Computer Methods and Programs in Biomedicine (2025)

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

Deep Learning Based Methods in Whole Slide Image Survival Prediction: A Survey

Jihao Li[a,1], Huhan Xie[a,b,1], Tong Xu[a,b], Xuewu Jiang[a], Tianzhao Zhong[a], Huaishui Yang[a], Mengye Lyu[a], Shaojun Liu[a,*]

[a]College of Health Science and Environmental Engineering, Shenzhen Technology University, Shenzhen 518118, China
[b]School of Applied Technology, Shenzhen Univerity, Shenzhen 518060, China

ARTICLE INFO

Article history:
Received 25th January 2025

Keywords:
Survival prediction
survival analysis
deep learning
whole slide image

ABSTRACT

The usage of gigapixel Whole-Slide Images (WSIs) in digital pathology is growing rapidly, and researchers are now exploring their potential in predicting survival rates among cancer patients. In this survey, we provide a comprehensive analysis of recent research on survival prediction using WSIs. We start by discussing the development and applications of WSI technology and the importance of survival prediction in cancer treatment. Next, we delve into the current methods of survival prediction in WSIs, especially focusing on deep learning techniques. We find that all the methods follow a similar workflow. It is comprised of three procedures: image feature extraction, image feature aggregation, as well as survival analysis and evaluation. For each procedure, we provide a review and discussion. In addition, we also notice and summarize the recent development in integrating multi-modality with WSIs for survival prediction. Moreover, in order to provide a comprehensive understanding of existing methods, we conduct experiments on five widely used datasets for ten state-of-the-art methods with the same experiment settings to ensure fairness. In the end, our survey discusses the challenges and future directions for WSI-based survival prediction. We believe our review provides a comprehensive summary that would serve as a valuable guide for future research on WSI survival prediction.

1. Introduction

Pathological analysis has been conventionally performed by human pathologists with a microscope to examine stained specimens on a slide. At the same time, there has been a growing trend towards capturing the complete slide with scanners and saving it as a digital Whole Slide Image (WSI) for future research and permanent recording [1]. With the advent of artificial intelligence and machine learning algorithms, computer-aided pathological analysis has become a significant research area. Consequently, WSI has emerged as a promising tool for developing and utilizing such algorithms [2].

Traditionally, pathological assessments are highly dependent on manual inspection and subjective interpretation of slides under a microscope. Although this method has been widely used for disease diagnosis and prognosis and is considered the best approach [3], it has some limitations. These limitations include the subjective divergence among pathologists and the huge labor of inspection. Moreover, the tissue sections usually contain extensive information beyond human perception [4], making it difficult to fully utilize the vast amount of data contained in tissue structures.

Fortunately, the emergence of WSI and the development of artificial intelligence, especially the advancement of computer-aided diagnosis, have brought about a new era in tissue analysis. The advent of WSI has provided researchers and clinicians with access to vast image databases, which present new opportunities for utilizing prognostic feature extraction, as well as ma-

*Corresponding author: Shaojun Liu.
e-mail: liusj14@tsinghua.org.cn (Shaojun Liu)
[1]These two authors contributed equally to this work.

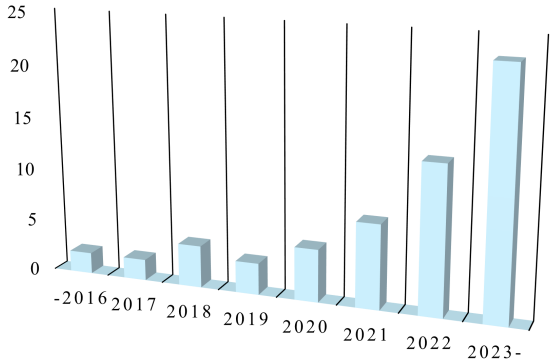

Fig. 1: A statistical chart of the publication years of all surveyed literature.

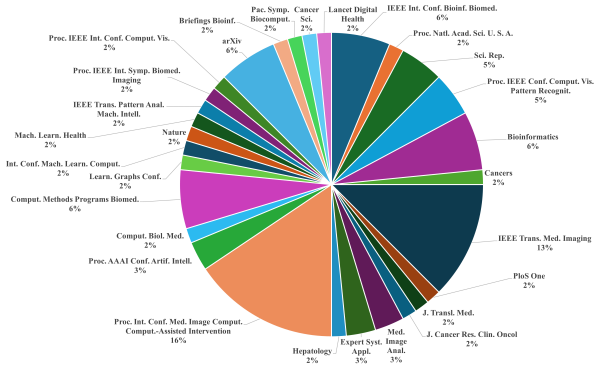

Fig. 2: A statistical chart of the publication sources of all surveyed literature.

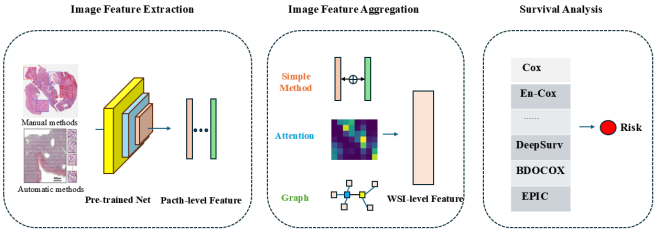

Fig. 3: General framework for survival prediction using WSIs. Typically, WSI is first partitioned into small patches. Each patch is then represented by a relatively low-dimension feature compared with the original RGB patch, compressing the data amount and integrating the information. Next, the patch features are aggregated to produce a comprehensive WSI-level feature. Finally, the survival prediction is made on the WSI-level feature.

chine and deep learning techniques to model and predict patient outcomes. The study of tissue morphology, cellular structures, and patterns within WSI can provide valuable insights into disease progression and prognosis. Recent advancements in artificial intelligence and the availability of vast WSI data samples have led to the development of deep learning algorithms for analyzing WSI. These algorithms can perform a wide range of tasks, including classification [5, 6, 7, 8, 9], segmentation [10, 11, 12], pathological grading [13, 14], and other computer-aided diagnosis tasks [15, 16].

Compared with cancer diagnosis, less attention is paid to using WSIs for predicting treatment response or survival rate [17]. The process of diagnosing cancer involves identifying or categorizing the disease on a particular slide. However, predicting a patient's prognosis is a more intricate task that depends on multiple internal factors, like the patient's antitumor immunity and physical well-being, as well as external factors, such as treatment. Despite the aforementioned obstacles, precise survival prediction is crucial in the medical field because it enables doctors to gain a better understanding of the patient's condition and plan treatment accordingly, ultimately improving their chances of survival and extending their lifespan. In essence, achieving a precise prognosis is essential to optimize treatment strategies and improve patient care.

On one hand, accurate survival prediction models can play a key role in guiding personalized treatment decisions. By stratifying patients according to predicted survival outcomes, clinicians could tailor interventions more effectively, opting for aggressive therapies, such as chemotherapy or surgery, for patients with poor prognoses while recommending conservative treatment for those with better outcomes, minimizing unnecessary risks and side effects. On the other hand, survival prediction tools could help identify biological subgroups within patient populations, offering deeper insight into the heterogeneity of certain types of cancer. This, in turn, can facilitate the development of targeted therapies or biomarker-driven treatment strategies. For example, computational survival models could reveal molecular or histological patterns associated with survival disparities, providing a basis for investigating the underlying biology of these subgroups.

It is widely accepted that WSI captures tissue sample heterogeneity and can provide rich information for disease modeling and treatment planning [18]. Through the analysis of WSI, researchers, and physicians can extract rich image features and use advanced machine learning and deep learning technologies to model and predict patient survival rates. The capability to extract abundant image features from digital pathology slide images—features provides an opportunity for more robust quantitative modeling of disease appearance and would potentially improve the predictions of disease aggressiveness and patient outcomes [19]. In addition, survival prediction using WSI can help clinical physicians uncover subtle patterns, bio-markers, and prognostic factors that may have been ignored in traditional pathology assessments. By enabling the high-resolution digitization of entire tissue slides, WSI has enhanced the efficiency and accuracy of pathological assessments and opened new avenues for predictive modeling and survival analysis.

There have been a few reviews on deep learning based pathological image analysis [20, 21, 22]. However, they either only summarize the applications of deep learning technology in pathological image analysis or focus on specific tasks such as classification and segmentation. To our knowledge, there have already been quite a few works on WSI-based survival prediction, but with no summary or comprehensive evaluation. To fill this gap, we comprehensively review WSI-based survival prediction over the past few decades and evaluate the state-of-the-art (SOTA) algorithms on several widely used datasets under the same experiment settings for fair comparison.

In this survey, we reviewed approximately 150 relevant papers, among which 64 focus specifically on WSI-based survival analysis. The initial literature search was conducted using

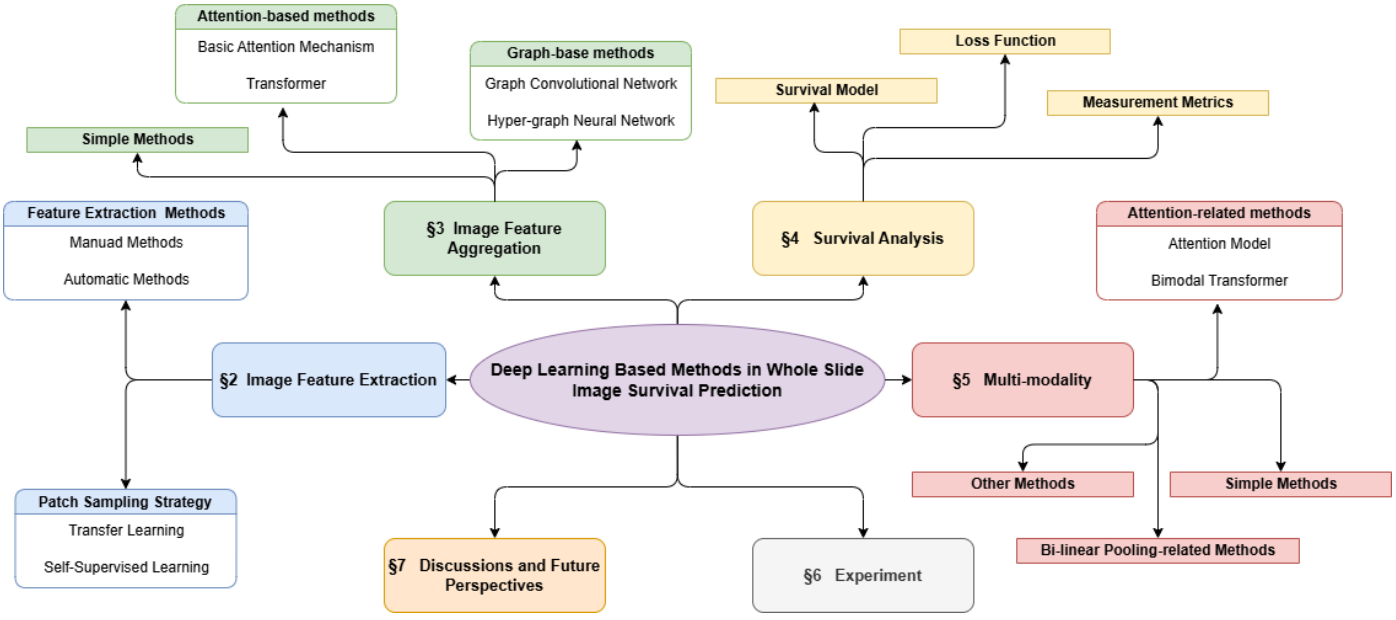

Fig. 4: Structure organization of the survey.

the search mode ("whole slide image" OR "pathological image") AND ("survival analysis" OR "survival prediction") on comprehensive literature search engines, i.e., Google Scholar, PubMed, and arXiv, with the search scope further expanded by citation tracing. Our analysis highlights that many of these methods are deeply influenced by advancements in computer vision (CV) and machine learning (ML), which have played a pivotal role in shaping modern survival analysis techniques. A significant number of these works have been published in prestigious journals and conferences, including Nature, the Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR) and the Proceedings of International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI). The histogram of publication year for all methods is summarized in Fig. 1, while detailed statistics regarding publication source are presented in Fig. 2. Among the reviewed papers, 60 were officially published in journals and conferences, and 4 are preprints.

Through the literature review, we find that there is a common framework for WSI-based survival prediction methods, mainly containing three procedures, as shown in Fig. 3.

The first stage is feature extraction, which includes local patch sampling and feature extraction. However, these densely or randomly sampled patches lack context awareness and lose the interaction between cell and tissue types, which is a predictive prognosis for patient survival. Therefore, in the second stage, many methods for feature aggregation are proposed using simple methods, attention models, or graph models. Then, in the third stage, the final survival analysis is conducted based on the aggregated features. These 64 methods are summarized according to these three stages and publication source in detail in Table 1. Additionally, we have noticed that there is a trend incorporating multi-modality data other than WSI for more comprehensive prediction. As the data is inherently heterogeneous for different modalities, the feature fusion for multi-modality data should be carefully considered. Therefore, we also provide the details about multi-modality feature fusion in the summary table 1 for clarification purposes.

In conclusion, the structure of this survey report is organized as shown in the Fig. 4, with the titles summarizing the main content of each chapter. We first go through the three procedures in Sections 2-4, summarizing existing methods and technologies. Then, the multi-modality is investigated in Section 5. Next, experiments are conducted on five widely used datasets for ten SOTA methods in Section 6. Then, the challenges, and future prospects of WSI-based survival prediction are discussed in Section 7. Finally, we conclude the survey in Section 8.

2. Image Feature Extraction

This section reviews the patch sampling strategies and feature extraction methods. This is the first procedure of the entire prediction framework. However, due to the lack of pixel-wise annotation in WSI, this procedure is also tricky.

2.1. Patch Sampling Strategy

Patch sampling is the first pre-processing step for feature extraction. Traditional models usually pre-select critical patches manually, or randomly sample subsets from the region of interest (ROI) as the input data. In contrast, newly proposed models utilize all patches in the gigabit pixel pathological images or adopt some filtering strategies to get vital patches. According to the strategy, patch sampling technologies can be divided into two subgroups: manual methods and automatic methods.

Table 1: Representative works of survival prediction based on WSIs. The image-only methods are summarized according to the three procedures, while the multi-modality methods are summarized according to image feature extractor, multi-modality feature fusion, and survival analysis since the image feature is fused with other modalities simultaneously. In addition, the publication source is also provided.

Method	Image Feature Extraction	Image Feature Aggregation	Multi-modality Feature Fusion	Survival Analysis Loss	Publication Source
DeepConvSurv [23]	manual method	Simple Method	———	NLPL	IEEE Int. Conf. Bioinf. Biomed.
SCNN [24]	manual method	Simple Method	———	NLPL	Proc. Natl. Acad. Sci. U. S. A.
Comprehensive analysis [25]	manual method	Simple Method	———	NLPL	Sci. Rep.
WSISA [26]	automatic method CellProfiler	Simple Method	———	NLPL + Regularization Loss	Proc. IEEE Conf. Comput. Vis. Pattern Recognit.
CapSurv [27]	automatic method VGG16	Simple Method	———	Survival Capsule Loss	IEEE Access
Deep survival analysis in breast cancer [28]	automatic method CNN	Simple Method	———	censored loss + the uncensored loss.	Bioinformatics
Dual Global Fusion [29]	automatic method HoVer-Net	Simple Method	———	NLPL	Cancers
BDOCOX [30]	automatic method Resnet128	Simple Method	———	NLPL+Ranking Loss	IEEE Trans. Med. Imaging
Deep learning-based survival prediction [31]	automatic method CNN	Simple Method	———	NLPL	PloS One
WSI-HSfeatures [32]	automatic method ResNet50	Simple Method	———	NLPL	J. Transl. Med.
Cancer Survival Prediction [33]	automatic method SSL (colorization and cross-channel)	Simple Method	———	consistency loss + contrastive loss	IEEE Trans. Med. Imaging
Deep Survival Analysis [34]	automatic method ResNet34	Simple Method	———	NLPL	Sci. Rep.
WDRNet [35]	automatic method CNN	Simple Method	———	NLPL	J. Cancer Res. Clin. Oncol
DeepAttnMISL [36]	automatic method VGG	Attention	———	NLPL + Regularization Loss	Med. Image Anal.
attMIL [37]	automatic method ResNet50	Attention	———	NLPL + Regularization Loss	Lancet Digital Health
EOCSA [38]	automatic method CellProfiler	Attention	———	NLPL + Regularization Loss	Expert Syst. Appl.
SCHMOWDER [39]	automatic method ResNet	Attention	———	NLPL	Hepatology

(Continued on the next page...)

