CA-MLIF: Cross-Attention and Multimodal Low-Rank Interaction Fusion Framework for Tumor Prognostic Prediction

Yajun An[1, 2*], Jiale Chen[1*], Huan Lin[3*], Zhenbing Liu[1], Siyang Feng[1], Hualong Zhang[1], Rushi Lan[1, 4†], Zaiyi Liu[1, 3, 5†], Xipeng Pan[1, 4†]

[1]School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin 541004, China
[2]School of Artificial Intelligence and Big Data, Chongqing College of Finance and Economics, Chongqing 402160, China
[3]Department of Radiology, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University; Guangzhou 510080, China
[4]Guangxi Key Laboratory of Image and Graphic Intelligent Processing, Guilin University of Electronic Technology, Guilin 541004, China
[5]Guangdong Provincial Key Laboratory of Artificial Intelligence in Medical Image Analysis and Application, Guangzhou 510080, China
rslan@guet.edu.cn, liuzaiyi@gdph.org.cn, pxp201@guet.edu.cn

Abstract

Cancer is a leading cause of death worldwide due to its aggressive nature and complex variability. Accurate prognosis is therefore challenging but essential for guiding personalized treatment and follow-up. Previous research often relied on single data sources, missing the opportunity to combine various types of patient information for more comprehensive survival predictions. To address these challenges, we propose a two-stage fusion method named Cross-Attention and Multimodal Low-Rank Interaction Fusion Framework (CA-MLIF). In the first stage, we propose a CA mechanism for real-time feature updates and cross-modal mutual learning to capture rich semantic information. In the second stage, we design a novel multimodal low-rank interaction fusion method for survival prediction. Specifically, we present modal attention mechanism (MAM) for feature filtration, low-rank multimodal fusion (LMF) for model complexity reduction, and optimal weight concatenation (OWC) for maximizing feature integration. Extensive experiments on two public datasets TCGA-GBMLGG and TCGA-KIRC, as well as a multicenter in-house lung adenocarcinoma (LUAD) dataset validate the effectiveness of CA-MLIF, which demonstrate that our method outperforms existing approaches in survival prediction under both pathology-gene fusion and CT-pathology fusion scenarios.

Introduction

Cancer is one of the most severe and deadly diseases worldwide. According to the 2022 global cancer statistics from the International Agency for Research on Cancer, there were approximately 20 million new cancer cases and 9.7 million cancer-related deaths globally (Bray et al. 2024). This underscores the urgent need for effective methods to accurately predict cancer survival rates and assist physicians in making informed treatment decisions.

The diagnosis and prognosis of tumors involve diverse data, including histopathological images, radiographic imaging, gene expression profiles, and clinical information, each with unique tumor characteristics. Pathology explores diseases from a morphological perspective (Yang et al. 2022), providing critical cellular features related to tumor invasion and ensuring effective treatment, making pathological evaluation the 'gold standard' for cancer diagnosis (Hoda and Cheng 2017). Genetic data, such as gene mutations and expression profiles, are also crucial for cancer diagnosis and prognosis (Cui et al. 2023). Additionally, radiographic images offer comprehensive anatomical details, aiding in precise lesion assessment and patient evaluation. However, relying on a single modality can miss important features. To address these limitations and fully utilize diverse information, multimodal medical image fusion has emerged.

In the field of image fusion, the most common fusion methods are splicing, tensor fusion, and attention mechanisms. Among them, a multimodal fusion framework based on multi-task correlation learning (MultiCoFusion) was proposed (Tan et al. 2022), which skillfully utilized the ResNet-152 deep neural network to mine the subtle features of pathology images and adopted a sparse graph convolutional network (SGCN) to extract the information of mRNA expression data. The multimodal low-rank interaction fusion framework integrating pathological images and genomic data (PG-MLIF) (Pan et al. 2024), addressed the two bottlenecks in the fusion process, i.e., insufficient fusion and inefficient fusion, by adopting a staged processing strategy, which significantly improves the accuracy of survival prognosis prediction. However, although this approach can effectively improve the performance of the model, it becomes challenging to realize end-to-end training because it involves the extraction of different modal features.

In addition, a multi-modal memory transformer network (MMTN) was proposed by (Cao et al. 2023), based on the attention mechanism, which exploited the cross-modal complementarity of medical vision and language for word pre-

*These authors contributed equally.
†Corresponding authors

diction and further improved the accuracy of the generated medical reports. Another approach was proposed by (Liu et al. 2023), which achieved accurate image fusion and segmentation by learning interaction features between multimodal images and introducing a full-time multimodal benchmark. Although the above attention mechanism fusion methods are able to highlight the key information in the input data and dynamically adjust the degree of attention to each feature, most of the fusion methods primarily focus on leveraging features from one type of data with favorable prognosis to integrate features from another data type for survival analysis. This unidirectional fusion leads to the existence of certain prerequisites for fusion.

