# SAMamba: Integrating State Space Model for Enhanced Multi-modal Survival Analysis

Wei Zhang
*Department of Data Science*
*City University of Hong Kong*
Hong Kong SAR, China
wzhang472-c@my.cityu.edu.hk

Tong Chen
*Department of Data Science*
*City University of Hong Kong*
Hong Kong SAR, China
tong.chen@my.cityu.edu.hk

Wenxin Xu
*Department of Data Science*
*City University of Hong Kong*
Hong Kong SAR, China
wenxinxu8-c@my.cityu.edu.hk

Xinyue Li(✉)
*Department of Data Science*
*City University of Hong Kong*
Hong Kong SAR, China
xinyueli@cityu.edu.hk

2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3503-8622-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/BIBM62325.2024.10822595

***Abstract*—Survival prediction represents a challenging ordinal regression task, involving modeling of intricate interactions among various data modalities. The recent evolution of state space models, Mamba in particular, has opened new vistas for effectively processing sequence data, including genomic profiles and gigapixel pathology Whole Slide Images (WSIs). In light of these advancements, we propose Survival Analysis Mamba (SAMamba), a novel approach that melds the Mamba framework with multi-modal survival prediction. Specifically, we propose a patch clustering layer to identify morphological prototypes from the extensive collection of patches within WSIs and employ gene set enrichment analysis to explore the biological associations between pathways and gene sets for enhanced and robust feature representation. Subsequently, we introduce Mamba structures to capture the intrinsic relationships within pathology WSIs and genomic profiles with linear computational complexity. Additionally, we utilize multi-modal attention to seamlessly integrate multi-modal data and design a self-attention pooling module to further refine insights from each data modality for enhanced survival outcome prediction. Extensive experiments on four public TCGA datasets are conducted to validate the effectiveness of our proposed SAMamba, using ablation studies, statistical analysis, and visualization. The experimental results demonstrate that our method achieves superior performance compared to state-of-the-art methods, highlighting the potential of the proposed SAMamba for multi-modal survival outcome prediction. Our code will be released at https://github.com/coffeeNtv/SAMamba.**

***Index Terms*—Computational Pathology, Survival Analysis, Multi-modal Learning, Mamba**

## I. Introduction

Survival prediction seeks to estimate the prognosis of patients by measuring the time span from either diagnosis or the onset of treatment to the occurrence of a particular event of interest, typically a critical endpoint such as death or the relapse of disease. It is essential in clinical oncology and broader medical practice as survival prediction enables clinicians to formulate tailored treatment strategies, assists patients in understanding their prognosis for more informed decision-making, and is crucial in clinical research for evaluating the efficacy of therapies.

Genomic data, with its intricately detailed representation of an individual's genetic blueprint, unearths underlying genetic predispositions and molecular pathways involved in disease progression, which may not be evident through visual examination alone. Concurrently, pathology images provide spatial context that reveals the extent of disease spread, tumor heterogeneity, and interactions between the tumor and its microenvironment, all of which are critical determinants of the disease trajectory.

With the advancement of multi-modal learning, research in survival prediction has shifted from relying on single-modality [1], [2] data to incorporating multiple modalities [3], [4]. Thus, the integration of genomic and pathological data has the potential to revolutionize survival prediction by uncovering correlations and patterns that might be missed when each is analyzed in isolation. Nevertheless, integrating genomic data and pathology images presents several challenges. Firstly, the sheer volume and complexity of both data types require substantial computational power and sophisticated algorithms to extract relevant prognostic features. Secondly, varying data acquisition and processing protocols may introduce discrepancies, necessitating a standardization process to ensure data from various sources can be accurately compared and integrated. Furthermore, the heterogeneity between structured genomic data and unstructured pathology images poses a barrier, making it difficult to assimilate and leverage insights from the two modalities effectively.

Many studies have made strides in addressing the challenges above. For genomic data, many studies [1], [5] focus on the covariates between the genetic and clinical attributes based on the Cox regression to extract relevant information for patient prognosis. Concurrently, due to the gigapixel of pathology images, Multiple Instance Learning (MIL) based models [2], [6], [7] are commonly applied in survival analysis where WSIs are formulated as bags. This line of research underscores the potential for utilizing either genomic or pathology data independently to forecast patient outcomes and paves the way for merging these two modalities. Notably, CLAM [8] and MACT [3] have laid the groundwork for integrative and comparative analyses with both pathology and genomic data and subsequently inspired a burgeoning interest in the field of multi-modal learning in survival prediction [4], [9], [10]. Among multi-modal learning research, the application of attention mechanisms has become progressively prevalent, especially the Transformer [11] architecture, which emerges as the predominant model of choice attributed to its proficiency in handling sequence data. Despite its strengths, Transformer-

based models encounter limitations. The intrinsic design of the attention mechanism, which processes the entire sequence collectively, can lead to a surge in computational complexity, scaling quadratically with the sequence length.

Recent advancements in state space models, such as Mamba [12], have revealed their considerable potential for efficiently handling sequence data such as language and genomics. These models, which maintain competitive performance alongside Transformer-based counterparts, boast scalability with linear computational complexity. Their efficacy also extends to the domain of computer vision for tasks including image segmentation [13], [14], classification [15], [16], object detection [15], [16], restoration [17], video understanding [18], [19] and point cloud analysis [20]. Building on these insights, we proposed Survival Analysis Mamba (SAMamba), a novel framework designed to harness both intrinsic and intertwined interactions from collections of WSIs and genomic attributes for survival analysis. Specifically, we apply the Mamba architecture to model the sequences present in bags of WSIs and the tabular genomic data with linear computational efficiency. Further, we propose a novel patch clustering layer to identify the morphological prototypes from lengthy bags of patches, which reduces the computational complexity significantly, and we use Gene Set Enrichment Analysis (GSEA) [21] to explore the potential biological mechanisms. Additionally, we employ a cross-modality attention strategy to capture the interconnections between genomic and pathological data, thereby enhancing the model's ability to make informed predictions of survival outcomes. Our major contributions are as follows:

- We propose SAMamba, a novel framework in survival prediction that incorporates the Mamba architecture for modeling the intrinsic association within lengthy bags of WSIs and tabular genetic attributes, and a multi-modal attention mechanism to examine the interplay between different modalities for cohesive feature representation.
- We propose a novel patch clustering layer to identify morphological prototypes from a vast number of patches within WSI, effectively reducing the computational complexity and enhancing the robustness of feature extraction from high-dimensional pathology images.
- We apply gene set enrichment analysis to enhance the interpretability and the utility of genomic data by using pathways from a large number of gene sets as our functional categories.
- Extensive experiments are conducted to demonstrate the effectiveness of our method among four public TCGA datasets, and the results indicate that our method achieved superior performance compared to state-of-the-art approaches, confirming the feasibility and improved accuracy of our SAMamba method in multi-modal survival prediction.

