MambaMIL: Enhancing Long Sequence Modeling with Sequence Reordering in Computational Pathology

Shu Yang[†1][0000−0002−1761−9286], Yihui Wang[†1][0009−0002−7606−9816], and Hao Chen[*1,2,3][0000−0002−8400−3780]

[1] Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong, China
[2] Department of Chemical and Biological Engineering, The Hong Kong University of Science and Technology, Hong Kong, China
[3] Division of Life Science, The Hong Kong University of Science and Technology, Hong Kong, China
syangcw@connect.ust.hk, ywangrm@connect.ust.hk, jhc@cse.ust.hk

Abstract. Multiple Instance Learning (MIL) has emerged as a dominant paradigm to extract discriminative feature representations within Whole Slide Images (WSIs) in computational pathology. Despite driving notable progress, existing MIL approaches suffer from limitations in facilitating comprehensive and efficient interactions among instances, as well as challenges related to time-consuming computations and overfitting. In this paper, we incorporate the Selective Scan Space State Sequential Model (**Mamba**) in Multiple Instance Learning (**MIL**) for long sequence modeling with linear complexity, termed as **MambaMIL**. By inheriting the capability of vanilla Mamba, MambaMIL demonstrates the ability to comprehensively understand and perceive long sequences of instances. Furthermore, we propose the Sequence Reordering Mamba (SR-Mamba) aware of the order and distribution of instances, which exploits the inherent valuable information embedded within the long sequences. With the SR-Mamba as the core component, MambaMIL can effectively capture more discriminative features and mitigate the challenges associated with overfitting and high computational overhead. Extensive experiments on two public challenging tasks across nine diverse datasets demonstrate that our proposed framework performs favorably against state-of-the-art MIL methods. The code is released at https://github.com/isyangshu/MambaMIL.

Keywords: Mamba · Computational Pathology · Whole Slide Images · Multiple Instance Learning.

[†] indicates the equal contribution.
[*] indicates the corresponding author.

1 Introduction

The digitalization of pathological images into Whole Slide Images (WSIs) has paved the way for computer-aided analysis in computational pathology [19,12,9]. However, employing deep learning methods for WSI analysis encounters unique challenges, primarily due to the high resolution of WSIs and the lack of pixel-level annotations. To address these issues, Multiple Instance Learning (MIL) [1,4] has arisen as an ideal solution, where each WSI is represented as a "bag" and partitioned into a sequence of tissue patches termed "instances".

The most widely used paradigm of MIL involves converting instances into low-dimensional features using pre-trained models [10,11,19], followed by aggregating these features into bag-level representations for subsequent analysis. Under this paradigm, MIL conceptualizes WSI analysis as a long sequence modeling problem, aiming to model the correlation between instances as well as overall contextual information within the entire bag to capture discriminative information. Despite the impressive performance, there remain several issues in existing MIL methods. Attention-based methods [12,14,15,23] primarily focus on instance-level information based on independent and identical distribution hypotheses. However, these methods neglect the contextual relationships among instances, resulting in inadequate representations of WSIs. Additionally, several methods [3,22,24,17] utilize transformers [18] for their capability to explore mutual-correlations between instances and model long sequences. Nonetheless, they face significant performance bottlenecks due to extensive computations and overfitting. Overall, the existing methods have limitations in comprehensively mining the contextual information within long sequences, which hinders performance.

Recently, Structured State Space Sequence (S4) [8] has been introduced as an efficient and effective architecture to address the bottleneck concerning long sequence modeling. Furthermore, Selective Scan Space State Sequential Model [7], namely Mamba, advances S4 in discrete data modeling by employing an input-dependent selection mechanism and a hardware-aware algorithm, which enables Mamba to achieve linear complexity without sacrificing global receptive fields. However, for inherently non-sequential visual data, the direct application of Mamba to a patchified and flattened image would inevitably lead to a constraint in the receptive fields. This limitation stems from the fact that Mamba solely permits interactions between each patch and previously scanned positions, precluding the estimation of relationships with unscanned patches. Unlike typical visual modalities [16,20,21], WSIs contain scattered and scarce positive patches that exhibit weak spatial correlation, which makes them highly suitable for leveraging the robust sequential modeling capabilities of Mamba. Recently, S4MIL [6] introduces the S4 model to WSI analysis as a multiple instance learner for instance sequences, which demonstrates the effectiveness of SSM in capturing long-range dependencies. Note that it directly adopts the S4 model without fully considering the unique characteristics of WSIs, resulting in sub-optimal results.