To address these challenges, we propose a two-stage end-to-end Cross-Attention and Multimodal Low-Rank Interaction Fusion Framework (CA-MLIF) for multimodal information integration. In the first stage, we develop a pair of encoder-decoder for efficient unimodal feature extraction and reconstruction. Based on this, we improve the Cross-Attention to facilitate deep interactions between different modal features, uncovering intrinsic connections hidden in heterogeneous information. In the second stage, we further optimize the feature interaction process by introducing low-rank constraints. In summary, our contributions are as follows:

- We propose a two-stage multimodal fusion framework named CA-MLIF, with cross-modal feature information interaction and feature fusion modules, interleaving Cross-Attention and low-rank tensor interaction fusion for more comprehensive survival prediction.
- We improve the Cross-Attention mechanism to enable bidirectional interaction between features, ensuring comprehensive and balanced information exchange. Additionally, this mechanism offers versatility across different multimodal data, significantly enhancing the practicality of the model for conducting survival prognosis analysis in various clinical settings.
- Extensive experiments on two publicly datasets TCGA-GBMLGG, TCGA-KIRC, and an in-house multi-center lung adenocarcinoma (LUAD) dataset demonstrate that our model significantly outperforms many state-of-the-art methods in pathology-gene fusion and CT-pathology fusion. The code is available at: https://github.com/980313/CA-MLIF.

Related Work

Multimodal Learning

Multimodal learning attracted significant attention from the research community. In computer vision, a strategy called CEN was proposed (Wang et al. 2020), combining channel exchange with multimodal learning for effective feature fusion. A self-supervised learning method, STiCA, was introduced (Patrick et al. 2021) for video and audio classification and retrieval, utilizing additional data to train the multimodal model, significantly improving performance. Another approach (Chen et al. 2021a) developed a model using a trusted triple-mapped middle ground to improve video and sound source localization accuracy. Although these methods lead to some improvements, they may limit the model's ability to learn rich semantic representations. Moreover, these natural-image-based approaches faced significant challenges in medical imaging. They often fell short of medical imaging requirements, failing to achieve the accuracy needed for detecting subtle pathological changes. Simple feature concatenation methods overlooked scale differences and contextual dependencies among various medical data modalities, resulting in inadequate feature interaction and integration. This constraint impedes capturing complex relationships crucial for accurate medical assessments.

Multimodal Fusion with Medical Images

Interest in multimodal medical image fusion has grown, as diverse imaging technologies capture complementary perspectives of organs or lesions. Recent methods focused on feature-level fusion (Lu, Shiradkar, and Liu 2021; Bala, Gupta, and Kumar 2022; Schneider et al. 2022) to extract and combine key features from multiple images for deeper analysis. Various multimodal fusion methods have emerged. MCAT (Chen et al. 2021b) was a multimodal co-attention transformer framework to capture genotype-phenotype interactions. GPDBN (Wang et al. 2021) employed a deep bilinear network for intra- and inter-modal tensor fusion of genetic and pathological data. Pathomic Fusion (Chen et al. 2020), an interpretable strategy for end-to-end fusion of histology images and genomic features for survival outcome prediction, modeling pairwise feature interactions with a gating-based attention mechanism.

Other attention-based fusion methods aimed to improve classification or prognostic performance (Guo et al. 2022; Li, Jia, and Qian 2021). For example, Ding et al. (Ding et al. 2024) developed a multimodal co-attention fusion network with online data augmentation, aiming to maximize data diversity and effectively perform dense multimodal interactions. CMTA (Zhou and Chen 2023) was a cross-modal translation and alignment framework exploring intrinsic cross-modal correlations. MOTCat (Xu and Chen 2023) was a common attention converter framework based on multimodal transmission, efficiently capturing interactions in the tumor microenvironment.

Although the above methods have achieved excellent results, they often process each modality separately before combining the results. These approaches may overlook inter-modal details and information, potentially leading to the loss of valuable information during fusion. And simple fusion based on the attention mechanism not only suffers from the limitation of unidirectional fusion, but also further increases the computational cost. Therefore, our goal is to achieve end-to-end multimodal fusion through cross-modal bi-directional mutual learning with low computational cost, so as to capture the rich semantic information in the data.

Method

We propose a novel multimodal fusion survival prognosis method based on Cross-Attention and Multimodal Low-Rank Interaction Fusion (MLIF) Framework, as shown in

