SurvMamba: State Space Model with Multi-grained Multi-modal Interaction for Survival Prediction

Ying Chen
School of Informatics, Xiamen University
Xiamen, China
cying2023@stu.xmu.edu.cn

Jiajing Xie
National Institute for Data Science in Health and Medicine, Xiamen University
Xiamen, China
xiejiajing@stu.xmu.edu.cn

Yuxiang Lin
National Institute for Data Science in Health and Medicine, Xiamen University
Xiamen, China
yuxianglin1218@gmail.com

Yuhang Song
School of Informatics, Xiamen University
Xiamen, China
songyh@stu.xmu.edu.cn

Wenxian Yang
Aginome Scientific
Xiamen, China
wx@aginome.com

Rongshan Yu[*]
School of Informatics, Xiamen University
Xiamen, China
rsyu@xmu.edu.cn

ABSTRACT

Multi-modal learning that combines pathological images with genomic data has significantly enhanced the accuracy of survival prediction. Nevertheless, existing methods have not fully utilized the inherent hierarchical structure within both whole slide images (WSIs) and transcriptomic data, from which better intra-modal representations and inter-modal integration could be derived. Moreover, many existing studies attempt to improve multi-modal representations through attention mechanisms, which inevitably lead to high complexity when processing high-dimensional WSIs and transcriptomic data. Recently, a structured state space model named Mamba emerged as a promising approach for its superior performance in modeling long sequences with low complexity. In this study, we propose Mamba with multi-grained multi-modal interaction (**SurvMamba**) for survival prediction. SurvMamba is implemented with a Hierarchical Interaction Mamba (HIM) module that facilitates efficient intra-modal interactions at different granularities, thereby capturing more detailed local features as well as rich global representations. In addition, an Interaction Fusion Mamba (IFM) module is used for cascaded inter-modal interactive fusion, yielding more comprehensive features for survival prediction. Comprehensive evaluations on five TCGA datasets demonstrate that SurvMamba outperforms other existing methods in terms of performance and computational cost.

CCS CONCEPTS

• **Applied computing → Life and medical sciences**.

KEYWORDS

Multi-modal learning, Survival Prediction, Multi-grained, Mamba

ACM Reference Format:
Ying Chen, Jiajing Xie, Yuxiang Lin, Yuhang Song, Wenxian Yang, and Rongshan Yu. 2024. SurvMamba: State Space Model with Multi-grained Multi-modal Interaction for Survival Prediction. In . ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Survival prediction evaluates patients' mortality risks, thereby enhancing the clinical decision-making process related to diagnosis and treatment planning [32]. For cancer patients, pathological images and genomic profiles provide critical and interconnected information for patient stratification and survival analysis [26]. For example, pathological images detail the tumor microenvironment, capturing the diversity of cancer cells and immune interactions [11]. At the same time, genomic profiles provide critical insights into cancer cell states and immune system factors that affect dynamic prognostic outcomes [24, 31]. Therefore, integrating pathological images and genomic data through multi-modal learning holds considerable promise to enhance the precision of cancer survival predictions.

Extensive efforts have been made in multi-modal survival analysis with histological whole slide images (WSIs) and transcriptomic data [14, 30, 34]. Among them, Multiple Instance Learning (MIL) [14, 25, 34] has emerged as an effective method to process the high-dimensional WSIs and transcriptomic data. In these MIL-based methods, each WSI is represented as a "bag" with numerous patches as instances, while transcriptomic data is organized into a "bag" with functional genomes (*i.e.*, genomic groups) as instances. Subsequently, fusion of histological and genomic features [5, 21] is conducted to predict survival outcomes.

Despite the complex, high-dimensional nature of WSIs and transcriptomic data, they exhibit significant inherent hierarchical structures. These structures stem from their fundamental biological functions and the pathology associated with diseases, as illustrated in Figure 1a. However, existing methods do not fully leverage these hierarchical structures in both WSIs and transcriptomic data [1, 15], which may lead to the following issues. The first issue regards to

[*]Corresponding author.

Conference'17, July 2017, Washington, DC, USA

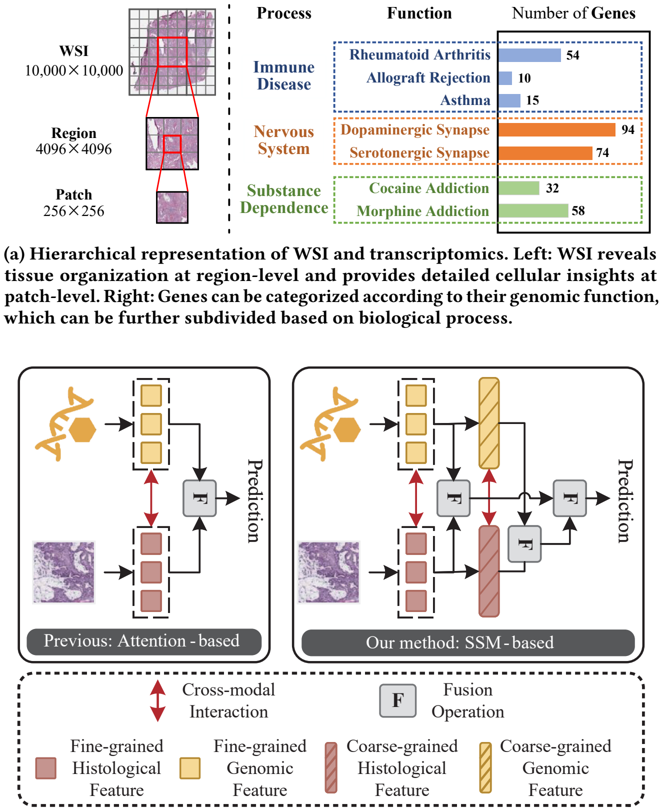
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

insufficient global representation. There are multi-level prognostic insights reflected in the hierarchical information. For example, fine-grained patch-level pathological images detail cell densities [36], while coarse-grained region-level images reveal tissue features and tumor-immune interactions [8]. Fine-grained function-level transcriptomics reveals the specific functionalities of gene sets [12], while coarse-grained process-level data focus on identifying those fundamental macroscopic biological processes [37]. Therefore, important global features for survival prediction could potentially be lost if a method focuses on fine-grained information only. Furthermore, there is limited cross-modality communications. Both WSIs and transcriptomic data contain a multitude of cross-modality communications at different hierarchical levels. For instance, adjacent patches/regions within a WSI tend to exhibit higher correlations [16], and functions/process sharing similar or related biological mechanisms demonstrate closer relationships [29], which cannot be easily revealed if cross-modality communications are only established at fine-grained features level, leading to inadequate intra-modal representations and and subsequently deficient inter-modal integration.

