# MULTI-STAGE BIDIRECTIONAL CROSS-ATTENTION MODEL FOR PREDICTING PROGNOSIS IN MULTIPLE PERITONEUM LESIONS WITH CLINICAL INFORMATION

*Haoshen Li*[1] *Jieyuan Cai*[2] *Yiyuan Wei*[2] *Zifan Chen*[1] *Heyun Chen*[1] *Jie Zhao*[3,6]
*Yanjie Shi*[2] *Bin Dong*[4,5,6] *Lei Tang*[2] *Xiaotian Zhang*[2] *Li Zhang*[1*]

[1] Center for Data Science, Peking University, China
[2] Peking University Cancer Hospital&Institute, China
[3] National Engineering Laboratory for Big Data Analysis and Applications, Peking University, China
[4] Beijing International Center for Mathematical Research, Peking University, China
[5] Center for Machine Learning Research, Peking University, China
[6] Peking University Changsha Institute for Computing and Digital Economy, China

## ABSTRACT

Peritoneal metastasis occurs when cancer cells spread from the primary tumor to the peritoneum, leading to morphological alterations that significantly impact patient survival. The specific changes across multiple peritoneal sites can effectively indicate a patient's risk level and prognosis. However, the considerable variation in peritoneal shape, size, and location poses significant challenges for feature extraction and prognostic analysis. Traditional multi-instance learning approaches typically fuse instance features extracted by the backbone network at the final stage, but their lack of intermediate interaction limits the ability to capture both common and distinct features. To address this limitation, we propose a Multi-Stage BiDirectional Cross-Attention framework (MSBDCA). Our approach enhances feature extraction and prognostic analysis by facilitating interaction among different peritoneal instances and incorporating clinical information. Experimental comparisons with five baseline models demonstrate our method's superiority, achieving improvements of 8.7% in Area Under Curve (AUC) and 5.9% in Concordance Index (C_index). These results suggest a promising direction for survival and prognostic analysis using peritoneal lesions. Our code is available at https://github.com/HaoshenLi/MSBDCA.

***Index Terms*—** Survival analysis, peritoneal metastasis analysis, deep learning, multi-instance learning

## 1. INTRODUCTION

Gastric cancer (GC) is a prevalent gastrointestinal malignant tumor and the second leading cause of cancer-related deaths globally [1]. Peritoneal metastasis occurs in up to 66% of patients with advanced GC and is associated with poor prognosis [2, 3, 4]. The morphological alterations in peritoneal structures induced by metastasis serve as crucial prognostic indicators, making peritoneal imaging analysis a valuable tool for survival prediction. However, as shown in Fig. 1, peritoneal lesions exhibit substantial heterogeneity in size, shape, and location, presenting significant challenges for effective modeling.

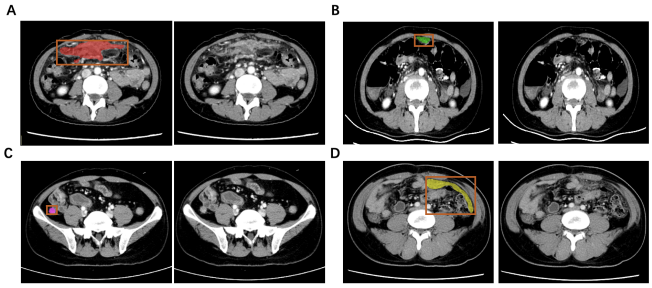


**Fig. 1**. Visualization of peritoneal lesion heterogeneity. A, B, C, D corresponds four sets of peritoneal lesions, with significant differences in size, shape and location.

In this work, we formulate the problem as a multi-instance learning task, where multiple peritoneal lesion images from the same patient form a bag, and the patient's survival outcome serves as the bag-level label. Traditional multi-instance learning approaches typically employ a backbone model to extract features from multiple instances, followed by feature fusion for bag-level predictions, as depicted in Fig. 2A. This architecture, however, lacks early-stage feature interaction among instances. When instances display significant heterogeneity, the model struggles to capture both shared patterns and distinctive characteristics, leading to a suboptimal bag-level predictions performance. This motivates us to explore more effective ways of feature interaction throughout the learning process.