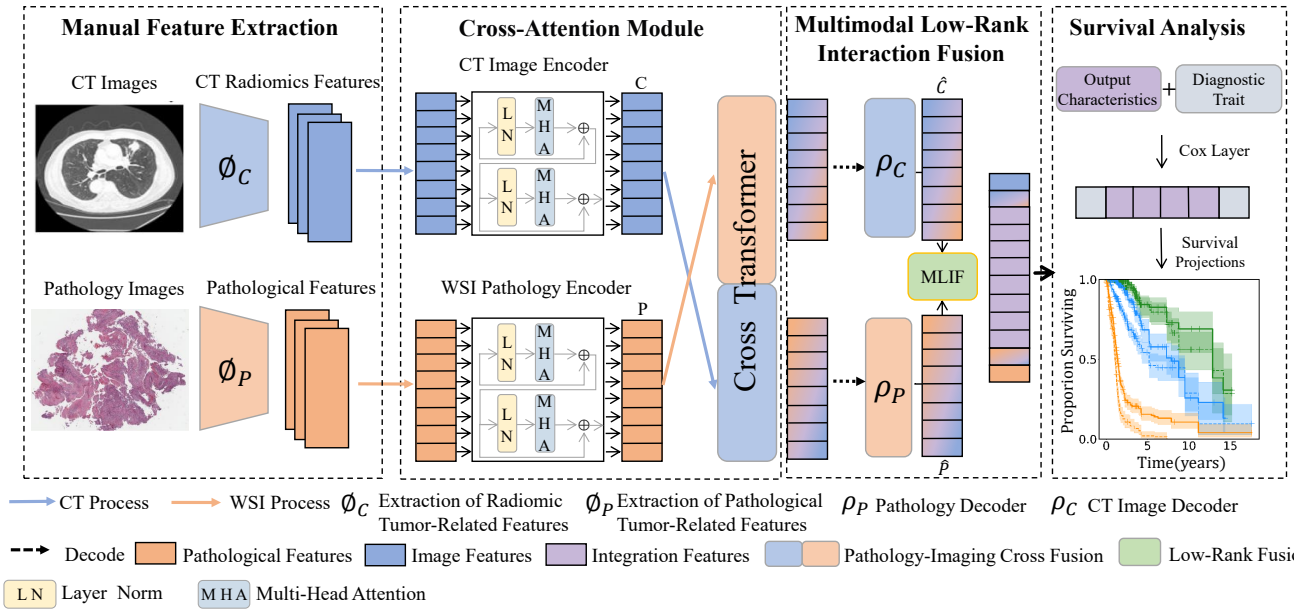

Figure 1: An overview of our CA-MLIF. It consists of two stages. In the first stage, a Cross-Attention is utilized to enhance the interaction between different features, achieving the extraction of complementary information across modalities and improving the modeling capability of multimodal data, thereby generating multimodal fusion features. In the second stage, a multimodal low-rank interaction fusion method is employed to promote the comprehensive integration of information between different modalities obtained in the first stage. Ultimately, this model integrates multimodal data end-to-end, revealing the correlations and complementarities between various data types, thereby improving patient treatment outcomes and survival rates.

Fig. 1. In the first stage, we construct a pair of encoder-decoders for each modality to process and extract features. We incorporate a Cross-Attention mechanism between the encoder-decoders to enhance feature interactions, thereby obtaining unimodal interaction features for fusion. In the second stage, a low-rank multimodal interaction method further integrates information from different modalities. Finally, the multimodal features are combined with clinical information to train survival prediction models.

Feature Extraction and Selection

For the in-house multi-center LUAD dataset, the feature extraction and selection processes for radiomics and pathology are conducted independently, tailored to the specific characteristics and consistency of each modality. For CT radiomics features, we extract 487 features, including texture, shape, density, and intensity characteristics. These features are combined into a single high-dimensional feature vector for further data analysis and modeling. For pathology features, we follow the previous works (Wang et al. 2022; Pan et al. 2023), to extract 783 pathological features from digital pathology images, including 58 from tumor regions, 290 from tumor epithelium, 290 from tumor stroma, and 155 from tumor cell nuclei. These features represent the tumor's characteristics in terms of morphology, tissue structure, and nuclear morphology. More details on feature extraction and selection are provided in the **Supplementary Material**.

Radiomics Encoder and Pathology Encoder

In the first stage, we focus on constructing and optimizing encoders for each modality to fully utilize the self-attention mechanism (Jamil, Jalil Piran, and Kwon 2023) to deeply explore and integrate intra-modal information. The goal of this stage is to extract modality-specific feature representations that are crucial for subsequent fusion and the final task. First, for the raw data of each modality, we design corresponding data processing steps (see **Supplementary Material**) to ensure data quality and consistency. Then, we construct a separate encoder for each modality. Within each encoder, we apply two self-attention layers to integrate information. These layers can simultaneously attend to different subspaces of the input data representations and generate a series of weighted feature vectors. These feature vectors capture not only the local characteristics of the data but also the global contextual information, enhancing the representational capacity of the encoders. By working in tandem, the encoders and self-attention mechanisms effectively extract and optimize important features within each modality, reducing redundant information and retaining task-critical information. The structure of the image encoder is identical to that of the pathology encoder.

We denote the pathological features as P and the CT radiomics features as C. For a given pathology $P = \{P_1, P_2, \cdots, P_K\}$, this study defines a learnable class token $p^{(0)}$ to collect information from all pathology sequences. The initial input representation of the pathology encoder is

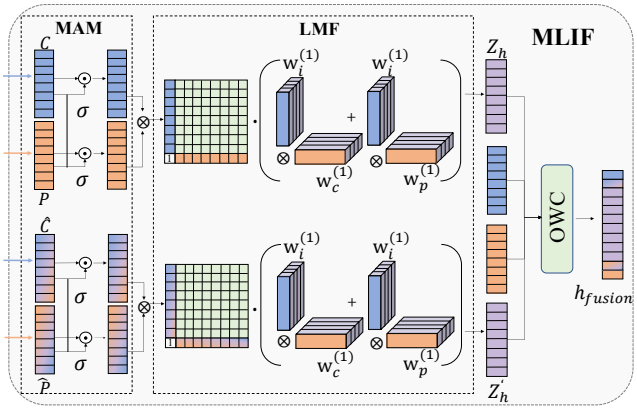

Figure 2: The framework of the proposed Multimodal Low-rank Interaction Fusion Module.

denoted as $P^{(0)} = \{p^{(0)}, p_1^{(0)}, p_2^{(0)}, \cdots, p_K^{(0)}\}$. The intra-modal representation of the pathology encoder is given by:

$$P^{(1)} = MSA(LN(P^{(0)})) + P^{(0)} \quad (1)$$

$$P^{(2)} = MSA(LN(P^{(1)})) + P^{(1)} \quad (2)$$

where MSA represents multi-head self-attention, LN indicates layer normalization, and the output $P^{(2)} = \{p^{(2)}, p_1^{(2)}, p_2^{(2)}, \cdots, p_K^{(2)}\} \in \mathbb{R}^{(k+1)\times d}$ is the intra-modal representation of the pathology. Similarly, the calculation process of $C^{(2)}$ is as same as $P^{(2)}$.