Although the integration of hierarchical structures into multi-modal survival prediction frameworks holds promise for yielding more comprehensive prognostic insights, it poses significant computational challenges, especially when implemented using attention mechanisms, given their high computational complexity. On the other hand, the Selective Structured State Space Model, known as Mamba [9], has demonstrated remarkable efficiency in long sequence modeling. Mamba effectively captures long-range dependencies and improves training and inference efficiency through a selection mechanism and a hardware-aware algorithm, hence providing an alternative to deal with high-dimensional data.

In this paper, we propose a Mamba-based survival prediction method with multi-grained multi-modal interaction, termed **SurvMamba**, which extracts multi-modal multi-grained information from hierarchical structure and facilitates efficient intra-modal and inter-modal interactions. Specifically, we design a novel Hierarchical Interaction Mamba (**HIM**) to efficiently capture the hierarchical characteristics of WSIs and transcriptomic data. HIM extracts coarse-grained features through the aggregation of fine-grained instances and enables efficient bidirectional interactions for multi-grained instances. In this way, SurvMamba can extract enhanced local information from fine-grained instances and global information from coarse-grained instances, resulting in more comprehensive intra-modal information to predict patient survival outcomes. Furthermore, we propose an Interaction Fusion Mamba (**IFM**) to facilitate interactions between histological and genomic features across different granularities, thereby providing refined intra-modal representations at both fine and coarse levels. Finally, these multi-grained features are adaptively integrated to formulate the final survival prediction. The main contributions of this paper can be summarized as follows:

- This work is the first attempt to introduce the Mamba model into multi-modal survival prediction, effectively processing high-dimensional WSIs and transcriptomic data with promising performance.



(a) Hierarchical representation of WSI and transcriptomics. Left: WSI reveals tissue organization at region-level and provides detailed cellular insights at patch-level. Right: Genes can be categorized according to their genomic function, which can be further subdivided based on biological process.

(b) Left: Attention-based methods focus solely on fine-grained information, resulting in high computational costs and could potentially miss crucial global information. Right: Our State Space Model (SSM)-based SurvMamba captures more comprehensive multi-modal information with efficient computation.

Figure 1: (a) Hierarchical structure of WSI and transcriptomic data. (b) Comparison of previous methods with SurvMamba.

- We propose a Hierarchical Interaction Mamba module to efficiently encode more comprehensive intra-modal representations at both fine-grained and coarse-grained levels from WSI and transcriptomic data.
- We introduce an Interaction Fusion Mamba module, designed to facilitate interaction and integration of histological and genomic features across various levels, thereby capturing multi-modal features from diverse perspectives.
- Extensive experiments are conducted on five public TCGA datasets, and results show that SurvMamba outperforms a variety of state-of-the-art methods with a smaller computational cost.

2 RELATED WORK

2.1 Multi-modal Survival Prediction

Survival outcome prediction, also known as time-to-event analysis [6], concentrates on probabilistic assessment of experiencing a specified event such as mortality in the clinical setting before a time under both uncensored and right-censored data. Right-censored data represents cases wherein the event of interest remains unobserved throughout the duration of the study. In the current state-of-the-art methods, estimating cancer patient survival heavily depends on physicians' assessment of histology and/or interpretation of genomic sequencing report [4]. Consequently, there is a growing interest in multi-modal learning methods that integrate histopathology and genomic data for survival prediction [2–5, 14, 30, 34]. Chen et al. [4] proposed a Multi-modal Co-Attention Transformer framework that identifies informative instances from pathological images using genomic features as queries. Qiu et al. [25] proposed PONET, a novel pathology-genomic deep model informed by biological pathway that integrates pathological images and genomic data to improve survival prediction. Jaume et al. [14] modeled interactions between biological pathway and histology patch tokens using a memory-efficient multi-modal Transformer for survival analysis. Zhang et al. [38] proposed a new framework named Prototypical Information Bottlenecking and Disentangling (PIBD), including Prototypical Information Bottleneck module for intra-modal redundancy and Prototypical Information Disentanglement module for inter-modal redundancy. Existing methods mainly regard multi-modal survival prediction as a fine-grained recognition task, focusing on histological patches and genomic functions to model fine-grained cross-modal relationships with local information. Neglecting coarser-grained information may result in overlooking some essential global features vital for accurately predicting patient survival. WSIs and genomic data exhibit an intrinsic hierarchical structure, where different granularity levels yield unique and critical insights into patient prognosis.

2.2 State Space Models

Recently, the State Space Models (SSMs) have shown significant effectiveness of state space transformation in capturing the dynamics and dependencies of language sequences. The structured state-space sequence model (S4) introduced in [10] is specifically designed to model long-range dependencies, offering the advantage of linear complexity. Various models including S5 [28], H3 [7] and GSS [23] have been developed base on S4, and Mamba [9] distinguishes itself by introducing a data-dependent SSM layer and a selection mechanism using parallel scan (S6). Compared to Transformers with quadratic-complexity attention, Mamba excels at processing long sequences with linear complexity. Existing models based on MIL for multi-modal survival prediction struggle with enabling effective and efficient interactions among vast numbers of instances [13, 19, 20, 27], burdened by significant computational requirements. The introduction of Mamba addresses these challenges by incorporating input-adaptive and global information modeling techniques that emulate self-attention functionalities while retaining linear complexity. This breakthrough diminishes the computational overhead and provides a potential multi-modal learning framework for survival prediction.

3 METHODOLOGY

In this section, we first describe the preliminaries of the state space model (SSM) and then provide an overview of our proposed SurvMamba (shown in Figure 2) and its core components.