## II. Methodology

In this section, we begin by providing an overview of the foundational concepts in the state space model and its variants, such as Mamba. Subsequently, we present the formulation of our task in survival prediction. Lastly, we introduce our method in detail, focusing on the aspects of feature representation and multi-modal learning, respectively.

### A. Fundamentals

**State Space Models.** State Space Models (SSMs) are often regarded as continuous linear time-invariant systems that transform an input signal $x(t) \in \mathbb{R}$ to an output signal $y(t) \in \mathbb{R}$ via implicit latent state $h(t) \in \mathbb{R}^N$. They are usually constructed as linear ordinary differential equations:

$$h'(t) = \mathbf{A}h(t) + \mathbf{B}x(t), \quad y(t) = \mathbf{C}h(t) \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{N\times N}$ and $\mathbf{B}, \mathbf{C} \in \mathbb{R}^N$ are evolution and projection parameters, respectively. Subsequently, in the Structured State Space Sequence model (S4) [22], the zero-order hold (ZOH) (Equation 2) [22] is applied to transform the "continuous parameters" $(\Delta, A, B)$ into "discrete versions" $(\bar{A}, \bar{B})$ to incorporate deep learning algorithms.

$$\bar{A} = \exp(\Delta A), \quad \bar{B} = (\Delta A)^{-1}(\exp(\Delta A) - I) \cdot \Delta B \tag{2}$$

The discretized system (from $(\Delta, A, B, C) \mapsto (\bar{A}, \bar{B}, C)$) can be expressed in a linear recurrent way:

$$h'(t) = \bar{\mathbf{A}}h(t) + \bar{\mathbf{B}}x(t), \quad y(t) = \mathbf{C}h(t) \tag{3}$$

As an alternative, Equation 3 can be further formulated as a global convolution:

$$\overline{\boldsymbol{K}} = \left(C\overline{\boldsymbol{B}}, C\overline{\boldsymbol{A}\boldsymbol{B}}, \ldots, C\overline{\boldsymbol{A}}^{l-1}\overline{\boldsymbol{B}}\right), \quad y = x * \overline{\boldsymbol{K}} \tag{4}$$

where $l$ is the length of the input sequence and $\bar{\boldsymbol{K}}$ represents a structured convolutional kernel.

**Mamba.** Regarding the Linear Time Invariance (LTI) systems in Equations 1 to 4 where parameters are invariant for all inputs, Mamba [12] proposed selection scan mechanism to address the constant dynamics in LTI systems. Parameters in Mamba, such as $B$ and $C$, are derived from inputs, which enable Mamba to have dynamic interactions along the input sequence.

### B. Survival Prediction Formulation

Survival prediction aims to estimate the time until an event of interest occurs, such as death. In this study, given the clinical information of patient $X = \{\mathcal{H}, \mathcal{G}, t, c\}$, our goal is to estimate the hazard function $f_{hazard}(T = t \mid T \geq t, X) \in [0, 1]$ which represents the instantaneous risk rate of the event occurring at time $t$. It is defined as follows:

$$f_{hazard}(T = t) = \lim_{\Delta t \to 0} \frac{\mathrm{P}(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \tag{5}$$

Specifically, $\mathcal{H}$ denotes the histological WSI, $\mathcal{G}$ denotes the genomic profile, $t$ is the overall survival time (in months), $c \in \{0, 1\}$ is the right censorship, where $c = 1$ indicates that death has not occurred within the observation period in this study. Instead of modeling the overall survival time of patients,

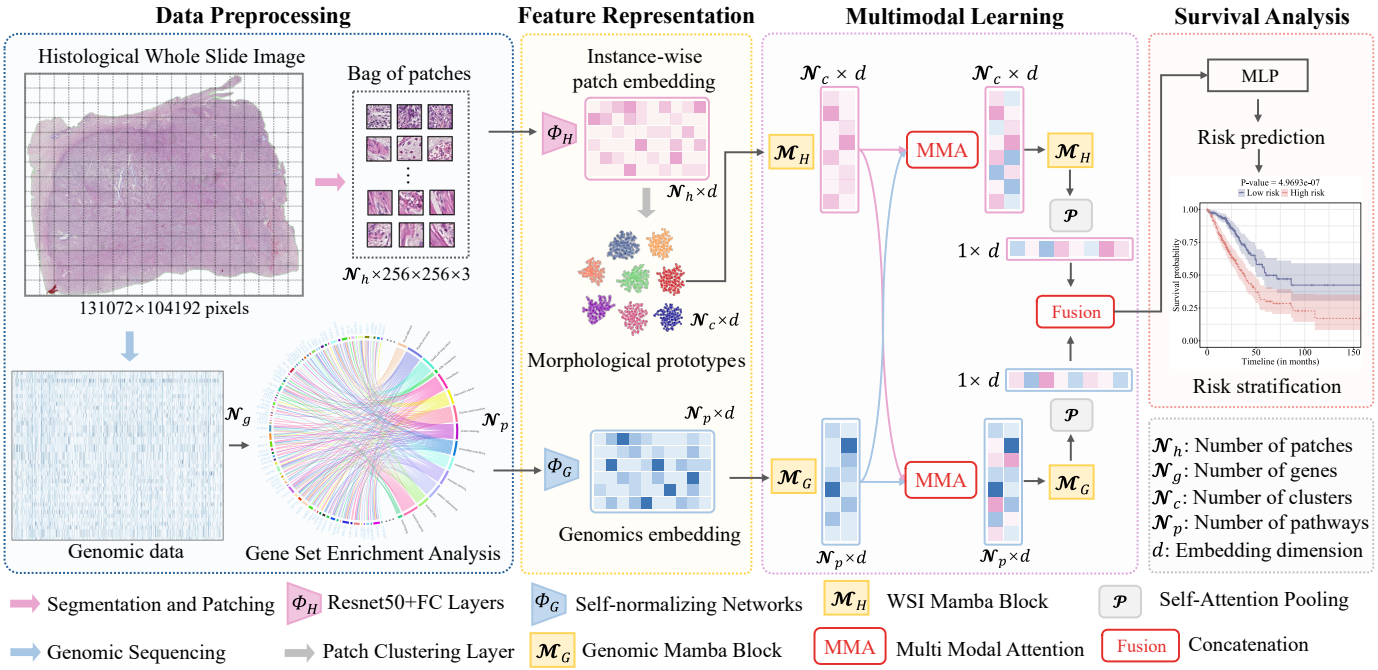


Fig. 1. Overview of SAMamba. Histological WSIs are segmented into patches and processed through a pre-trained ResNet50 model with a patch clustering layer to generate morphological prototypes. For genomic data, gene set enrichment analysis is utilized to identify enriched pathways, which are subsequently processed by a self-normalizing network to obtain genomic embeddings. Mamba blocks and multi-modal attention mechanisms are employed to model intra-modality associations and inter-modality interactions. Finally, self-attention pooling and feature fusion, followed by an MLP, are used to predict survival risk (Zoom in for better view).