Multimodal Cross-Attention

After obtaining unimodal augmented feature information, we construct a Cross-Attention (CA) module to augment the information interactions.The CA module is designed to explore potential cross-modal correlations and interactions, and to convey complementary information between multimodal data. In this section, the traditional CA structure (Fukui et al. 2016) has been improved. The original self-attention operation encodes the input vectors as query Q, key K, and value V. The global attention map is computed by matrix multiplication QK^T, where QK^T is of size $\mathbb{R}^{N\times N}$, resulting in high memory usage. We improve this by multiplying the global context vector K^T by V, reducing the computational load. As shown in Fig. A3 of **Supplementary Material**, the improved CA structure includes multiple CA layers. Within each CA layer, the representations of each modality are influenced by other modalities. Each CA layer is able to cross and fuse features between different modalities. This study retains two branches and designs a CA mechanism that globally exchanges information between the two branches, learning the weight distribution among different features. This ensures that the final feature representation contains features of their respective modalities and includes interaction information between them. The specific process of the CA is as follows:

$$G_C = K_C^T V_C \qquad G_P = K_P^T V_P \quad (3)$$

$$U_C = X_C^{inter} sm(G_P) \quad U_P = X_P^{inter} sm(G_C) \quad (4)$$

where G, U, and $sm(\cdot)$ represent the global context vector, interaction features, and Softmax operation, respectively. To achieve attention from different representation subspaces, the multi-head mechanism is retained, with the number of heads matching the main structure of the Transformer.

At this stage, the features of these two modules will exchange their information through a symmetrical dual-path structure. The input features are first linearly encoded in the multimodal CA module to generate residual vectors X^{res} and interaction vectors X^{inter}.

The output of the CA process is obtained by multiplying the interaction vector with the context vector from another modality path. Specifically, the interaction vector of one modality is embedded into the K and V values of the feature vector of the other modality. To achieve attention from different representation subspaces, the multi-head mechanism is retained, with the number of heads matching the main structure of the Transformer. Finally, the participating result vector U is concatenated with the another modality residual vector X^{res}, enabling more comprehensive fusion of information across different modalities.

Feature Fusion Module

In the second phase, we utilize the interaction-rich features obtained in the first phase for fine-grained feature fusion. To this end, we refine the Multimodal Low-Rank Interaction Fusion (MLIF) module, which is inspired by (Pan et al. 2024). The module consists of three parts, namely Modal Attention Mechanism (MAM), Low-rank Multimodal Fusion (LMF) and Optimal Weight Concatenation (OWC).

To harness the full potential of these disparate modalities, we present MAM that dynamically determines the contribution of each modality to feature expression, adapting to the intrinsic importance of each modality. This adaptive allocation of attention enables the network to effectively emphasize or suppress specific modalities based on their relevance to the task, enhancing its ability to interpret and utilize information from multiple sources.

The primary goal of our research is to integrate single-modality features into compact features suitable for downstream tasks. By introducing low-rank factors, LMF captures significant interactions through the parallel decomposition of low-rank weight tensors and input tensors. In contrast to traditional tensor fusion networks techniques, LMF fusion stands out by avoiding the explicit creation of weights to capture interactions. Moreover, LMF fusion exhibits linear scalability in modalities, reducing model parameters.

Additionally, after extracting and integrating deep multimodal data via LMF to get the final vector, PG-MLIF devises an innovative OWC feature fusion module. The core of this strategy lies in the adaptive dynamic weight adjustment mechanism, which not only intelligently balances the information contributions between different modalities, but also dynamically adjusts the weight allocation in the fusion process according to the specific characteristics of the data and the contextual environment, thus significantly improving the generalization ability and prediction accuracy of the

fusion model. To realize OWC, we construct a small neural network model that deeply fuse the efficient information processing capability of feedforward networks with the nonlinear mapping advantage of Sigmoid activation functions. This network architecture is specifically trained to learn and optimize the optimal weights for each modality or feature combination. Through an iterative optimization process, the network is able to automatically adjust the weighting parameters to maximize the overall performance of the combined model, ensuring that the fused feature representations are both comprehensive and accurate.

Unlike PG-MLIF (Pan et al. 2024), this improvement transcends simple dynamic weighting. We adopt a more effective fusion strategy, which first utilizes the CA Mechanism to deeply mine and extract the complementary information between modalities, and generate new unimodal feature representations that are rich in interaction properties. Subsequently, these newly generated unimodal features are fused twice to further enrich the diversity. Finally, our fusion framework takes the original unimodal feature fusion results, the new unimodal feature fusion results obtained based on the CA mechanism, as well as the original unimodal features for OWC operations, and performs comprehensive evaluation and integration through the adaptive dynamic weight assignment mechanism. This process not only makes full use of the original information of each modality, but also fully exploits the potential correlation and complementary advantages between the data through multiple fusion and weight optimization, and constructs a more complete, robust and efficient multimodal feature representation. Our fusion strategy strengthens model performance of survival analysis. The final feature fusion can be expressed as:

$$h_{fusion} = MLP\Big(LMF(P,C) \oplus LMF\Big(\hat{P},\hat{C}\Big) \oplus P \oplus C\Big) \tag{5}$$

where $\hat{P}$ and $\hat{C}$ are new enhanced features obtained after CA processing, respectively, $\oplus$ denotes the OWC operation, and $LMF(\cdot)$ is the low-rank multimodal fusion operation.

