sequence-based tasks.Inspired by it, Petar have presented graph attention networks (GATs) [86], novel convolution-style neural networks that operate on graph-structured data, leveraging masked self-attentional layers. Experiments have shown that GAT outperforms traditional graph convolution methods in both direct learning and inductive learning tasks. The TEA graph [87] adopts a graph attention network (GAT) model structure, which utilizes attention scores in GNN to learn contextual features in heterogeneous tumor environments.We adapt the supernode method [88] to WSI to compress and represent the gigapixel-sized image into memory-efficient graph structures.Using a Graph Attention Network (GAT) to learn contextual features within the tumor environment. It can effectively handle pathological features with varying backgrounds, such as immune cells, and their interactions with the surrounding environment.TEA-graph uses a GAT with positional embeddings to extract the context features around the superpatch by aggregating the neighbourhoods of the superpatch with different attention scores.

In this article, we propose a new hybrid graph convolutional network (HGCN) [77]equipped with an online masked automatic encoder for multimodal cancer survival prediction, which mainly includes data from pathological sections, clinical records, and genome maps. This network can effectively utilize complementary information from different patterns and is robust in the absence of patterns. The key idea of our framework is to model multimodal medical data as a full graph structure to facilitate inter modal and intra modal interactions, and to construct masking signal models to simulate missing modal scenarios to support modal interactions in the inference process. We propose a new hybrid graph convolutional network, which consists of the same graph convolutional layer and a well-designed hypergraph convolutional layer with hyperedge

mixing modules, to better facilitate inter modal and intra modal interactions in multimodal graphs. We have designed an online masked automatic encoder example to handle missing modal scenarios. This paradigm cleverly utilizes the inherent dependencies of multiple modalities learned through the transformer, thereby generating missing hyperedges during the model inference process.

This article proposes a multi hypergraph based learning framework called "HGSurvNet" [89]to achieve effective high-order global representation of all slide histopathological images (WSI) for survival prediction. inspired by DeepGraphSurv, they use two important types of information, phenotype (visual appearance) and topology information, for the multi-hypergraph learning. In the latent feature space, two patches with similar feature vectors may be connected by a common hyperedge. In contrast, in the image space two neighboring patches on a common topological path may be connected by the same hyperedge. We name the sub-hypergraph generated in the above two spaces as phenotype-wise sub-hypergraph and topology-wise sub-hypergraph. In the last step, we concatenate the two sub-hypergraph and form a combined hypergraph incidence matrix H from Htop and Hphe.In the final step, we connect the two subgraphs and form a composite hypergraph incidence matrix H from Htop and Hphe. This completes the initialization of multiple hypergraphs

In their proposed HIGT framework [23], a WSI pyramid is constructed as a hierarchical graph. Their Hierarchical Interaction GNN and Hierarchical Interaction ViT block have the capability to capture both local and global features. The Bidirectional Interaction module within the latter allows nodes from different levels to interact. Finally, their Fusion block aggregates both coarse-grained and fine-grained features to generate the slide-level prediction. The paper presents a novel Hierarchical Interaction Graph-Transformer (HIGT) for Whole Slide Image (WSI) analysis, which combines Graph Neural Network and Transformer architectures to learn both short-range local information and long-range global representation of the WSI pyramids. The HIGT framework abstracts each WSI as a hierarchical graph, where feature embeddings from multi-resolution patches serve as nodes and the edges represent spatial and scaling relationships. The framework includes hierarchical graph convolution blocks to learn short-range relationships among graph nodes, pooling operations to aggregate local context, a Separable Self-Attention-based Hierarchical Interaction Transformer to learn long-range relationships, and a fusion block to aggregate features from different levels of WSI pyramids for slide-level prediction. The paper also analyzes the computation cost of the proposed methods and demonstrates that the model maintains promising prediction results while reducing computational cost and model size effectively.