we measure the probability of the total survival time longer than $t$ with the cumulative distribution function defined as:

$$f_{survival}(T \geq t, X) = \prod_{u=1}^{t}(1 - f_{hazard}(T = u \mid T \geq u, X)). \tag{6}$$

### C. Feature Representation

**Pathology images embedding.** Building upon the MIL paradigm, we consider the pathological WSIs as collections of permutation-invariant instances. In particular, we adopt the procedure established by CLAM [8], which starts by segmenting the tissue within WSIs and dividing it into non-overlapping $256 \times 256$ pixel patches. These patches are then processed through a pre-trained ResNet50 network [23], which projects them into a feature space of 1024 dimensions. Subsequently, these features are passed through fully connected layers to synthesize a set of patch features within WSIs. We denote the $i_{th}$ WSI as $\mathcal{H}^{(i)} = \{h_j^{(i)}\}_{j=1}^{\mathcal{N}_h} \in \mathbb{R}^{\mathcal{N}_h \times d}$, where $\mathcal{N}_h$ is the total number of instances within the bag of WSI, and $d$ represents the dimensionality of each instance feature. Given that the original WSIs contain gigapixels, each WSI includes a vast and diverse number of patches. Loading all these patches during training is memory-intensive and impractical. This necessitates the identification of representative patches from the WSI for feature selection and reduction. As illustrated in Figure 2, many patches within WSI share similar features and tend to cluster together. Based on these observation, we design a Patch Clustering Layer (PCL) to select the morphological prototypes from WSIs and to unify the number of patch features among all WSIs. Our PCL is described in Algorithm 1, where $B$ and $\mathcal{N}_c$ are the batch size and the number of clusters in PCL, respectively.

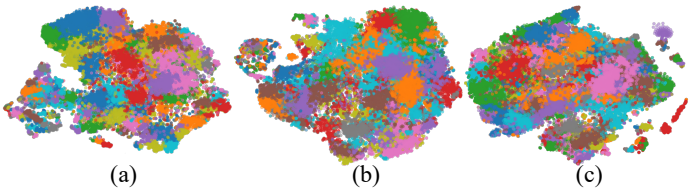


Fig. 2. The t-SNE visualizations of patch features within BRCA WSIs. (a-c) are three different examples to illustrate clustered features and the rationale behind the patch clustering layer design.

**Algorithm 1** Patch Clustering Layer (PCL)

**Input:** $x$: Bags of patches from WSIs # $(B, \mathcal{N}_h, d)$
**Output:** $y$: Feature clusters from patches # $(B, \mathcal{N}_c, d)$
1: Initialize cluster centers $\in \mathbb{R}^{\mathcal{N}_c \times d}$
2: $d_{i,j} \leftarrow \|x_i - \text{centers}_j\|$ # The L2 distances, $(B, d, \mathcal{N}_c)$
3: indices $\leftarrow \min(d, \text{dim} = 1)$ # Find nearest features to cluster centers
4: $y \leftarrow \text{gather}(x, 1, \text{indices})$ # Gather the nearest features
5: **return** $y$

**Genomics embedding.** The genomic profiles, including the mutation statuses, copy number variations, and levels of transcripts (quantified through bulk RNA-Seq), are typically represented as single-value measurements, which offer a limited view if incorporated into the survival analysis task alone. To enhance the interpretability and the utility of the tabular genomics data, we apply GSEA [21], [24] to segment genomic data into many pathways. Specifically, we use the Kyoto Encyclopedia of Genes and Genomes (KEGG) [25] as our reference pathway database, and we perform enrichment analysis to identify enriched pathways based on hypergeometric distribution with False Discovery Rate (FDR) correction for multiple testing. Pathways are considered statistically significant based on a threshold of FDR $\leq 0.05$. Subsequently, these pathways are processed through a Self-Normalizing Network (SNN) [26] to derive a more informative feature representation of the genomic data. For the $i_{th}$ patient, its genomic representation is denoted as $\mathcal{G}^{(i)} = \{g_j^{(i)}\}_{j=1}^{\mathcal{N}_p} \in \mathbb{R}^{\mathcal{N}_p \times d}$ , where $d$ represents the dimension of each pathway, and $\mathcal{N}_p$ denotes the total number of pathways.

### *D. Multi-modal Learning*

**Mamba Block.** Considering the heterogeneity in pathology WSIs and genomic profiles, we incorporate Mamba, an emerging architecture that excels in modeling sequence data, to investigate the intrinsic correlation within each modality. The pseudocode of our Mamba blocks is shown in Algorithm 2, where $T_L$ denotes the output from the $L_{th}$ mamba block, $B$,$\mathcal{N}$, and $d$ denote the batch size, sequence length, and dimension of input tokens, respectively.