3.1 Preliminaries

The SSM-based models, *i.e.*, structured state-space sequence models (S4), and Mamba, have emerged as promising architectures for modeling long sequences with linear complexity. With four parameters $(\mathbf{\Delta}, \mathbf{A}, \mathbf{B}, \mathbf{C})$, they map an input stimulus $x(t) \in \mathbb{R}$ to an output response $y(t) \in \mathbb{R}$ through an intermediate latent state $h(t) \in \mathbb{R}^N$. This process can be illustrated in the following equations:

$$\begin{aligned} h'(t) &= \mathbf{A}h(t) + \mathbf{B}x(t) \\ y(t) &= \mathbf{C}h(t) \end{aligned} \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{N \times N}$ denotes the evolution parameter, and $\mathbf{B} \in \mathbb{R}^N$, $\mathbf{C} \in \mathbb{R}^N$ represent projection parameters. The models exploit a timescale parameter $\mathbf{\Delta}$ to convert continuous parameters $\mathbf{A}$, $\mathbf{B}$ into their discrete counterparts $\bar{\mathbf{A}}$, $\bar{\mathbf{B}}$, according to:

$$\begin{aligned} \bar{\mathbf{A}} &= \exp(\mathbf{\Delta}\mathbf{A}) \\ \bar{\mathbf{B}} &= (\mathbf{\Delta}\mathbf{A})^{-1}(\exp(\mathbf{\Delta}\mathbf{A}) - I) \cdot \mathbf{\Delta}\mathbf{B}. \end{aligned} \tag{2}$$

Subsequently, these discrete parameters enable the reformulation of Eq. (1) in a recurrent format, facilitating efficient autoregressive inference:

$$\begin{aligned} h_t &= \bar{\mathbf{A}}h_{t-1} + \bar{\mathbf{B}}x_t, \\ y_t &= \mathbf{C}h_t. \end{aligned} \tag{3}$$

Finally, the output is derived through global convolution. M is the length of the input sequence x and $\bar{\mathbf{K}} \in \mathbb{R}^M$ is a structured convolutional kernel:

$$\begin{aligned} \bar{\mathbf{K}} &= (\mathbf{C}\bar{\mathbf{B}}, \mathbf{C}\bar{\mathbf{A}}\bar{\mathbf{B}}, \dots, \mathbf{C}\bar{\mathbf{A}}^{M-1}\bar{\mathbf{B}}), \\ \mathrm{y} &= \mathrm{x} * \bar{\mathbf{K}}. \end{aligned} \tag{4}$$

3.2 Overview and Problem Formulation

As illustrated in Figure 2, we propose a novel state space model with multi-grained multi-modal interaction named **SurvMamba** for survival prediction. Clinical data for each patient is encapsulated in a quadruple $X_i = (I_i, G_i, c_i, t_i)$, where I_i represent the set of WSIs, G_i refers to the set of transcriptomics, $c_i \in \{0, 1\}$ is the censoring status and $t_i \in \mathbb{R}^+$ denotes overall survival time (in months). Our objective is to employ WSI I_i and transcriptomic data G_i to estimate hazard functions $f^i_{hazard}(t)$, which represent the probability of a death event occurring in a brief interval following time t for the i-th patient.

To capture more comprehensive multi-modal representations, we exploit the inherent hierarchical structure of WSI and transcriptomic data to extract multi-grained information, as will be described in Sec 3.3, subsequently enabling efficient interaction and fusion among them through Mamba-based modules. Following [14, 25, 34], we adopt a MIL framework to learn pathological and genomic representations. We first formulate each WSI and transcriptomics as a "bag", extracting features of fine-grained instances (Patch or Function) from the Pathology Encoder and the Genomics Encoder in groups. Then, with bidirectional Mamba in dual-level MIL framework, the Hierarchical Interaction Mamba (**HIM**) module aggregates fine-grained instances into coarse-grained instances (Region or Process) and enables efficient intra-modal interactions

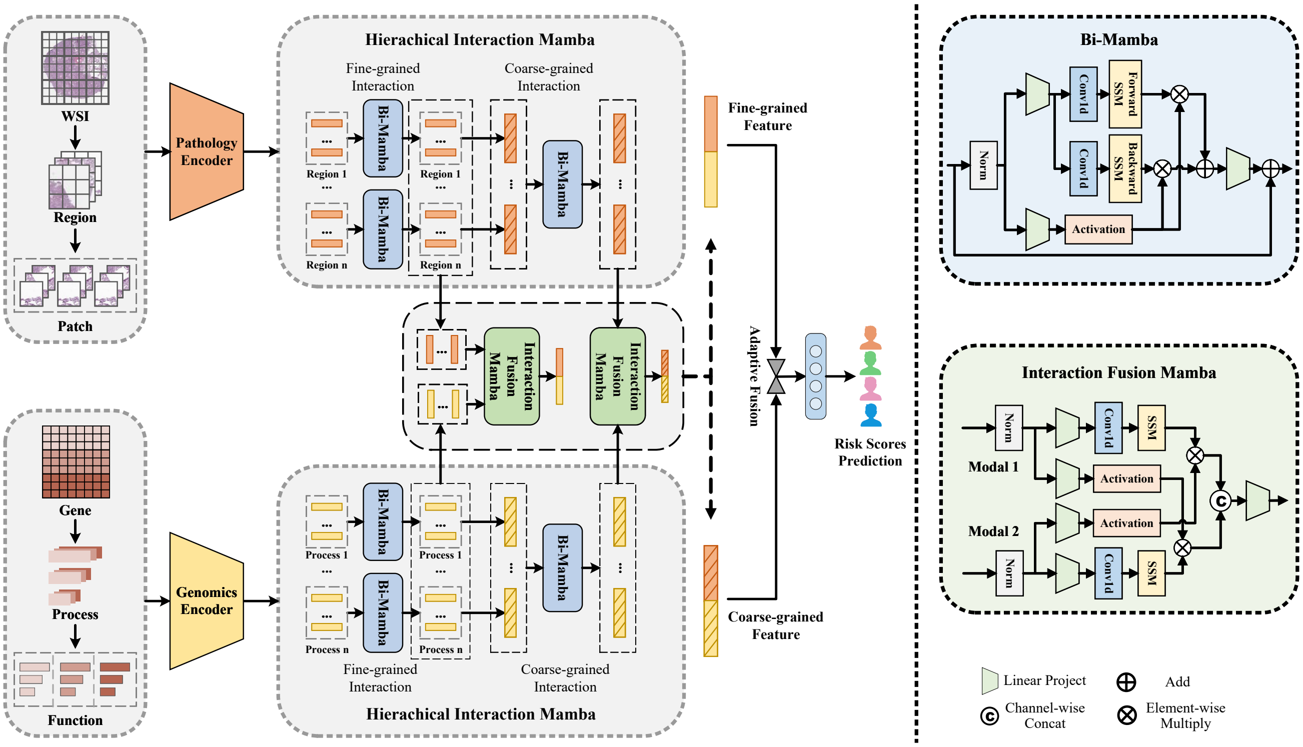


Figure 2: Overview of SurvMamba architecture. WSIs and transcriptomics are illustrated in a three-layer structure, comprising WSI/Region/Patch for WSIs and Gene/Process/Function for transcriptomics, respectively. Initially, Pathology and Genomics Encoders extract fine-grained features in groups. Then, the Hierarchical Interaction Mamba (HIM) module enhances intra-modal feature interactions across various granularities to improve unimodal representation. Meanwhile, the Interaction Fusion Mamba (IFM) module enables inter-modal integration across different levels for more comprehensive multi-modal representations. Finally, multi-grained multi-modal features are adaptively integrated to predict survival risk scores.

at different granularities. The details of this module will be described in Sec. 3.4. Further, the Interaction Fusion Mamba (**IFM**) module facilitates multi-grained interactions between histological and genomic features to derive fine-grained and coarse-grained inter-modal representations, as will be described in Sec. 3.5. Finally, multi-grained multi-modal features will be adaptively fused to predict a hazard function and get the survival risk scores, as will be described in Sec. 3.6.