Previous studies primarily tackled survival prediction by directly applying a regression model to each patient's WSI, often overlooking the relative ranking order among patients. Apart from individual data predictions, it's worth noting that ranking order becomes even more critical when comparing different datasets. The paper proposes a ranking-based survival prediction method called RankSurv [11], which considers the ranking information during the learning process. It uses a hypergraph representation to conduct hazard prediction on each whole-slide image (WSI) and conducts a ranking-based prediction process using pairwise survival data. This study presents two notable innovations. First, they introduce the use of a hypergraph structure to represent hierarchical information. In this approach, they create the hypergraph signal matrix X (SIZE) and the hyperedges incidence matrix H (SIEZE). Hypergraph spectral convolutional layers are employed for training. After multiple layers of spectral convolution, the N hyperedges can correspond to N patterns of pathogenic factors. To enhance the supervision of ranking prediction, they introduce a BCR loss function and a survival-specific likelihood function for the prediction process. It uses Bayesian optimization criterion to make the concordance index learnable and supervisable . This method has been evaluated on three public carcinoma datasets (LUSC, GBM, and NLST) and has shown significant improvements over state-of-the-art methods in terms of quantitative results.

However, directly modeling WSI as a graph is computationally complex, and the lack of priori knowledge leads to the difficulty in learning the best graph representation. In WU's research, a multi-instance learning (MIL) framework called DeepGCNMIL [90]was introduced for survival analysis of gigapixel Whole-Slide Images (WSIs). Initially, the ResNet50 pretrained model from ImageNet was employed to extract the morphological features of patches, which were subsequently clustered using the K-Means method. Within the context of a graph structure, a Graph Neural Network (GNN) was applied, treating each feature patch in every cluster as a node. The GNN subnetworks were embedded into a deep learning architecture, consisting of three layers of Graph Convolutional Neural Networks paired with ReLU activation functions, and a final addition of global pooling layers, such as average pooling. Relationships between feature patches were established through edge connections between nodes, and feature information was forwarded to generate the entire graph's feature vector. This approach effectively modeled relationships between patches, extracting meaningful representations of phenotypes for survival prediction. Finally, multiple heads and aggregated phenotype features were introduced to the WSI representation, facilitating prognostic risk assessment. It is more suitable for situations where manual ROI labeling is lacking in large-scale cancer datasets, and can simultaneously process prognosis predictions for cancer patients with different numbers and sizes of entire slide images. Moreover, this model can be easily transplanted to other cancer datasets and applied to other tumor types.

Due to the massive pixel data in a single WSI, fully exploiting cell-level structural information from the gi-

gapixel WSI is challenging. Most current studies address the issue by selecting a limited number of image patches to create a graph-based model, such as a hypergraph. However, the scale of this sampling is a crucial bottleneck because it represents a fundamental obstacle to expanding the sample size for transductive learning. In the proposed method, the "Large Hypergraph Factorization Neural Network (b-HGFN)" [91]is introduced to extract higher-order representations from full slide images (WSI). This method consists of two main components: a "large hypergraph factorization neural network" and a "multi-level sorting survival prediction". Initially, this method used pre trained networks for sampling, using each sampled patch as a vertex in the hypergraph. The hyperedges of a graph are constructed based on the distance between the visual features of each pair of vertices (such as Euclidean distance), which is determined using the K-nearest neighbor (KNN) method. A significant innovation of this study is the introduction of a factorization module in constructing the correlation matrix H. This factorization module allows for the training of large hypergraphs with a large number of vertices, thereby establishing relationships between "vertices belonging to hyperedges" and "hyperedges containing vertices". In addition, this method defines a low dimensional hypergraph decomposition Laplace matrix.Due to dimensional decomposition, b-HGFN can effectively handle densely sampled patches on a large scale, while the dimensionality reduction hypergraph convolution layer can generate high-order global features that encapsulate the essence of each WSI.The focus of the 'Multi level Ranking Survival Prediction' component is to use a regression layer to calculate the hazard score for each WSI, which has a fully connected neural network using a negative Cox-log partial likelihood loss function. In order to consider ranking information, a multi-level loss function was designed and an overall ranking supervision module (NDCGLoss2) was introduced to effectively utilize global information.