The specific fusion operation in the second stage is shown in Fig. 2. Additionally, in the fusion stage, this study imposes alignment constraints on the cross-modal representation. This constraint is achieved by maximizing the similarity between different modal representations. Specifically, in the first stage, we extract feature representations from the input data of each modality. These feature representations are then sent to the CA module for further processing. The similarity between different modal feature representations is maximized during this process to ensure their semantic consistency. This benefit is to promote information exchange and fusion between different modalities, thereby improving the overall model performance and robustness. We use the L1 norm to measure the distance between cross-modal representation and intra-modal representation:

$$\mathcal{L}_{sim} = \frac{1}{d}(\|P - \hat{P}\|_1 + \|C - \hat{C}\|_1) \tag{6}$$

After the second stage of training, we design and train an end-to-end neural network to predict the survival prognosis of patients by integrating multimodal features and feature vectors from clinical data and using multilayer perceptron as a regression prediction model. During training, the network optimizes the loss function related to survival time through the backpropagation algorithm, updating the weights of each modal feature in real-time and promoting their interaction. We conduct data splitting, hyperparameter adjustment, and model evaluation to ensure the model's performance on the training and test sets. The real survival data of patients are then used as labels for training, and the performance of the multimodal fusion prognosis network is finally validated on the test set. The prediction results of the multimodal fusion prognosis network are regarded as the final prognosis results. The Cox survival analysis loss function $\mathcal{L}_{sur}$ is used for the survival prediction, we set α as 0.5 according to (Zhou and Chen 2023), and the total loss function of the CA-MLIF framework is as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{sur} + \alpha\mathcal{L}_{sim} \tag{7}$$

Experiments

Datasets

To validate our CA-MLIF model, we collected glioma and clear cell renal cell carcinoma datasets from TCGA (Chen et al. 2020). We selected 1505 diagnostic tissue slice images from 769 patients of TCGA-GBMLGG, each with survival outcomes and histopathological grading labels. For clear cell renal cell carcinoma in TCGA-KIRC, we manually extracted 1,251 regions of interest from diagnostic section images of 417 ccRCC patients. For genomic analysis, RNA-seq expression data for TCGA-GBMLGG and TCGA-KIRC were obtained from the cBioPortal database (Cerami et al. 2012; Gao et al. 2013). Using DESeq2, we identified the top 240 prognostic genes, in addition to 79 copy number variations and 1 mutation status, which together comprise a total of 320 genomic elements per patient.

Additionally, the in-house LUAD dataset includes CT radiomics images, pathological hematoxylin and eosin-stained whole-slide images, and related clinical information from 1,047 clinical cases. All cases were strictly screened based on inclusion and exclusion criteria. More details are available in the **Supplementary Material**.

Experimental Details and Evaluation Metrics

In practice, we use the Monte Carlo 15-fold cross-validation method and randomly partition the data into the training and test set (according to the ID number) with a ratio of 8:2. The Adam optimizer is used during the training process, with an initial learning rate of 10^{-3}. We utilize the OWC strategy for real-time dynamic parameter adjustment to ensure that the network could fully learn the intrinsic patterns of the data. The mini-batch is set to 16 to effectively use computing resources for model training. We conducted all experiments by PyTorch and trained on NVIDIA GeForce RTX 3080Ti GPU.

We use the C-index (Steck et al. 2007), a standard metric in survival analysis, to assess the model's performance across the three datasets. The C-index can be calculated us-

Model	CT	Path_Tu	Path_Ep	Path_St	Path_Nu	Path_All	CT+Path
C-index	0.688	0.645	0.621	0.618	0.637	0.667	**0.719**

Table 1: C-index of single-modal and CA-MLIF survival prediction models on multi-center LUAD datasets.

ing the following formula:

$$C - index = \frac{\sum_{i,j} 1_{t_j<t_i} \cdot 1_{\hat{Y}_j^{(1)}>\hat{Y}_i^{(1)}} \cdot \delta_j}{\sum_{i,j} 1_{t_j<t_i} \cdot \delta_j} \quad (8)$$

the C-index ranges from 0 to 1. If the C-index is 0.5, the model's prediction is ineffective. Values above 0.5 indicate effective prediction, with higher values representing better predictive performance.

Multimodal Survival Prediction

In this experiment, in addition to implementing end-to-end cross-modal fusion, survival prediction models were trained using CT images and pathology data separately as inputs. The experimental results are as shown in Tab. 1. Path_Tu, Path_Ep, Path_St, Path_Nu, and Path_All represent the pathological tumor region features, pathological epithelial tissue features, pathological stroma features, pathological nucleus features, and fused features of CA-MLIF model, respectively. Specifically, the C-index for CT images and pathology used to train the single-modal prognosis models were 0.688 and 0.667, respectively. Using the CA-MLIF model for multimodal data fusion training, the performance improved by 3.1% and 5.2%, respectively. This striking contrast highlights significant advantages of multimodal data fusion in survival prediction tasks. Furthermore, to further explore the potential of pathology data in survival prediction, this paper conducted separate survival analyses for different pathological features. It was found that when the pathological tumor region features were the focus of the analysis, the prediction results were the most prominent. This finding provides a solid basis for subsequent model optimization and further demonstrates the potential complementarity and synergy between different modalities in multimodal data fusion.