To address the challenges posed by limited multi-instance interaction, we propose a multi-stage interaction approach,

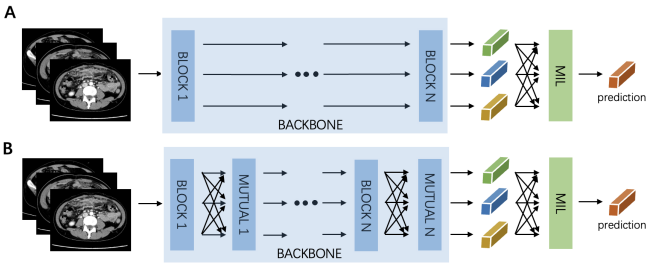


**Fig. 2**. Panel A shows the classic multi-instance learning framework, which only fuses the instance features after the last layer of the feature extractor; Panel B shows our framework, that more interactions are added after each stage block.

as shown in Fig. 2B. Our method performs multi-instance feature fusion and interaction after each stage block in the backbone network. Specifically, we introduce a bag-level global instance embedding and implement BiDirectional Cross-Attention (BDCA) interaction between this global embedding and multi-instance features after each block. This design enables more comprehensive multi-scale feature interaction and fusion. Additionally, recognizing the crucial role of clinical features in prognostic prediction, we incorporate bidirectional cross-attention interaction between multi-instance image features and clinical features at each stage. The final prediction combines the global instance embedding, multi-instance image features, and clinical embedding through a fully connected (FC) layer to predict survival outcomes. Experimental results validate the effectiveness of our approach.

Our key contributions are summarized as follows:

- We propose MSBDCA, a novel multi-instance learning framework that enhances feature interaction among instances through multi-stage processing.
- We develop two complementary interaction mechanisms: BiDirectional Cross-Attention (BDCA) for multi-instance image features and an integrated approach incorporating clinical features.
- We demonstrate through extensive experiments that our method achieves superior performance in prognostic analysis, with both interaction mechanisms significantly contributing to improved outcomes.

## 2. METHOD

### 2.1. Overall Framework

The overall framework of our proposed Multi-Stage BiDirectional Cross-Attention (MSBDCA) framework is illustrated in Fig. 3. Our framework processes multiple peritoneal lesions, represented as $\mathbf{X}^l \in \mathbb{R}^{N\times D\times H\times W}$, where N denotes the number of instances and (D, H, W) represents the 3D shape. We integrate clinical information $\mathbf{X}^c \in \mathbb{R}^{5\times C1}$, including age, sex, Lauren type, immunotherapy status, and target therapy status. Additionally, we introduce a learnable embedding to represent the global instance feature, represented as $\mathbf{X}^g \in \mathbb{R}^{1\times C2}$.

The backbone network extracts image features in multiple stages using architectures such as VGG [5], ResNet [6], or ViT [7]. After each block in the feature extractor, we incorporate a BiDirectional Cross-Attention (BDCA) module to facilitate interaction among multi-instance features, global instance embedding, and clinical embedding. Let $\mathbf{B}_i, \mathbf{BDCA}_i, \mathbf{X}_i^l, \mathbf{X}_i^c, \mathbf{X}_i^g$ denote the feature extractor block, mutual module, instance features, clinical embedding, and global instance embedding at stage $i$, respectively. The operations at the $i$-th stage can be expressed as:

$$\begin{aligned}\mathbf{X}_{i'}^l, \mathbf{X}_{i'}^c, \mathbf{X}_{i'}^g &= \mathbf{BDCA}_i(\mathbf{B}_i(\mathbf{X}_i^l), \mathbf{X}_i^c, \mathbf{X}_i^g)\\ \mathbf{X}_{i+1}^l &= \mathbf{X}_i^l + \mathbf{X}_{i'}^l\\ \mathbf{X}_{i+1}^c &= \mathbf{X}_i^c + \mathbf{X}_{i'}^c\\ \mathbf{X}_{i+1}^g &= \mathbf{X}_i^g + \mathbf{X}_{i'}^g\end{aligned}$$

After the interaction module, the information of all three types of features is further enhanced. Finally, assuming there are T stages, we concatenate $(\mathbf{X}_T^l, \mathbf{X}_T^c, \mathbf{X}_T^g)$ and feed them through a dual-layered fully connected (FC) predictor to estimate the survival risk. The network outputs a binary prediction $[p_0, p_1]$, where 0 and 1 indicate high and low survival risk, respectively. We optimize this binary classification task using cross-entropy loss.