3.3 Hierarchical Multi-modal Representations

WSIs and transcriptomics exhibit hierarchical structure with multi-level prognostic insights. In this study, we harness this hierarchical nature by representing WSIs and transcriptomics through a three-layer structure. Specifically, WSI is formulated with structures of WSI-, region-, and patch-level, while transcriptomic data are structured across gene-, function-, and process-level, respectively.

For an original WSI I, we split it into M non-overlapping regions R with size of $4,096 \times 4,096$, and each region is further split into N non-overlapping patches P with size of 256×256, where $I = \{R_1, R_2, \ldots, R_M\}$, and $R_m = \{P_{m1}, P_{m2}, \ldots, P_{mN}\}, 1 \leq m \leq M$. We extract the patch-level features with a frozen pre-trained Pathology Encoder $f_I(\cdot)$, resulting in patch-level token sequence T^I_m in R_m, $T^I_m = \{f_I(P_{m1}), f_I(P_{m2}), \ldots, f_I(P_{mN})\}$.

Given a set of transcriptomics measurements, it can be mapped into J different biological processes S. The j-th process S_j contains K_j genomic functions F, where $G = \{S_1, S_2, \ldots, S_J\}$, and $S_j = \{F_{j1}, F_{j2}, \ldots, F_{jK_j}\}, 1 \leq j \leq J$. The number of genomic functions contained within each process varies. We encode genomic functions with Genomics Encoder $f_G(\cdot)$ which contains multilayer perceptrons (MLPs) with learnable weights, resulting in function-level token sequence T^G_j in S_j, $T^G_j = \{f_G(F_{j1}), f_G(F_{j2}), \ldots, f_G(F_{jK_j})\}$.

3.4 Hierarchical Interaction Mamba

Structural WSI and transcriptomic data exhibit certain dependencies within local or global features. To capture these dependencies, we develop the HIM module. In inspired by [39], this module is designed to learn intra-model features by integrating a bidirectional Mamba (Bi-Mamba) mechanism within a dual-level MIL framework. A comprehensive overview of the Bi-Mamba is provided in Algorithm 1. Our approach aims to model correlations across features of varying granularities effectively. For the first level MIL, the HIM module regards fine-grained (*i.e.*, patch- and function-level) features T_i^I and T_j^G as instances, using Bi-Mamba with shared parameters between groups (i.e., region or process) to model fine-grained histological and genomic long sequences, resulting in enhanced fine-grained features $T_m^{I\,\prime}$ and $T_j^{G\prime}$:

$$\begin{aligned} T_m^{I\,\prime} &= \textbf{Bi-Mamba}(T_m^I) \\ T_j^{G\prime} &= \textbf{Bi-Mamba}(T_j^G). \end{aligned} \tag{5}$$

Subsequently, with the second level MIL, features $T_m^{I\,\prime}$ and $T_j^{G\prime}$ are pooling into coarse-grained (*i.e.*, region- and process-level) features, and Bi-Mamba is then re-applied to facilitate interactions among them to get improved coarse-grained features T^I and T^G. Through this dual-level deep interaction mechanism for each modality, the HIM ensures a thorough and efficient integration of local fine-grained and global coarse-grained features, obtaining enhanced and comprehensive intra-modal histological and genomic features.

$$\begin{aligned} T^I &= \textbf{Bi-Mamba}([\mathrm{Pool}(T_1^{I\prime}), \mathrm{Pool}(T_2^{I\prime}), \ldots, \mathrm{Pool}(T_M^{I\,\prime})]) \\ T^G &= \textbf{Bi-Mamba}([\mathrm{Pool}(T_1^{G\prime}), \mathrm{Pool}(T_2^{G\prime}), \ldots, \mathrm{Pool}(T_J^{G\prime})]). \end{aligned} \tag{6}$$

Algorithm 1 Bi-Mamba

```
Input: token sequence $T_{l-1} : (B, M, D)$
Output: token sequence $T_l : (B, M, D)$
$T'_{l-1} : (B, M, D) \leftarrow \mathrm{Norm}(T_{l-1})$
$x : (B, M, E) \leftarrow \mathrm{Linear}^x(T'_{l-1})$
$z : (B, M, E) \leftarrow \mathrm{Linear}^z(T'_{l-1})$
for o in {forward, backward} do
    $x'_o : (B, M, E) \leftarrow \mathrm{SiLU}(\mathrm{Conv1d}_o(x))$
    $B_o : (B, M, N) \leftarrow \mathrm{Linear}_o^B(x'_o)$
    $C_o : (B, M, N) \leftarrow \mathrm{Linear}_o^C(x'_o)$
    $\Delta_o : (B, M, E) \leftarrow \log(1 + \exp(\mathrm{Linear}_o^\Delta(x'_o) + \mathrm{Parameter}_o^\Delta))$
    $\overline{A_o} : (B, M, E, N) \leftarrow \Delta_o \otimes \mathrm{Parameter}_o^A$
    $\overline{B_o} : (B, M, E, N) \leftarrow \Delta_o \otimes B_o$
    $Y_o : (B, M, E) \leftarrow \mathrm{SSM}(\overline{A_o}, \overline{B_o}, C_o)(x'_o)$
end for
$Y'_{\mathrm{forward}} : (B, M, E) \leftarrow Y_{\mathrm{forward}} \odot \mathrm{SiLU}(z)$
$Y'_{\mathrm{backward}} : (B, M, E) \leftarrow Y_{\mathrm{backward}} \odot \mathrm{SiLU}(z)$
$T_l : (B, M, D) \leftarrow \mathrm{Linear}^T(Y'_{\mathrm{forward}} + Y'_{\mathrm{backward}}) + T_{l-1}$
return $T_l$
```