The paper [92] presents a structural learning approach aimed at capturing potential patch correlations and generating adaptive and sparse structures. This approach allows for optimal patch selection for feature aggregation and efficient GCN training, eliminating the need for predefined, fixed graphs tailored to the task.The proposed method introduces a scalable graph convolutional network called GraphLSurv designed for survival prediction based on gigabit-pixel full-slide images (WSI). This framework comprises three key components: WSI preprocessing, survival perception structure learning, and GCN-based survival prediction.In a broader context, the author conducts a review of deep learning methods relevant to gigabit-pixel WSIs, categorizing them into two groups: graph-based and graph-free approaches. GraphLSurv stands out by generating adaptive and sparse structures for patches, allowing it to dynamically capture and adjust potential patch correlations.

Given the problems posed by pre-fixed or densely-connected patch structures,this paper [92] proposes a structure learning method to construct adaptive and sparse patch correlations. A simple way to construct a patch structure is to connect each patch with its k nearest neighbors (k-NN) or adjacencies. This article proposes a structural learning strategy to generate adaptive and sparse structures, using a patch similarity learning technique that can optimize connection patches and directly calculate edge connectivity without relying on previous computational models used in DeepGraphSurv. This method can be summarized as calculating the cosine distance for each patch to measure patch similarity. This outputs a symmetric matrix, where the size of each element represents the strength of the edge connection. Afterwards, if the connection strength of AL is less than the threshold (x), We mask its elements to 0. This operation produces a sparse structure that can filter unnecessary connections, thereby reducing noise correlation in AL.

This article proposes a new multi instance learning framework H2-MIL [93] for full slide image analysis based on graphical neural networks. It utilizes heterogeneous graphics of different resolutions to simulate the characteristics and spatial scale relationships of multi resolution patches. Firstly, use heterogeneous graphs with resolution attributes to represent WSI with multiple resolution levels. Then input it into the H2MIL network, which consists of multiple RAConv modules and IHPool modules, forming the core of the network. This part constitutes the learning process of graph attention convolutional networks, where nodes continuously aggregate information from their neighbors to extract dense but differentiated feature representations. This helps to mine hierarchical semantic information for WSI analysis. Finally, based on the learned representation, a WSI level classifier is used for prediction. This framework includes a resolution aware attention convolution (raconV) module for learning compact discriminative representations from graphs and addressing the heterogeneity of node neighbors at different resolutions. It also integrates the Iterative Hierarchy Pooling (IH Pooling) module, which can gradually aggregate heterogeneous graphs based on the scaling relationships of different nodes, thereby exploring structured information related to WSI and tasks.

The author proposes a new graph neural network (GNN) based model, called SlideGraph+ [94] for directly predicting HER2 state from full slide images of conventional hematoxylin and Eosin (H&E) glass slides. This model captures the overall organization and structure of an organization, unlike traditional patch based methods with limited visual context, which can manipulate graphics at the entire slide level. The framework proposed by this network consists of four steps: extracting features from local regions of the entire slide image, spatial clustering to group similar image blocks into clusters, generating graph representations based on these clusters to capture cellular and morphological topology, and using graph neural networks to predict receptor states at the graph node level and slide level. This article also introduces the DAB density regression model, which directly predicts

DAB intensity from HE stained images and may not require IHC staining when evaluating HER2 expression.

IV. CONCLUSION AND DISCUSSIONS

A conclusion section is not required. Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions.

A. Challenges

B. Future Prospects

APPENDIX AND THE USE OF SUPPLEMENTAL FILES

Appendices, if needed, appear before the acknowledgment. If an appendix is not critical to the main message of the manuscript and is included only for thoroughness or for reader reference, then consider submitting appendices as supplemental materials. Supplementary files are available to readers through IEEE Xplore® at no additional cost to the authors but they do not appear in print versions. Supplementary files must be uploaded in ScholarOne as supporting documents, but for accepted papers they should be uploaded as Multimedia documents. Refer readers to the supplementary files where appropriate within the manuscript text using footnotes. [1]

ACKNOWLEDGMENT

The preferred spelling of the word “acknowledgment” in American English is without an “e” after the “g.” Use the singular heading even if you have many acknowledgments. Avoid expressions such as “One of us (S.B.A.) would like to thank . . . .” Instead, write “F. A. Author thanks . . . .” In most cases, sponsor and financial support acknowledgments are placed in the unnumbered footnote on the first page, not here.