Motivated by the above observations, we propose an efficient and effective benchmark MIL model (MambaMIL) with the following contributions: (1) We

incorporate the Mamba framework in MIL to tackle the challenges associated with long sequence modeling and overfitting, marking the first application of Mamba in computational pathology. (2) We propose the Sequence Reordering Mamba (SR-Mamba) aware of the order and distribution of instances, which excels at capturing long-range dependencies among scattered positive instances. As the core component of MambaMIL, SR-Mamba is tailored to learn the correlations between instances in both sequential ordering and transpositional ordering, which significantly enhances the capability of the original Mamba to capture more discriminative features. (3) To investigate the effectiveness of MambaMIL, we conduct comprehensive experiments including overall comparison and ablation studies on two challenging tasks across nine datasets, which demonstrates that MambaMIL can achieve superior performance against the state-of-the-art.

2 Method

In this section, we start by presenting the preliminaries associated with State Space Models. Subsequently, we elaborate on the MambaMIL and its core component: Sequence Reordering Mamba (SR-Mamba).

2.1 Preliminaries

Inspired by State Space Models [13], structured state space sequence (S4) models have emerged as a promising architecture for effective long sequence modeling. S4 models are defined with four parameters $(\triangle, A, B, C)$ as linear time-invariant systems, which map stimulation $x(t) \in \mathbb{R}^L$ to response $y(t) \in \mathbb{R}^L$ though an implicit latent state $h(t) \in \mathbb{R}^N$. The entire progress can be formulated as,

$$h'(t) = Ah(t) + Bx(t), \quad y(t) = Ch(t), \tag{1}$$

where $A \in \mathbb{R}^{N\times N}$ refers to evolution parameter. $B \in \mathbb{R}^{N\times 1}$ and $C \in \mathbb{R}^{N\times 1}$ present projection parameters. S4 models utilize a timescale parameter $\triangle$ to transform the continuous parameters A, B to the discrete parameters $\bar{A}$, $\bar{B}$,

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

After transforming the parameters, we can utilize the discrete parameters to re-frame the Eq. 1 in the recurrent mode for efficient autoregressive inference,

$$h_t = \bar{A}h_{t-1} + \bar{B}x_t, \quad y_t = Ch_t. \tag{3}$$

Alternatively, the models can also compute output through convolutional mode for efficient parallelizable training,

$$\bar{K} = (C\bar{B}, C\bar{A}\bar{B}, ..., C\bar{A}^{M-1}\bar{B}), \quad y = x * \bar{K}. \tag{4}$$

Mamba further integrates selection mechanisms into S4 models to make the parameters be functions of the input with the efficient hardware-aware parallel algorithm. Therefore, Mamba can conduct effective and efficient long sequence modeling by selectively propagating or forgetting information along the sequence based on the current token.

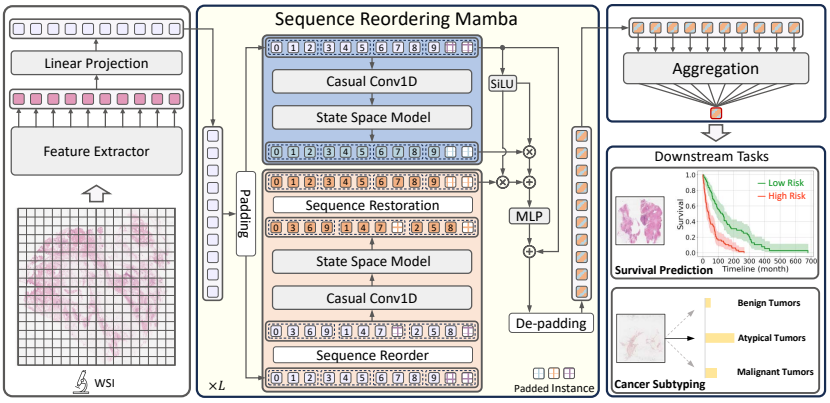

Fig. 1: Overview of MambaMIL. Given a set of patches cropped from a slide, we sequentially utilize Feature Extractor, Linear Projection, stacked SR-Mamba modules and Aggregation for WSI analysis.