### 2.2. BiDirectional Cross-Attention module (BDCA)

The BDCA module facilitates interaction between multi-instance features and both global instance embedding and clinical embedding. As shown in Fig. 3, we first transform the multi-instance feature maps $\mathbf{X}_i^l \in \mathbb{R}^{N\times C\times D\times H\times W}$ into a 2D token sequence $\mathbf{X}_{i''}^l \in \mathbb{R}^{N\times CDHW}$, where $N$ represents the instance number and $L = CDHW$ denotes the feature length of each instance. The BDCA module then manages interactions between the multi-instance features $\mathbf{X}_{i''}^l$ and both global instance embedding $\mathbf{X}_i^g$ and clinical embedding $\mathbf{X}_i^c$. After obtaining queries, keys, and values for all three feature types through linear projection, the interaction process is formulated as:

$$\begin{aligned}\mathbf{X}_{i'}^l &= \mathbf{CA}(q2, k1, v1) + \mathbf{CA}(q2, k3, v3)\\ \mathbf{X}_{i'}^c &= \mathbf{CA}(q1, k2, v2)\\ \mathbf{X}_{i'}^g &= \mathbf{CA}(q3, k2, v2)\end{aligned}$$

where $\mathbf{CA}$ denotes the Cross-Attention mechanism. This bidirectional cross-attention design enables comprehensive interaction and mutual enhancement between instance features and clinical features.

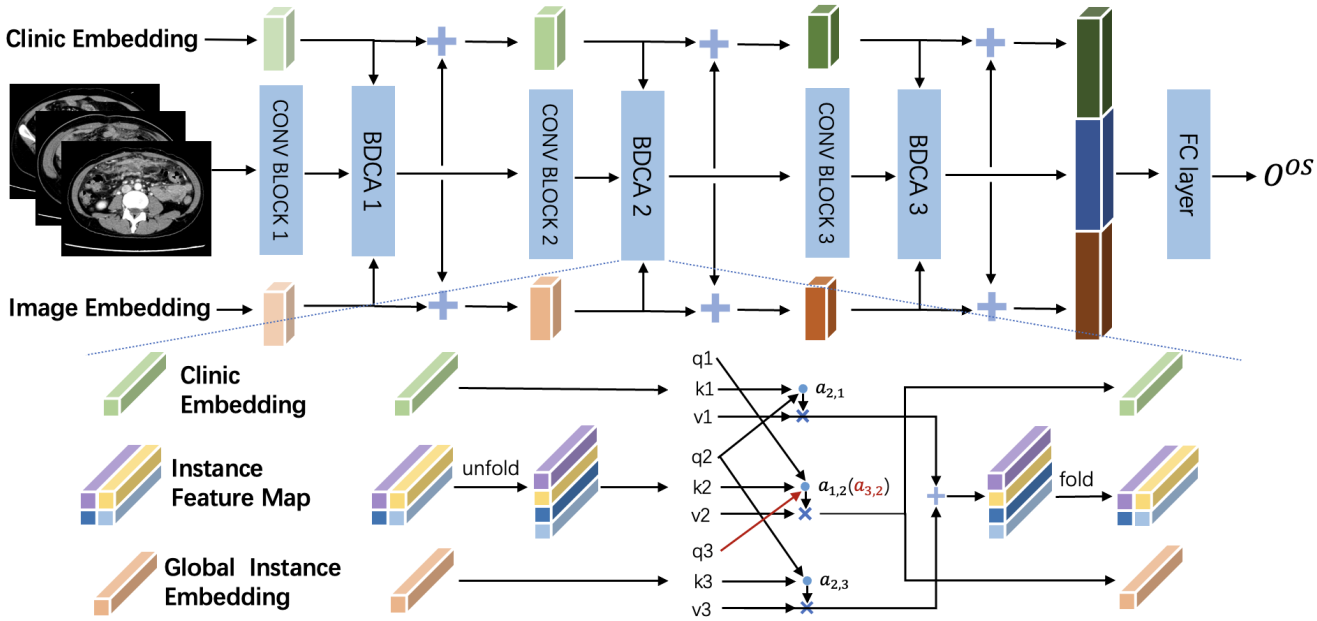


**Fig. 3**. Illustration of our Framework. After each conv block in feature extractor, our proposed BiDirectional Cross Attention module (BDCA) conduct interaction based on instance feature map with clinic embedding and global instance embedding. Finally, we concatenate features from these three parts and obtain the survival prediction $O^{OS}$ through the FC layer.

## 3. EXPERIMENT

### 3.1. Dataset Collection

Our dataset comprises 156 patients from Peking University Cancer Hospital, with CT scans of median size $512 \times 512 \times 96$ voxels and resolution $0.713 \times 0.713 \times 5$mm. Using one-year survival as the threshold, we categorize patients into low-risk (survival $>$ 1 year, label 0) and high-risk (survival $\leq$ 1 year, label 1) groups. The dataset contains 64 low-risk, 82 high-risk, and 10 censored cases. Each patient has 1-5 peritoneal lesions and associated clinical data (age, sex, Lauren type, immunotherapy status, and target therapy status). We randomly split the dataset into training, validation, and test sets with a 7:1:2 ratio.