3.5 Interaction Fusion Mamba

To facilitate cross-modal feature interaction and fusion at different granularities, we introduce an IFM module (Algorithm 2). In IFM, we project features from two modalities and employ gating mechanisms to encourage complementary feature learning from each other while suppressing redundant features. After that, multi-modal features are concentrated to form a fused representation. Cross-modality communications are established at fine-grained and coarse-grained level via cascaded IFM. Through this block, we can obtain fine-grained fused features H_f from features $T^{I\prime}$ and $T^{G\prime}$, and get coarse-grained fused features H_c from features T^I and T^G, as follows:

$$\begin{aligned} H_f &= \textbf{IFM}([T^{I\prime}, T^{G\prime}]) \\ H_c &= \textbf{IFM}([T^I, T^G]). \end{aligned} \tag{7}$$

Algorithm 2 The IFM module

```
Input: token sequence $T_{l-1}^{a1} : (B, M, D), T_{l-1}^{a2} : (B, M, D)$
Output: token sequence $T_l : (B, M, D)$
for o in {a1, a2} do
    $T_{l-1}^o : (B, M, D) \leftarrow \mathrm{Norm}(T_{l-1}^o)$
    $x'_o : (B, M, E) \leftarrow \mathrm{SiLU}(\mathrm{Conv1d}_o(x))$
    $B_o : (B, M, N) \leftarrow \mathrm{Linear}_o^B(x'_o)$
    $C_o : (B, M, N) \leftarrow \mathrm{Linear}_o^C(x'_o)$
    $\Delta_o : (B, M, E) \leftarrow \log(1 + \exp(\mathrm{Linear}_o^\Delta(x'_o) + \mathrm{Parameter}_o^\Delta))$
    $\overline{A_o} : (B, M, E, N) \leftarrow \Delta_o \otimes \mathrm{Parameter}_o^A$
    $\overline{B_o} : (B, M, E, N) \leftarrow \Delta_o \otimes B_o$
    $Y_o : (B, M, E) \leftarrow \mathrm{SSM}(\overline{A_o}, \overline{B_o}, C_o)(x'_o)$
end for
$z1 : (B, M, E) \leftarrow \mathrm{Linear}^z(T_{l-1}^{a1})$
$z2 : (B, M, E) \leftarrow \mathrm{Linear}^z(T_{l-1}^{a2})$
$Y^{a1} : (B, M, E) \leftarrow Y^{a1} \odot \mathrm{SiLU}(z2)$
$Y^{a2} : (B, M, E) \leftarrow Y^{a2} \odot \mathrm{SiLU}(z1)$
$T_l : (B, M, D) \leftarrow \mathrm{Linear}^T(\mathrm{Cat}(Y^{a1}, Y^{a2}))$
return $T_l$
```

3.6 Survival Prediction

Recognizing the varying significance of features at different granularities for survival prediction, we employ an adaptive fusion strategy to fuse features across granular levels. Utilizing learned hyper-parameter denoted as α, we subsequently derive the final feature set for prognostication with H_f and H_c, as follows:

$$H = \alpha H_f + (1 - \alpha) H_c. \tag{8}$$

Survival prediction estimates the risk probability of an outcome event before a specific time. For the final multi-modal feature H^i of i-th patient, we use NLL loss [35] as the loss function for optimizing survival prediction, following previous works [34]:

$$\begin{aligned} L_{\mathrm{surv}}\left(\{H^i, t^i, c^i\}_{i=1}^{N_D}\right) = &-\sum_{i=1}^{N_D} c^i \log(f_{\mathrm{surv}}^i(t|H^i)) \\ &+ (1 - c^i) \log(\{1 - f_{\mathrm{surv}}^i(t-1|H^i)\}) \\ &+ (1 - c^i) \log(\{1 - f_{\mathrm{hazard}}^i(t|H^i)\}) \end{aligned} \tag{9}$$

where N_D represents the number of samples in the training set, $f_{\mathrm{surv}}^i(t|H^i) = P(T = t | T \geq t, H^i)$ denotes the hazard function characterizing the probability of death, and $f_{\mathrm{surv}}^i(t|H^i) = \prod_{k=1}^{t}(1 - f_{\mathrm{hazard}}^i(k|H^i))$ is defined as the survival function, representing the probability of survival up to the time point t.

4 EXPERIMENTS

In this section, we first describe the datasets and evaluation metrics employed in our study. The experimental results indicate that our approach surpasses contemporary methods in both performance

Table 1: c-index (mean ± std) over five TCGA datasets. G. and H. refer to genomic modality (transcriptomics) and histological modality (WSI), respectively. The best results and the second-best results are highlighted in bold and in underline. Cat refers to concatenation, KP refers to Kronecker product.

Model	G.	H.	BRCA	BLCA	COADREAD	UCEC	LUAD	Overall
SNN	✓		0.606±0.011	0.610±0.038	0.617±0.025	0.610±0.032	0.589±0.031	0.606
SNNTrans	✓		0.621±0.032	0.611±0.010	0.635±0.044	0.592±0.017	0.602±0.044	0.612
ABMIL		✓	0.613±0.033	0.588±0.033	0.624±0.050	0.618±0.014	0.604±0.043	0.609
CLAM-SB		✓	0.605±0.062	0.602±0.031	0.598±0.036	0.576±0.043	0.586±0.033	0.593
CLAM-MB		✓	0.611±0.041	0.609±0.010	0.611±0.036	0.589±0.023	0.612±0.022	0.606
TransMIL		✓	0.628±0.015	0.604±0.045	0.627±0.425	0.601±0.030	0.626±0.030	0.617
ABMIL (Cat)	✓	✓	0.633±0.029	0.621±0.062	0.637±0.030	0.632±0.014	0.611±0.049	0.627
ABMIL (KP)	✓	✓	0.661±0.042	0.656±0.028	0.660±0.032	0.649±0.030	0.642±0.020	0.654
CLAM-MB (Cat)	✓	✓	0.628±0.047	0.619±0.032	0.614±0.021	0.601±0.031	0.610±0.032	0.614
CLAM-MB (KP)	✓	✓	0.655±0.045	0.633±0.027	0.651±0.053	0.637±0.021	0.629±0.061	0.641
TransMIL (Cat)	✓	✓	0.651±0.039	0.631±0.031	0.636±0.026	0.622±0.043	0.641±0.033	0.636
TransMIL (KP)	✓	✓	0.671±0.021	0.656±0.038	0.661±0.034	0.649±0.036	0.652±0.054	0.658
MCAT	✓	✓	0.670±0.032	0.669±0.026	0.667±0.025	0.660±0.032	0.682±0.042	0.670
MOTCat	✓	✓	0.692±0.036	0.688±0.029	0.669±0.042	0.692±0.024	0.687±0.046	0.686
SurvPath	✓	✓	0.713±0.025	0.707±0.014	0.683±0.022	0.720±0.026	0.684±0.025	0.701
SurvMamba	✓	✓	**0.737±0.014**	**0.720±0.027**	**0.697±0.018**	**0.731±0.012**	**0.702±0.020**	**0.717**

and computational efficiency. Further, we conduct ablation studies to examine the influence of critical components. Finally, from a statistical perspective, we utilize Kaplan-Meier survival curves and the Logrank test to illustrate the effectiveness of survival analysis.