**Algorithm 2** Mamba Block

**Input:** input sequence $T_{L-1}$ # (B, $\mathcal{N}$, d)
**Output:** output sequence $T_L$ # (B, $\mathcal{N}$, d)

$T'_{L-1} \leftarrow \text{Norm}(T_{L-1})$ # (B, $\mathcal{N}$, d)
$x, z \leftarrow \text{Linear}^x(T'_{L-1}), \text{Linear}^z(T'_{L-1})$ # (B, $\mathcal{N}$, E)
$x' \leftarrow \text{SiLU}(\text{Conv1d}(x))$ # (B, $\mathcal{N}$, E)
$B, C \leftarrow \text{Linear}^B(x'), \text{Linear}^C(x')$ # (B, $\mathcal{N}$, N)
$\Delta \leftarrow \log(1 + \exp(\text{Linear}^\Delta(x') + \text{Parameter}^\Delta))$ #(B,$\mathcal{N}$,E)
$\bar{A} \leftarrow \Delta \otimes \text{Parameter}^A$ # Outer product, (B, $\mathcal{N}$, E, N)
$\bar{B} \leftarrow \Delta \otimes B$ # (B, $\mathcal{N}$, E, N)
$y \leftarrow \text{SSM}(\bar{A}, \bar{B}, C)(x')$ # (B, $\mathcal{N}$, E)
$y' \leftarrow y \odot \text{SiLU}(z)$ # Element-wise product, (B, $\mathcal{N}$, E)
$T_L \leftarrow \text{Linear}^T(y') + T_{L-1}$ # (B, $\mathcal{N}$, d)
Return $T_L$

**Multi-modal Attention.** To capture the associations between histological and genomic data, Multi-Modal Attention (MMA) is applied to account for the interaction between different modalities for further feature refinement:

$$\text{MMA}(\mathcal{G}_{en}, \mathcal{H}_{en}, \mathcal{H}_{en}) = \text{Softmax}\left(\frac{\mathcal{G}_{en}\mathcal{H}_{en}^T}{\sqrt{d_h}}\right)\mathcal{H}_{en} \quad (7)$$

$$\text{MMA}(\mathcal{H}_{en}, \mathcal{G}_{en}, \mathcal{G}_{en}) = \text{Softmax}\left(\frac{\mathcal{H}_{en}\mathcal{G}_{en}^T}{\sqrt{d_g}}\right)\mathcal{G}_{en} \quad (8)$$

where $d_g$ and $d_h$ are dimension of $\mathcal{G}_{en}$ (Encoded gene feature) and $\mathcal{H}_{en}$ (Encoded histological image feature), respectively.

**Self Attention Pooling.** We design a Self Attention Pooling (SAP) operation to condense the inherent correlations within each modality and reduce the dimensionality of the refined features for feature fusion. The pseudocode of SAP is demonstrated in Algorithm 3.

**Algorithm 3** Self Attention Pooling (SAP)

**Input:** $x$: input features # $(B, \mathcal{N}, d)$
**Output:** $x_{pool}$: pooling features # $(B, d)$

Initialization of query $\in \mathbb{R}^{1\times d}$
$x_{\text{proj}} \leftarrow \text{fc}(x)$ # Projection of $x$, $(B, \mathcal{N}, d)$
att_scores $\leftarrow \text{matmul}(x_{\text{proj}}, \text{query}^T)$ # Compute attention scores, $(B, \mathcal{N}, 1)$
att_weights $\leftarrow \text{Softmax}(\text{att\_scores})$ # Compute attention weights using Softmax
$x_{pool} \leftarrow \text{matmul}(\text{att\_weights}, x)$ # Compute weighted sum, $(B, d)$
**return** $x_{pool}$

Our SAMamba framework is presented in Algorithm 4. Initially, a patch clustering layer is employed to extract representative patch features. Subsequently, refined pathway features and patch features are input into Mamba blocks to explore the intrinsic attributes within each modality. Multi-modal attention is then utilized to capture the interplay between modalities, accompanied by a self-attention pooling module for feature condensing. Finally, concatenation is applied for multi-modalfeature fusion. The risk score for each case is predicted using a multilayer perceptron (MLP).

**Algorithm 4** The SAMamba Algorithm

**Input:** $\mathcal{H}$: Bags of WSI patches # (B, $\mathcal{N}_h$, d)
**Input:** $\mathcal{G}$: Genomics # (B, $\mathcal{N}_p$, d)
**Output:** Risk scores

$\mathcal{H}_c \leftarrow PCL(\mathcal{H})$# Clustering for WSI patches, Alg. 1
$\mathcal{H}_{en}, \mathcal{G}_{en} \leftarrow \mathcal{M}_H(\mathcal{H}_c), \mathcal{M}_G(\mathcal{G})$ # Mamba encoding
$\mathcal{H}_{att} \leftarrow \text{MMA}(\mathcal{G}_{en}, \mathcal{H}_{en}, \mathcal{H}_{en})$ # Multi-modal attention for WSI patches, Eq. 7
$\mathcal{G}_{att} \leftarrow \text{MMA}(\mathcal{H}_{en}, \mathcal{G}_{en}, \mathcal{G}_{en})$ # Multi-modal attention for genomics, Eq. 8
$\mathcal{H}_{de}, \mathcal{G}_{de} \leftarrow \mathcal{M}_H(\mathcal{H}_{att}), \mathcal{M}_G(\mathcal{G}_{att})$ # Mamba decoding
$\mathcal{H}_{pool}, \mathcal{G}_{pool} \leftarrow \mathcal{P}(\mathcal{H}_{de}), \mathcal{P}(\mathcal{G}_{de})$ # SAP, Alg. 3
$\mathcal{F} \leftarrow \mathcal{H}_{pool} \oplus \mathcal{G}_{pool}$ # multi-modal concatenation
Risk scores $\leftarrow \text{MLP}(\mathcal{F})$ # Risk prediction by MLP
**return** Risk scores

### *E. Objective function*

We apply the Negative Log-Likelihood loss (NLL) in previous work [3], [27] for survival prediction as our object function:

$$\mathcal{L}_{\text{survival}} = -\sum_{i=1}^{K} c_i \cdot \log\left(f_{\text{survival}}\left(t_i \mid \mathcal{F}_i\right)\right) - \sum_{i=1}^{K} (1-c_i) \cdot \log\left(f_{\text{survival}}\left(t_i - 1 \mid \mathcal{F}_i\right)\right) - \sum_{i=1}^{K} (1-c_i) \cdot \log\left(f_{\text{hazard}}\left(t_i \mid \mathcal{F}_i\right)\right) \tag{9}$$

where $K$ refers to the length of the dataset, $f_{hazard}$ is the hazard function defined in Equation 5, $f_{survival}$ is the cumulative distribution function defined in Equation 6. $\mathcal{F}$ is the integrated multi-modal featured shown in Algorithm 4, $c$ indicates the censorship.

## III. Experiments

### A. Datasets

The Cancer Genome Atlas (TCGA)[1] is a landmark cancer genomics program that represents a comprehensive and publicly accessible resource for the cancer research community, providing a multi-dimensional view of the genomics changes and their corresponding WSIs across various cancer types. In this study, four datasets from TCGA are used to validate the efficacy of our proposed SAMamba model, i.e. bladder urothelial carcinoma (BLCA) ($n$ =412), breast invasive carcinoma (BRCA) ($n$ =1093), lung adenocarcinoma (LUAD) ($n$ =514), and uterine corpus endometrial carcinoma (UCEC) ($n$ =560).