Table 1 (Continued)

Method	Image Feature Extraction	Image Feature Aggregation	Multi-modality Feature Fusion	Survival Analysis Loss	Publication Source
SeTranSurv [40]	automatic method SSL (SimCLR)	Transformer	———	NLPL	Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention
ESAT [41]	automatic method ResNet34	Transformer	———	NLPL+ Regularization Loss	Proc. AAAI Conf. Artif. Intell.
AdvMIL[42]	automatic method ResNet34	GNN	———	NLPL	Med. Image Anal.
DSCA [43]	automatic method ResNet50	Transformer	———	censored cross-entropy loss	Expert Syst. Appl.
HVTSurv [44]	automatic method ResNet50	Transformer	———	censored cross-entropy loss	Proc. AAAI Conf. Artif. Intell.
MHAttnSurv [45]	automatic method ResNet18	Transformer	———	NLPL	Comput. Biol. Med.
Surformer [46]	automatic method ResNet50	Transformer	———	NLPL	Comput. Methods Programs Biomed.
GNNSurvivalRankLoG [47]	automatic method Shuffle Net	GNN	———	Pairwise Ranking Loss	Learn. Graphs Conf.
DeepGraphSurv [48]	automatic method VGG16	GNN	———	NLPL	Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention
Patch-GCN [49]	automatic method ResNet50	GNN	———	NLPL	Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention
HistoFL [49]	automatic method ResNet50	Attention	———	NLPL	Med. Image Anal.
DeepGCNMIL [50]	automatic method ResNet50	GNN	———	NLPL	Int. Conf. Mach. Learn. Comput.
SlideGraph+ [51]	automatic method ResNet50	GNN	———	NLPL	Med. Image Anal.
GraphLSurv [52]	automatic method ResNet50	GNN	———	NLPL+Dirichlet energy Loss	Comput. Methods Programs Biomed.
TEA Graph [53]	automatic method EfficientNet (pre-trained on ImageNet)	GNN	———	NLPL	Nature
CoADS [54]	automatic method ResNet50	GNN	———	NLPL	Comput. Methods Programs Biomed.

(Continued on the next page...)

Table 1 (Continued)

Method	Image Feature Extraction	Image Feature Aggregation	Multi-modality Feature Fusion	Survival Analysis Loss	Publication Source
Hyper-adac [55]	automatic method SSL (SimCLR)	HGNN	———	NLPL	Mach. Learn. Health
HGSurvNet [56]	automatic method ResNet	HGNN	———	MSE loss + NLPL + BCR loss	IEEE Trans. Pattern Anal. Mach. Intell.
RankSurv [57]	automatic method ResNet34	HGNN	———	BCR loss	Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention
b-HGFN [58]	automatic method VGG16	HGNN	———	BCRLoss + NDCGLoss2	IEEE Trans. Med. Imaging
HIPT [59]	automatic method ResNet50	Transformer	———	survival cross-entropy loss	Proc. IEEE Conf. Comput. Vis. Pattern Recognit.
HGT [60]	automatic method ResNet50	Transformer	———	NLPL	Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention
Lung Cancer Survival Prediction [61]	automatic method (minimum-model)	———	Sparsity analysis	NLPL	Proc. IEEE Int. Symp. Biomed. Imaging
DeepCorrSurv [62]	automatic method CellProfiler	———	Simple Fusion	NLPL	Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention
PathOmics [63]	automatic method ResNet50	———	Simple Fusion	NLPL	Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention
GPMKL [64]	automatic method CellProfiler	———	other method	NLPL	Comput. Methods Programs Biomed.
OSCCA [65]	automatic method	———	other method	NLPL	Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention
AMMASurv [66]	automatic method ResNet18	———	Bimodal Transformer	NLPL	IEEE Int. Conf. Bioinf. Biomed.
OMMFS [67]	automatic method	———	other method	NLPL	IEEE Trans. Med. Imaging
Gc-Splem [68]	automatic method ResNet50	———	Attention Model	NLPL	IEEE Int. Conf. Bioinf. Biomed.
Pathomic Fusion [69]	automatic method VGG19	———	Attention Model	NLPL	IEEE Trans. Med. Imaging

(Continued on the next page...)

Table 1 (Continued)

Method	Image Feature Extraction	Image Feature Aggregation	Multi-modality Feature Fusion	Survival Analysis Loss	Publication Source
GPDBN [70]	automatic method CellProfiler	———	Bi-linear pooling	censored cross-entropy loss	Bioinformatics
MCAT [71]	automatic method ResNet50	———	Attention Model	NLPL	Proc. IEEE Int. Conf. Comput. Vis.
MotCat [72]	automatic method ResNet34	———	Attention Model	NLPL	ICCV
HFBSurv [73]	automatic method CellProfiler	———	Bi-linear pooling	NLPL + Regularization Loss	Bioinformatics
PORPOISE [74]	automatic method Hover-Net	———	Bi-linear pooling	NLPL	Cancer Cell
HGCN [75]	automatic method KimiaNet	———	Bimodal Transformer	NLPL	IEEE Trans. Med. Imaging
ponet [76]	automatic method SSL (pre-trained ViT)	———	Bi-linear pooling	NLPL + Regularization Loss	arXiv
Ada-RSIS [77]	automatic method	———	other method	NLPL	Briefings Bioinf.
MultimodalPrognosis [78]	automatic method SqueezeNet	———	Simple Method	NLPL	Bioinformatics
PAGE-Net [79]	automatic method CNN	———	Simple Method	NLPL	Pac. Symp. Biocomput.
MultiSurv [80]	automatic method Resnet	———	Simple Method	NLPL	Sci. Rep.
PG-TFNet [81]	manual method Resnet	———	Bimodal Transformer	NLPL + Regularization Loss	IEEE Int. Conf. Bioinf. Biomed.
MultiDeepCox-SC [82]	automatic method CellProfiler	———	Simple Method	NLPL	Cancer Sci.
MOME [83]	automatic method Resnet	———	Attention Model	NLPL	Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention
SURVPATH [84]	automatic method SSL	———	Attention Model	NLPL	Proc. IEEE Conf. Comput. Vis. Pattern Recognit.
TITAN [85]	automatic method ResNet	———	Attention Model	NLPL	arXiv
Graph Attention-Based Fusion[86]	automatic method SSL	———	Attention Model	NLPL	IEEE Trans. Med. Imaging
HMCAT[87]	manual method VGG	———	Attention Model	NLPL	IEEE Trans. Med. Imaging

2.1.1. Manual Methods

Due to computational limitations, traditional methods typically make survival predictions on manually annotated ROIs. Assuming some WSIs contain cancer tissue, pathologists need to locate and annotate all cancer cells, which makes the annotations difficult and time-consuming. The model in [62] uses a core sample set from the UT MD Anderson Cancer Center. It utilizes manually annotated ROIs to locate potential tumor areas in pathological images, laying the foundation for subsequent steps. SCNN [24] and DeepAttnMISL [36] are trained by first annotating the ROI and selecting patches from it. However, these methods only collect a small portion of patches from the WSI. Most importantly, the artificial annotations demand a heavy economic burden and selection bias.

Alternatively, patches can also be randomly sampled from WSIs, as investigated in [23], [26], [57], [36], [88], and [89]. Unfortunately, such methods may miss key ROIs with a relatively high probability, leading to severe category imbalance and less representativeness. Moreover, random sampling techniques might result in sampling from vacant areas, thereby introducing extra noise to the model.

2.1.2. Automatic Methods

To tackle the disadvantages of the extra noise introduced by the blank areas, *i.e.*, background areas, MSFN [90] proposes a new patch sampling strategy based on image entropy. This strategy automatically and efficiently extracts patches with strong discriminative characteristics by pruning pathologists' manually annotated ROIs and calculating the image information entropy for each patch. Patches characterized by low image entropy usually contain too large blank regions, lacking predictive value for the model, and potentially introducing additional noise. Conversely, patches with high image entropy are rich in content, providing ample information and exhibiting high predictive value. Therefore, image information entropy is used as the standard to select representative sampling patches.

In WSI, automatic methods like Otsu or hysteresis are used to segment pathological tissues, which can identify tissue locations and reduce the number of patches. The model proposed in [25] can identify tumor regions as ROIs for hematoxylin and eosin (H&E) stained pathology images using the predicted likelihood of image patches, where each patch is tagged as the category with the highest probability. In this way, some tumor-related features can be extracted as the descriptors of the ROIs, including area, perimeter, convex area, filled area, major axis length, minor axis length, and so on. As a supplementary method, researchers use high-throughput cell image analysis tools like CellProfiler [91] to automatically extract predefined cell features such as size, shape, intensity, and texture [92]. However, these handcrafted features are often inherently limited and relatively redundant [93], therefore, may not contain precise and enough prognostic information.

The operations employed in the literature, including contrast normalization, morphological operations, and specific patch scoring systems, play a crucial role in the reduction of the candidate patch pool and can contribute to automated patch localization. These processes aid in enhancing the efficiency of downstream tasks by focusing on relevant patches. However, verifying whether each region has the same label as the WSI typically requires professional knowledge, and the preprocessing involves selecting the optimal algorithm, requiring a certain degree of human intuition. Li *et al.* [32] train a Resnet50 to distinguish cell patch from stroma patch. This can alleviate the demand for patch quality assessment, but unfortunately, the accuracy is also controversial. Furthermore, borders between tissue types are often ambiguous, leading to inconsistencies among pathologists. The high variability of tissue morphology makes it difficult to cover all possible examples during annotation. These drawbacks make the aforementioned automatic methods strongly biased by expert-defined annotations and make them difficult to learn comprehensively.

In order to avoid the annotation burden and selection bias of experts, recent researchers have applied weakly supervised methods, such as the multiple instance learning (MIL) method which can train deep learning models to explore the relationships within WSIs, making the patch sampling fully automatic without any handcrafted rules. For MIL tasks, the samples are organized as bags, each containing a set of instances but with only a single label. Then the training aims to supervise the model to predict the labels for each bag. This is quite suitable for WSI analysis and the applications can be divided into two categories [7].

The first category involves an instance-level algorithm, where a CNN is initially trained by assigning pseudo labels to individual instances according to bag-level labels. Subsequently, the algorithm selects the first k instances with the highest cancer probabilities for aggregation. Nevertheless, this approach demands a substantial quantity of WSI since only a limited number of instances in each slide can actively contribute to the training process. The second category is the embedding level algorithm. This algorithm maps each patch in the entire slide to a fixed-length embedding and then aggregates all feature embeddings through operators such as maximum pooling. The final result is then reported to obtain the ultimate slide-level diagnosis.

Recently, CLAM (clustering-constrained-attention multiple-instance learning) [94] is proposed, according to the aforementioned second MIL sampling strategy. It has been widely adopted in many other researches since then.

2.2. Feature Extraction Methods

Feature extraction can be primarily categorized into transfer learning and self-supervised learning. Transfer learning utilizes pre-trained models, which are trained with fully supervised classification tasks on either natural images or pathological images. Self-supervised learning (SSL), on the other hand, involves pre-training models with unlabeled data using pretext tasks such as image reconstruction or coloring. These approaches offer effective means of feature extraction for WSI analysis.

2.2.1. Transfer Learning

Using too little training data to perform supervised learning often leads to the over-fitting problem with poor generalization

performance. This is especially notable for deep learning, as the capacity of deep learning models is usually much larger than traditional models. In this case, learning usually does not start from scratch but starts with a model pre-trained on another similar and relevant task. This learning method is called transfer learning, and it has achieved promising results in the field of WSI analysis [95]. Typically, the final fully connected layer in the pre-trained network is replaced by a new sub-network suitable for the target task. It can greatly alleviate the over-fitting problem for WSI analysis due to the lack of well-annotated WSIs [96].

A common strategy is to extract features through models pre-trained on natural images, especially the ImageNet dataset. For example, ResNet employed in DSCA [43], EfficientNet in DTMIL [97], and many other models in [98, 99] are all pre-trained on ImageNet to serve as the feature extractors.

Although pre-trained models using general images have shown some success in analyzing tissue pathology images, it is important to recognize the inherent differences between microscopic pathology images and natural images. These differences include variations in microscopic structures, semantic features, and resistance to transformations such as rotation and scaling. Therefore, if pre-trained on datasets with a significant number of pathological images, the models have the potential to outperform those pre-trained on ImageNet. Diao *et al.* [29] utilize a HoVer Net [100] pre-trained on the PanNuke dataset [101], an open pan-cancer nuclei segmentation and detection dataset. Li *et al.* [32] use the NCT-CRC-HE-100K dataset to train a patch-based ResNet-50 tool for slide image analysis. Multiple works have demonstrated the effectiveness of directly using pre-trained models for survival prediction on the TCGA dataset without the need for fine-tuning [49, 43].

2.2.2. *Self-Supervised Learning*

In the aforementioned studies, patch-level feature extraction typically relies on either manually crafted features or pre-trained cellular neural networks on ImageNet without any fine-tuning. These pre-trained CNN models are not tailored for diverse WSI and may result in sub-optimal performance.Recently, SSL techniques have shown remarkable results on several standard computer vision tasks, providing a new strategy for pre-training neural networks with unlabeled data [102].

Usually, SSL methods use pretext tasks to create surrogate supervision for training, where inputs and labels come from raw data without human annotations. In survival prediction from WSIs, designing and training CNN-based feature extractors can be treated as the pretext task, and survival prediction is the downstream task. KimiaNet [103], pre-trained via image retrieval task on a collection of the variable multi-organ open image repository The Cancer Genome Atlas Program (TCGA), is among the most welcome pre-trained models for histological image analysis [104, 6]. Muhammad *et al.* [105] use image reconstruction as a pretext task to train convolutional networks and apply it to survival prediction based on WSI. The work in [106] and Divide-and-Rule [88] introduce the coloring task as a pretext task.

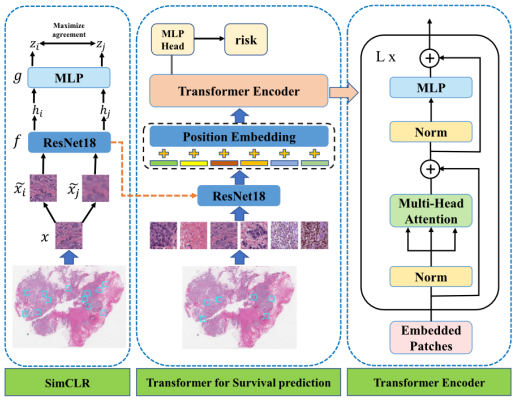

Fig. 5: Framework of SeTranSurv. The left SimCLR is for training a feature extraction model. The middle part is the Transformer encoder. The right part is the detailed process of the encoder block. The image is directly from [40].