REFERENCES

[1] Olivier Gevaert, Frank De Smet, Dirk Timmerman, Yves Moreau, and Bart De Moor. Predicting the prognosis of breast cancer by integrating clinical and microarray data with bayesian networks. Bioinformatics, 22(14):e184–e190, 2006.

[2] Anant Madabhushi and George Lee. Image analysis and machine learning in digital pathology: Challenges and opportunities. Medical image analysis, 33:170–175, 2016.

[3] Bingzhong Jing, Tao Zhang, Zixian Wang, Ying Jin, Kuiyuan Liu, Wenze Qiu, Liangru Ke, Ying Sun, Caisheng He, Dan Hou, et al. A deep survival analysis method based on ranking. Artificial intelligence in medicine, 98:1–9, 2019.

[4] Harald Steck, Balaji Krishnapuram, Cary Dehing-Oberije, Philippe Lambin, and Vikas C Raykar. On ranking in survival analysis: Bounds on the concordance index. Advances in neural information processing systems, 20, 2007.

[5] Patrick J Heagerty and Yingye Zheng. Survival model predictive accuracy and roc curves. Biometrics, 61(1):92–105, 2005.

[6] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907, 2016.

[1]Supplementary materials are available in the supporting documents/multimedia tab. Further instructions on footnote usage are in the Footnotes section on the next page.

[7] Richard J Chen, Ming Y Lu, Jingwen Wang, Drew FK Williamson, Scott J Rodig, Neal I Lindeman, and Faisal Mahmood. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. IEEE Transactions on Medical Imaging, 41(4):757–770, 2020.

[8] Michael R Lamprecht, David M Sabatini, and Anne E Carpenter. Cellprofiler™: free, versatile software for automated biological image analysis. biotechniques, 42(1):71–75, 2007.

[9] Xinliang Zhu, Jiawen Yao, and Junzhou Huang. Deep convolutional neural network for survival analysis with pathological images. In 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pages 544–547. IEEE, 2016.

[10] Xinliang Zhu, Jiawen Yao, Feiyun Zhu, and Junzhou Huang. Wsisa: Making survival prediction from whole slide histopathological images. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7234–7242, 2017.

[11] Donglin Di, Shengrui Li, Jun Zhang, and Yue Gao. Ranking-based survival prediction on histopathological whole-slide images. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 428–438. Springer, 2020.

[12] Jiawen Yao, Xinliang Zhu, Jitendra Jonnagaddala, Nicholas Hawkins, and Junzhou Huang. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. Medical Image Analysis, 65:101789, 2020.

[13] Christian Abbet, Inti Zlobec, Behzad Bozorgtabar, and Jean-Philippe Thiran. Divide-and-rule: self-supervised learning for survival analysis in colorectal cancer. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23, pages 480–489. Springer, 2020.

[14] Jiawen Yao, Xinliang Zhu, and Junzhou Huang. Deep multi-instance learning for survival prediction from whole slide images. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22, pages 496–504. Springer, 2019.

[15] Pooya Mobadersany, Safoora Yousefi, Mohamed Amgad, David A Gutman, Jill S Barnholtz-Sloan, José E Velázquez Vega, Daniel J Brat, and Lee AD Cooper. Predicting cancer outcomes from histology and genomics using convolutional networks. Proceedings of the National Academy of Sciences, 115(13):E2970–E2979, 2018.

[16] Jiawen Yao, Xinliang Zhu, Feiyun Zhu, and Junzhou Huang. Deep correlational learning for survival prediction from multi-modality data. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 406–414. Springer, 2017.

[17] Shidan Wang, Alyssa Chen, Lin Yang, Ling Cai, Yang Xie, Junya Fujimoto, Adi Gazdar, and Guanghua Xiao. Comprehensive analysis of lung cancer pathology images to discover tumor shape and boundary features that predict survival outcome. Scientific reports, 8(1):10393, 2018.

[18] Callum Christopher Mackenzie, Muhammad Dawood, Simon Graham, Mark Eastwood, et al. Neural graph modelling of whole slide images for survival ranking. In Learning on Graphs Conference, pages 48–1. PMLR, 2022.

[19] Ozan Ciga, Tony Xu, and Anne Louise Martel. Self supervised contrastive learning for digital histopathology. Machine Learning with Applications, 7:100198, 2022.