2.2 Overview of MambaMIL

To efficiently capture the comprehensive contextual information within long sequences of instances, we introduce a novel approach, MambaMIL, by integrating the Mamba framework into MIL, as illustrated in Fig. 1. Inheriting the attributes of Mamba, MambaMIL enables each instance to interact with any of the previously scanned instances through a compressed hidden state, effectively reducing the computation complexity.

Specifically, given a WSI, we partition the tissue regions into a sequence of L patches $\{p_1, p_2, ..., p_L\}$, followed by mapping all the patches into instance features $X \in \mathbb{R}^{L \times D}$ by Feature Extractor, where D refers to the feature dimension. Subsequently, the input X is passed through Linear Projection to reduce the dimension. The output is then fed into a series of stacked SR-Mamba modules, which are responsible for modeling long sequences. Finally, we utilize the Aggregation module to obtain bag-level representations for downstream tasks.

2.3 Sequence Reordering Mamba

To tackle the restricted receptive fields, we devise the Sequence Reordering Mamba (SR-Mamba) aware of the order and distribution of instances, which exploits the inherent valuable information embedded within the instances. As illustrated in Fig. 1, considering the scattered and scarce positive patches, we establish parallel SSM-based branches upon vanilla Mamba to enhance long sequence modeling. SR-Mamba models two long sequences with distinct sequence orderings, each associated with a unique compressed hidden state, facilitating the learning of more discriminative features.

In detail, given instance features $X \in \mathbb{R}^{L \times D}$, we first partition the sequence of instances into non-overlapping segments of size R, and obtain $N = L/R$ segments

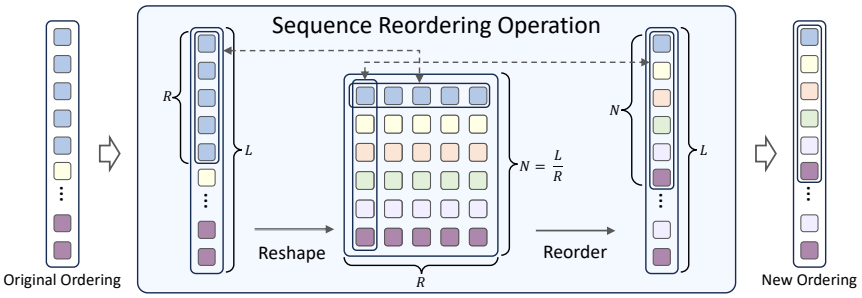

Fig. 2: Illustration of Sequence Reordering Operation.

from the entire sequence. For sequences whose lengths are not divisible by R, we pad them with zeros for subsequent reordering. Then the X is fed into two independent branches. For the first branch, we preserve the original ordering of X, which is fed to the subsequent Casual Convolution Layer and State Space Model (SSM) for sequence modeling. The entire process can be formulated as:

$$X' = \text{Norm}(X), \quad Y = \text{SSM}(\text{SiLU}(\text{Conv1D}(\text{Linear}(X')))). \tag{5}$$

Then the X is also used to generate the gating value for Y obtained from SSM,

$$Z = \text{SiLU}(\text{Linear}(X')), \quad X'' = Z \odot Y. \tag{6}$$

For the second branch, we propose a Sequence Reordering operation as the core component of SR-Mamba. Specifically, the input instance features are reshaped into a 2-D feature map, $X \in \mathbb{R}^{L \times D} \rightarrow X' \in \mathbb{R}^{R \times N \times D}$. We then sample instances from each non-overlapping segment successively along the second dimension of X', which can be regarded as a permutation and rearrangement operation. By performing this, we generate the instance features X_r with the new ordering, which can be utilized to embed more discriminative features by the inherent position-sensitive characteristic of Mamba. The entire Sequence Reordering operation is depicted in Fig. 2. Then we utilize the subsequent Casual Convolution Layer and State Space Model to model X_r,

$$X'_r = \text{Norm}(X'_r), \quad Y_r = \text{SSM}(\text{SiLU}(\text{Conv1D}(\text{Linear}(X'_r)))). \tag{7}$$

For the enhanced X'_r, we rearrange the sequences into the original ordering through partitioning and permutation operations, and gate the feature by Z,

$$Y'_r = \psi(Y_r), \quad X''_r = Z \odot Y'_r, \tag{8}$$

where ψ denotes Sequence Restoration operation. After modeling the long sequences with distinct orderings, we can obtain two discriminative instance features X'' and X''_r, and aggregate them to obtain X_{output}. We devise the aggregation operation as an element-wise addition of the two features,

$$X_{\text{output}} = \text{Linear}(X'' + X''_r) + X. \tag{9}$$

Table 1: Survival Prediction results on seven main datasets.