### 3.2. Data Pre-Processing

For peritoneum images, considering the significant size differences among the peritoneum, we adopt a crop size of (16, 144, 144) to accommodate most peritoneal lesions. To address the limited dataset size, we implement extensive data augmentation, including random rotation, flipping, intensity scaling, intensity shifting, and Gaussian noise. In addition, we introduce a global instance embedding to interact with multi-instance features.

The clinical data, on the other hand, undergoes a structured encoding process to be mapped to their respective embeddings: age is binarized (under/over 60); sex is encoded as male/female; Lauren type is categorized into intestinal, diffuse, and mixed types; and both immunotherapy and targeted therapy status are binarized as presence or absence.

### 3.3. Implementation Details

Given the dataset size constraints, we adapt the VGG11 [5] architecture by reducing the number of blocks and feature dimensions. Our backbone consists of three stages, with instance feature dimensions ranging from 8 to 64 during downsampling. Both clinic embedding dimension ($C1$) and global instance embedding dimension ($C2$) are set to 576. In the cross-attention module at BDCA, we employ a singular attention head and a singular layer. We train the proposed model and all compared baselines up to 1000 epochs, with each batch comprising 8 samples. We use SGD optimizer with a learning rate of $5 \times 10^{-5}$. All codes are implemented with PyTorch (Version 1.12.1) [8], and our computational environment is powered by a single NVIDIA RTX3090 GPU with 24GB memory.

### 3.4. Baselines

We have devised several classic multi-instance learning methods to serve as comparative baselines, including MaxPooling [9], MeanPooling [9], ABMIL [10], GAMIL [10] and TransMIL [11]. All the methods use the same backbone.

To ensure a rigorous and comprehensive evaluation, we adopt two key metrics in survival analysis: Area Under Curve

| Methods | AUC (%) | C_index (%) |
|---|---|---|
| MaxPooling [9] | 66.2±10.29 | 59.6±6.19 |
| MeanPooling [9] | 66.8±10.84 | 57.5±5.87 |
| ABMIL [10] | 66.6±10.08 | 58.5±6.70 |
| GAMIL [10] | 65.1±10.45 | 59.4±6.02 |
| TransMIL [11] | 62.3±10.53 | 56.1±6.43 |
| MSBCDA (ours) | **75.5**±9.11 | **65.5**±6.02 |

**Table 1**. Performance evaluation of our method against five baseline models. **Bold** indicates the best performance, while underlined signifies the second-best.

| Methods | AUC (%) | C_index (%) |
|---|---|---|
| MeanPooling [9] | 66.8±10.84 | 57.5±5.87 |
| Clinic | 64.4±10.68 | 58.6±6.84 |
| BDCA (Image) | 70.9±9.87 | 63.1±7.31 |
| BDCA (Image + Clinic) | **75.5**±9.11 | **65.5**±6.02 |

**Table 2**. Ablation study on our mutual module BDCA, **Bold** indicates the best performance, while underlined signifies the second-best.

(AUC) and Concordance Index (C_index). Each metric undergoes 5000 rounds of bootstrapping, with results presented as means along with their respective standard deviations.

### 3.5. Experimental Results

The comparative performance of our proposed method against five established baseline models is detailed in Tab.1. Our method outperforms all baseline models across all evaluation metrics. Specifically, our method shows an improvement of 8.7% in AUC, 5.9% in C_index over the next best model. In summary, these empirical results unequivocally establish the superiority of our proposed approach over existing state-of-the-art models. We attribute this to the bidirectional cross attention mutual among instance features and between image features and clinic features.

### 3.6. Ablation Studies

We conduct ablation studies to evaluate the impact of our BDCA module on instance features alone (BDCA (Image)) and with clinical features (BDCA (Image + Clinic)), as shown in Tab. 2. We compare against MeanPooling [9] and a clinic-only baseline that simply averages the embeddings of multiple clinical features and then classifies through a linear layer.

As shown in Tab.2, results demonstrate that both image and clinical features contribute substantially to prognostic analysis. When applying BDCA only to the image instances, it can be observed that the AUC increased by 4.1% and the C_index increased by 5.6% compared to MeanPooling [9].