Comparison Experiment

The comparison methods in this experiment include GPDBN (Wang et al. 2021), Pathomic Fusion (Chen et al. 2020), MCAT (Chen et al. 2021b), MOTCat (Xu and Chen 2023), CMTA (Zhou and Chen 2023), and PG-MLIF (Pan et al. 2024). Pathomic Fusion, GPDBN, MCAT, MOTCat, and PG-MLIF methods all primarily integrate pathological information based on genomics. CMTA was the previous best-performing multimodal fusion method. Compared to CMTA, this model improved performance by 4.5% on GBMLGG, 0.3% on KIRC, and 2.6% on the multi-center LUAD dataset. Tab. 2 shows the experimental results of all methods on three datasets, indicating that the method proposed in this section consistently outperforms other state-of-the-art multimodal learning methods.

Notably, in the KIRC dataset, the performance improvement of the proposed model is not significant. This result may be due to the relatively small size of the KIRC dataset, with only 417 samples. Due to the limitation of sample size, the model may only partially learn the inherent rules and features of the data during training, leading to suboptimal performance on this dataset. To further improve the model's performance on the KIRC dataset, this study needs to take a series of measures in future work, such as increasing the sample size of the dataset, optimizing the model structure and parameters, and introducing more feature engineering. These methods are expected to continue the research and improve the model's performance on the KIRC dataset, and to accurately reveal the underlying rules and patterns of the data.

Method	Datasets		
	GBMLGG	KIRC	MC LUAD
GPDBN('21)	$0.812_{0.015}$	$0.712_{0.038}$	$0.657_{0.024}$
MCAT('21)	$0.835_{0.023}$	$0.708_{0.032}$	$0.672_{0.032}$
Pathomic Fusion('20)	$0.878_{0.009}$	$0.719_{0.031}$	$0.684_{0.033}$
MOTCat('23)	$0.867_{0.028}$	$0.721_{0.026}$	$0.678_{0.038}$
CMTA('23)	$0.863_{0.012}$	$0.729_{0.029}$	$0.693_{0.027}$
PG-MLIF('24)	$0.895_{0.007}$	$0.728_{0.026}$	$0.691_{0.028}$
Ours	$\mathbf{0.908}_{0.011}$	$\mathbf{0.732}_{0.028}$	$\mathbf{0.719}_{0.021}$

Table 2: Comparison of CA-MLIF with other multimodal methods on three datasets. The "MC LUAD" denotes the multi-center LUAD dataset.

Method	FT	C-index	P-value
w/o CA	H+C	$0.696_{0.024}$	4.6426e-07
w/o Alignment Constraint	H+C	$0.703_{0.031}$	6.2170e-05
w/o Second Fusion	H+C	$0.708_{0.027}$	6.4282e-08
CA-MLIF	H+C	$\mathbf{0.719}_{0.021}$	**3.8907e-09**

Table 3: Indicators of concordance between CA-MLIF survival prediction and ablation experiments in multi-center LUAD dataset. The FT denotes the Feature Type, "H+C" denotes that both histology features and CT features are used in the experiment simultaneously.

Ablation Study

To evaluate the efficiency of CA-MLIF during the fusion stage, ablation experiments were designed to assess the impact of each module and its constraints on overall performance. Modules were removed sequentially while keeping the model's functionality. The results, based on the multi-center LUAD dataset, are shown in Tab. 3. The ablation study involved three steps: **(1) CA Module:** This module

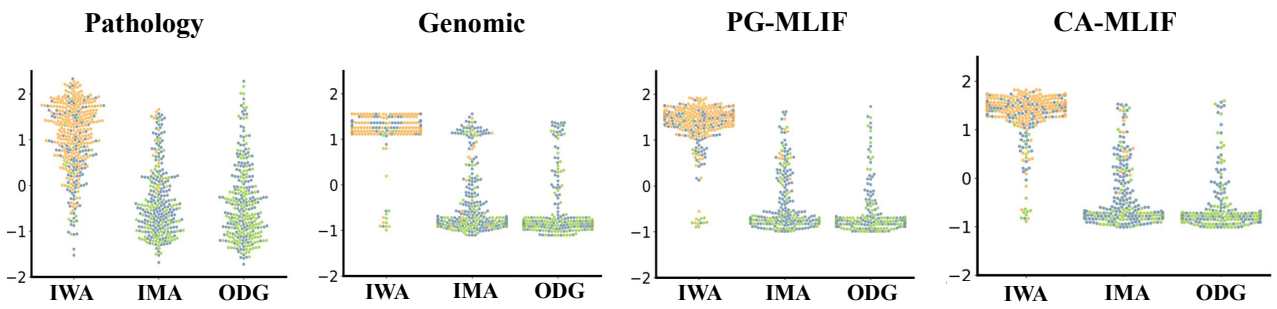

Figure 3: Predictive risk distribution of GBMLGG patients. The four columns illustrate the visualization results for pathology, genomic data, PG-MLIF, and CA-MLIF, respectively. The vertical axis denotes the Hazard (Z-Score), while the horizontal axis labels IWA, IMA, and ODG, corresponding to IDH-wt astrocytoma, IDH-mut astrocytoma, and oligodendroglioma, respectively. Orange, green, and blue dots represent Grades IV, II, and III, respectively.