Method \ Dataset	BLCA	BRCA	COADREAD	KIRC	KIRP	LUAD	STAD	MEAN
ResNet-50								
Max-Pooling	0.531±0.055	0.570±0.047	0.555±0.090	0.616±0.038	0.530±0.105	0.553±0.085	0.577±0.072	0.562
Mean-Pooling	0.595±0.067	0.602±0.057	0.592±0.109	0.660±0.039	0.691±0.073	0.602±0.045	0.595±0.059	0.620
ABMIL [12]	0.565±0.060	0.612±0.059	0.624±0.046	0.677±0.057	0.707±0.099	0.626±0.054	0.629±0.061	0.635
CLAM-MB [15]	0.571±0.009	0.633±0.035	0.601±0.023	0.596±0.003	0.679±0.037	0.608±0.018	0.582±0.014	0.610
DSMIL [14]	0.593±0.018	0.609±0.060	0.628±0.059	0.682±0.042	0.722±0.085	0.624±0.057	0.609±0.057	0.638
DTFDMIL [23]	0.552±0.053	0.626±0.037	0.638±0.034	0.687±0.075	0.724±0.102	0.623±0.048	0.619±0.073	0.638
TransMIL [17]	0.623±0.037	0.632±0.029	0.624±0.014	0.684±0.052	0.747±0.082	0.641±0.049	0.629±0.020	0.654
S4MIL [6]	0.624±0.018	0.641±0.057	0.608±0.049	0.691±0.039	0.689±0.061	0.622±0.026	0.613±0.044	0.641
MambaMIL	**0.652±0.028**	**0.675±0.065**	**0.671±0.066**	**0.721±0.045**	**0.748±0.094**	**0.653±0.059**	**0.639±0.076**	**0.680**
PLIP								
Max-Pooling	0.540±0.050	0.611±0.053	0.599±0.070	0.645±0.045	0.620±0.154	0.565±0.076	0.578±0.044	0.594
Mean-Pooling	0.599±0.039	0.603±0.060	0.674±0.064	0.669±0.065	0.766±0.063	0.617±0.048	0.603±0.052	0.647
ABMIL [12]	0.571±0.041	0.607±0.036	0.641±0.013	0.643±0.077	0.772±0.065	0.570±0.066	0.573±0.037	0.625
CLAM-MB [15]	0.600±0.029	0.619±0.025	0.628±0.031	0.597±0.022	0.722±0.063	0.603±0.026	0.593±0.020	0.623
DSMIL [14]	0.589±0.052	0.613±0.033	0.640±0.048	0.673±0.048	0.768±0.074	0.565±0.074	0.601±0.059	0.636
DTFDMIL [23]	0.568±0.040	0.616±0.020	0.625±0.061	0.702±0.034	0.772±0.096	0.624±0.032	0.624±0.032	0.647
TransMIL [17]	0.586±0.059	0.611±0.065	0.620±0.031	0.673±0.030	0.798±0.063	0.622±0.036	0.630±0.067	0.649
S4MIL [6]	0.625±0.023	0.614±0.051	0.657±0.065	0.695±0.026	0.799±0.055	0.635±0.056	0.637±0.063	0.666
MambaMIL	**0.677±0.053**	**0.651±0.029**	**0.698±0.063**	**0.715±0.049**	**0.805±0.051**	**0.652±0.027**	**0.653±0.253**	**0.693**

Distinct from the original Mamba, we maintain the sequential ordering and distribution, while generating new ordering of the instances from a global perspective for feature re-embedding. Building upon the vanilla Mamba, SR-Mamba is tailored to robustly comprehend and perceive lengthy sequences of instances that are partitioned from WSIs. Built on stacked SR-Mamba modules, MambaMIL is capable of modeling long-range dependencies with linear complexity, resulting in effective model generalization.