Apart from pretext tasks, the method named SimCLR in [107] is one of the SSL methods using contrastive learning. Despite utilizing unlabeled digital histopathology datasets, this method possesses outstanding feature extraction capability comparable to a fully supervised model. This promising capability is achieved by huge training data. SeTranSurv, the model proposed in [40] applies SimCLR to train the initial feature extraction model ResNet-18 to get a specialized model for survival prediction, as shown in Fig. 5. The framework comprises the following major processes. A data augmentation module randomly transforms the original image x into two differently augmented images $\widetilde{x}_i$ and $\widetilde{x}_j$. Then, the augmented images undergo a ResNet18 encoder f to extract representative features. The consistency is maximized using the contrastive loss [108] over the features, *i.e.*, distinguishing $\widetilde{x}_j$ from $\{\widetilde{x}_k\}_{k\neq i}$ for a given $\widetilde{x}_i$. This strategy has also been employed in the image feature extraction of these works [55, 6, 33].

Instead of relying solely on single image data, current research methods such as cTransPath[109], MI-Zero[110], and PLIP[111] adopt a cross-modal contrastive learning strategy by pairing images with textual descriptions. These approaches effectively extract powerful visual and linguistic features by jointly modeling pathological images and their corresponding text descriptions. After pretraining on large-scale datasets, these models demonstrate outstanding performance and exhibit remarkable cross-modal adaptability.

2.3. *Summary*

The choice between manual and automatic methods depends on the available resources as well as the balance between detailed annotation and scalability. Manual methods offer intense supervision, while they are resource-intensive and require annotation biases. Automatic methods provide scalability, while they may face challenges in handling ambiguous tissue boundaries and extracting effective features. Combining the strengths of both approaches could be a promising direction for improving the efficiency and accuracy of survival prediction models based on WSIs.

As for the feature extraction ability, models pre-trained on diverse pathological images, like HoVer Net [100] and KimiaNet

[103], showcases potential improvements over those pre-trained on ImageNet. Transfer learning offers practicality and robustness through leveraging existing knowledge, while SSL methods exhibit a capacity for learning informative representations directly from unlabeled data. In other words, SSL methods have emerged as a promising alternative, particularly in scenarios where patch-level annotations are unavailable. The ability of these methods to extract meaningful features without relying on labeled data proves its feasibility for survival prediction from WSIs.

The choice between these methods may depend on factors such as the availability of corresponding training data and the specific requirements of the survival prediction task. Certainly, these methods can also be integrated. Specifically, both the labeled and unlabeled data could be utilized to obtain better results in a two-stage framework. The first stage involves pretraining the model on a large amount of generic unlabeled data, allowing the model to learn general features that are applicable across different domains. In the second stage, the model can be fine-tuned on a small amount of labeled data from the specific domain in a downstream task, improving its performance on the specific task. Overall, the choice and combination of these feature extraction strategies contribute to advancements in the field of WSI-based survival prediction, addressing the challenges related to data scarcity, cost, and accessibility.

3. Image Feature Aggregation

DeepConvSurv [23] is the first method using CNN for WSI survival analysis, focusing on small patches. This is an end-to-end survival prediction method, combining CNNs with traditional Cox models. WSI contains a vast number of local patches, but individual patches can not provide enough information to predict the overall survival risk of a patient accurately. Therefore, aggregation of information is necessary. In this section, we will discuss different methods of feature aggregation, including simple methods, Attention-based methods, Graph-based methods. Although some methods use a combination of these techniques, we will focus on their primary framework for clear classification and analysis purposes.

3.1. Simple methods

In image feature aggregation, a widely used method is to aggregate feature vectors from different sources like different magnifications. This can be achieved through simple operations such as concatenating and weighted summation.

WSISA [26] is the first method developed for survival prediction based on the whole WSIs. To utilize the discriminative survival information of the whole WSI, WSISA first extracts hundreds of patches from each WSI through adaptive sampling and then clusters them into different categories. Then, WSISA suggests training an aggregation model for patient-level prediction, using cluster-level DeepConvSurv model [23] to predict survival risks. After that, it selects clusters with better performance than random conditions and calculates the contribution of each cluster based on its number of patches. Through a simple weighted sum, it gets the aggregated features. Li *et al.* [32] combine the features of different regions via concatenation to predict patient-level risks. Another example is [31], where the patch features are aggregated via average pooling. In the study of [28], the authors fuse RGB color images with nuclear, tumor, and lymphocyte (NTL) data and analyze them using composite images from six channels. As for feature aggregation, they simply concatenate the features to form a feature vector.

Another study [35] proposes a weakly supervised learning-based dual resolution deep learning network termed as WDR-Net. The network extracts features from coarsely labeled patches using MIL and introduces a dual stream network structure to fully utilize global information. It simultaneously inputs patches at two magnifications into the feature extractor to generate respective features. These features are then concatenated and passed through a fully connected (FC) layer to predict the likelihood of cancer for each patch pair. A similar dual stream network strategy is also proposed by Diao *et al.* [29] based on DeepMixer [112]. In the model, one module focuses on extracting global features from WSI maps and the other module focuses on extracting contextual features between different regions cropped from embedded maps. Then, the model concatenates these two types of features to predict the patient's prognosis.

Some methods use fully connected Layer instead of simple calculations to aggregate features. Tang *et al.* [27] propose CapSurv, a variant of the capsule network [113] specifically designed for feature aggregation and WSI-based survival analysis. The model utilizes the VGG-16 network to extract discriminative patch features related to cancer tissue from WSIs and then uses K-means clustering. Similarly, it trains a DeepConvSurv model [23] based on these clusters having the best prediction performance. Finally, it uses a network with capsule layers and fully connected Layer to aggregate the features of clusters. Similarly, BDOCOX [30], SCNN [24] and Tu *et al.* [114] also use convolutional layers and fully connected layers to aggregate features.

3.2. Attention-based methods

Attention mechanisms help the network focus on task-relevant portions while ignoring irrelevant or redundant ones [115]. This is achieved by assigning different allocation weights, *i.e.*, attention, to different input features. Since the information contained in a WSI is huge in total amount but thin over space, the attention mechanism can better aggregate contextual information and achieve better results than simple methods. In this section, we will introduce models that incorporate basic attention modules and models utilizing the latest Transformer models.

3.2.1. Basic Attention Mechanism

In the model proposed by Saillard *et al.* [39], a SCHMOWDER framework is introduced, whose diagram is illustrated in Fig. 6.

First, they use annotations provided by pathologists to train the upper branch, to recognize tiles as either tumor or non-tumor tissues. Then, by assigning a tumoral score to each tile and applying an attention mechanism to these scores, the upper

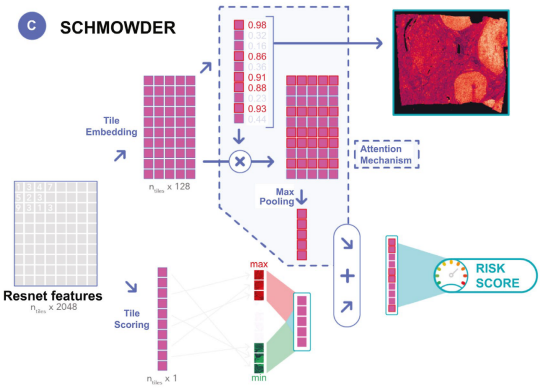

Fig. 6: Framework of SCHMOWDER. For the upper supervised branch, the attention mechanism is employed to generate representative features for survival prediction. The image is directly from [39].

branch generates a representation of tiles with a high probability of being tumoral. The lower branch is weakly supervised and only generates a small number of tiles either with the highest or the lowest risks, which is the most useful for survival prediction. The representations of the two branches are merged by some attention calculation to generate survival risk as output.

In DeepAttnMISL [36], the authors incorporate attention mechanisms for feature aggregation into their previous work DeepMISL [89], and claim that attention mechanisms significantly improve the performance of the model.

In the study of [38], Liu *et al.* introduce the EOCSA framework for survival analysis of Epithelial ovarian cancer based on WSIs. They develop a prediction model called DeepConvAttentionSurv (DCAS) using a convolutional block attention module [116], demonstrating excellent feature extraction ability and solid performance. The DCAS model adopts a cluster selection strategy to eliminate redundant information effectively. Unlike other deep survival prediction models, they integrate spatial and channel attention modules in the DCAS model to capture tumor-related information.

Moreover, inspired by modern feature pyramid networks [117], a dual-stream model with cross-attention (DSCA) is introduced in [43]. The dual-stream module is designed to process low-resolution and high-resolution patches separately. The module allows for efficient learning of hierarchical WSI representations. In the high-resolution stream, a square pooling layer is introduced to dramatically reduce the number of high-resolution patches. Therefore, the pooling operation significantly reduces computational costs during network training. This square pooling operation is implemented using a cross-attention mechanism. In detail, high-resolution patches are pooled under the guidance of global low-resolution patches. This approach effectively addresses the potential mismatch between features of different resolutions, enabling the seamless merging of dual-stream features.

3.2.2. Transformer

Transformers are a state-of-the-art deep learning model originally developed for natural language processing but now widely adopted across diverse fields, including medical image analysis. Their core strength lies in the self-attention mechanism, which enables them to identify dependencies between all elements in the input, regardless of their position. This makes Transformers particularly effective for capturing long-range interactions and integrating global context. In medical imaging, Transformers are used to analyze high-dimensional data such as whole-slide images, allowing the model to focus on critical regions while considering global context. Furthermore, their ability to process multimodal data, such as combining histological and genetic information, makes them powerful tools for survival analysis, where integrating heterogeneous data sources is often necessary for accurate predictions. Their scalability and adaptability have positioned them at the forefront of deep learning research in medicine.

Many researchers have been inspired by the remarkable success of the Vision Transformer (ViT) [118] in various computer vision tasks. They have implemented survival prediction models based on ViT, which is different from typical patch-based deep learning processing paradigms. The transformer comprises position encodings and self-attention modules, which can easily restore the WSI spatial information through certain position encodings. For each unit in the input sequence, the self-attention mechanism can determine the attention weight according to the similarity with other units.

Transformers can be expensive in terms of computation and storage costs, especially for WSIs due to their extremely high resolution. The main advantage of the Transformer model is that it can obtain a context-sensitive representation of each element in the input sequence. However, it requires attention calculation between each element pair, therefore, longer input sequences result in more computation. Fortunately, when it comes to predictive information extraction tasks, the only important thing is the global representation of the input sequence. This implies that the computationally intensive attention calculation for the input sequences can be simplified. One feasible approach is to retain only the relevant attention between global queries and local keywords. This modification ensures that the computational cost is linearly proportional to the input length.

In [45], a new method has been proposed for cancer survival prediction called MHAttnSurv, illustrated in Fig. 7. First, a ResNet model is used to extract features from randomly selected WSI patches. Next, the feature map is projected into values and keys. The value and key matrix, along with a learnable query vector, are then split into several chunks. The attention process works in each chunk to explore and identify the most significant regions. The result from each attention map is concatenated for survival prediction. This multi-head attention framework for cancer survival prediction has shown promising results in experiments.

SeTranSurv [40] is a method employed to reduce computational complexity. In this approach, the quantity is initially reduced through random sampling, and subsequently, the ViT architecture is utilized for feature aggregation and prediction. The encoding block dimension equals the number of blocks sampled from WSI. To further reduce the computational complexity of ViT, ESAT [41] uses Nystrom-based linear transformers [119] instead of the conventional self-attention transformers, leading

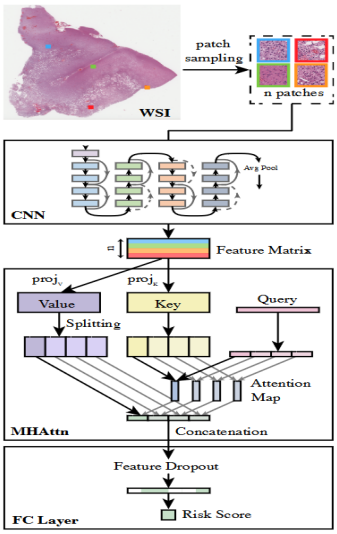

Fig. 7: Overview of the model structure of MHAttnSurv. For each WSI, n patches are randomly sampled. 4 heads are used in the multi-head attention part. The image is directly from [45]

to lower memory usage. Li *et al.* [97] apply the deformable transformer [120] to decrease the computational complexity of the self-attention operation.

Some researchers have modified the attention mechanism in conventional Transformers. HVTSurv [44] uses a hierarchical visual transformer structure for patient survival prediction. This model reduces computational complexity by employing a window attention mechanism and captures local features using a new spatial information embedding method. The model processes data hierarchically from local to global levels and uses a random window masking strategy to effectively encode patient-level contextual and hierarchical information. In [46], Wang *et al.* propose a novel neural network called pattern-perceptive survival transformer (Surformer). They introduce a ratio-reserved cross-attention module (RRCA) to detect both global and local features simultaneously. To achieve this, they employ a learnable global prototype p_{global} and multiple local prototypes p_{locals}. They quantify the patches correlated to each p_{locals} in the form of ratio factors (RFs). The ratio information is then embedded in the feature space to enhance the representation. As a quantification index, the RFs have different statistical distributions for high-risk and low-risk patients and can benefit the interpretation of the model.

Existing methods are typically based on a single size of local patch. However, different sizes of local patches correspond to different pathological features. For example, 16×16 patches capture fine-grained features such as cells, 256×256 patches reveal inter-cellular interactions, and 4096×4096 patches depict macroscopic interactions within an organization. These features are complementary to each other and can be aggregated for better representation. Inspired by this, HIPT [59] extracts visual markers over patches of different sizes and integrates the features via a Transformer. This method adjusts the self-attention of the transformer to permutation equivariant layers, simulating visual concept dependencies.

3.3. Graph-based methods

Graphs are mathematical models that can represent connections between pairs of elements. They play a crucial role in capturing contextual information. In a WSI context, graphs are particularly effective in illustrating relationships among individual patches based on spatial proximity or correlation. Unlike traditional neural networks that work on regular grids, graph convolutional networks (GCNs) excel in handling irregular, structured data, making them ideal for modeling relationships in histopathology images or patient datasets. Previous research efforts have only focused on aggregating images through patch-based methodologies. However, these approaches may overlook essential contextual relationships between image patches and their neighboring elements, leading to inaccurate predictions [121]. Recent advancements include graph-based methods that go beyond mere feature extraction and patch aggregation, exploring the topology of local patches over the entire WSI. To fully exploit the multi-scale and heterogeneous information inherent in WSIs, researchers have proposed network architectures that leverage hierarchical features and spatial structures at different resolutions. This section intends to introduce models that use graph GCNs and hyper-graph neural networks.

3.3.1. Graph Convolutional Network

In the study by [26], researchers acknowledge the crucial importance of topological relations among pathological patches in medical tasks. Graphs are commonly employed to represent these topological structures, and the existing methods for constructing graphs on WSI are mainly based on patches and cells.

Graph Construction based on Patches. For graphs with patches as vertices, the vertex features are generated by the feature extractor. Given a set of sampled patch images $\mathbf{P} = \{\mathbf{P}_i\}$ from WSI, the blocks in margin areas are discarded before the graph construction because these bolocks often have few cells. Therefore, the cardinality $\|\mathbf{P}\|$ differs by WSI, and the constructed graphs for WSIs are usually of different sizes.