[20] Hakim Benkirane, Maria Vakalopoulou, Stergios Christodoulidis, Ingrid-Judith Garberis, Stefan Michiels, and Paul-Henry Cournède. Hyper-adac: Adaptive clustering-based hypergraph representation of whole slide images for survival analysis. In Machine Learning for Health, pages 405–418. PMLR, 2022.

[21] Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 14318–14328, 2021.

[22] Abtin Riasatian, Morteza Babaie, Danial Maleki, Shivam Kalra, Mojtaba Valipour, Sobhan Hemati, Manit Zaveri, Amir Safarpoor, Sobhan Shafiei, Mehdi Afshari, et al. Fine-tuning and

training of densenet for histopathology image representation using tcga diagnostic slides. Medical Image Analysis, 70:102032, 2021.

[23] Ziyu Guo, Weiqin Zhao, Shujun Wang, and Lequan Yu. Higt: Hierarchical interaction graph-transformer for whole slide image analysis. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 755–764. Springer, 2023.

[24] Tianling Liu, Ran Su, Changming Sun, Xiuting Li, and Leyi Wei. Eocsa: Predicting prognosis of epithelial ovarian cancer with whole slide histopathological images. Expert Systems with Applications, 206:117643, 2022.

[25] S Woo, J Park, JY Lee, and I So Kweon. Cbam: convolutional block attention module. in proceedings of the european conference on computer vision (eccv): 3-19, 2018.

[26] Charlie Saillard, Benoit Schmauch, Oumeima Laifa, Matahi Moarii, Sylvain Toldo, Mikhail Zaslavskiy, Elodie Pronier, Alexis Laurent, Giuliana Amaddeo, Hélène Regnault, et al. Predicting survival after hepatocellular carcinoma resection using deep learning on histological slides. Hepatology, 72(6):2000–2013, 2020.

[27] Ziwang Huang, Hua Chai, Ruoqi Wang, Haitao Wang, Yuedong Yang, and Hejun Wu. Integration of patch features through self-supervised learning and transformer for survival analysis on whole slide images. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24, pages 561–570. Springer, 2021.

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[29] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 9729–9738, 2020.

[30] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.

[31] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[32] Yifan Shen, Li Liu, Zhihao Tang, Zongyi Chen, Guixiang Ma, Jiyan Dong, Xi Zhang, Lin Yang, and Qingfeng Zheng. Explainable survival analysis with convolution-involved vision transformer. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pages 2207–2215, 2022.

[33] Madhusudan Verma. Beyond nystr\" omformer–approximation of self-attention by spectral shifting. arXiv preprint arXiv:2103.05638, 2021.

[34] Huidong Liu and Tahsin Kurc. Deep learning for survival analysis in breast cancer with whole slide image data. Bioinformatics, 38(14):3629–3637, 2022.

[35] Ellery Wulczyn, David F Steiner, Zhaoyang Xu, Apaar Sadhwani, Hongwu Wang, Isabelle Flament-Auvigne, Craig H Mermel, Po-Hsuan Cameron Chen, Yun Liu, and Martin C Stumpe. Deep learning-based survival prediction for multiple cancer types using histopathology images. PloS one, 15(6):e0233678, 2020.

[36] Lei Fan, Arcot Sowmya, Erik Meijering, and Yang Song. Cancer survival prediction from whole slide images with self-supervised learning and slide consistency. IEEE Transactions on Medical Imaging, 2022.

[37] Pei Liu, Bo Fu, Feng Ye, Rui Yang, and Luping Ji. Dsca: A dual-stream network with cross-attention on whole-slide image pyramids for cancer prognosis. Expert Systems with Applications, 227:120280, 2023.

[38] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE/CVF international conference on computer vision, pages 568–578, 2021.

[39] Songhui Diao, Pingjun Chen, Eman Showkatian, Rukhmini Bandyopadhyay, Frank R Rojas, Bo Zhu, Lingzhi Hong, Muhammad Aminu, Maliazurina B Saad, Morteza Salehjahromi, et al. Automated cellular-level dual global fusion of whole-slide imaging for lung adenocarcinoma prognosis. Cancers, 15(19):4824, 2023.