3 Experiments

3.1 Datasets and Evaluation Metrics

To verify the effectiveness of our proposed MambaMIL, we conduct thorough experiments on nine datasets using two distinct feature sets: ResNet-50 [10] pre-trained with ImageNet [5] and PLIP [11] pre-trained with OpenPath.

Survival Prediction. We conduct comprehensive experiments on seven public cancer datasets (**BLCA**, **BRCA**, **COADREAD**, **KIRC**, **KIRP**, **LUAD**, and **STAD**) from **TCGA**, containing WSIs annotated with survival outcomes. To reduce the impact of data split on model evaluation, we implement a 5-fold cross-validation approach, partitioning the data into training and validation subsets in a 4:1 ratio. We use the cross-validated Concordance Index (C-Index), along with its standard deviation (std), to assess the model's effectiveness.

Cancer Subtyping. We perform comparative experiments on two public challenging datasets: **BRACS** [2] and **NSCLC**. To ensure the robust evaluation of comparison experiments, we employ 10-fold Monte Carlo cross-validation, which partitions the data into training, validation, and testing sets with a ratio of 8:1:1. Additionally, for fair comparisons with existing methods, we also perform experiments on the official split of the BRACS dataset, marked as $\star$ in Table 2.

Table 2: Cancer Subtyping results on two main datasets.

Method \ Dataset	BRACS-7⋆		BRACS-7		NSCLC-2		MEAN	
	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC
ResNet-50								
Max-Pooling	0.630	0.241	0.707±0.053	0.389±0.066	0.943±0.019	0.869±0.017	0.760	0.500
Mean-Pooling	0.658	0.299	0.729±0.039	0.396±0.060	0.913±0.041	0.837±0.037	0.767	0.511
ABMIL [12]	0.715	0.230	0.765±0.041	0.393±0.084	0.938±0.025	0.864±0.036	0.806	0.495
CLAM-MB [15]	0.729	0.379	0.780±0.043	0.457±0.073	0.933±0.027	0.851±0.022	0.814	0.563
DSMIL [14]	0.751	0.333	0.768±0.045	0.452±0.059	0.940±0.024	0.880±0.023	0.820	0.555
DTFDMIL [23]	0.753	0.390	0.758±0.057	0.448±0.049	0.928±0.055	0.835±0.031	0.813	0.558
TransMIL [17]	0.613	0.310	0.699±0.040	0.363±0.073	0.937±0.019	0.846±0.044	0.750	0.506
S4MIL	0.718	0.356	0.760±0.028	0.422±0.095	0.914±0.036	0.829±0.039	0.797	0.536
MambaMIL	**0.773**	**0.460**	**0.804±0.028**	**0.506±0.050**	**0.959±0.027**	**0.891±0.044**	**0.845**	**0.619**
PLIP								
Max-Pooling	0.652	0.230	0.720±0.035	0.365±0.072	0.941±0.020	0.869±0.025	0.771	0.488
Mean-Pooling	0.649	0.333	0.744±0.030	0.454±0.053	0.924±0.020	0.849±0.017	0.772	0.545
ABMIL [12]	0.699	0.333	0.797±0.038	0.487±0.074	0.944±0.015	0.867±0.034	0.813	0.562
CLAM-MB [15]	0.693	0.264	0.780±0.038	0.469±0.073	0.944±0.018	0.864±0.033	0.806	0.532
DSMIL [14]	0.667	0.333	0.771±0.037	0.478±0.079	0.933±0.020	0.860±0.022	0.790	0.557
DTFDMIL [23]	0.697	0.368	0.799±0.039	0.486±0.040	0.945±0.023	0.839±0.059	0.814	0.564
TransMIL [17]	0.688	0.345	0.705±0.028	0.328±0.070	0.928±0.021	0.848±0.035	0.774	0.506
S4MIL [6]	0.676	0.299	0.776±0.046	0.469±0.062	0.935±0.019	0.856±0.027	0.796	0.541
MambaMIL	**0.718**	**0.379**	**0.803±0.040**	**0.498±0.073**	**0.947±0.020**	**0.870±0.037**	**0.822**	**0.582**

Following the standard setting, we adopt the Area Under Curve (AUC) and Accuracy (ACC) metrics along with their standard deviation (std) for evaluation, which provides a reliable assessment which is less sensitive to class imbalance.