### B. Evaluation

In our study, we conducted a 5-fold cross-validation for each cancer type and used the cross-validated concordance index (C-index) [33] to measure the predictive accuracy of a risk score in survival analysis. It assesses the proportion of all usable patient pairs where the predictions and outcomes are concordant. Mathematically, for a set of $n$ independent patient pairs, the C-index is defined as:

$$C - index = \frac{1}{N} \sum_{i}^{n} \sum_{j>i}^{n} \left(\mathbb{I}(t_i < t_j) \cdot \mathbb{I}(r_i > r_j) \cdot (1 - c_i)\right) \tag{10}$$

Where $t_i$ and $t_j$ are the observed survival times of patients $i$ and $j$, with $i \neq j$, $r_i$ and $r_j$ are predicted risk scores, $\mathbb{I}(\cdot)$ is the indicator function, which equals 1 when its argument is true and 0 otherwise. $N$ is the number of all possible pairs of patients where $t_i \neq t_j$, $c_i$ refers to the right censorship.

### C. Implementation

For the pathological WSIs data, we apply a pre-trained ResNet50 [23] to derive the 1024-dim feature as CLAM [8], then fully connected layers are used to obtain 256-dim feature embeddings. For genomic attributes, SNN [8], [26] is employed to extract their features into 256 dimensions. We adopt the RAdam optimizer in our training process with an initial learning rate set at 2e-4 and weight decay of 1e-5. The number of clusters $\mathcal{N}_c$ in the PCL is 256, and the number of pathways $\mathcal{N}_p$ ranges from 188 to 285 for different cancer types. Our architecture includes a single Mamba block for both the encoding and decoding of each modality data. We set the batch size as one due to variability in bag sizes across different WSIs and trained our model for 30 epochs with one NVIDIA A800 GPU.

[1]TCGA datasets can be found at https://portal.gdc.cancer.gov

### D. Experimental Results

We compare our method with single-modal and multi-modal benchmark methods as follows:

**(1) Single-modal methods.** For genomic data, we adopt the SNN [3], [26], DeepSurv [5], [9] and CoxRegression [9], [28] as compared models. Additionally, for the pathology WSIs, we assess the performance of our model by comparing it with five State-Of-The-Art (SOTA) methods: AttentionMIL [7], DeepAttnMISL [2], DeepGraphConv [30], DeepAttnMISL [2], CLAM [8] and Patch-GCN [6].

**(2) Multi-modal methods.** For integrating multimodality data for survival predictions, we compare our SAMamba with four multi-modal SOTA methods, including MCAT [3], PORPOISE [10], MGCT [9], MOTCat [31], CMTA [32] and Survpath [4].

The results shown in Table I indicate that our method outperforms both the single-modal and multi-modal SOTA methods in BRCA (70.05%), UCEC (71.27%) and LUAD (69.3%) datasets, and achieves competitive performance in BLCA (66.06%) dataset, indicating the effectiveness of integrating multi-modal data, and the great potential of Mamba structure in multi-modal survival analysis. We observe that CMAT outperforms our approach on the BLCA dataset. Upon analysis, this superior performance might attribute to CMAT's utilization of all patches within WSIs during training. In contrast, our method is limited to a fixed number of patches due to computational resource constraints. This difference in patch utilization likely contributes to the better predictive performance of CMTA for BLCA. Note that we use experimental results from original studies to ensure fair comparisons. Results for Survpath [4] are only available for BRCA and BLCA datasets among the datasets we compared.

### E. Ablation Studies

To further validate the effectiveness of our SAMamba model in survival prediction, we conduct three ablation experiments with respect to genomics data, pathology WSIs, and feature fusion, respectively. In the first experiment, we train three SAMamba models with different genomic data categorization strategies, namely the gene family [3], hallmark [4] and GSEA [21], to explore the effectiveness of different methods for genomic data integration. Concurrently, we investigate the performance of our patch clustering layer with three settings: (1) Randomly select $\mathcal{N}_c$ features to simulate the absence of the PCL; (2) Using the average feature of each cluster as the feature of interest; (3) Taking the nearest feature to each clustering center as the representative feature. Furthermore,

TABLE I

RESULTS COMPARED WITH DIFFERENT BENCHMARK METHODS WITH DIFFERENT MODALITY WHERE THE BEST RESULTS ARE **HIGHLIGHTED**.