The DeepGraphSurv model [48] emerges as a pioneering approach in utilizing GCN for survival prediction based on WSIs. The authors advocate for the appropriateness of intermediate patch-wise features in constructing a graph. They integrate global topology features with local patch features of WSIs through spectral convolution, serving as a central component of the entire architecture. As illustrated in Fig. 8, DeepGraphSurv treats each patch as a node, where the feature is extracted with a VGG-16 model pre-trained on ImageNet. This framework combines local patch features with global topological structures through convolution, allowing for the simultaneous learning of both local and global representations of the entire slide images. Similar patch-based graph convolution model strategies also include Patch-GCN [49]. The difference between the two models is that the graph for patch-GCN is constructed in the physical space but that of DeepGraphSurv is in the latent space. Consequently, this method can utilize spatial convolution that performs local neighborhood aggregation functions similar to CNN.

The previous methods all treat the generated patches directly as nodes, resulting in a large-scale graph. A good strategy to

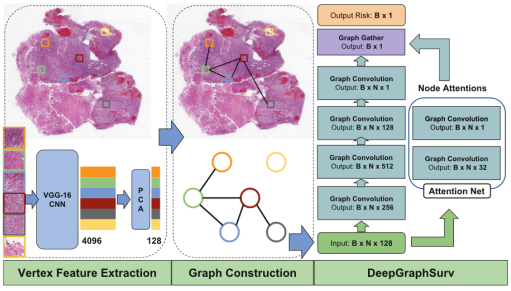

Fig. 8: Framework of DeepGraphSurv. Randomly selected patches are input into the VGG-16 feature extractor and principle component analysis is employed to generate node features. Then, the graph with 6 nodes is used to get the survival scores. The image is directly from [48].

simplify a graph is feature clustering. Methods like DeepGCN-MIL [50], SlideGraph+ [51], and NAGAN [122] employ spatial clustering methods, such as K-means, to group adjacent image blocks with similar features into clusters. Then, based on these clusters, these models generate graph representations to capture the cellular and morphological topology of WSIs. Finally, the graph constructed by the entire WSI will be used as input for the graph neural network. GraphLSurv [52] is also a good structural learning method aimed at capturing potential patch correlations and generating adaptive and sparse structures. This approach can reduce the computational complexity of feature aggregation and improve the efficiency of GCN by selecting the optimal patches.

Inspired by graph attention networks (GATs) [123], the TEA graph [53] adopts a GAT model structure. It uses a GAT with positional embeddings to extract the context features around the super-patch by aggregating the neighborhoods of the super-patch with different attention scores. Lee *et al.* adopt the super-node method [124] to compress WSIs and represent the gigapixel-sized image into memory-efficient graph structures. It can effectively handle pathological features with varying backgrounds, such as immune cells, and their interactions with the surrounding environment. Based on the extracted spatial topology, the hierarchical graph convolutional layer in HGT[60]progressively aggregates patch-level features into tissue-level features, thereby learning the topological features of variant microenvironments ranging from fine-grained (e.g., cells) to coarse-grained (e.g., necrosis, epithelium, etc.). Another attention-based approach is the cross-attention-based dual-space graph model (CoADS) in [54]. It is a dual-space GCN model that combines both the physical space and the latent space. The model uses cross-attention to effectively predict the overall survival of cancer patients, overcoming the limitations of distance-based GCN in node representation. Moreover, the model particularly focuses on exploring semantic and structural information in the tumor microenvironment.

Graph Construction based on Cells. The patch-based methods have become the main method for constructing graphs, but there are also works that model cells into graphs. As for this strategy, accurate nuclear segmentation is necessary. With segmented cell nuclei at hand, Pathomic Fusion [69], which is illustrated in Fig. 9, hypothesizes that adjacent cells will have the most significant cell-cell interactions. Therefore, the approach limits the adjacency matrix to its K nearest neighbors. Consequently, the constructed graph will be sparse, reducing the following computational burden.

3.3.2. Hyper-graph Neural Network

Hyper-graphs are the natural representation of a broad range of systems where multiple-to-multiple relationships exist among their interacting parts. Specifically, a hyper-graph is a generalization of a graph where a hyper-edge allows for the connection of an arbitrary number of nodes [125]. In many computer vision tasks, hyper-graphs have been used to model high-order correlations among data. For instance, hyper-graph neural network (HGNN) [126] utilizes hyper-graph convolutions to more effectively capture higher-order data correlations for representation learning. In light of their strong description power of complex data, hyper-graph and HGNN are now also used in survival analysis based on WSIs.

RankSurv [57] is the first work employing hyper-graph structures to represent hierarchical information. In this method, they create a hyper-graph and use spectral convolutional layers to aggregate the features. HGSurvNet [56] employs multiple hyper-graphs, using high-order global representations of WSIs for predicting survival. It first generates phenotype and topology sub-graphs based on phenotype (visual appearance) and topology information and then combines the two sub-graphs together. B-HGFN [58] introduces a large-scale hyper-graph decomposition neural network for extracting high-order representations from WSIs. It uses a pre-trained network to sample the patches and constructs hyper-edges based on the visual feature distance between patches (such as Euclidean distance). This method also defines a low dimensional hyper-graph Laplacian matrix, effectively handling patches with large-scale dense sampling.

3.4. Summary

The image feature aggregation method in survival prediction has advantages in obtaining comprehensive patient information and improving prediction accuracy. These methods overcome the limitations of individual patches and achieve the aggregation of features from patch level to WSI level.

Simple methods usually cannot dig into the underlying relationship between local patch features. CNN has strong feature extraction capabilities, but it may face challenges in capturing spatial relationships and processing different data types. Attention and graph-based models are good at capturing contextual information and relevance but may encounter more computation. Moreover, incorporating the WSI pyramid can benefit the aggregation for more discriminative WSI-level feature representation.

A worrying issue is the substantial size of the datasets, which means that the inputs of these models include all patches and may result in a significant computational burden. Simplifying the model or selecting patches with distinctive features poses a challenging topic for exploration. More and more hybrid models that combine the advantages of CNN, transformer architecture, and graph structure have demonstrated excellent performance. The current problem focuses on addressing the chal-

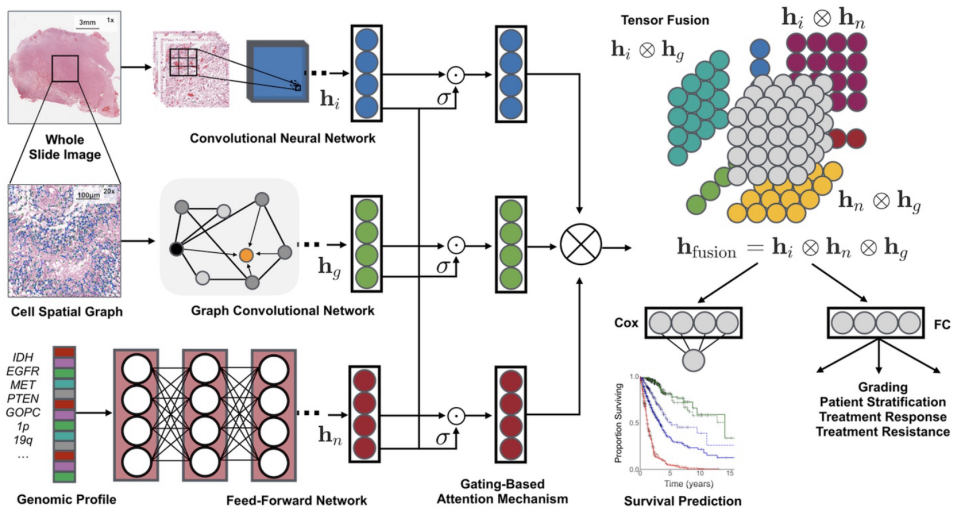

Fig. 9: Framework of Pathomic Fusion. Histological features are derived through CNNs and GCNs using the cell spatial graphs. At the same time, Genomic features are extracted with fully connected networks. Then, the model utilizes a gating-based attention mechanism in each modality, followed by the Kronecker product to model pairwise feature interactions across modalities. The image is directly from [69].

lenges of achieving efficient feature aggregation and resolving the scalability of the input data.

4. Survival Analysis

The aim of survival analysis is to estimate the duration time from the diagnosis to the event of interest which is usually the death due to a certain disease. In survival analysis for cancer, each patient can be classified into two categories: censored patients and uncensored patients. Censored patients refer to those for whom no mortality events are observed during the follow-up period. Therefore, their actual survival time is longer than the recorded data. On the other hand, uncensored patients indicate that their recorded survival time is the exact time from initial diagnosis to death.

In this chapter, we will present the classic Cox survival model and the related loss functions to train the models. Additionally, we will introduce commonly used metrics for evaluating model performance.

4.1. Survival Model

In survival analysis, the observation time of one patient is either a survival time or a censored time. An instance in the survival data is usually represented as $(\boldsymbol{x}, t, \delta)$ where $\boldsymbol{x}$ is the feature vector, t is the observed time, δ is the indicator with 1 for an uncensored instance and 0 for a censored instance. The survival function $S(t|\boldsymbol{x}) = Pr(T \geq t|\boldsymbol{x})$ is used to identify the probability of being still alive at time t. Then the hazard function is defined based on $S(t|\boldsymbol{x})$ as:

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

$$= \lim_{\Delta t \to 0} \frac{S(t|\boldsymbol{x}) - S(t + \Delta t|\boldsymbol{x})}{\Delta t}, \tag{2}$$

which assesses the instantaneous rate of death at time t.

In traditional survival modeling, it is assumed that the time duration follows an unknown distribution. The Kaplan-Meier method is used to estimate the survival function from the observation of a group. Then the log-rank test is used to compare whether two or more observations are significantly different in statistics. The Cox Proportional Hazards (CPH) model, on the other hand, is used to examine the effects of covariates on the hazard function[127]. Among popular modeling methods, the Cox proportional risk model stands out and has been applied in many survival prediction models [30], because it focuses on modeling risk rather than survival functions:

$$h(t|\boldsymbol{x}) = h_0(t)e^{\boldsymbol{\beta}^T \boldsymbol{x}}, \tag{3}$$

The expression involves the time variable t, covariates $\boldsymbol{x}$ of dimension p, a vector of regression parameters $\boldsymbol{\beta}$ and the baseline hazard $h_0(t)$. The risk function, which is also known as the regression function, is denoted as $f(x) = \boldsymbol{\beta}^T \boldsymbol{x}$ in equation (3).

Several regularization methods have been proposed in the literature based on the Cox model, including LASSO-COX [128] and En-Cox [129]. In addition, the studies in [23, 24] have combined the traditional Cox model with CNN to demonstrate end-to-end prediction models, which have become a powerful tool in solving computer vision problems. Compared with traditional survival analysis, Katzman *et al.* introduce DeepSurv [130], a fully connected deep neural network designed for capturing nonlinear relationships between covariates and the risk function. This model replaces the exponent $\boldsymbol{\beta}^T \boldsymbol{x}$ in the traditional Cox model with a nonlinear fully connected network. It is a nonlinear extension of CPH, avoiding issues related to standard linear CPH.

4.2. Loss Functions

The most famous loss function is the negative log partial likelihood (NLPL). It is employed in CPH, DeepSurv [130], and

many other models, optimizing the sample ranking. It can be expressed as the following equation.

$$NLPL = -\sum_{i=1}^{n} \delta_i \left(r(\boldsymbol{x}_i) - \log \sum_{j \in R(t_i)} e^{r(\boldsymbol{x}_j)} \right), \tag{4}$$

where n represents the patient number, t_i corresponds to the censored time or event time for patient i and δ_i indicates whether the patient is censored ($\delta_i = 0$) or with event ($\delta_i = 1$). $\boldsymbol{x}_i$ denotes the covariates of patient i. $r(\cdot)$ is the survival risk predictor, either linear by the traditional Cox model or non-linear by deep learning models. $R(t_i)$ is the set of at-risk patients whose obversed time $t_j \geq t_i$.

BDOCOX [30] proposes a novel ranking loss, defined as equation (5), and combines it with the NLPL loss to supervise the model training.

$$RankLoss = \sum_{i=1}^{n} \delta_i \left(\sum_{j \in R(t_i)} \max\left(0, 1 - e^{(r(\boldsymbol{x}_i) - r(\boldsymbol{x}_j))}\right) \right), \tag{5}$$

EPIC-Survival [131] is a new approach for analyzing histological sections using the traditional Cox loss and the End-to-end Part Learning framework. Specifically, the NLPL is combined with the clustering function, which is based on minimizing the distance between embeddings of sampled tiles and embeddings of their assigned centroids:

$$Loss = NLPL + \lambda_c \sum_{i=1}^{N} \|z_i - c_i\|^2, \tag{6}$$

where z_i is the embedding of randomly sampled tiles, c_i is the centroid assigned during the previous training epoch to the WSI from which z_i is sampled, and λ_c is a weighting parameter.

RankDeepSurv [132] employs an innovative loss function known as censored cross-entropy loss. The loss function is a combination of extended mean square error loss and paired sorting loss. The paired sorting loss is based on survival data sorting information. This loss has also demonstrated excellent performance in other survival prediction models such as DSCA [43], HVTSurv [44], and GPDBN [70]. Some ranking models encounter challenges in achieving adequate discrimination, especially when handling highly similar cases. This can result in errors in ranking predictions. RankSurv [57] introduces a Bayesian-based strategy called Bayesian Concordance Readjust (BCR) to further fine-tune ranking predictions. Di *et al.* [58] utilize a combination of the aforementioned BCRLoss and NDCGLoss2 from the existing LambdaLoss framework[133], providing a more comprehensive oversight of the overall ranking information.

To address the challenges posed by high-dimensional data, regularization methods have been introduced. These regularization methods exhibit sparsity properties, aiding in the prevention of overfitting. Consequently, a series of regularization techniques have been proposed to enhance the performance of the high-dimensional Cox model [134]. In PONET [76], HFB-Surv [73], and SuperCGGM [135], l_1 regularization is applied to model parameters. In PG-TFNet [81], l_2 regularization is employed.

By introducing a threshold of survival time, the survival prediction problem can be transformed into a classification problem. For instance, the GPDBN [70] serves as a framework designed for the classification of patients into two categories, those with long-term survival and those with shorter-term survival. In this situation, the binary cross entropy is a common objective function for survival prediction.

4.3. Measurement Metrics

To evaluate the predictive performance in survival analysis, the standard evaluation metric used by the ordinary model is the concordance index (C-index) [136]. It is a widely accepted measure for model assessment in survival prediction [137] and can be expressed as:

$$C_{index} = \frac{\sum_{i=1}^{n} \delta_i \left(\sum_{j \neq i | j \in R(t_i)} \mathcal{I}\left[r(\boldsymbol{x}_i) \geq r(\boldsymbol{x}_j) \right] \right)}{\sum_{i=1}^{n} \delta_i \left(\sum_{j \neq i | j \in R(t_i)} 1 \right)}. \tag{7}$$

Here, $r(\boldsymbol{x}_i)$ is the derived survival risk of patient i by a model and $\mathcal{I}[.]$ denotes the indicator function. From equation (7), it can be easily found that C-index measures the proportion of correctly ranked patient pairs. The C-index ranges from 0 to 1, where a larger C-index indicates better prediction performance, and vice versa. Particularly, 0 represents the worst condition where none is right, 1 is the best where all is right, and 0.5 is the value for a random guess.