3.2 Implementation Details

We present the experimental results of our MambaMIL on nine datasets, in comparison to the following methods: **(1)** conventional pooling methods, including Mean Pooling and Max Pooling; **(2)** ABMIL [12] and three distinct variants, including CLAM-MB [15], DSMIL [14] and DTFDMIL [23]; **(3)** the Transformer-based TransMIL [17]; **(4)** the SSM-based S4MIL [6]. Following common settings, we adopt the same data pre-processing as in the CLAM [15] and set a learning rate of 2×10^{-4} for these methods to ensure optimal results and enable fair comparisons. In contrast, to mitigate the randomness introduced by atomic operations in the SR-Mamba module during back-propagation, we implement distinct learning rates for training for different datasets, detailed hyper-parameters can be found in the Appendix. The special adjustment aims to diminish the effect of gradient disparities on convergence, thereby ensuring stability and reproducibility.

3.3 Comparison Results

Survival Prediction. As presented in Table 1, we conduct comparison experiments with two distinct feature settings on seven TCGA cancer datasets. The results demonstrate that MambaMIL achieves the best performance on all benchmarks compared to the state-of-the-art methods. Under the two feature sets, MambaMIL outperforms the second-best performance method by 2.6% and 2.7% on mean performance across all seven datasets.

Table 3: Performance comparisons with different variations of Mamba.

Method \ Dataset	BLCA	BRCA	COADREAD	KIRC	KIRP	LUAD	STAD	MEAN
				ResNet-50				
Mamba	0.622±0.053	0.664±0.034	0.650±0.066	0.700±0.058	0.734±0.062	0.643±0.027	0.621±0.056	0.662
Bi-Mamba	0.647±0.024	**0.675±0.065**	0.662±0.058	0.690±0.048	0.737±0.052	0.628±0.059	0.622±0.068	0.665
SR-Mamba	**0.652±0.028**	0.673±0.063	**0.671±0.066**	**0.721±0.064**	**0.748±0.094**	**0.653±0.059**	**0.639±0.076**	**0.680**

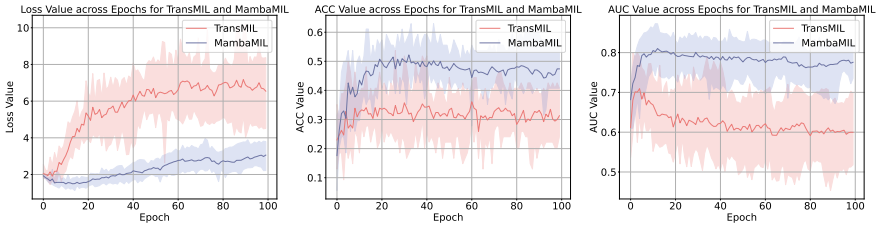

Fig. 3: The performance comparison between TransMIL and our proposed MambaMIL on the BRCAS validation set throughout the training process.

Cancer Subtyping. Table 2 shows experimental results on two datasets, encompassing both binary and multiple classification tasks. Compared to the state-of-the-art, our proposed MambaMIL demonstrates outstanding performance, attaining an AUC of 80.4% on the BRACS dataset and 95.9% on the NSCLC dataset. Notably, MambaMIL employs the same aggregation module as ABMIL but significantly outperforms it, with significant improvements of 3.9% and 2.1% in terms of AUC for BRACS and NSCLC datasets, respectively.

3.4 Ablation Study

To assess the effectiveness of SR-Mamba, we conduct extensive experiments to compare the performance of different variations of Mamba block: the vanilla Mamba [7], Bidirectional Mamba (BiMamba) [25] and Our SR-Mamba, on survival prediction datasets. For a fair comparison of each specific dataset, we utilize the same setting to train these variants. As shown in Table 3, SR-Mamba surpasses the performance of Mamba and Bi-Mamba, which demonstrates the effectiveness of sequence reordering. Meanwhile, overfitting poses a substantial challenge in applying MIL methods for WSI analysis, especially for transformer-based methods like TransMIL. As illustrated in Fig. 3, during the training process, TransMIL displays clear signs of overfitting on the validation set, characterized by a significant increase in validation loss alongside decreases in both the ACC and the AUC metrics. In contrast, MambaMIL exhibits stable performance across the evaluation period, showcasing its strong ability to alleviate overfitting. This capability originates from the more discriminative representations extracted from various sequence orderings, akin to the effects of data augmentation, which significantly enhances model robustness.