| | Method | BRCA (↑) | BLCA (↑) | UCEC (↑) | LUAD (↑) | Mean (↑) |
|---|---|---|---|---|---|---|
| Genes | SNN [3], [26] (NIPS 2017) | 0.466 ± 0.058 | 0.541 ± 0.016 | 0.493 ± 0.096 | 0.539 ± 0.069 | 0.5098 |
| | DeepSurv [5], [9] (BMC MRM 2018) | 0.598 ± 0.054 | 0.567 ± 0.049 | 0.577 ± 0.058 | 0.608 ± 0.026 | 0.5875 |
| | CoxRegression [9], [28] (JMLR 2018) | 0.568 ± 0.077 | 0.591 ± 0.041 | 0.464 ± 0.099 | 0.574 ± 0.042 | 0.5493 |
| | SAMamba (Ours) | **0.667 ± 0.0221** | **0.6001 ± 0.0286** | **0.6768 ± 0.0519** | **0.6167 ± 0.0179** | **0.6402** |
| WSIs | Deep sets [3], [29] (NIPS 2017) | 0.500 ± 0.000 | 0.500 ± 0.000 | 0.500 ± 0.000 | 0.496 ± 0.008 | 0.4990 |
| | AttentionMIL [3], [7] (ICML 2018) | 0.564 ± 0.050 | 0.536 ± 0.038 | 0.625 ± 0.057 | 0.559 ± 0.060 | 0.5710 |
| | DeepGraphConv [6], [30] (MICCAI 2018) | 0.574 ± 0.044 | 0.499 ± 0.057 | 0.659 ±0.056 | 0.552 ± 0.058 | 0.5710 |
| | DeepAttnMISL [2], [3] (MIA 2020) | 0.524 ± 0.043 | 0.504 ± 0.042 | 0.597 ± 0.059 | 0.548 ± 0.050 | 0.5433 |
| | CLAM [8], [9] (NBE 2021) | 0.578 ± 0.032 | 0.565 ± 0.027 | 0.609 ± 0.082 | 0.582 ± 0.072 | 0.5835 |
| | Patch-GCN [6] (MICCAI 2021) | 0.580 ± 0.025 | 0.560 ± 0.034 | 0.629 ± 0.052 | 0.585 ± 0.012 | 0.5885 |
| | SAMamba (Ours) | **0.596 ± 0.0274** | **0.652 ± 0.0364** | **0.673 ± 0.0444** | **0.643 ± 0.0181** | **0.6410** |
| Multimodal | Deep sets [3], [29] (NIPS 2017) | 0.521 ± 0.079 | 0.604 ± 0.042 | 0.598 ± 0.077 | 0.616 ± 0.027 | 0.5848 |
| | AttentionMIL [3], [7] (ICML 2018) | 0.551 ± 0.077 | 0.605 ± 0.045 | 0.614 ± 0.052 | 0.563 ± 0.050 | 0.5833 |
| | DeepAttnMISL [2], [3] (MIA 2020) | 0.545 ± 0.071 | 0.611 ± 0.049 | 0.615 ± 0.020 | 0.595 ± 0.061 | 0.5915 |
| | MCAT [3] (ICCV 2021) | 0.580 ± 0.069 | 0.624 ± 0.034 | 0.622 ± 0.019 | 0.620 ± 0.032 | 0.6115 |
| | PORPOISE [10] (Cancer Cell 2022) | 0.583 ± 0.048 | 0.644 ± 0.036 | 0.688 ± 0.096 | 0.616 ± 0.063 | 0.6328 |
| | MGCT [9] (BIBM 2023) | 0.608 ± 0.026 | 0.640 ± 0.039 | 0.645 ± 0.039 | 0.596 ± 0.078 | 0.6223 |
| | MOTCat [31] (ICCV 2023) | 0.673 ± 0.006 | 0.683 ± 0.026 | 0.675 ± 0.04 | 0.67 ± 0.038 | 0.6753 |
| | CMTA [32] (ICCV 2023) | 0.6679 ± 0.0434 | **0.691 ± 0.0426** | 0.6975 ± 0.0409 | 0.6864 ± 0.0359 | 0.6857 |
| | Survpath [4] (CVPR 2024) | 0.640 ± 0.093 | 0.628 ± 0.073 | - | - | - |
| | SAMamba (Ours) | **0.7005 ± 0.0265** | 0.6606 ± 0.017 | **0.7127 ± 0.0328** | **0.693 ± 0.0309** | **0.6917** |

TABLE II

ABLATION STUDIES ON GENE TOKEN, PATCH CLUSTERING LAYER, FUSION, AND BACKBONE BLOCKS. THE BEST RESULTS ARE **HIGHLIGHTED**.

| Ablations | Method | BRCA (↑) | BLCA (↑) | UCEC (↑) | LUAD (↑) | Mean (↑) |
|---|---|---|---|---|---|---|
| Gene token | Gene family | 0.6777 ± 0.0384 | **0.6723 ± 0.036** | 0.6593 ± 0.0458 | 0.6781 ± 0.0407 | 0.6719 |
| | Hallmark | 0.6931 ± 0.0344 | 0.668 ± 0.0303 | 0.6803 ± 0.0383 | 0.6637 ± 0.0459 | 0.6763 |
| | GSEA (Ours) | **0.7005 ± 0.0265** | 0.6606 ± 0.017 | **0.7127 ± 0.0328** | **0.693 ± 0.0309** | **0.6917** |
| PCL | w/o | 0.673 ± 0.0423 | 0.6473 ± 0.0216 | 0.6751 ± 0.0543 | 0.684 ± 0.023 | 0.6699 |
| | Average | 0.6771 ± 0.0193 | 0.6497 ± 0.0415 | 0.6451 ± 0.0632 | 0.6832 ± 0.0393 | 0.6638 |
| | Top (Ours) | **0.7005 ± 0.0265** | **0.6606 ± 0.017** | **0.7127 ± 0.0328** | **0.693 ± 0.0309** | **0.6917** |
| Fusion | Bilinear | 0.6081 ± 0.0395 | 0.6087 ± 0.0072 | 0.6795 ± 0.0404 | 0.6379 ± 0.0385 | 0.6336 |
| | Concat (Ours) | **0.7005 ± 0.0265** | **0.6606 ± 0.017** | **0.7127 ± 0.0328** | **0.693 ± 0.0309** | **0.6917** |
| Backbone | Transformer | 0.6667 ± 0.0319 | 0.6472 ± 0.0303 | 0.6851 ± 0.0505 | 0.6877 ± 0.0208 | 0.6717 |
| | Mamba (Ours) | **0.7005 ± 0.0265** | **0.6606 ± 0.017** | **0.7127 ± 0.0328** | **0.693 ± 0.0309** | **0.6917** |

TABLE III

COMPUTATIONAL COMPARISON OF SAMAMBA.

| Types | FLOPs (M) | # Para (M) | Time (s) |
|---|---|---|---|
| Mamba | 283.116 | 1.752 | 591 |
| Transformer | 570 | 2.108 | 874 |

we also evaluate different multi-modal feature fusion methods, such as bilinear pooling and concatenation, to determine the most effective fusion strategy. Additionally, we substitute all Mamba blocks with Transformer blocks in SAMamba and calculate the floating point operations (FLOPs), the number of parameters for both Mamba and Transformer blocks, and the inference time across the LUAD cohort in Table III to examine their performance and computational efficiency. The results in Table II indicate that the integration of multi-modal data leads to notable enhancements compared to using single-modality data in survival prediction. Notably, we obtained the best performance in the BLCA dataset when applying the gene family strategy. However, we choose the categorization strategy for genomic data based on the average C-index value of four cancer types in Table II, indicating that GSEA can be effective in SAMamba. We also found that using the features nearest to each clustering center boosts the performance of our model more than other approaches. Moreover, simple concatenation emerges as the superior method in our task among the fusion strategies tested. Further, the results in Table II and II indicate that our SAMamba utilizing Mamba blocks outperforms the Transformer-based SAMamba in both prediction performance and computational efficiency.