aims to explore the potential correlation and interaction between multimodal data. After removing this module and conducting the same fusion using the original features, the performance on the multi-center lung cancer data dropped by 2.3%. This significant performance decline indicates that simply relying on the original features is insufficient during the cross-modal fusion process. We also need to fully utilize the information closely related to disease diagnosis. **(2) Alignment Constraint:** To ensure the quality of the information after the CA, this experiment applied an alignment constraint $\mathcal{L}_{sim}$ to the cross-modal representation. If this loss item is removed during model optimization, the performance on the multi-center LUAD dataset drops by 1.6%. This indicates that alignment helps maintain information consistency between modalities, reducing information loss and enhancing the model's performance. **(3) Second Fusion:** To ensure effective fusion, this experiment combines enhanced features with the original features for fusion, followed by optimal weight concatenation (OWC). Without the second fusion, the performance on the multi-center LUAD dataset drops by 1.1%. This result indicates that the second fusion allows the model to fully learn and utilize the complementary specific information in the two features. The original features provide the foundation and details of single-modal data, while the enhanced features capture the potential correlation and interaction between modalities through the CA module. The combination of the two further improves the accuracy and reliability of the diagnosis. The results were thoroughly validated through multiple experimental runs, showing statistically significant improvements.

Visualization Results

From Fig. 3, we observe that compared to other methods, CA-MLIF more accurately allocates patients into three high-density clusters, aligning with the WHO subtyping paradigm. The Kaplan-Meier (K-M) curves in Fig. A5 of **Supplementary Material** demonstrate actual grades with anticipated risk categories. It is observed that the K-M curves for intermediate and low-risk patients interweave, indicating a challenge in stratifying low-to-intermediate-risk patients. In contrast, the K-M curves for the intermediate-to-high-risk group display clear distinctions, affirming the challenge in distinguishing WHO Grade II and III from Grade III and IV. Notably, the CA-MLIF in Fig. A5 distinctly segregates low, medium, and high-risk patients, effectively alleviating the confounding of intermediate-to-low-risk patients. This vividly underscores the superior performance of the CA-MLIF model.

In Fig. A6(a) of **Supplementary Material**, a comparison is made between the actual grades and anticipated risk categories within the same K-M curves. We observe that the stratification of high and low-risk patients by CA-MLIF is more distinct, demonstrating the superior performance of our proposed models. Similarly, upon comparing the risk distribution of KIRC patients with shorter and longer survival times (Fig. A6(b) in **Supplementary Material**), CA-MLIF and PG-MLIF excel in stratifying patients into groups with distinct short and prolonged survival, revealing a bimodal risk prediction plot. Here, the survival risk is quantified using the Hazard Ratio, computed through the semi-parametric Cox proportional hazard model.

Conclusion

This paper proposes a framework named CA-MLIF that integrates end-to-end learning strategies, aiming to optimize the performance of survival prognosis analysis. The core of the framework lies in its unique CA mechanism, which enables bidirectional interaction at the feature level across modal boundaries, eliminating the limitations inherent in traditional unidirectional fusion methods. End-to-end training improves the modeling of multimodal data and captures rich semantic information. Model parameters are optimized by backpropagation to minimize task-specific loss functions. The innovative fusion strategy focuses on using LMF with quadratic adaptive weight allocation for fusion, effectively solving the problems of low fusion efficiency and insufficient interaction between features. Extensive experiments demonstrate that our proposed method outperforms other multimodal fusion networks.

Acknowledgments

This work was supported in part by the Guangxi Natural Science Foundation (No. 2024GXNSFFA010014) and the National Natural Science Foundation of China (Nos. 82360356, 62172120 and 82272075)

References

Bala, N.; Gupta, R.; and Kumar, A. 2022. Multimodal biometric system based on fusion techniques: a review. *Information Security Journal: A Global Perspective*, 31(3): 289–337.

Bray, F.; Laversanne, M.; Sung, H.; Ferlay, J.; Siegel, R. L.; Soerjomataram, I.; and Jemal, A. 2024. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 74(3): 229–263.

Cao, Y.; Cui, L.; Zhang, L.; Yu, F.; Li, Z.; and Xu, Y. 2023. MMTN: Multi-modal memory transformer network for image-report consistent medical report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 277–285.

Cerami, E.; Gao, J.; Dogrusoz, U.; Gross, B. E.; Sumer, S. O.; Aksoy, B. A.; Jacobsen, A.; Byrne, C. J.; Heuer, M. L.; Larsson, E.; et al. 2012. The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discovery*, 2(5): 401–404.

Chen, H.; Xie, W.; Afouras, T.; Nagrani, A.; Vedaldi, A.; and Zisserman, A. 2021a. Localizing visual sounds the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16867–16876.

Chen, R. J.; Lu, M. Y.; Wang, J.; Williamson, D. F.; Rodig, S. J.; Lindeman, N. I.; and Mahmood, F. 2020. Pathomic fusion: An integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Transactions on Medical Imaging*, 41(4): 757–770.