4 Conclusion

In this paper, we introduce a novel Mamba-based MIL method (MambaMIL) to tackle the challenges associated with long sequence modeling and overfitting, marking the first application of the Mamba framework in computational pathology. Our approach, based on the specially designed Sequence Reordering Mamba module, enables the effective leveraging of intrinsic global information contained within the long sequences of instances. The experimental results on nine benchmarks demonstrate that MambaMIL benefits from long sequence modeling and outperforms existing competitors. Given the excellent performance of MambaMIL, we anticipate its application can be extended to other modalities in computational pathology, including genomics, pathology reports, and clinical data. This expansion would enable the leveraging of multi-modal information for effective and accurate diagnosis, prognosis, and therapeutic-response prediction.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (No. 62202403), Hong Kong Innovation and Technology Fund (No. PRP/034/22FX), and Research Grants Council of the Hong Kong Special Administrative Region, China (No. R6003-22).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Amores, J.: Multiple instance classification: Review, taxonomy and comparative study. Artificial intelligence **201**, 81–105 (2013)
2. Brancati, N., Anniciello, A.M., Pati, P., Riccio, D., Scognamiglio, G., Jaume, G., De Pietro, G., Di Bonito, M., Foncubierta, A., Botti, G., et al.: Bracs: A dataset for breast carcinoma subtyping in h&e histology images. Database **2022**, baac093 (2022)
3. Chen, R.J., Lu, M.Y., Weng, W.H., Chen, T.Y., Williamson, D.F., Manz, T., Shady, M., Mahmood, F.: Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4015–4025 (2021)
4. Chen, Z., Chi, Z., Fu, H., Feng, D.: Multi-instance multi-label image classification: A neural approach. Neurocomputing **99**, 298–306 (2013)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255. Ieee (2009)
6. Fillioux, L., Boyd, J., Vakalopoulou, M., Cournède, P.H., Christodoulidis, S.: Structured state space models for multiple instance learning in digital pathology. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 594–604. Springer (2023)
7. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
8. Gu, A., Goel, K., Ré, C.: Efficiently modeling long sequences with structured state spaces. arXiv preprint arXiv:2111.00396 (2021)

9. Guo, Z., Ma, J., Xu, Y., Wang, Y., Wang, L., Chen, H.: Histgen: Histopathology report generation via local-global feature encoding and cross-modal context interaction. arXiv preprint arXiv:2403.05396 (2024)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
11. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., Zou, J.: A visual–language foundation model for pathology image analysis using medical twitter. Nature medicine **29**(9), 2307–2316 (2023)
12. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International Conference on Machine Learning. pp. 2127–2136. PMLR (2018)
13. Kalman, R.E.: A new approach to linear filtering and prediction problems (1960)
14. Li, B., Li, Y., Eliceiri, K.W.: Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 14318–14328 (2021)
15. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering **5**(6), 555–570 (2021)
16. Ma, J., Li, F., Wang, B.: U-mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv preprint arXiv:2401.04722 (2024)
17. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in Neural Information Processing Systems **34**, 2136–2147 (2021)
18. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
19. Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Huang, J., Yang, W., Han, X.: Transpath: Transformer-based self-supervised learning for histopathological image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 186–195. Springer (2021)
20. Xing, Z., Ye, T., Yang, Y., Liu, G., Zhu, L.: Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. arXiv preprint arXiv:2401.13560 (2024)
21. Xu, R., Yang, S., Wang, Y., Du, B., Chen, H.: A survey on vision mamba: Models, applications and challenges. arXiv preprint arXiv:2404.18861 (2024)
22. Xu, Y., Chen, H.: Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 21241–21251 (October 2023)
23. Zhang, H., Meng, Y., Zhao, Y., Qiao, Y., Yang, X., Coupland, S.E., Zheng, Y.: Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 18802–18812 (2022)
24. Zhou, F., Chen, H.: Cross-modal translation and alignment for survival analysis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21485–21494 (2023)
25. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: Efficient visual representation learning with bidirectional state space model. arXiv preprint arXiv:2401.09417 (2024)