Besides, the Kaplan-Meier estimator is used to illustrate the survival rate trend and to assess the effectiveness of different models. The survival rate adheres to the subsequent relationship at the k-th time point (for uncensored samples):

$$S(t_k) = S(t_{k-1})(1 - \frac{e_k}{r_k}). \tag{8}$$

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 intuitively evaluate the survival model, the entire dataset can be divided into high-risk and low-risk groups based on predicted relative risk score quantiles used as a threshold. Model performance is evaluated by comparing the Kaplan-Meier curves of these two sub-groups. A greater degree of separation between the curves indicates 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.

Alternatively, the survival prediction problem can be simplified into a binary classification task by distinguishing high-risk and low-risk patients with a threshold. Positive samples include uncensored patients with survival time under the threshold, while negative samples comprise patients with survival/censored time exceeding the threshold. Thus, the model performance can be evaluated by plotting the Receiver Operating Characteristic (ROC) curve and calculating the Area Under the ROC Curve (AUC). Since treating survival analysis as a binary classification is pretty rough, this metric is rarely used in practice.

4.4. Summary

The Cox survival model, a cornerstone in survival analysis, considers each patient's observation as either a survival or censored time. The Cox model is particularly popular in survival analysis, as it focuses on modeling the risk rather than survival functions. It has been extended through various regularization methods. Deep learning approaches, exemplified by DeepSurv [132], introduce nonlinear relationships between covariates and risk functions, offering a departure from traditional linear models. Other models like RankDeepSurv [132] and EPIC Surveillance [131] combine deep learning with survival analysis, integrating innovative loss functions and clustering techniques to enhance predictive performance.

The assessment of model performance involves various metrics. The C-index measures the ratio of correctly ordered pairs to all possible ranking pairs. The Kaplan-Meier estimator provides a visual representation of survival trends. Additionally, models are evaluated through ROC curves and AUC when transformed into binary classification tasks. In WSI survival analysis, the design objective of the loss function is to accurately predict survival time or probability. Choosing appropriate metrics and considering the integration of multiple tasks can enhance the performance of the model.

5. Multi-modality

One limitation of some existing survival models is that they initially focus on a single 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 characteristics that can describe the same event in tumor growth and symptoms. Therefore, it would be helpful to combine and fuse various types of data, such as pathological images, genotypic information, and clinical data, to explain and understand the complex symptoms and heterogeneity of cancer. By doing so, we can develop customized treatments that will improve survival predictions.

Multi-modal fusion is a critical research point in multi-modal studies, involving the fusion of information extracted from different modalities into a comprehensive multi-modal representation. This also constitutes a challenging aspect of existing research. Therefore, in this section, we focus on different fusion operations employed in existing methods. According to the fusion strategies, these methods are mainly divided into four types [138]: simple methods, attention-related methods, bi-linear pooling-related methods, and other methods.

5.1. Simple Methods

In deep learning, a widely used method is to merge feature vectors from different information sources like various modalities. This can be achieved through simple operations such as concatenation and weighted average. These methods are commonly known as simple fusion.

This simple operation results in almost no interactions between the features of different modalities, and the fused feature

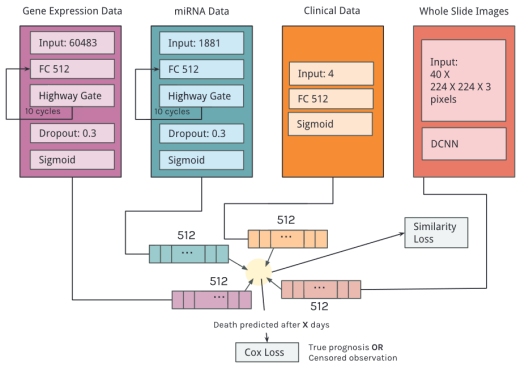

Fig. 10: Framework of Multimodalprognosis. Fully connected layers with sigmoid activations are utilized for clinical data, deep highway networks for genomic data, and the SqueezeNet architecture for WSI images. The generated features are concatenated together as a unified representation to predict the final survival risk. The image is directly from [78].

is usually further processed by subsequent simple network layers. In the DeepCorSurv model [62], the fusion is conducted by concatenating the output vectors of the image branch and the molecular branch and is then fed to the following layer for further refinement. MultiSurv [80] designs a data fusion layer, where the input data comes from the NCI's Genomic Data Commons (GDC) with six different data patterns. Specifically, the system uses a deep learning sub-model for each data modality to generate feature representations, and then the feature vectors are combined into a single fusion vector using the data fusion layer, which serves as input for the following modules. In the PAGE-Net model [79], clinical patient data is directly introduced into a demographic layer. It is then combined with genomic features and survival discriminative features in the final

As for the aforementioned simple operations, this type of multi-modal fusion is considered to lack learnability. Another feasible approach is to embed genetic or clinical data into a more complicated network for training. In the papers [24] and [82], the models directly utilize genomic variables and other information as the input of a fully connected network. hidden layer. Based on a prior study [112], HGCN [75] uses two Multi-layer Perceptron (MLP) layers to combine hyper-edges from token-wise and channel-wise modalities for inter-modal interaction. Additionally, it is feasible that different modalities of data are pre-processed for the final fusion, and the fusion layer contains a loss function related to the modality fusion. In the MultimodalPrognosis model proposed in [78], clinical, genomic, and WSI image data are first processed separately, as shown in Fig. 10. PathOmics [63] develops an unsupervised data fusion strategy by minimizing the mean square error (MSE) loss and mapping images and genomes into the same space. It deploys a concatenated layer to obtain a fused multimodal feature representation. For clinical data, the method uses fully connected layers with sigmoid activations. For genomic data, deep highway networks [139] are utilized. For WSI images, the SqueezeNet architecture [140] is employed. Then, inspired by [141], the generated feature vectors are fused by maximizing the cosine similarity between feature represen-

tations for views from the same object and minimizing it for views from different objects. To ensure stability, a margin-based hinge-loss formulation is employed, penalizing different-object feature representations only if they fall within a margin of the same-object representations. This compels different views of individual patient information to have similar feature vectors, preventing mode collapse where all features predict the exact same vector for all patients. The fused feature is then fed into a Cox model for overall survival prediction.

5.2. Attention-related methods

5.2.1. Attention Model

Inspired by methods in visual question answering, MCAT [71] introduces a new approach to let histology patches attend to genes in the survival prediction. This method introduces a Genome Guided Common Attention (GCA) layer to learn the dense common attention mapping between WSIs and bag representations in genomics, enabling the visualization of multi-modal interactions. These interactions are visualized as attention heatmaps at the WSI level for each genomic embedding. The subsequent fusion procedure simply concatenates WSI-level bag representations with genomic features. Additionally, the GCA layer reduces the effective sequence length of WSI bags. This reduction makes it possible for more advanced feature fusion techniques using self-attention to enable supervision with entire WSIs.

As an improvement on MCAT [71], MotCat [72] proposes an Optimal Transport (OT) based Co-Attention. Based on collaborative attention, OT is used to calculate the optimal matching flow for identifying information instances with global structural consistency. After aggregating the selected instances, features from the two modalities are linked for survival prediction. In GC-SPLeM [68], cross-attention is used to fuse gene information and image information, and finally, concatenation is used to form the final feature embeddings. SURVPATH [84] efficiently captures dense interactions between pathways and histology patches through a sparse multimodal attention mechanism, while avoiding the memory bottleneck of patch-to-patch self-attention computation, enabling cross-modal fusion and efficient modeling. Pathomic Fusion [69] and MoME [83] use a gating-based attention mechanism to fuse the multi-modality features, assigning different attention scores to different modalities.

5.2.2. Bimodal Transformer

The Bimodal Transformer is a kind of dual-stream transformer architecture based on the attention mechanism inspired by BERT [142]. The architecture has two streams of embedding representations that handle different input information, such as images and text. These two streams interact through a shared attention layer, which integrates information from different streams for comprehensive feature extraction and modeling. This design allows for handling information from multiple modalities effectively.

The AMMASurv [66] method effectively integrates WSI features and genomic features based on an architecture similar to BERT. Asymmetrical Multi-Modal Attention (AMMA) serves as the core of the model, which is shown in Fig. 11. The model is inspired by [143] and optimizes multi-modal data fusion. This method overcomes the limitations of the traditional full transformer self-attention mechanism by highlighting the varying importance between different modalities. It achieves effective updating of gene expression nodes through WSI feature guidance, thus improving the efficiency and accuracy of information integration. Specifically, the design of the structure ensures that genomic features do not interfere with image features, letting WSI features dominate the fusion with genomic features.

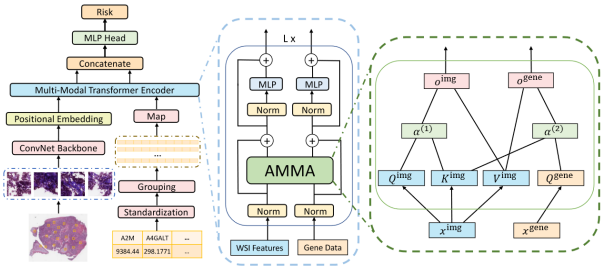

Fig. 11: Framework of AMMASurv. Firstly, the image data and genomic data are extracted and pre-processed separately. Then the extracted features are fused via a Transformer encoder with an asymmetrical multi-modal attention as the core module, where only the image feature can affect the genomic feature uniaxially. The image is directly from [66].

Other methods based on Bimodal Transformers, such as SurvPath [144], use self-attention-based transformers to capture pairwise similarities between histological patch tokens and biological pathway tokens. In PG-TFNet [81], a multi-modal data fusion module based on a cross-attention transformer is employed. This module fuses pathological and genomic features by exchanging and merging their representations from two separate branches. Chen *et al.* propose a novel graph-transformer architecture in [145]. It utilizes a converter architecture to globally aggregate WSI node features and uses clinical data with genomic data as additional feature markers.

5.3. Bi-linear Pooling-related Methods

Bi-linear pooling is a technique that combines visual feature vectors and text feature vectors to create a joint representation space. This is achieved by computing their outer product, which takes into account all the interactions between the elements of the two vectors.

The Kronecker product method is one of the most commonly used bi-linear pooling methods, which can unfold two input features separately and then perform dot product operations. PORPOISE [74] calculate the Kronecker product to model pairwise feature interactions between histological and molecular features. Wang *et al.* [70] present a new framework called Genomic and Pathological Deep Bilinear Network (GPDBN) that integrates genomic data and pathological images for breast cancer prognosis prediction. The GPDBN framework includes a Cross-Modal Bilinear Feature Encoding module (Inter BFEM) and two Internal Modal Bilinear Feature Encoding modules (Intra BFEMs). These modules are designed to facilitate the efficient exchange of information between pathological images and

genomic data, as well as within each modality. However, the introduction of a substantial number of parameters in the Kronecker product may result in elevated computational costs and pose a potential risk of overfitting.

To overcome the limitations of the Kronecker product, HFBSurv [73] extends the GPDBN framework by adopting a decomposed bi-linear model. This model presents a hierarchical multi-modal fusion approach that gradually integrates information from different levels using a decomposed bi-linear model, thus reducing computational complexity.

Besides, PONET proposed in [76] uses the multi-modal factorized bi-linear pooling method instead of an original bi-linear model. They use uni-modal, bimodal, and tri-modal fusion modules to enable feature interactions inside a single modality, between two modalities, and among three modalities, respectively. This allows for better exploration of information, compared with HFBSurv[73].

5.4. Other Methods

In addition to the aforementioned methods, some other approaches have also yielded promising results. For instance, multi-kernel learning is employed for merging different types of data in GPMKL [64], and joint training is adopted in AdaRSIS [77]. Through multi-kernel learning or joint training, the layers responsible for advanced feature extraction can be adjusted, enabling the model to adapt to different tasks. These approaches are particularly valuable when dealing with multi-modal data or information from diverse sources. They facilitate a more comprehensive capture of complex structures and relationships within the data.

One additional method for multi-modal fusion involves analyzing features of multi-modal data for sparsity. This method achieves selective integration of information by choosing or reweighting crucial features. Such analysis aids in identifying task-relevant, significantly correlated features, and it enables more effective capture of relevant information when fusing multi-modal data. Zhu *et al.* [61] utilize preprocessed genetic data to select representative features for survival prediction. They employ the Sparse partial correlation estimation (SPACE) [146] for feature selection. This is followed by the principal component regression model for feature fusion. Based on the generalized sparse canonical correlation analysis framework [147], OSCCA [65] is proposed as another sparsity correlation analysis method.

5.5. Summary

Integrating information from various modalities through multi-modal fusion improves the accuracy of survival prediction. Innovative methods such as MCAT [71] and PG-TFNet [81] showcase ongoing exploration of new techniques.

However, there are still some concerns to address. Firstly, some methods, particularly those relying on attention mechanisms, may introduce complexity that requires careful hyperparameter tuning and model training. This could result in increased computational demands. Secondly, the availability and quality of multi-modal data often degrade the effectiveness of these models. In some studies, samples with incomplete data are discarded, resulting in insufficient training data. Insufficient or noisy data can have a negative impact on the ability of models to generalize and perform well. Therefore, improving the adaptability and efficiency of multi-modal learning models to incomplete data is essential. Furthermore, deep learning fusion models, such as attention-related methods, may lack interpretability, making it difficult to comprehend the reasoning behind predictions.

In summary, multi-modal fusion is a critical aspect of survival prediction models in multi-modal scenes. Attention-related methods, bi-linear pooling, and other techniques have been explored, each with its advantages and challenges. The choice of fusion method depends on the specific characteristics of the data and the desired balance between complexity and interpretability. However, handling incomplete or noisy data and ensuring model interpretability remain ongoing challenges in the field of multi-modal survival prediction.

6. Experiment

As investigated in Sections 2-5, survival prediction models are often complex and require careful consideration, as any changes can have a significant impact on their performance. Therefore, to provide a comprehensive comparison of existing methods, we conduct experiments on five widely used public datasets for ten SOTA methods.

6.1. Dataset

The five publicly available datasets include TCGA-BLCA (Bladder Urothelial Carcinoma) with 373 patients, TCGA-BRCA (Breast invasive carcinoma) with 956 patients, TCGA-LUAD (Lung Adenocarcinoma) with 453 patients, TCGA-GBMLGG (Glioblastoma & Lower Grade Glioma) with 569 patients, TCGA-UCEC (Uterine Corpus Endometrial Carcinoma) with 480 patients. All the datasets are comprised of H&E WSIs and other modality data. Since the data for other modalities differs across the datasets, we only use the WSIs in our experiment. For all five datasets, 256×256 patches at a magnification of $\times 10$ are sampled, and a ResNet-50 is used as the feature extractor.

6.2. Comparison Methods

The comparison methods are selected based on the following two rules to ensure a fair and comprehensive comparison. Firstly, there should be publicly available code to make sure that the method is properly implemented. Secondly, each strategy should at least have one representative method. Since the feature extraction The selected SOTA methods include WSISA[26], Patch-GCN [49], MCAT [71], HistoFL [148], PORPOISE [74], AdvMIL [42], DSCA [43], GraphLSurv [52], HGCN [75], TEA [53]. MCAT [71] and HGCN [75] are not solely reliant on image data; they also incorporates other modality data, such as clinical and genomic data. However, for a fair comparison, we focus exclusively on image data, deliberately omitting other modalities. For the implementation of AdvMIL [42], we employ Patch-GCN [49] as the backbone architecture. In the case of HistoFL [148], it is executed on a single machine

Table 2: Survival prediction performance measured by 5-fold cross-validation. The same data splits and loss functions are employed in all methods for fair comparisons. The best value is in **bold** and the second best value is underlined.