4.1 Datasets and Settings

Datasets. To demonstrate the performance of our proposed method, we conducted a series of experiments using five public cancer datasets from The Cancer Genome Atlas (TCGA)[1], which include paired diagnostic WSIs and transcriptomic data alongside verified survival outcomes. The datasets encompass Breast Invasive Carcinoma (BRCA) with 869 cases, Bladder Urothelial Carcinoma (BLCA) with 359 cases, Colon and Rectum Adenocarcinoma (COADREAD) with 296 cases, Uterine Corpus Endometrial Carcinoma (UCEC) with 480 cases, and Lung Adenocarcinoma (LUAD) with 453 cases. Regarding transcriptomic data, we identify 352 unique genomic functions and 42 biological processes, as cataloged in the Kyoto Encyclopedia of Genes and Genomes database (KEGG)[2]. More details about the dataset are provided in Supplementary Material.

Evaluation Metrics. The models are evaluated using the concordance index (c-index), where a higher value indicates better performance. This index quantifies the proportion of all possible pairs of observations for which the model accurately predicts the sequence of actual survival outcomes.

Implementation. For WSI preprocessing, we segmented each WSI into regions of 4096 × 4096 pixels at a magnification level of 20×, subsequently subdividing these regions into patches of 256 × 256 pixels. The Rectified Adam (RAdam) optimizer was utilized to facilitate model optimization, with a batch size set to 1, a learning rate of 2×10^{-4}, and a weight decay parameter of 5×10^{-3}. The Pathology Encoder [33] generates 768-dimensional embeddings that are projected to 512 dimensions. The Genomics Encoder comprises a two-layer feed-forward network designed to produce genomic tokens with 512 dimensions. All computational experiments were conducted on an NVIDIA-A800 GPU. To enhance the robustness of model training, 5-fold cross-validation was applied on all models.

4.2 Comparisons with the State-of-the-Art

To perform a comprehensive comparison with our method, we implemented and evaluated some latest survival prediction methods. We compared our method against the unimodal baselines and the multi-modal SOTA methods. Table 1 shows the experimental results of all methods on all five TCGA datasets.

Unimodal baselines. For transcriptomic data, we specifically implemented SNN [17] and SNNTrans [17, 18]. SNN takes the concatenated genomic profiles as a feature vector and predicts the survival outcomes. SNNTrans initially categorizes the genomic data according to the functions and then utilizes TransMIL [27] to predict the overall survival. For histology, we compared the SOTA MIL methods ABMIL [13], CLAM [22], TransMIL [27].

Multi-modal baselines. We compared SOTA methods for multi-modal survival outcome prediction with the previous set-based network architectures (ABMIL, CLAM, TransMIL) with concatenation (Cat) and Kronecker product (KP), two common late fusion mechanisms to integrate bag-level WSI features and genomic features as multi-modal baselines. We also compared MCAT [4], MOTCat [34] and SurvPath [14] three SOTA methods for multi-modal survival outcome prediction.

Unimodal v.s. Multi-modal. Compared with all unimodal methods, our proposed SurvMamba achieved the highest performance in

[1] http://www.cancer.gov/tcga
[2] https://www.genome.jp/kegg/

all five datasets, indicating effective integration of multi-modal features in our method. In comparison with methods for genomic data, SNN and SNNTrans, SurvMamba outperformed them on all benchmarks, with overall c-index performance increases of 11.1% and 10.5%, respectively. Against the pathology baselines, SurvMamba improved on all the pathology-based unimodal approaches, with performance improvements in the overall c-index ranging 10.0% to 12.4%, demonstrating the merit of integrating histopathology and genomic features. The comparison results also highlight the benefits of utilizing multi-modality in survival prediction.

Multi-modal SOTA v.s. SurvMamba. SurvMamba outperforms all multi-modal approaches with an overall c-index performance increase ranging from 1.6% to 10.3%. In comparative analyses within each dataset, SurvMamba achieved the highest c-index performance in four out of five cancer benchmarks, demonstrating its potential as a general method for survival prediction task. When compared with MIL-based multi-modal methods with different fusion methods, SurvMamba achieved a performance increase in overall c-index ranging from 5.9% to 10.3%, highlighting the effectiveness of the proposed multi-modal integration method. Besides, our model also outperforms other SOTA multi-modal learning methods by a obvious margin, including MCAT, MOTCat, and SurvPath, showcasing its outstanding ability for multi-modal learning for prognosis.

Table 2: Computational complexity analysis on a single NVIDIA A800 GPU.

Model	Params	FLOPs	c-index
TransMIL (KP)	2339.20K	6.01G	0.658
MCAT	4869.64K	24.796G	0.670
MOTCat	4869.64K	24.796G	0.686
SurvPath	854.15K	5.19G	0.701
SurvMamba	398.85K	2.51G	0.717

Computational Cost Comparison. Existing frameworks for multi-modal survival prediction are based on the attention mechanism to model the intra- and inter-relationships, while our method leverages the state space model. To validate the computational efficiency of our method, we compared SurvMamba with some multi-modal SOTA methods which show the top 5 on the overall c-index of the five TCGA datasets. The computational complexity is evaluated using model parameters (Params) and floating-point operation count (FLOPs). Params evaluates the network's scale, while FLOPs assess the model's complexity. As shown in Table 2, we report the overall c-index on the TCGA dataset, Params, and FLOPs for different SOTA models. Compared to TransMIL, the model size and GPU memory of SurvMamba are reduced by 82.9% and 58.2% with a 5.9% better overall c-index. Compared to the second best method SurvPath, SurvMamba also decreases by 53.3% and 51.6% in model size and GPU memory with a 1.6% better c-index. These results show that our model achieves superior performance compared to the state-of-the-art methods with reduced computational cost.