[40] Zhuchen Shao, Yang Chen, Hao Bian, Jian Zhang, Guojun Liu, and Yongbing Zhang. Hvtsurv: hierarchical vision transformer for patient-level survival prediction from whole slide image. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, pages 2209–2217, 2023.

[41] Kan Wu, Houwen Peng, Minghao Chen, Jianlong Fu, and Hongyang Chao. Rethinking and improving relative position encoding for vision transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 10033–10041, 2021.

[42] Yushan Zheng, Jun Li, Jun Shi, Fengying Xie, Jianguo Huai, Ming Cao, and Zhiguo Jiang. Kernel attention transformer for histopathology whole slide image analysis and assistant cancer diagnosis. IEEE Transactions on Medical Imaging, 2023.

[43] Shuai Jiang, Arief A Suriawinata, and Saeed Hassanpour. Mhattnsurv: Multi-head attention for survival prediction using whole-slide pathology images. Computers in Biology and Medicine, 158:106883, 2023.

[44] Yan-Jun Li, Hsin-Hung Chou, Peng-Chan Lin, Meng-Ru Shen, and Sun-Yuan Hsieh. A novel deep learning-based algorithm combining histopathological features with tissue areas to predict colorectal cancer survival from whole-slide images. Journal of Translational Medicine, 21(1):731, 2023.

[45] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. Advances in neural information processing systems, 27, 2014.

[46] Xinliang Zhu, Jiawen Yao, Xin Luo, Guanghua Xiao, Yang Xie, Adi Gazdar, and Junzhou Huang. Lung cancer survival prediction from pathological images and genetic data—an integration study. In 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), pages 1173–1176. IEEE, 2016.

[47] Jie Peng, Pei Wang, Nengfeng Zhou, and Ji Zhu. Partial correlation estimation by joint sparse regression models. Journal of the American Statistical Association, 104(486):735–746, 2009.

[48] Dongdong Sun, Ao Li, Bo Tang, and Minghui Wang. Integrating genomic data and pathological images to effectively predict breast cancer clinical outcome. Computer methods and programs in biomedicine, 161:45–53, 2018.

[49] Robert Tibshirani. The lasso method for variable selection in the cox model. Statistics in medicine, 16(4):385–395, 1997.

[50] Yi Yang and Hui Zou. A cocktail algorithm for solving the elastic net penalized cox's regression in high dimensions. Statistics and its Interface, 6(2):167–173, 2013.

[51] John D Kalbfleisch and Ross L Prentice. The statistical analysis of failure time data. John Wiley & Sons, 2011.

[52] Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. Random survival forests. 2008.

[53] Andreas Mayr and Matthias Schmid. Boosting the concordance index for survival data–a unified framework to derive and evaluate biomarker combinations. PloS one, 9(1):e84483, 2014.

[54] Eric Bair, Trevor Hastie, Debashis Paul, and Robert Tibshirani. Prediction by supervised principal components. Journal of the American Statistical Association, 101(473):119–137, 2006.

[55] Anika Cheerla and Olivier Gevaert. Deep learning with multimodal representation for pancancer prognosis prediction. Bioinformatics, 35(14):i446–i454, 2019.

[56] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway networks. arXiv preprint arXiv:1505.00387, 2015.

[57] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. arXiv preprint arXiv:1602.07360, 2016.

[58] Wei Shao, Jun Cheng, Liang Sun, Zhi Han, Qianjin Feng, Daoqiang Zhang, and Kun Huang. Ordinal multi-modal feature selection for survival analysis of early-stage renal cancer. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 648–656. Springer, 2018.

[59] Hady Ahmady Phoulady, Dmitry B Goldgof, Lawrence O Hall, and Peter R Mouton. Nucleus segmentation in histology images with hierarchical multilevel thresholding. In Medical Imaging

2016: Digital Pathology, volume 9791, pages 280–285. SPIE, 2016.

[60] Wei Shao, Zhi Han, Jun Cheng, Liang Cheng, Tongxin Wang, Liang Sun, Zixiao Lu, Jie Zhang, Daoqiang Zhang, and Kun Huang. Integrative analysis of pathological images and multi-dimensional genomic data for early-stage cancer prognosis. IEEE transactions on medical imaging, 39(1):99–110, 2019.