	Method	Mean C-Index (Standard Deviation)					Number of parameters
		BLCA	BRCA	LUAD	GBMLGG	UCEC	
Simple Methods	WSISA [26]	0.5106 ± 0.03	0.5116 ± 0.03	0.5209 ± 0.02	0.5421 ± 0.04	0.4865 ± 0.02	0.7M
Attention-based Methods	MCAT [71]	0.5828 ± 0.01	0.5685 ± 0.05	**0.6275 ± 0.05**	0.8037 ± 0.02	0.6265 ± 0.04	3.5M
	HistoFL [148]	**0.6244 ± 0.02**	0.5978 ± 0.06	0.5879 ± 0.04	0.8014 ± 0.03	0.6205 ± 0.03	0.8M
	PORPOISE [74]	0.6168 ± 0.04	0.5927 ± 0.05	0.6223 ± 0.04	**0.8090 ± 0.01**	0.6279 ± 0.03	3.7M
	DSCA [43]	0.5888 ± 0.05	0.6034 ± 0.03	0.5767 ± 0.04	0.7621 ± 0.01	0.6279 ± 0.03	6.3M
Graph-Based Methods	Patch-GCN [49]	0.5857 ± 0.03	0.6050 ± 0.05	0.6097 ± 0.03	0.8061 ± 0.03	0.6245 ± 0.03	1.3M
	TEA [53]	0.5303 ± 0.02	0.5421 ± 0.01	0.5296 ± 0.02	0.7962 ± 0.03	0.6130 ± 0.02	2.1M
	GraphLSurv [52]	0.5932 ± 0.05	0.6039 ± 0.02	0.5094 ± 0.03	0.7296 ± 0.01	0.6168 ± 0.04	0.3M
	HGCN [75]	0.5668 ± 0.06	**0.6157 ± 0.02**	0.5067 ± 0.02	0.7948 ± 0.01	0.6364 ± 0.06	10.8M
	AdvMIL [42]	0.5904 ± 0.03	0.6012 ± 0.02	0.5260 ± 0.03	0.7537 ± 0.02	**0.6417 ± 0.04**	1.6M

to maintain computational simplicity. Regarding PORPOISE [74], our experiments utilize settings similar to those employed in the MCAT [71] framework. All the parameters are appropriately set according to their original paper.

6.3. Implementation Details

For a fair comparison, we employ the same survival loss function NLPL, and the same image feature extractor ResNet-50. As for dataset partition, we conduct a stratified 5-fold split to create 4:1 partitions for training and validation.

The performance is evaluated with C-Index across the 5 splits. Additionally, in order to determine whether the patient stratification obtained by predicted risks is statistically significant, we use the log-rank test on Kaplan-Meier curves. Specifically, we use the median value of predicted risks to partition the high-risk and low-risk sub-groups for each method on the validation folds and create Kaplan-Meier curves. All the experiments are conducted on a single Nvidia A100 80GB GPU using PyTorch.

6.4. Results

Quantitative comparisons of C-Index are shown in Table 2, and the corresponding log-rank test results of Kaplan-Meier curves are shown in Fig. 12.

As shown in Table 2, the performances on different datasets are quite different for all methods. The performances on the TCGA-GBMLGG dataset are better than those on other datasets for all models, where PORPOISE [74] achieves the best result. As for the TCGA-BLCA dataset, HistoFL [148] performs the best. When it comes to the TCGA-BRCA dataset, HGCN [75] has the best performance. MCAT [71] and AdvMIL [42] have the best results on TCGA-LUAD and TCGA-UCEC datasets, respectively.

As shown in Fig. 12, the Kaplan-Meier curves differ in appearance and the log-rank p-value for different methods and different datasets. PORPOISE [74] and MCAT [71] are the best models across all the five datasets, with the most significant differentiation of high/low-risk. Moreover, all methods can achieve good performance in the TCGA-GBMLGG dataset, indicating that this is an easier dataset. This coincides with the quantitative results in Table 2, where all methods get the best results for TCGA-GBMLGG. Patch-GCN [49] has the third-best differentiation in the datasets TCGA-LUAD and TCGA-UCEC. AdvMIL [42] has the third-best differentiation in the dataset TCGA-BRCA, and HistoFL [148] is the third-best in the dataset TCGA-BLCA.

6.5. Interpretability Comparison

Although deep learning models have achieved remarkable performance, they usually lack interpretability. This poses a significant obstacle to validation and acceptance. Fortunately, for survival analysis, there is an alternative choice to interpret the model predictions. Heatmap, showing tissue regions with different survival risks in different colors, can help pathologists identify high-risk tissue regions more easily and is a practical tool in histologic pattern analysis. The procedure to generate heatmaps is explained as follows.

First, the CLAM tool is employed to obtain effective patches for each WSI. Only valid patches are eligible for feature extraction, and the resolution, patch size, and settings mentioned earlier remain consistent throughout the process. Their coordinates information is recorded for positioning the predicted risk back to the heatmap. Then, the patches are fed to the trained models for risk prediction, and the predicted risks are arranged back to the heatmap according to their coordinates. For MIL methods, the local patch feature can be directly fed to the survival prediction subnetwork to produce the survival risk, consequently generating the heatmap easily. However, the computation within GNN poses a challenge in accurately estimating the risk value for an individual patch. This difficulty arises because the network takes the composed sub-graph as a whole and produces the final risk value for the whole sub-graph. To address this issue, we assume that each patch in a sub-graph contributes equally to the estimated risk, therefore, each patch has the same risk as the whole sub-graph. Specifically, the adjacency matrix for the whole graph with a dimension of $n \times n$ is divided into several 3×3 sub-matrices. The original graph construction method is maintained, where any nodes that cannot be regrouped into sub-matrices of size 3×3 are treated as separate nodes.

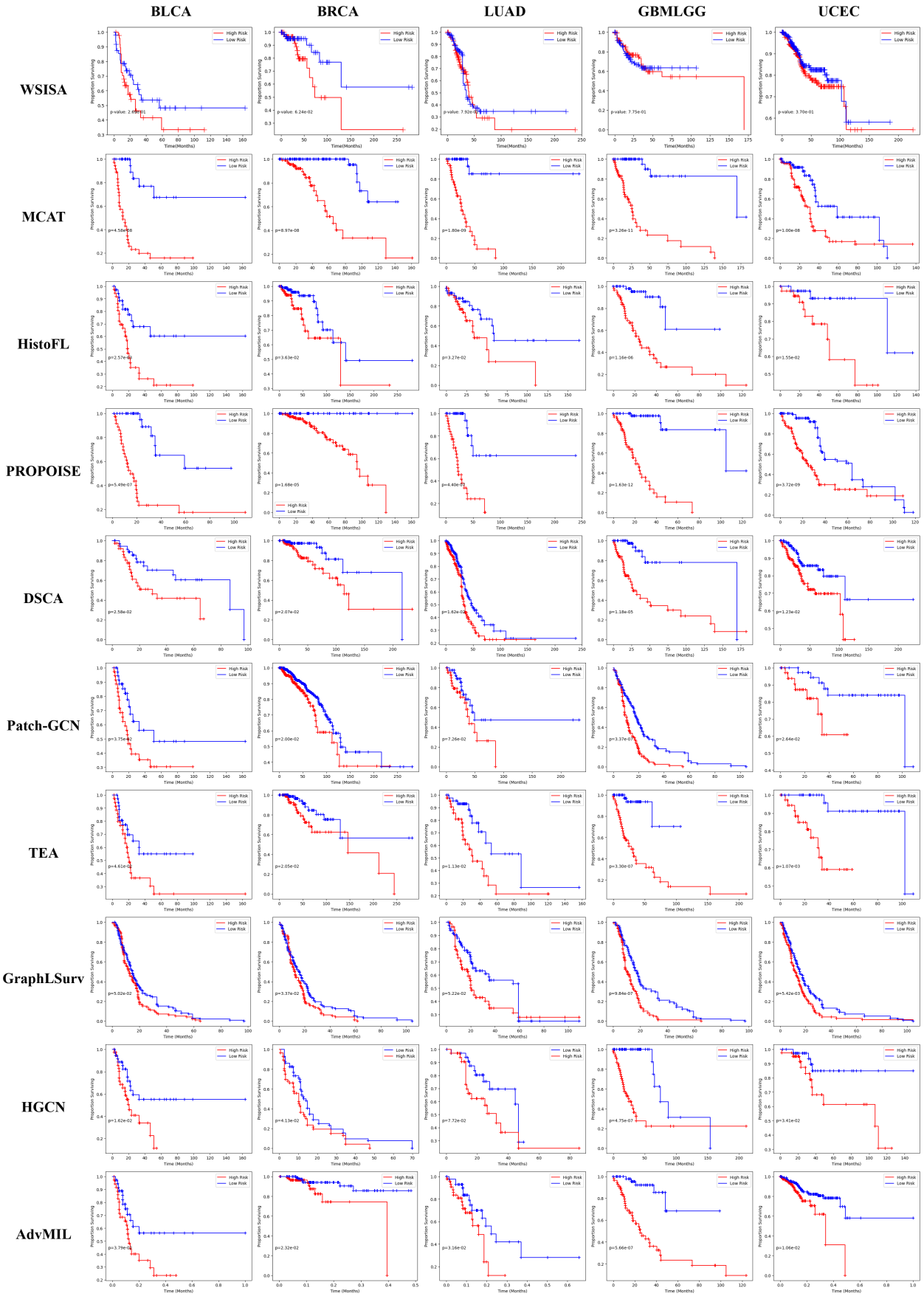

Fig. 12: Kaplan-Meier curves and log-rank test results of the methods for different datasets. Patients are partitioned into low-risk (blue curve) and high-risk (red curve) sub-groups based on the utilization of the median risk score as a threshold.

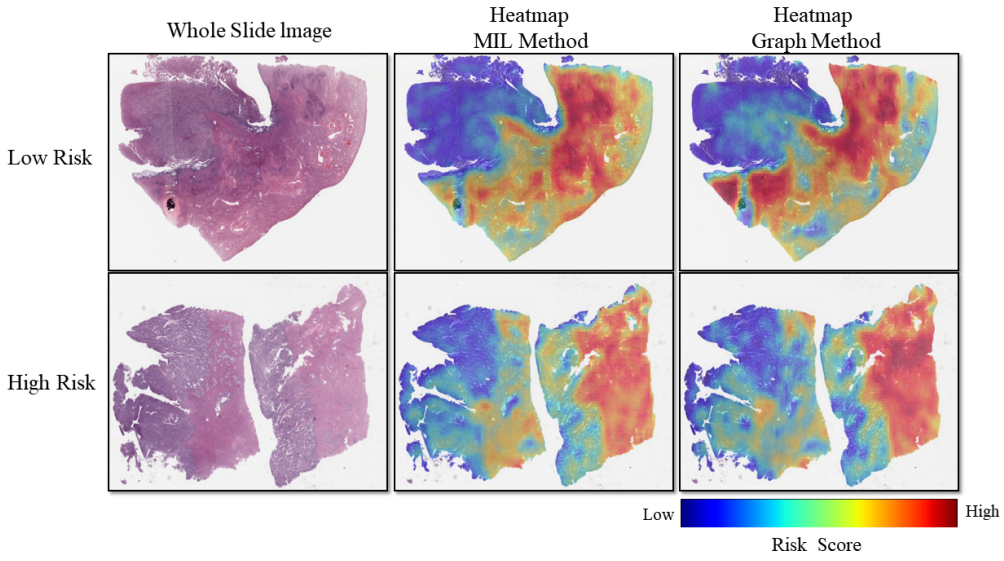

Fig. 13: Heatmaps of a MIL method MCAT [71] and a graph-based method Patch-GCN [49] on low-risk WSI and high-risk WSI. The heatmpas are normalized in WSI-level for better visualization. Heatmaps for the MIL approach seem refined, at a patch level. In contrast, those for the graph-based method are coarse, at a local region level.

According to the aforementioned procedure, we generate heatmaps of representative high-risk and low-risk examples using the MIL and graph-based methods respectively. The example heatmaps are shown in the Fig. 13 where the heatmaps are normalized in WSI-level for better visualization. MCAT [71] and Patch-GCN [49] are employed for example of the MIL and graph-based methods respectively. It can be found that heatmaps for the MIL approach are refined, at a patch level. In contrast, those for the graph-based method seems coarse, at a local region level. For the graph-based method, high-risk zones form larger and more continuous areas. Moreover, the colors used to depict these regions are usually darker, leading to a strong contrast between high and low-risk areas. If the graph-based approach is applied in clinical scenarios, it may not be as precise as MIL methods since the regions identified by GNN are usually larger. As for the reason, it is due to the inherent characteristics of the methods for heatmap generation. If a patch node involved in a sub-graph contains suspected cancerous tissue, the heatmap generation method may elevate risk values for other patches in the same sub-graph. As a result, the corresponding region of the heatmap becomes darker in color. In other words, the dark-colored blocks within a region do not always imply a higher risk due to the generation method of heatmap for graph-based methods. On the other hand, the GNN approach is more effective in distinguishing the boundaries between high-risk and low-risk regions because of the high contrast.

To improve the heatmap of graph-based methods for histologic risk assessment, several potential improvements can be considered. These include refining precision, improving granularity, adjusting contrast, normalizing size, integrating with MIL, calibrating risk indicators, including contextual information, and conducting robustness testing. By strategically addressing these aspects, heatmap methods can be tailored to provide more accurate, interpretable, and context-aware visualizations for risk assessment in pathology.

7. Discussions and Future Perspectives

In this section, we discuss the challenges and opportunities that are common in survival prediction based on WSI, including limited labels, ultra-high resolution, model interpretability, and integration with multi-modality data.

7.1. Limited Labels

Obtaining a sufficient number of labeled images is crucial for training patch feature extractors. The annotation entails manual delineation of ROI (such as abnormal or malignant tumors) in all WSIs by pathologists [149]. However, this requires a huge amount of labor and time. Consequently, there are quite limited fine-labeled data.

As investigated in Section 2, researchers often use models pre-trained on natural images as feature extractors. However, since histological images and natural images have different properties, it would be better to fine-tune the models pre-trained on ImageNet with histological image datasets, either under full supervision or self-supervision. There are two fine-tuning strategies from recent works in the computer vision society that can be borrowed. Developing a solution based on variational information bottleneck, the method presented in [150] addresses the challenges of fine-tuning and computational limitations. This is achieved through its incorporation of minimal sufficient statistics and attribution properties. Another helpful fine-tuning strategy involves teacher and student models. In the approach discussed in [151], the student model is smaller and is trained using pseudo-labeled examples, which are provided by the teacher model that is pre-trained on manually labeled images. After this initial training, the student model is fine-tuned using locally available annotations from a separate dataset.

As an alternative solution to limited labeled data, researchers have used MIL to handle WSIs. MIL has some limitations when dealing with large or imbalanced datasets. These limitations can result in over-fitting problems, especially when there is limited labeled data available. There have been several trails to tackle this problem. Lu *et al.* [152] introduce a two-stage semi-supervised approach with contrastive predictive coding to alleviate the over-fitting issues of MIL models. DS-MIL [6] develops a dual-stream architecture with trainable distance metrics. This architecture successfully extracts effective MIL representations through self-supervised contrastive learning, thereby reducing the high memory cost associated with large bags. DTFD-MIL [5] introduces the concept of pseudo-bags to build a two-layer MIL framework that effectively utilizes intrinsic features. Shao *et al.* [7] propose TransMIL, a Transformer-based MIL model that efficiently handles various classification tasks while ensuring good interpretability and visualization.