4.3 Ablation Study

In this ablation study, we conducted some extra experiments to investigate the effectiveness of the key components of SurvMamba. Furthermore, we compared the performance of unidirectional and bidirectional Mamba designs in our method.

Ablation of components. We conducted ablation studies on five TCGA cancer datasets to validate the effectiveness of the proposed modules. Detailed experimental setups are as follows:

(A) **None**: All fine-grained histological or genomic features are fed as an input vector into the Bi-Mamba block. These unimodal features are then concatenated to predict survival outcomes.
(B) **Fine-grained HIM**: Utilizes shared Bi-Mamba in HIM to refine fine-grained features across specific groups, enhancing these features for prediction without considering coarse-grained information.
(C) **Multi-grained HIM**: Extends Model (B) by incorporating coarse-grained information and integrating multi-grained data for prediction.
(D) **Multi-grained HIM + Fine-grained IFM**: Model (C) with IFM for fine-grained features.
(E) **Multi-grained HIM + Coarse-grained IFM**: Model (C) with IFM for coarse-grained features.
(F) **Multi-grained HIM + Multi-grained IFM**: Model (C) with IFM for multi-grained features (*i.e.*, SurvMamba).

Table 3 illustrates that Model B, by grouping and processing fine-grained features rather than directly learning from extensive sequences of these features like Model A, more effectively identifies nuanced local relationships. This methodological shift results in a c-index increase from 0.674 to 0.689. The introduction of coarse-grained information, representing broader characteristics in Model C further enhances the c-index to 0.695. These outcomes highlight the advantage of a hierarchical approach in harnessing multi-grained information for prognostic purposes. Improving upon Model C, the application of IFM to integrate and interact with fine- or coarse-grained features resulted in a c-index of 0.701 and 0.689, respectively. This demonstrates the IFM module's effectiveness in integrating multi-modal features. Our proposed SurvMamba model, facilitating both intra-modal and inter-modal interactions and integrations across different granularity levels, delivers a notable c-index of 0.717. Results of the ablation study illustrate the benefits of integrating fine and coarse-grained features and the promising SSM architecture for survival outcome prediction.

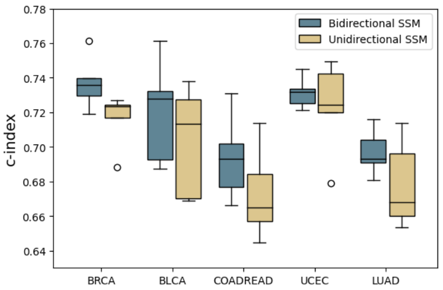
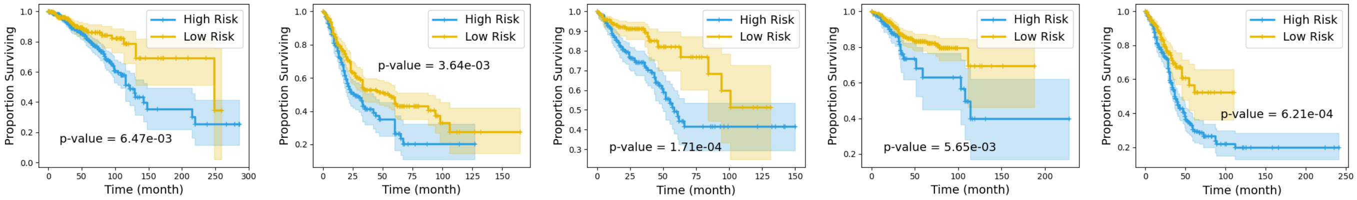


Figure 3: Ablation study on the bidirectional design in our proposed SurvMamba.

Table 3: We compare the effects of different components on the performance (c-index) of SurvMamba on 5 datasets.

Model	HIM		IFM		BRCA	BLCA	COADREAD	UCEC	LUAD	Overall
	Fine-grained	Coarse-grained	Fine-grained	Coarse-grained						
A					0.688±0.023	0.667±0.031	0.660±0.031	0.686±0.015	0.668±0.047	0.674
B	✓				0.708±0.011	0.683±0.065	0.666±0.022	0.717±0.023	0.674±0.309	0.689
C	✓	✓			0.707±0.007	0.691±0.030	0.681±0.030	0.719±0.020	0.675±0.033	0.695
D	✓	✓	✓		0.711±0.081	0.700±0.034	0.684±0.036	0.721±0.017	0.689±0.016	0.701
E	✓	✓		✓	0.709±0.010	0.696±0.013	0.687±0.030	0.720±0.010	0.679±0.019	0.698
F	✓	✓	✓	✓	0.737±0.014	0.720±0.027	0.697±0.018	0.731±0.012	0.702±0.020	0.717



(a) BRCA (b) BLCA (c) COADREAD (d) UCEC (e) LUAD

Figure 4: Kaplan-Meier survival curves of SurvMamba on five TCGA cancer datasets, where all patients are stratified into a low-risk group (yellow) and a high-risk group (blue) according to predicted risk scores. Shared areas within two groups refer to the confidence intervals, and a p-value of less than 0.05 indicates significant statistical difference.

Impact of Bidirectional and Unidirectional SSM. Figure 3 demonstrates that integrating bidirectionality into state space models significantly enhances prognostic predictions, as indicated by a higher median c-index and a narrower interquartile range. This effect is particularly pronounced in the BRCA dataset, where the boxplots show minimal overlap, suggesting a robust improvement. This improvement is likely due to the model's enhanced ability to capture complex, bidirectional dependencies within histological and genomic features.

4.4 Survival Analysis

To further validate the effectiveness of SurvMamba for survival analysis, we divided all patients into a low-risk group and a high-risk group based on the median value of the predicted risk scores generated by SurvMamba. Then, we utilize Kaplan-Meier analysis to visualize the survival events of all patients. Meanwhile, we also employ the Logrank test (p-value) to measure the statistical significance between the low-risk group and the high-risk group. A p-value of less than 0.05 indicates statistical significance. As shown in Figure 4, patients in the low-risk and high-risk groups are stratified clearly on all datasets, demonstrating the prognostic value of SurvMamba in predicting patient outcomes and guiding treatment decisions.

5 CONCLUSION

In this paper, we proposed a novel state space model with multi-grained multi-modal interaction, termed SurvMamba, for survival prediction from WSIs and transcriptomic data. SurvMamba utilizes multi-grained information from hierarchical structures in WSIs and transcriptomic data. We introduced a HIM module that facilitates efficient interactions between intra-modal features at different granularity levels, thereby enriching the unimodal representations. Furthermore, we introduced an IFM module to integrate inter-modal features across various levels, capturing more comprehensive multi-modal features for survival analysis. The experimental results demonstrate SurvMamba's superiority in both performance and computational efficiency, underlining its potential for clinical utilization, such as informing diagnostic and treatment choices for cancer patients.