### *F. Visualizations with Statistical Analysis and Interpretation*

We apply the Kaplan-Meier curve to visualize the survival experience of a cohort over time in Figure 3. The curve represents the probability that a patient will survive beyond a certain time point, assuming that the event of interest is death. Each step downward in the Kaplan-Meier curve indicates the occurrence of an event, and a horizontal line to the next

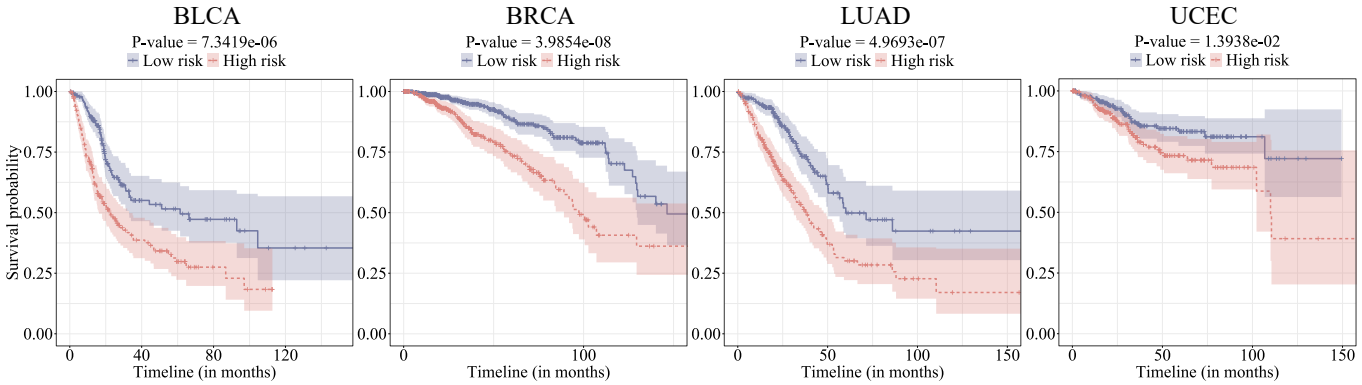


Fig. 3. Patient stratification with Kaplan-Meier curves and log-rank tests among different cancers (Zoom in for better view).

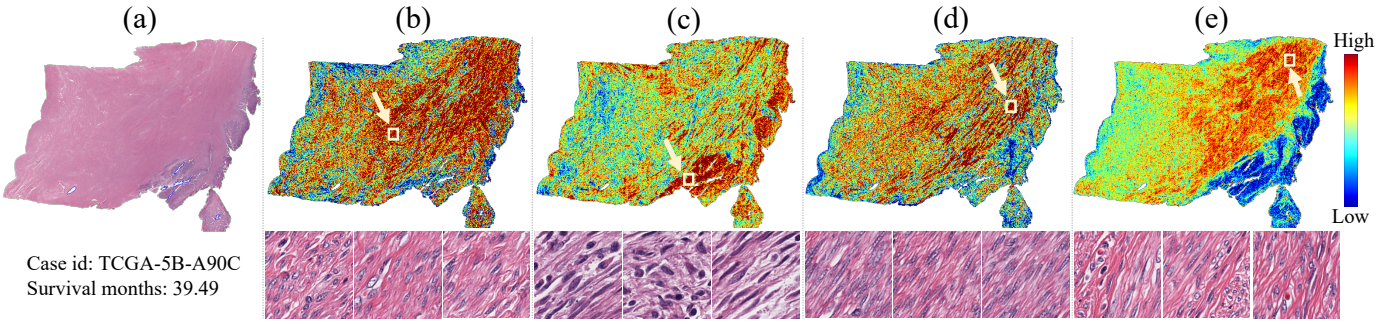


Fig. 4. Co-attention maps of pathways on WSI with the highly associated patches in UCEC. (a) The WSI. (b) Focal adhesion. (c) PD-L1 expression and PD-1 checkpoint pathway in cancer. (d) FoxO signaling pathway. (e) TGF-beta signaling pathway.

event signifies that subjects have survived up to that point. Specifically, we divide samples into high-risk and low-risk groups based on their median risk values, sort these samples by their survival times in ascending order, and calculate the survival probabilities for each risk group over the survival timeline to draw the curve. Subsequently, the log-rank test is performed to assess the statistical significance of differences between the two survival curves. As shown in Figure 3, the p-values in our log-rank test among four datasets are all below 0.05, indicating that the difference in survival between the groups is considered statistically significant.

Further, we present the co-attention maps of pathways and WSIs in Figure 4. Specifically, we selected four pathways in three functional categories regarding to the biological mechanisms of uterine corpus endometrial carcinoma. Namely, the PD-L1 expression and PD-1 checkpoint pathway in cancer [34] (Human diseases), TGF-beta signaling pathway [35] and FoxO signaling pathway [36] (Signal transduction), and Focal adhesion [37], [38] (Cellular processes). We highlight patches corresponding to highly correlated regions within the WSIs to illustrate the patterns and associated locations where different pathways may exert their influence on the WSIs.

## IV. Conclusion

This paper proposes SAMamba, a pioneering study incorporating the Mamba structure for enhanced multi-modal survival prediction. Specifically, we propose a novel patch clustering layer to extract representative features from a vast number of patches within whole slide images, effectively reducing the computational complexity and enhancing the robustness of feature extraction from high-dimensional pathology images. Then, we apply gene set enrichment analysis to explore the biological mechanisms between pathways and large gene sets, providing a deeper understanding of the genomic contributions to survival outcomes. Additionally, we incorporate Mamba structures to capture intrinsic relationships within pathology WSIs and genomic profiles with linear computational complexity, enabling efficient and scalable analysis of multi-modal data. Furthermore, we utilize a multi-modal attention mechanism to seamlessly integrate data from different modalities and design a self-attention pooling module to extract the most informative features from each data modality, ensuring a cohesive and comprehensive analysis for more accurate survival predictions. Extensive experiments are conducted on four public TCGA datasets to demonstrate the effectiveness of our proposed SAMamba through ablation studies, statistical analysis, and visualizations. The experiment results show that our method achieves superior performance to state-of-the-art methods, indicating the feasibility and clinical utility of the proposed SAMamba for multi-modal survival outcome prediction.

## REFERENCES


[1] J.-y. Liang, D.-s. Wang, H.-c. Lin, X.-x. Chen, H. Yang, Y. Zheng, and Y.-h. Li, "A novel ferroptosis-related gene signature for overall survival prediction in patients with hepatocellular carcinoma," *International journal of biological sciences*, vol. 16, no. 13, p. 2430, 2020.

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

[3] R. J. Chen, M. Y. Lu, W.-H. Weng, T. Y. Chen, D. F. Williamson, T. Manz, M. Shady, and F. Mahmood, "Multimodal co-attention transformer for survival prediction in gigapixel whole slide images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4015–4025.