[61] Jun Cheng, Jie Zhang, Yatong Han, Xusheng Wang, Xiufen Ye, Yuebo Meng, Anil Parwani, Zhi Han, Qianjin Feng, and Kun Huang. Integrative analysis of histopathological images and genomic data predicts clear cell renal cell carcinoma prognosis. Cancer research, 77(21):e91–e100, 2017.

[62] Daniela M Witten and Robert J Tibshirani. Extensions of sparse canonical correlation analysis with applications to genomic data. Statistical applications in genetics and molecular biology, 8(1), 2009.

[63] Jie Hao, Sai Chandra Kosaraju, Nelson Zange Tsaku, Dae Hyun Song, and Mingon Kang. Page-net: interpretable and integrative deep learning for survival analysis using histopathological images and genomic data. In Pacific Symposium on Biocomputing 2020, pages 355–366. World Scientific, 2019.

[64] Jie Hao, Youngsoon Kim, Tejaswini Mallavarapu, Jung Hun Oh, and Mingon Kang. Cox-pasnet: pathway-based sparse deep neural network for survival analysis. In 2018 IEEE international conference on bioinformatics and biomedicine (BIBM), pages 381–386. IEEE, 2018.

[65] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. Advances in neural information processing systems, 30, 2017.

[66] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. arXiv preprint arXiv:1702.01992, 2017.

[67] Ruoqi Wang, Ziwang Huang, Haitao Wang, and Hejun Wu. Ammasurv: asymmetrical multi-modal attention for accurate survival analysis with whole slide images and gene expression data. In 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pages 757–760. IEEE, 2021.

[68] Zhiqin Wang, Ruiqing Li, Minghui Wang, and Ao Li. Gpdbn: deep bilinear network integrating both genomic data and pathological images for breast cancer prognosis prediction. Bioinformatics, 37(18):2963–2970, 2021.

[69] Luís A Vale-Silva and Karl Rohr. Long-term cancer survival prediction using multimodal deep learning. Scientific Reports, 11(1):13505, 2021.

[70] Richard J Chen, Ming Y Lu, Wei-Hung Weng, Tiffany Y Chen, Drew FK Williamson, Trevor Manz, Maha Shady, and Faisal Mahmood. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 4015–4025, 2021.

[71] Zhilong Lv, Yuexiao Lin, Rui Yan, Zhenghe Yang, Ying Wang, and Fa Zhang. Pg-tfnet: transformer-based fusion network integrating pathological images and genomic data for cancer survival analysis. In 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pages 491–496. IEEE, 2021.

[72] Fulong Yu, Fei Quan, Jinyuan Xu, Yan Zhang, Yi Xie, Jingyu Zhang, Yujia Lan, Huating Yuan, Hongyi Zhang, Shujun Cheng, et al. Breast cancer prognosis signature: linking risk stratification to disease subtypes. Briefings in bioinformatics, 20(6):2130–2140, 2019.

[73] Zijian Ding, Songpeng Zu, and Jin Gu. Evaluating the molecule-based prediction of clinical drug responses in cancer. Bioinformatics, 32(19):2891–2895, 2016.

[74] Richard J Chen, Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Jana Lipkova, Zahra Noor, Muhammad Shaban, Maha Shady, Mane Williams, Bumjin Joo, et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. Cancer Cell, 40(8):865–878, 2022.

[75] Yuzhang Xie, Guoshuai Niu, Qian Da, Wentao Dai, and Yang Yang. Survival prediction for gastric cancer via multimodal learning of whole slide images and gene expression. In 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pages 1311–1316. IEEE, 2022.

[76] Lin Qiu, Aminollah Khormali, and Kai Liu. Deep biological pathway informed pathology-genomic multimodal survival prediction. arXiv preprint arXiv:2301.02383, 2023.

[77] Wentai Hou, Chengxuan Lin, Lequan Yu, Jing Qin, Rongshan Yu, and Liansheng Wang. Hybrid graph convolutional network with online masked autoencoder for robust multimodal cancer survival prediction. IEEE Transactions on Medical Imaging, 2023.

[78] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 16000–16009, 2022.