REFERENCES

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

[2] Richard J Chen, Ming Y Lu, Muhammad Shaban, Chengkuan Chen, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. 2021. 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.* Springer, 339–349.

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

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

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

[6] G Kleinbaum David and Klein Mitchel. 2012. Survival analysis: a Self-Learning text.

[7] Daniel Y Fu, Tri Dao, Khaled K Saab, Armin W Thomas, Atri Rudra, and Christopher Ré. 2022. Hungry hungry hippos: Towards language modeling with state space models. *arXiv preprint arXiv:2212.14052* (2022).

[8] Yu Fu, Alexander W Jung, Ramon Viñas Torne, Santiago Gonzalez, Harald Vöhringer, Artem Shmatko, Lucy R Yates, Mercedes Jimenez-Linan, Luiza Moore, and Moritz Gerstung. 2020. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nature cancer* 1, 8 (2020), 800–810.

[9] Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752* (2023).

[10] Albert Gu, Karan Goel, and Christopher Ré. 2021. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396* (2021).

[11] Cheng-Peng Gui, Yu-Hang Chen, Hong-Wei Zhao, Jia-Zheng Cao, Tian-Jie Liu, Sheng-Wei Xiong, Yan-Fei Yu, Bing Liao, Yun Cao, Jia-Ying Li, et al. 2023. Multimodal recurrence scoring system for prediction of clear cell renal cell carcinoma outcome: a discovery and validation study. *The Lancet Digital Health* 5, 8 (2023), e515–e524.

[12] Robert Ietswaart, Benjamin M Gyori, John A Bachman, Peter K Sorger, and L Stirling Churchman. 2021. GeneWalk identifies relevant gene functions for a biological context using network representation learning. *Genome biology* 22 (2021), 1–35.

[13] Maximilian Ilse, Jakub Tomczak, and Max Welling. 2018. Attention-based deep multiple instance learning. In *International conference on machine learning*. PMLR, 2127–2136.

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

[15] Minoru Kanehisa and Susumu Goto. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* 28, 1 (2000), 27–30.

[16] Jing Ke, Yiqing Shen, Yizhou Lu, Yi Guo, and Dinggang Shen. 2023. Mine local homogeneous representation by interaction information clustering with unsupervised learning in histopathology images. *Computer Methods and Programs in Biomedicine* 235 (2023), 107520.

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

[18] David G Kleinbaum and Mitchel Klein. 1996. *Survival analysis a self-learning text.* Springer.

[19] Bin Li, Yin Li, and Kevin W Eliceiri. 2021. 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*. 14318–14328.

[20] Hao Li, Ying Chen, Yifei Chen, Wenxian Yang, Bowen Ding, Yuchen Han, Liansheng Wang, and Rongshan Yu. 2024. Generalizable Whole Slide Image Classification with Fine-Grained Visual-Semantic Interaction. *arXiv preprint arXiv:2402.19326* (2024).

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

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

[23] Harsh Mehta, Ankit Gupta, Ashok Cutkosky, and Behnam Neyshabur. 2022. Long range language modeling via gated state spaces. *arXiv preprint arXiv:2206.13947* (2022).

[24] Pedro Milanez-Almeida, Andrew J Martins, Ronald N Germain, and John S Tsang. 2020. Cancer prognosis with shallow tumor RNA sequencing. *Nature medicine* 26, 2 (2020), 188–192.

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

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

[27] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. 2021. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems* 34 (2021), 2136–2147.

[28] Jimmy TH Smith, Andrew Warrington, and Scott W Linderman. 2022. Simplified state space layers for sequence modeling. *arXiv preprint arXiv:2208.04933* (2022).

[29] Bogna J Smug, Krzysztof Szczepaniak, Eduardo PC Rocha, Stanislaw Dunin-Horkawicz, and Rafał J Mostowy. 2023. Ongoing shuffling of protein fragments diversifies core viral functions linked to interactions with bacterial hosts. *Nature Communications* 14, 1 (2023), 7460.

[30] Vaishnavi Subramanian, Tanveer Syeda-Mahmood, and Minh N Do. 2021. Multimodal fusion using sparse CCA for breast cancer survival prediction. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 1429–1432.

[31] Mengsha Tong, Yuxiang Lin, Wenxian Yang, Jinsheng Song, Zheyang Zhang, Jiajing Xie, Jingyi Tian, Shijie Luo, Chenyu Liang, Jialiang Huang, et al. 2023. Prioritizing prognostic-associated subpopulations and individualized recurrence risk signatures from single-cell transcriptomes of colorectal cancer. *Briefings in Bioinformatics* 24, 3 (2023), bbad078.

[32] Khoa A Tran, Olga Kondrashova, Andrew Bradley, Elizabeth D Williams, John V Pearson, and Nicola Waddell. 2021. Deep learning in cancer diagnosis, prognosis and treatment selection. *Genome Medicine* 13 (2021), 1–17.

[33] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Junzhou Huang, Wei Yang, and Xiao Han. 2021. Transpath: Transformer-based self-supervised learning for histopathological image classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*. Springer, 186–195.

[34] Yingxue Xu and Hao Chen. 2023. Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 21241–21251.

[35] Shekoufeh Gorgi Zadeh and Matthias Schmid. 2020. Bias in cross-entropy-based training of deep survival networks. *IEEE transactions on pattern analysis and machine intelligence* 43, 9 (2020), 3126–3137.

[36] Daiwei Zhang, Amelia Schroeder, Hanying Yan, Haochen Yang, Jian Hu, Michelle YY Lee, Kyung S Cho, Katalin Susztak, George X Xu, Michael D Feldman, et al. 2024. Inferring super-resolution tissue architecture by integrating spatial transcriptomics with histology. *Nature Biotechnology* (2024), 1–6.

[37] Xi Zhang, Yining Hu, and David Roy Smith. 2021. Protocol for HSDFinder: Identifying, annotating, categorizing, and visualizing duplicated genes in eukaryotic genomes. *STAR protocols* 2, 3 (2021), 100619.

[38] Yilan Zhang, Yingxue Xu, Jianqi Chen, Fengying Xie, and Hao Chen. 2024. Prototypical Information Bottlenecking and Disentangling for Multimodal Cancer Survival Prediction. *arXiv preprint arXiv:2401.01646* (2024).

[39] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. 2024. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417* (2024).