[4] G. Jaume, A. Vaidya, R. Chen, D. Williamson, P. Liang, and F. Mahmood, "Modeling dense multimodal interactions between biological pathways and histology for survival prediction," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[5] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, "Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network," *BMC medical research methodology*, vol. 18, pp. 1–12, 2018.

[6] R. J. Chen, M. Y. Lu, M. Shaban, C. Chen, T. Y. Chen, D. F. Williamson, and F. 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*. Springer, 2021, pp. 339–349.

[7] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 2127–2136.

[8] M. Y. Lu, D. F. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature Biomedical Engineering*, vol. 5, no. 6, pp. 555–570, 2021.

[9] M. Liu, Y. Liu, H. Cui, C. Li, and J. Ma, "Mgct: Mutual-guided cross-modality transformer for survival outcome prediction using integrative histopathology-genomic features," in *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2023, pp. 1306–1312.

[10] R. J. Chen, M. Y. Lu, D. F. Williamson, T. Y. Chen, J. Lipkova, Z. Noor, M. Shaban, M. Shady, M. Williams, B. Joo *et al.*, "Pan-cancer integrative histology-genomic analysis via multimodal deep learning," *Cancer Cell*, vol. 40, no. 8, pp. 865–878, 2022.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[12] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[13] J. Ruan and S. Xiang, "Vm-unet: Vision mamba unet for medical image segmentation," *arXiv preprint arXiv:2402.02491*, 2024.

[14] Z. Wang *et al.*, "Mamba-unet: Unet-like pure visual mamba for medical image segmentation," *arXiv preprint arXiv:2402.05079*, 2024.

[15] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision mamba: Efficient visual representation learning with bidirectional state space model," *arXiv preprint arXiv:2401.09417*, 2024.

[16] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, and Y. Liu, "Vmamba: Visual state space model," *arXiv preprint arXiv:2401.10166*, 2024.

[17] Y. Shi, B. Xia, X. Jin, X. Wang, T. Zhao, X. Xia, X. Xiao, and W. Yang, "Vmambair: Visual state space model for image restoration," *arXiv preprint arXiv:2403.11423*, 2024.

[18] K. Li, X. Li, Y. Wang, Y. He, Y. Wang, L. Wang, and Y. Qiao, "Videomamba: State space model for efficient video understanding," *arXiv preprint arXiv:2403.06977*, 2024.

[19] G. Chen, Y. Huang, J. Xu, B. Pei, Z. Chen, Z. Li, J. Wang, K. Li, T. Lu, and L. Wang, "Video mamba suite: State space model as a versatile alternative for video understanding," *arXiv preprint arXiv:2403.09626*, 2024.

[20] J. Liu, R. Yu, Y. Wang, Y. Zheng, T. Deng, W. Ye, and H. Wang, "Point mamba: A novel point cloud backbone based on state space model with octree-based ordering strategy," *arXiv preprint arXiv:2403.06467*, 2024.

[21] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander *et al.*, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences*, vol. 102, no. 43, pp. 15 545–15 550, 2005.

[22] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," in *The International Conference on Learning Representations (ICLR)*, 2022.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[24] S. X. Ge, D. Jung, and R. Yao, "Shinygo: a graphical gene-set enrichment tool for animals and plants," *Bioinformatics*, vol. 36, no. 8, pp. 2628–2629, 2020.

[25] M. Kanehisa and S. Goto, "Kegg: kyoto encyclopedia of genes and genomes," *Nucleic acids research*, vol. 28, no. 1, pp. 27–30, 2000.

[26] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks," *Advances in neural information processing systems*, vol. 30, 2017.

[27] S. G. Zadeh and M. Schmid, "Bias in cross-entropy-based training of deep survival networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 9, pp. 3126–3137, 2020.

[28] H. Kvamme, Ø. Borgan, and I. Scheel, "Time-to-event prediction with neural networks and cox regression," *Journal of machine learning research*, vol. 20, no. 129, pp. 1–30, 2019.

[29] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov, and A. J. Smola, "Deep sets," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.

[30] R. Li, J. Yao, X. Zhu, Y. Li, and J. Huang, "Graph cnn for survival analysis on whole slide pathological images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 174–182.

[31] Y. Xu and H. Chen, "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*, 2023, pp. 21 241–21 251.

[32] F. Zhou and H. Chen, "Cross-modal translation and alignment for survival analysis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21 485–21 494.

[33] F. E. Harrell Jr, K. L. Lee, and D. B. Mark, "Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors," *Statistics in medicine*, vol. 15, no. 4, pp. 361–387, 1996.

[34] S. Song, H. Gu, J. Li, P. Yang, X. Qi, J. Liu, J. Zhou, Y. Li, and P. Shu, "Identification of immune-related gene signature for predicting prognosis in uterine corpus endometrial carcinoma," *Scientific Reports*, vol. 13, no. 1, p. 9255, 2023.

[35] W. H. Bae, J. Y. Hwang, W. K. Hur, M. Nam, Y. Choi, L. Kim, Y. H. Lee, W. Cheng, E. Kim, E. Yu *et al.*, "Tgf-$\beta$ signaling pathway-related gene mutations are associated with increased neoantigen counts, enhanced cytolytic activity, and improved survival outcomes in tp53-mutated endometrial carcinoma," *Cancer Research*, vol. 81, no. 13_Supplement, pp. 452–452, 2021.

[36] J. Jiang, C. Zhang, J. Wang, Y. Zhu, X. Wang, and P. Mao, "Knockdown of prom2 enhances paclitaxel sensitivity in endometrial cancer cells by regulating the akt/foxo1 pathway," *Anti-Cancer Agents in Medicinal Chemistry (Formerly Current Medicinal Chemistry-Anti-Cancer Agents)*, vol. 23, no. 19, pp. 2127–2134, 2023.

[37] C. Yan, L. He, Y. Ma, J. Cheng, L. Shen, R. K. Singla, and Y. Zhang, "Establishing and validating an innovative focal adhesion-linked gene signature for enhanced prognostic assessment in endometrial cancer," *Reproductive Sciences*, pp. 1–13, 2024.

[38] P. Lei, H. Wang, L. Yu, C. Xu, H. Sun, Y. Lyu, L. Li, and D.-L. Zhang, "A correlation study of adhesion g protein-coupled receptors as potential therapeutic targets in uterine corpus endometrial cancer," *International Immunopharmacology*, vol. 108, p. 108743, 2022.