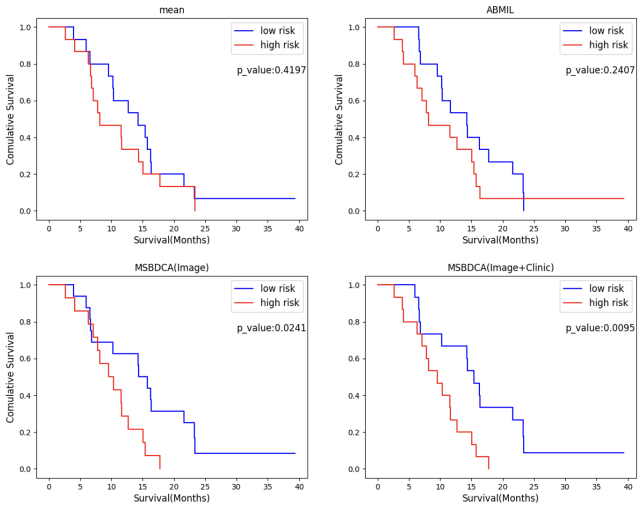


**Fig. 4**. Kaplan-Meier analysis among two baseline methods, and our proposed module, with p_value displayed on top-right position of the figure.

Further incorporation of clinical features yields additional gains, achieving 75.5% AUC and 65.5% C-index. Compared to using only image data or only clinical data, this represents an improvement of 8.7% in AUC and 6.9% in C_index, respectively. This underscores the efficacy and suitability of our proposed method.

### 3.7. Kaplan-Meier Analysis

To validate our method's ability to stratify patient risk, we perform Kaplan-Meier (KM) analysis by dividing patients into low-risk and high-risk groups based on the median prediction score. As shown in Fig. 4, while MeanPooling [9] and MaxPooling [9] show limited stratification ability, our MSBDCA framework achieves significant differentiation between risk groups ($p < 0.05$). This demonstrates our model has superior prognostic capability both qualitatively and quantitatively.

## 4. CONCLUSION

In this study, we have introduced the Multi-Stage BiDirectional Cross-Attention framework (MSBDCA) for prognostic analysis using multiple peritoneal lesions. Our framework demonstrates significant effectiveness through both multi-instance image feature interaction and clinical feature integration. MSBDCA achieves substantial improvements over classic multi-instance learning methods, not only advancing the state of the art in peritoneal-based prognostic analysis but also highlighting the crucial role of instance interaction in multi-instance learning. These findings provide valuable insights for future research in both medical image analysis and multi-instance learning applications.

## 5. COMPLIANCE WITH ETHICAL STANDARDS

This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by The Peking University Cancer Hospital Ethics Committee.

## 6. ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (12090022 to B.D., 11831002 to B.D., 81801778 to L.Z.), Clinical Medicine Plus X-Young Scholars Project of Peking University (PKU2023LCXQ041 to L.Z.).

## 7. REFERENCES

[1] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal, "Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 68, no. 6, pp. 394–424, 2018.

[2] Irene Thomassen, Yvette R van Gestel, Bert van Ramshorst, Misha D Luyer, Koop Bosscha, Simon W Nienhuijs, Valery E Lemmens, and Ignace H de Hingh, "Peritoneal carcinomatosis of gastric origin: a population-based study on incidence, survival and risk factors," *International journal of cancer*, vol. 134, no. 3, pp. 622–628, 2014.

[3] Kazumasa Fujitani, Han-Kwang Yang, Junki Mizusawa, Young-Woo Kim, Masanori Terashima, Sang-Uk Han, Yoshiaki Iwasaki, Woo Jin Hyung, Akinori Takagane, Do Joong Park, et al., "Gastrectomy plus chemotherapy versus chemotherapy alone for advanced gastric cancer with a single non-curable factor (regatta): a phase 3, randomised controlled trial," *The Lancet Oncology*, vol. 17, no. 3, pp. 309–318, 2016.

[4] National Comprehensive Cancer Network et al., "Nccn clinical practice guidelines in oncology," *http://www.nccn. org/professionals/physician_gls/PDF/occult. pdf*, 2008.

[5] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, et al., "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[8] Adam Paszke, Sam Gross, Francisco Massa, et al., "Pytorch: An imperative style, high-performance deep learning library," in *NIPS*, 2019, vol. 32.

[9] Maximilian Ilse, Jakub M Tomczak, and Max Welling, "Deep multiple instance learning for digital histopathology," in *Handbook of Medical Image Computing and Computer Assisted Intervention*, pp. 521–546. Elsevier, 2020.

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

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