Improving MIL methods necessitates a focus on developing more flexible and robust model structures. These structures should be able to adapt to complex WSI data. Integrating advanced techniques, such as self-supervised learning and contrastive learning, can enhance model performance and generalization. Therefore, further optimizing MIL for WSI-based survival prediction is a promising direction.

7.2. Ultra-High Resolution

The ultra-high resolution of the entire WSI, reaching up to one billion pixels, poses a significant challenge [153]. As a common practice, WSI is often cut into smaller patches, typically with a size of 256×256 [154]. However, this processing strategy encounters too many patches, bringing about two challenges. The first is that a large number of patches may increase the model's scale, especially in models based on transformers, GNNs, or a combination of both. The second challenge is how to aggregate patch-level features into WSI-level features. Therefore, efficient and effective feature aggregation for both local and global levels is an important research direction.

Patch-based methods aim to leverage all non-overlapping tiled patches, albeit at the expense of increased computation and memory overhead. This approach may result in challenges such as class imbalance and slow training. Since patches are extremely smaller than the original WSI, random sampling may increase class imbalance [155]. In order to tackle this problem, numerous techniques use clustering to choose significant patches. Nevertheless, the success of clustering approaches in selecting important patches could be affected by the differences in data distribution and properties among various datasets and tasks. Some work is based on optimizing the model structure for training acceleration. For instance, Li *et al.* [97] apply the deformable transformer [120] to reduce the computational complexity of the self-attention operation.

As explained in Section **??**, multi-scale information is of vital importance in diagnosis and prognosis in clinical situations. Therefore, it is crucial for innovative approaches to integrate data of varying resolutions and improve the handling of large or imbalanced image bags during model training. Actually, there have already been several trials. HIGT [104] employs graph neural networks and Transformers for a unified framework, enabling the learning of both short-term local and long-term global information in WSI pyramids. H2-MIL [156] constructs a heterogeneous graph. The graph involves the heterogeneity and spatial scale relationships of multi-resolution patches, thus capturing image information more comprehensively. Further in-depth research is required to investigate these strategies and approaches.

7.3. Model Interpretability

Currently, there are still issues with limited interpretability in survival analysis algorithms. In digital pathology, interpretability refers to a model's capability to localize suspected target regions. For instance, a good survival prediction model should have the ability to detect the presence of tumor areas in a WSI. This interpretability also involves whether these ROIs can be understood by humans. In other words, if pathologists can meaningfully interpret the prompted high-risk regions as relevant to tumors, the model is considered interpretable[157].

It is challenging to extract understandable prognostic features from survival prediction models. On one hand, high-resolution WSI of pathological tissues offers rich and valuable information, on the other hand, the interpretability of computational pathology methods for survival prediction is relatively weak. This makes it difficult to justify the decision-making process reasonably.

When it comes to predicting a patient's likelihood of survival, there is often a gap in interpretability compared to cancer diagnosis and classification methods. This is because these two tasks have distinct goals. Cancer diagnosis and classification are mainly concerned with the detection and differentiation of cancer. In contrast, survival analysis requires the integration of instances and global features from tumors and surrounding tissues to accurately evaluate a patient's survival risk. Moreover, most current methods follow the MIL assumption, labeling a "bag" as positive if it contains at least one positive instance, otherwise negative. However, these methods cannot effectively explain the correlation between global and local features, limiting the interpretability of these methods in survival analysis tasks.

7.4. Integration with Multi-Modality Data

Multi-modal survival prediction is an approach that utilizes diverse data sources (modalities) to forecast patient survival outcomes. These modalities include various medical imaging, clinical records, and molecular biology data. The method typically involves the integration of information from different sources. Each modality requires feature extraction or selection to capture the most predictive information.

In this process, there are several challenges. For instance, in survival analysis, we may not have complete information for a certain modality, and the presence of missing information and heterogeneity can pose difficulties in handling multi-modality data. Another issue is how to preserve crucial information from each modality during feature fusion, avoiding information loss, which constitutes a significant challenge.

Currently, some methods have attempted to address the issue of incomplete data. For example, HGCN [75] employs an online masked autoencoder paradigm to complete the missing data and GC-SPLeM [68] attempts to address the issue with a GCN. More approaches are making efforts to better integrate information from different modalities. This may involve some simple methods [62] [80] as well those employing complicated deep neural networks [66] [144]. We believe that evaluating the performance of multimodal models typically requires considering the contributions and interactions between different patterns, but currently, there is little research on this aspect.

7.5. Summary

Limited labeled data in WSI poses challenges for training feature extractors. Manual annotation of ROIs is time-consuming, resulting in a scarcity of fine-labeled data. Pre-training models on natural images and fine-tuning with histological datasets are common approaches, with recent strategies like variational information bottleneck and teacher-student models showing promise. Besides, MIL is utilized as an alternative solution for handling WSI data, but with the threats of over-fitting and imbalanced datasets. Further optimization of MIL remains a promising direction.

The ultra-high resolution of WSI poses challenges, often leading to the common practice of resizing or cutting them into smaller patches. However, handling a large number of patches presents challenges in model scalability and effective feature aggregation. Patch-based methods, while leveraging all patches, may suffer from class imbalance and slow training. Exploring efficient strategies for information exchange, especially in multi-scale data, remains a promising avenue for research.

Survival analysis algorithms in digital pathology face challenges in interpretability, particularly in localizing target regions such as tumor areas within WSIs. Current methods, usually based on the MIL assumption, struggle to provide effective explanations, further hindering the model interpretability for survival analysis tasks.

Multi-modal survival prediction involves intricate tasks such as data integration. With advancements in data science and deep learning, multi-modal approaches play an increasingly crucial role in survival prediction. Incomplete data and efficient but effective fusion are the main problems to be further explored.

8. conclusion

This survey comprehensively summarizes and outlines the current mainstream methods for survival analysis based on WSI. We find the majority of existing methods for survival analysis follow a three-step framework, including image feature extraction, image feature aggregation, and survival analysis. Additionally, we provide a brief review of methods that utilize multi-modal data. Next, we conduct replication experiments of ten SOTA methods on five widely-used different datasets, evaluating and discussing the current approaches. Finally, we discuss the current issues and possible solutions based on WSI survival analysis. Our aim is to provide a summary of the current mainstream methods based on WSI survival analysis, as well as a faithful guidance for future researchers.

Declaration of Competing Interest

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

Acknowledgement

This work was supported by the National Natural Science Foundation of China [grant number 62301332] and the Natural Science Foundation of Top Talent of SZTU [grant number GDRC202117].

Author Contribution

Jihao Li: Writing - Original draft, Review&Editing, Software and Conceptualization.

Huhan Xie: Writing - Original Draft, Review&Editing and Conceptualization.

Tong Xu: Investigation and Data curation.

Xuewu Jiang: Resources and Supervision.

Tianzhao Zhong: Data curation and Visulization.

Huaishui Yang: Data curation and Formal analysis.

Mengye Lyu: Writing - Review&Editing and Supervision.

Shaojun Liu: Writing - Review&Editing, Conceptualization, Supervision, Project administration and Funding acquisition.

Data Availability

All the data used in this research are public from The Cancer Genome Atlas Program (TCGA) and can be downloaded from the Genomic Data Commons Data Portal at `https://portal.gdc.cancer.gov/`.

References

[1] Liron Pantanowitz. Digital images and the future of digital pathology. *Journal of pathology informatics*, 1, 2010.

[2] Neeta Kumar, Ruchika Gupta, and Sanjay Gupta. Whole slide imaging (wsi) in pathology: current perspectives and future directions. *Journal of digital imaging*, 33(4):1034–1040, 2020.

[3] Vinay Kumar, Abul K Abbas, Nelson Fausto, and Jon C Aster. *Robbins and Cotran pathologic basis of disease, professional edition e-book*. Elsevier health sciences, 2014.

[4] Adam Yala, Peter G Mikhael, Fredrik Strand, Gigin Lin, Kevin Smith, Yung-Liang Wan, Leslie Lamb, Kevin Hughes, Constance Lehman, and Regina Barzilay. Toward robust mammography-based models for breast cancer risk. *Science Translational Medicine*, 13(578):eaba4373, 2021.

[5] Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E Coupland, and Yalin Zheng. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18802–18812, 2022.

[6] 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.

[7] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang. TransMIL: Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classification. *Advances in Neural Information Processing Systems*, 34:2136–2147, December 2021.

[8] Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre L Moreira, Narges Razavian, and Aristotelis Tsirigos. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature medicine*, 24(10):1559–1567, 2018.

[9] Pierre Courtiol, Charles Maussion, Matahi Moarii, Elodie Pronier, Samuel Pilcer, Meriem Sefta, Pierre Manceron, Sylvain Toldo, Mikhail Zaslavskiy, Nolwenn Le Stang, Nicolas Girard, Olivier Elemento, Andrew G. Nicholson, Jean-Yves Blay, Françoise Galateau-Sallé, Gilles Wainrib, and Thomas Clozel. Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nature medicine*, 25(10):1519–1525, 2019.

[10] Hao Chen, Xiaojuan Qi, Lequan Yu, and Pheng-Ann Heng. Dcan: deep contour-aware networks for accurate gland segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2487–2496, 2016.

[11] Korsuk Sirinukunwattana, Josien P. W. Pluim, Hao Chen, Xiaojuan Qi, Pheng-Ann Heng, Yun Bo Guo, Li Yang Wang, Bogdan J. Matuszewski, Elia Bruni, Urko Sanchez, Anton Böhm, Olaf Ronneberger, Bassem Ben Cheikh, Daniel Racoceanu, Philipp Kainz, Michael Pfeiffer, Martin Urschler, David R. J. Snead, and Nasir M. Rajpoot. Gland segmentation in colon histology images: The glas challenge contest. *Medical image analysis*, 35:489–502, 2017.

[12] Faisal Mahmood, Daniel Borders, Richard J Chen, Gregory N McKay, Kevan J Salimian, Alexander Baras, and Nicholas J Durr. Deep adversarial training for multi-organ nuclei segmentation in histopathology images. *IEEE transactions on medical imaging*, 39(11):3257–3267, 2019.

[13] Yanning Zhou, Simon Graham, Navid Alemi Koohbanani, Muhammad Shaban, Pheng-Ann Heng, and Nasir Rajpoot. Cgc-net: Cell graph convolutional network for grading of colorectal cancer histology images. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019.

[14] Ashwin Raju, Jiawen Yao, Mohammad MinHazul Haq, Jitendra Jonnagaddala, and Junzhou Huang. Graph attention multi-instance learning for accurate colorectal cancer staging. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23*, pages 529–539. Springer, 2020.

[15] Manan Shah, Dayong Wang, Christopher Rubadue, David Suster, and Andrew Beck. Deep learning assessment of tumor proliferation in breast cancer histological images. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 600–603. IEEE, 2017.

[16] Peter D Caie, Arran K Turnbull, Susan M Farrington, Anca Oniscu, and David J Harrison. Quantification of tumour budding, lymphatic vessel density and invasion through image analysis in colorectal cancer. *Journal of translational medicine*, 12(1):1–12, 2014.

[17] Pingjun Chen, Maliazurina B. Saad, Frank R. Rojas, Morteza Salehjahromi, Muhammad Aminu, Rukhmini Bandyopadhyay, Lingzhi Hong, Kingsley Ebare, Carmen Behrens, Don L. Gibbons, Neda Kalhor, John V. Heymach, Ignacio I. Wistuba, Luisa M. Solis Soto, Jianjun Zhang, and Jia Wu. Cellular architecture on whole slide images allows the prediction of survival in lung adenocarcinoma. In *International Workshop on Computational Mathematics Modeling in Cancer Analysis*, pages 1–10. Springer, 2022.

[18] Arka Bhowmik and Sarah Eskreis-Winkler. Deep learning in breast imaging. *BJR— Open*, 4:20210060, 2022.

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

[20] Daisuke Komura and Shumpei Ishikawa. Machine learning methods for histopathological image analysis. *Computational and structural biotechnology journal*, 16:34–42, 2018.

[21] Weiming Hu, Xintong Li, Chen Li, Rui Li, Tao Jiang, Hongzan Sun, Xinyu Huang, Marcin Grzegorzek, and Xiaoyan Li. A state-of-the-art survey of artificial neural networks for whole-slide image analysis: from popular convolutional neural networks to potential visual transformers. *Computers in Biology and Medicine*, 161:107034, 2023.

[22] Shidan Wang, Donghan M Yang, Ruichen Rong, Xiaowei Zhan, and Guanghua Xiao. Pathology image analysis using segmentation deep learning algorithms. *The American journal of pathology*, 189(9):1686–1698, 2019.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] Bo Tang, Ao Li, Bin Li, and Minghui Wang. Capsurv: Capsule network for survival analysis with whole slide pathological images. *IEEE Access*, 7:26022–26030, 2019.

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

[29] Songhui Diao, Pingjun Chen, Eman Showkatian, Rukhmini Bandyopadhyay, Frank R. Rojas, Bo Zhu, Lingzhi Hong, Muhammad Aminu, Maliazurina B. Saad, Morteza Salehjahromi, Amgad Muneer, Sheeba J. Sujit, Carmen Behrens, Don L. Gibbons, John V. Heymach, Neda Kalhor, Ignacio I. Wistuba, Luisa M. Solis Soto, Jianjun Zhang, Wenjian Qin, and Jia Wu. Automated cellular-level dual global fusion of whole-slide imaging for lung adenocarcinoma prognosis. *Cancers*, 15(19):4824, 2023.

[30] Wei Shao, Tongxin Wang, Zhi Huang, Zhi Han, Jie Zhang, and Kun Huang. Weakly supervised deep ordinal cox model for survival prediction from whole-slide pathological images. *IEEE Transactions on Medical Imaging*, 40(12):3739–3747, 2021.

[31] 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.

[32] 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.

[33] 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.

[34] Suzanne C Wetstein, Vincent MT de Jong, Nikolas Stathonikos, Mark Opdam, Gwen MHE Dackus, Josien PW Pluim, Paul J van Diest, and Mitko Veta. Deep learning-based breast cancer grading and survival analysis on whole-slide histopathology images. *Scientific reports*, 12(1):15102, 2022.

[35] Yan Xu, Liwen Jiang, Wenjing Chen, Shuting Huang, Zhenyu Liu, and Jiangyu Zhang. Computer-aided detection and prognosis of colorectal cancer on whole slide images using dual resolution deep learning. *Journal of Cancer Research and Clinical Oncology*, 149(1):91–101, 2023.

[36] 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.

[37] Xiaofeng Jiang, Michael Hoffmeister, Hermann Brenner, Hannah Sophie Muti, Tanwei Yuan, Sebastian Foersch, Nicholas P West, Alexander Brobeil, Jitendra Jonnagaddala, Nicholas Hawkins, et al. End-to-end prognostication in colorectal cancer by deep learning: a retrospective, multicentre study. *The Lancet Digital Health*, 6(1):e33–e43, 2024.

[38] 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.

[39] Charlie Saillard, Benoit Schmauch, Oumeima Laifa, Matahi Moarii, Sylvain Toldo, Mikhail Zaslavskiy, Elodie Pronier, Alexis Laurent, Giuliana Amaddeo, Hélène Regnault, Daniele Sommacale, Marianne Ziol, Jean-Michel Pawlotsky, Sébastien Mulé, Alain Luciani, Gilles Wainrib, Thomas Clozel, Pierre Courtiol, and Julien Calderaro. Predicting survival after hepatocellular carcinoma resection using deep learning on histological slides. *Hepatology*, 72(6):2000–2013, December 2020.