[79] Luan Tran, Xiaoming Liu, Jiayu Zhou, and Rong Jin. Missing modalities imputation via cascaded residual autoencoder. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1405–1414, 2017.

[80] Yanbei Liu, Lianxi Fan, Changqing Zhang, Tao Zhou, Zhitao Xiao, Lei Geng, and Dinggang Shen. Incomplete multi-modal representation learning for alzheimer's disease diagnosis. Medical Image Analysis, 69:101953, 2021.

[81] Ting Wei, Xin Yuan, Ruitian Gao, Luke Johnston, Jie Zhou, Yifan Wang, Weiming Kong, Yujing Xie, Yue Zhang, Dakang Xu, et al. Survival prediction of stomach cancer using expression data and deep learning models with histopathological images. Cancer Science, 114(2):690, 2023.

[82] Joshua Levy, Christian Haudenschild, Clark Barwick, Brock Christensen, and Louis Vaickus. Topological feature extraction and visualization of whole slide images using graph neural networks. In BIOCOMPUTING 2021: Proceedings of the Pacific Symposium, pages 285–296. World Scientific, 2020.

[83] Ruoyu Li, Jiawen Yao, Xinliang Zhu, Yeqing Li, and Junzhou Huang. Graph cnn for survival analysis on whole slide pathological images. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 174–182. Springer, 2018.

[84] Richard J Chen, Ming Y Lu, Muhammad Shaban, Chengkuan Chen, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. Whole slide images are 2d point clouds: Context-aware survival prediction using patch-based graph convolutional networks. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24, pages 339–349. Springer, 2021.

[85] Guohao Li, Matthias Muller, Ali Thabet, and Bernard Ghanem. Deepgcns: Can gcns go as deep as cnns? In Proceedings of the IEEE/CVF international conference on computer vision, pages 9267–9276, 2019.

[86] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al. Graph attention networks. stat, 1050(20):10–48550, 2017.

[87] Yongju Lee, Jeong Hwan Park, Sohee Oh, Kyoungseob Shin, Jiyu Sun, Minsun Jung, Cheol Lee, Hyojin Kim, Jin-Haeng Chung, Kyung Chul Moon, et al. Derivation of prognostic contextual histopathological features from whole-slide images of tumours via graph deep learning. Nature Biomedical Engineering, pages 1–15, 2022.

[88] Natalie Stanley, Roland Kwitt, Marc Niethammer, and Peter J Mucha. Compressing networks with super nodes. Scientific reports, 8(1):10892, 2018.

[89] Donglin Di, Changqing Zou, Yifan Feng, Haiyan Zhou, Rongrong Ji, Qionghai Dai, and Yue Gao. Generating hypergraph-based high-order representations of whole-slide histopathological images for survival prediction. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(5):5800–5815, 2022.

[90] Fei Wu, Pei Liu, Bo Fu, and Feng Ye. Deepgcnmil: Multi-head attention guided multi-instance learning approach for whole-slide images survival analysis using graph convolutional networks. In 2022 14th International Conference on Machine Learning and Computing (ICMLC), pages 67–73, 2022.

[91] Donglin Di, Jun Zhang, Fuqiang Lei, Qi Tian, and Yue Gao. Big-hypergraph factorization neural network for survival prediction from whole slide image. IEEE Transactions on Image Processing, 31:1149–1160, 2022.

[92] Pei Liu, Luping Ji, Feng Ye, and Bo Fu. Graphlsurv: A scalable survival prediction network with adaptive and sparse structure learning for histopathological whole-slide images. Computer Methods and Programs in Biomedicine, 231:107433, 2023.

[93] Wentai Hou, Lequan Yu, Chengxuan Lin, Helong Huang, Rongshan Yu, Jing Qin, and Liansheng Wang. H^ 2-mil: Exploring hierarchical representation with heterogeneous multiple instance learning for whole slide image analysis. In Proceedings of the AAAI conference on artificial intelligence, volume 36, pages 933–941, 2022.

[94] Wenqi Lu, Michael Toss, Muhammad Dawood, Emad Rakha, Nasir Rajpoot, and Fayyaz Minhas. Slidegraph+: Whole slide image level graphs to predict her2 status in breast cancer. Medical Image Analysis, 80:102486, 2022.