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

Cui, C.; Yang, H.; Wang, Y.; Zhao, S.; Asad, Z.; Coburn, L. A.; Wilson, K. T.; Landman, B. A.; and Huo, Y. 2023. Deep multimodal fusion of image and non-image data in disease diagnosis and prognosis: a review. *Progress in Biomedical Engineering*, 5(2): 022001.

Ding, S.; Li, J.; Wang, J.; Ying, S.; and Shi, J. 2024. Multimodal co-attention fusion network with online data augmentation for cancer subtype classification. *IEEE Transactions on Medical Imaging*, 43(11): 3977–3989.

Fukui, A.; Park, D. H.; Yang, D.; Rohrbach, A.; Darrell, T.; and Rohrbach, M. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*.

Gao, J.; Aksoy, B. A.; Dogrusoz, U.; Dresdner, G.; Gross, B.; Sumer, S. O.; Sun, Y.; Jacobsen, A.; Sinha, R.; Larsson, E.; et al. 2013. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science Signaling*, 6(269): pl1–pl1.

Guo, S.; Wang, L.; Chen, Q.; Wang, L.; Zhang, J.; and Zhu, Y. 2022. Multimodal MRI image decision fusion-based network for glioma classification. *Frontiers in Oncology*, 12: 819673.

Hoda, S. A.; and Cheng, E. 2017. Robbins basic pathology.

Jamil, S.; Jalil Piran, M.; and Kwon, O.-J. 2023. A comprehensive survey of transformers for computer vision. *Drones*, 7(5): 287.

Li, J.; Jia, C.; and Qian, C. 2021. Progressive breast cancer diagnosis model based on multi-classifier and multi-modal fusion. *International Journal of Machine Learning and Computing*, 11(6).

Liu, J.; Liu, Z.; Wu, G.; Ma, L.; Liu, R.; Zhong, W.; Luo, Z.; and Fan, X. 2023. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8115–8124.

Lu, C.; Shiradkar, R.; and Liu, Z. 2021. Integrating pathomics with radiomics and genomics for cancer prognosis: A brief review. *Chinese Journal of Cancer Research*, 33(5): 563.

Pan, X.; An, Y.; Lan, R.; Liu, Z.; Liu, Z.; Lu, C.; and Yang, H. 2024. PG-MLIF: Multimodal low-rank interaction fusion framework integrating pathological images and genomic data for cancer prognosis prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 347–357. Springer.

Pan, X.; Cheng, J.; Hou, F.; Lan, R.; Lu, C.; Li, L.; Feng, Z.; Wang, H.; Liang, C.; Liu, Z.; et al. 2023. SMILE: Cost-sensitive multi-task learning for nuclear segmentation and classification with imbalanced annotations. *Medical Image Analysis*, 88: 102867.

Patrick, M.; Huang, P.-Y.; Misra, I.; Metze, F.; Vedaldi, A.; Asano, Y. M.; and Henriques, J. F. 2021. Space-time crop & attend: Improving cross-modal video representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10560–10572.

Schneider, L.; Laiouar-Pedari, S.; Kuntz, S.; Krieghoff-Henning, E.; Hekler, A.; Kather, J. N.; Gaiser, T.; Froehling, S.; and Brinker, T. J. 2022. Integration of deep learning-based image analysis and genomic data in cancer pathology: A systematic review. *European Journal of Cancer*, 160: 80–91.

Steck, H.; Krishnapuram, B.; Dehing-Oberije, C.; Lambin, P.; and Raykar, V. C. 2007. On ranking in survival analysis: Bounds on the concordance index. *Advances in Neural Information Processing Systems*, 20.

Tan, K.; Huang, W.; Liu, X.; Hu, J.; and Dong, S. 2022. A multi-modal fusion framework based on multi-task correlation learning for cancer prognosis prediction. *Artificial Intelligence in Medicine*, 126: 102260.

Wang, Y.; Huang, W.; Sun, F.; Xu, T.; Rong, Y.; and Huang, J. 2020. Deep multimodal fusion by channel exchanging.

Advances in Neural Information Processing Systems, 33: 4835–4845.

Wang, Y.; Pan, X.; Lin, H.; Han, C.; An, Y.; Qiu, B.; Feng, Z.; Huang, X.; Xu, Z.; Shi, Z.; et al. 2022. Multi-scale pathology image texture signature is a prognostic factor for resectable lung adenocarcinoma: A multi-center, retrospective study. *Journal of Translational Medicine*, 20(1): 595.

Wang, Z.; Li, R.; Wang, M.; and Li, A. 2021. GPDBN: Deep bilinear network integrating both genomic data and pathological images for breast cancer prognosis prediction. *Bioinformatics*, 37(18): 2963–2970.

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

Yang, J.; Ju, J.; Guo, L.; Ji, B.; Shi, S.; Yang, Z.; Gao, S.; Yuan, X.; Tian, G.; Liang, Y.; et al. 2022. Prediction of HER2-positive breast cancer recurrence and metastasis risk from histopathological images and clinical information via multimodal deep learning. *Computational and Structural Biotechnology Journal*, 20: 333–342.

Zhou, F.; and Chen, H. 2023. Cross-modal translation and alignment for survival analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21485–21494.