[40] Ziwang Huang, Hua Chai, Ruoqi Wang, Haitao Wang, Yuedong Yang, and Hejun Wu. Integration ofhuang2021integration 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.

[41] 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.

[42] Pei Liu, Luping Ji, Feng Ye, and Bo Fu. Advmil: Adversarial multiple instance learning for the survival analysis on whole-slide images. *Medical Image Analysis*, 91:103020, 2024.

[43] 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.

[44] 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.

[45] 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.

[46] Zhikang Wang, Qian Gao, Xiaoping Yi, Xinyu Zhang, Yiwen Zhang, Daokun Zhang, Pietro Liò, Chris Bain, Richard Bassed, Shanshan Li, Yuming Guo, Seiya Imoto, Jianhua Yao, Roger J. Daly, and Jiangning Song. Surformer: An interpretable pattern-perceptive survival transformer for cancer survival prediction from histopathology whole slide images. *Computer Methods and Programs in Biomedicine*, 241:107733, 2023.

[47] Callum Christopher Mackenzie, Muhammad Dawood, Simon Graham, Mark Eastwood, and Fayyaz Ul Amir Afsar Minhas. Neural graph modelling of whole slide images for survival ranking. In *Learning on Graphs Conference*, pages 48–1. PMLR, 2022.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

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

[54] Lu Zhao, Runping Hou, Haohua Teng, Xiaolong Fu, Yuchen Han, and Jun Zhao. Coads: Cross attention based dual-space graph network for survival prediction of lung cancer using whole slide images. *Computer Methods and Programs in Biomedicine*, 236:107559, 2023.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16144–16155, 2022.

[60] Wentai Hou, Yan He, Bingjian Yao, Lequan Yu, Rongshan Yu, Feng Gao, and Liansheng Wang. Multi-scope analysis driven hierarchical graph transformer for whole slide image based cancer survival prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 745–754. Springer, 2023.

[61] 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.

[62] 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.

[63] Kexin Ding, Mu Zhou, Dimitris N Metaxas, and Shaoting Zhang. Pathology-and-genomics multimodal transformer for survival outcome prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 622–631. Springer, 2023.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] 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.

[69] 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.

[70] 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.

[71] 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.

[72] Yingxue Xu and Hao Chen. Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction. *arXiv preprint arXiv:2306.08330*, 2023.

[73] Ruiqing Li, Xingqi Wu, Ao Li, and Minghui Wang. Hfbsurv: hierarchical multimodal fusion with factorized bilinear models for cancer survival prediction. *Bioinformatics*, 38(9):2587–2594, 2022.

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

[75] 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.

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

[77] Zhangxin Zhao, Qianjin Feng, Yu Zhang, and Zhenyuan Ning. Adaptive risk-aware sharable and individual subspace learning for cancer survival analysis with multi-modality data. *Briefings in Bioinformatics*, 24(1):bbac489, 2023.

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

[79] 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.

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

[81] 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.

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

[83] Conghao Xiong, Hao Chen, Hao Zheng, Dong Wei, Yefeng Zheng, Joseph JY Sung, and Irwin King. Mome: Mixture of multimodal experts for cancer survival prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 318–328. Springer, 2024.

[84] Guillaume Jaume, Anurag Vaidya, Richard J Chen, Drew FK Williamson, Paul Pu Liang, and Faisal Mahmood. Modeling dense multimodal interactions between biological pathways and histology for survival prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11579–11590, 2024.

[85] Tong Ding, Sophia J Wagner, Andrew H Song, Richard J Chen, Ming Y Lu, Andrew Zhang, Anurag J Vaidya, Guillaume Jaume, Muhammad Shaban, Ahrong Kim, et al. Multimodal whole slide foundation model for pathology. *arXiv preprint arXiv:2411.19666*, 2024.

[86] Yi Zheng, Regan D Conrad, Emily J Green, Eric J Burks, Margrit Betke, Jennifer E Beane, and Vijaya B Kolachalama. Graph attention-based fusion of pathology images and gene expression for prediction of cancer survival. *IEEE transactions on medical imaging*, 2024.

[87] Zhe Li, Yuming Jiang, Mengkang Lu, Ruijiang Li, and Yong Xia. Survival prediction via hierarchical multimodal co-attention transformer: A computational histology-radiology solution. *IEEE Transactions on Medical Imaging*, 42(9):2678–2689, 2023.

[88] 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.

[89] 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.

[90] Le Li, Yong Liang, Mingwen Shao, Shanghui Lu, Shuilin Liao, and Dong Ouyang. Self-supervised learning-based multi-scale feature fusion network for survival analysis from whole slide images. *Computers in Biology and Medicine*, 153:106482, 2023.

[91] 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.

[92] Siteng Chen, Ning Zhang, Liren Jiang, Feng Gao, Jialiang Shao, Tao Wang, Encheng Zhang, Hong Yu, Xiang Wang, and Junhua Zheng. Clinical use of a machine learning histopathological image signature in diagnosis and survival prediction of clear cell renal cell carcinoma. *International journal of cancer*, 148(3):780–790, 2021.

[93] Juan C. Caicedo, Sam Cooper, Florian Heigwer, Scott Warchal, Peng Qiu, Csaba Molnar, Aliaksei S. Vasilevich, Joseph D. Barry, Harmanjit Singh Bansal, Oren Kraus, Mathias Wawer, Lassi Paavolainen, Markus D. Herrmann, Mohammad Rohban, Jane Hung, Holger Hennig, John Concannon, Ian Smith, Paul A. Clemons, Shantanu Singh, Paul Rees, Peter Horvath, Roger G. Linington, and Anne E. Carpenter. Data-analysis strategies for image-based cell profiling. *Nature methods*, 14(9):849–863, 2017.

[94] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021.

[95] Masayuki Tsuneki, Makoto Abe, and Fahdi Kanavati. A deep learning model for prostate adenocarcinoma classification in needle biopsy whole-slide images using transfer learning. *Diagnostics*, 12(3):768, 2022.

[96] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016.

[97] Hang Li, Fan Yang, Yu Zhao, Xiaohan Xing, Jun Zhang, Mingxuan Gao, Junzhou Huang, Liansheng Wang, and Jianhua Yao. Dt-mil: deformable transformer for multi-instance learning on histopathological image. 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 206–216. Springer, 2021.

[98] Brady Kieffer, Morteza Babaie, Shivam Kalra, and Hamid R Tizhoosh. Convolutional neural networks for histopathology image classification: Training vs. using pre-trained networks. In *2017 seventh international conference on image processing theory, tools and applications (IPTA)*, pages 1–6. IEEE, 2017.

[99] Zixuan Ye, Yunxiang Zhang, Yuebin Liang, Jidong Lang, Xiaoli Zhang, Guoliang Zang, Dawei Yuan, Geng Tian, Mansheng Xiao, and Jialiang Yang. Cervical cancer metastasis and recurrence risk prediction based on deep convolutional neural network. *Current Bioinformatics*, 17(2):164–173, 2022.

[100] Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical image analysis*, 58:101563, 2019.

[101] Jevgenij Gamper, Navid Alemi Koohbanani, Ksenija Benet, Ali Khuram, and Nasir Rajpoot. Pannuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification. In *Digital Pathology: 15th European Congress, ECDP 2019, Warwick, UK, April 10–13, 2019, Proceedings 15*, pages 11–19. Springer, 2019.

[102] Jiashu Xu. A review of self-supervised learning methods in the field of medical image analysis. *International Journal of Image, Graphics and Signal Processing (IJIGSP)*, 13(4):33–46, 2021.

[103] Abtin Riasatian, Morteza Babaie, Danial Maleki, Shivam Kalra, Mojtaba Valipour, Sobhan Hemati, Manit Zaveri, Amir Safarpoor, Sobhan Shafiei, Mehdi Afshari, Maral Rasoolijaberi, Milad Sikaroudi, Mohd Adnan, Sultaan Shah, Charles Choi, Savvas Damaskinos, Clinton J. V. Campbell, Phedias Diamandis, Liron Pantanowitz, Hany Kashani, Ali Ghodsi, and H. R. Tizhoosh. Fine-tuning and training of densenet for histopathology image representation using tcga diagnostic slides. *Medical Image Analysis*, 70:102032, 2021.

[104] 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.

[105] Hassan Muhammad, Carlie S. Sigel, Gabriele Campanella, Thomas Boerner, Linda M. Pak, Stefan Büttner, Jan N. M. IJzermans, Bas Groot Koerkamp, Michael Doukas, William R. Jarnagin, Amber L. Simpson, and Thomas J. Fuchs. Unsupervised subtyping of cholangiocarcinoma using a deep clustering convolutional autoencoder. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22*, pages 604–612. Springer, 2019.

[106] Lei Fan, Arcot Sowmya, Erik Meijering, and Yang Song. Learning visual features by colorization for slide-consistent survival prediction from 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 592–601. Springer, 2021.

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

[108] 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.

[109] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical image analysis*, 81:102559, 2022.

[110] Ming Y Lu, Bowen Chen, Andrew Zhang, Drew FK Williamson, Richard J Chen, Tong Ding, Long Phi Le, Yung-Sung Chuang, and Faisal Mahmood. Visual language pretrained multiple instance zero-shot transfer for histopathology images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19764–19775, 2023.

[111] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature Medicine*, pages 1–10, 2023.

[112] Ilya O. Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021.

[113] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. *Advances in neural information processing systems*, 30, 2017.

[114] Chao Tu, Denghui Du, Tieyong Zeng, and Yu Zhang. Deep multi-dictionary learning for survival prediction with multi-zoom histopathological whole slide images. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2023.

[115] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

[116] 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.

[117] 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.

[118] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv*, October 2020.

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

[120] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

[121] 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.

[122] Yonghang Guan, Jun Zhang, Kuan Tian, Sen Yang, Pei Dong, Jinxi Xiang, Wei Yang, Junzhou Huang, Yuyao Zhang, and Xiao Han. Node-aligned graph convolutional network for whole-slide image representation and classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18813–18823, 2022.

[123] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. *stat*, 1050(20):10–48550, 2017.

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

[125] Alessia Antelmi, Gennaro Cordasco, Mirko Polato, Vittorio Scarano, Carmine Spagnuolo, and Dingqi Yang. A survey on hypergraph representation learning. *ACM Computing Surveys*, 56(1):1–38, 2023.

[126] Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3558–3565, 2019.

[127] Hongyuan Wang, Fuyong Xing, Hai Su, Arnold Stromberg, and Lin Yang. Novel image markers for non-small cell lung cancer classification and survival prediction. *BMC bioinformatics*, 15:1–12, 2014.

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

[129] 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.

[130] Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1):1–12, 2018.

[131] Hassan Muhammad, Chensu Xie, Carlie S Sigel, Michael Doukas, Lindsay Alpert, William R Jarnagin, Amber Simpson, and Thomas J Fuchs. Epic-survival: End-to-end part inferred clustering for survival analysis, featuring prognostic stratification boosting. *arXiv preprint arXiv:2101.11085*, 2021.

[132] Bingzhong Jing, Tao Zhang, Zixian Wang, Ying Jin, Kuiyuan Liu, Wenze Qiu, Liangru Ke, Ying Sun, Caisheng He, Dan Hou, Linquan Tang, Xing Lv, and Chaofeng Li. A deep survival analysis method based on ranking. *Artif. Intell. Med.*, 98:1–9, July 2019.

[133] Xuanhui Wang, Cheng Li, Nadav Golbandi, Michael Bendersky, and Marc Najork. The lambdaloss framework for ranking metric optimization. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pages 1313–1322, 2018.

[134] Hong-Kun Jiang and Yong Liang. The l1/2 regularization network cox model for analysis of genomic data. *Computers in Biology and Medicine*, 100:203–208, 2018.

[135] Xinliang Zhu, Jiawen Yao, Guanghua Xiao, Yang Xie, Jaime Rodriguez-Canales, Edwin R Parra, Carmen Behrens, Ignacio I Wistuba, and Junzhou Huang. Imaging-genetic data mapping for clinical outcome prediction via supervised conditional gaussian graphical model. In *2016 IEEE international conference on bioinformatics and biomedicine (BIBM)*, pages 455–459. IEEE, 2016.

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

[137] 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.

[138] Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing*, 14(3):478–493, 2020.

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

[140] 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.

[141] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, 2005.

[142] 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.

[143] Shaowei Yao and Xiaojun Wan. Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4346–4350, 2020.

[144] Guillaume Jaume, Anurag Vaidya, Richard Chen, Drew Williamson, Paul Liang, and Faisal Mahmood. Modeling dense multimodal interactions between biological pathways and histology for survival prediction. *arXiv preprint arXiv:2304.06819*, 2023.

[145] Yihang Chen, Weiqin Zhao, and Lequan Yu. Transformer-based multimodal fusion for survival prediction by integrating whole slide images, clinical, and genomic datachen2023transformer. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2023.

[146] 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.

[147] 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.

[148] Ming Y Lu, Richard J Chen, Dehan Kong, Jana Lipkova, Rajendra Singh, Drew FK Williamson, Tiffany Y Chen, and Faisal Mahmood. Federated learning for computational pathology on gigapixel whole slide images. *Medical image analysis*, 76:102298, 2022.

[149] Hamid Reza Tizhoosh and Liron Pantanowitz. Artificial intelligence and digital pathology: challenges and opportunities. *Journal of pathology informatics*, 9(1):38, 2018.

[150] Honglin Li, Chenglu Zhu, Yunlong Zhang, Yuxuan Sun, Zhongyi Shui, Wenwei Kuang, Sunyi Zheng, and Lin Yang. Task-specific fine-tuning via variational information bottleneck for weakly-supervised pathology whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7454–7463, 2023.

[151] Niccolo Marini, Sebastian Otálora, Henning Müller, and Manfredo Atzori. Semi-supervised training of deep convolutional neural networks with heterogeneous data and few local annotations: An experiment on prostate histopathology image classification. *Medical image analysis*, 73:102165, 2021.

[152] Ming Y Lu, Richard J Chen, Jingwen Wang, Debora Dillon, and Faisal Mahmood. Semi-supervised histology classification using deep multiple instance learning and contrastive predictive coding. *arXiv preprint arXiv:1910.10825*, 2019.

[153] Byron Smith, Meyke Hermsen, Elizabeth Lesser, Deepak Ravichandar, and Walter Kremers. Developing image analysis pipelines of whole-slide images: Pre-and post-processing. *Journal of Clinical and Translational Science*, 5(1):e38, 2021.

[154] Kayhan Basak, Kutsev Bengisu Ozyoruk, and Derya Demir. Whole slide images in artificial intelligence applications in digital pathology: challenges and pitfalls. *Turkish Journal of Pathology*, 1(1), 2023.

[155] Neofytos Dimitriou, Ognjen Arandjelović, and Peter D Caie. Deep learning for whole slide image analysis: an overview. *Frontiers in medicine*, 6:264, 2019.

[156] Wentai Hou, Lequan Yu, Chengxuan Lin, Helong Huang, Rongshan Yu, Jing Qin, and Liansheng Wang. Hˆ 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.

[157] Thomas E Tavolara, Ziyu Su, Metin N Gurcan, and M Khalid Khan Niazi. One label is all you need: Interpretable ai-enhanced histopathology for oncology. In *Seminars in Cancer Biology*. Elsevier, 2023.