PG-MLIF: Multimodal Low-rank Interaction Fusion Framework Integrating Pathological Images and Genomic Data for Cancer Prognosis Prediction

Xipeng Pan[1], Yajun An[1], Rushi Lan[2] (✉), Zhenbing Liu[1], Zaiyi Liu[1,3], Cheng Lu[1,3], and Huihua Yang[1,4] (✉)

[1] School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin, Guangxi 541004, China
[2] International Joint Research Laboratory of Spatio-temporal Information and Intelligent Location Services, Guilin University of Electronic Technology, Guilin 541004, China
rslan@guet.edu.cn
[3] Department of Radiology, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou, Guangdong 510080, China
[4] School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China
yhh@bupt.edu.cn

Abstract. Precise prognostication can assist physicians in developing personalized treatment and follow-up plans, which help enhance the overall survival rates. Recently, enormous amount of research rely on unimodal data for survival prediction, not fully capitalizing on the complementary information available. With this deficiency, we propose a Multimodal Low-rank Interaction Fusion Framework Integrating Pathological images and Genomic data (PG-MLIF) for survival prediction. In this framework, we leverage the gating-based modality attention mechanism (MAM) for effective filtering at the feature level and propose the optimal weight concatenation (OWC) strategy to maximize the integration of information from pathological images, genomic data, and fused features at the model level. The model introduces a parallel decomposition strategy called low-rank multimodal fusion (LMF) for the first time, which simplifies the complexity and facilitates model contribution-based fusion, addressing the challenge of incomplete and inefficient multimodal fusion. Extensive experiments on the public dataset of GBMLGG and KIRC demonstrate that our PG-MLIF outperforms state-of-the-art survival prediction methods. Additionally, we significantly stratify patients based on the hazard ratios obtained from training the two types of datasets, and the visualization results were generally consistent with the true grade classification. The code is available at: https://github.com/panxipeng/PG-MLIF.

Keywords: Pathological images analysis · Multi-genomics data · Multimodal learning · Low-rank interaction fusion · Survival prediction.

1 Introduction

As a highly aggressive disease, cancer presents challenges in accurate prognosis prediction due to tumor heterogeneity and biological complexity [3]. Therefore, accurate prognosis prediction is a formidable challenge. In this context, assessing tumor progression and accurately predicting prognosis is crucial for physicians to make correct decisions. Previous studies have primarily relied on morphological information from pathological images [18,14] or genetic information [10] for single-modal prognosis prediction [15]. However, single-modal data capture features with a more homogeneous dimension and lack information from different levels. Recent experiments have demonstrated that effective multi-modal data fusion can lead to more accurate patient prognosis predictions [12]. Tumor biology research involves various data types, such as pathological images, genomic data, and clinical information, each offering unique tumor biological characteristics. Pathological tissue images are considered crucial for cancer diagnosis, providing nuclei morphological attributes related to tumor invasion, while genomic data, including gene mutations and expression features, contribute to cancer diagnosis and prognosis prediction [5,9].

Research on cancer prognosis prediction is advancing with the application of multi-modal data and deep learning-based fusion methods in survival analysis. Wang et al. [17] implemented a deep bilinear network integrating genes and pathologies for fusion within and between modalities, and Tan et al. [16] proposed a multimodal fusion framework based on multi-task correlation learning, with the core fusion method of vector concatenation. However, due to complex relationships between modalities, simple feature selection methods may not fully capture diverse information, necessitating more comprehensive fusion techniques. Another challenge arises from the curse of dimensionality when fusing data from different modalities, requiring more computational resources and complex models. For example, Chen et al. [6] and Gordon et al. [2] captured cross-modal interactions with tensor-based Kronecker products, increasing feature dimensionality and overfitting risk. Chen et al. [7] developed a sophisticated feature aggregation strategy using a co-attentive transformer approach in multi-instance learning to capture genotype-phenotype interactions for prognostic prediction. Lu et al. [13] similarly used a fusion strategy with Transformer for Glioma patients for tumor grading and survival analysis. Zhou et al. [20] unified multimodal Transformers for patient triaging. Despite challenges, judicious fusion leverages the complementarity of multi-modal information to enhance survival prognosis prediction.

Given the challenges above, we propose a Multi-modal Low-rank Interaction Fusion Framework Integrating Pathological images and Genomic data(PG-MLIF). The framework use pathological images and multi-genomic data, where pathological images focus on morphological changes at the tissue level, and genomic data pay more attention to genetic information and variation at molecular levels, which complement each other at different levels. Our technical contributions in PG-MLIF are three folds: (i) The MLIF framework consists of a gating-based modality attention mechanism (MAM), low-rank multimodal fu-

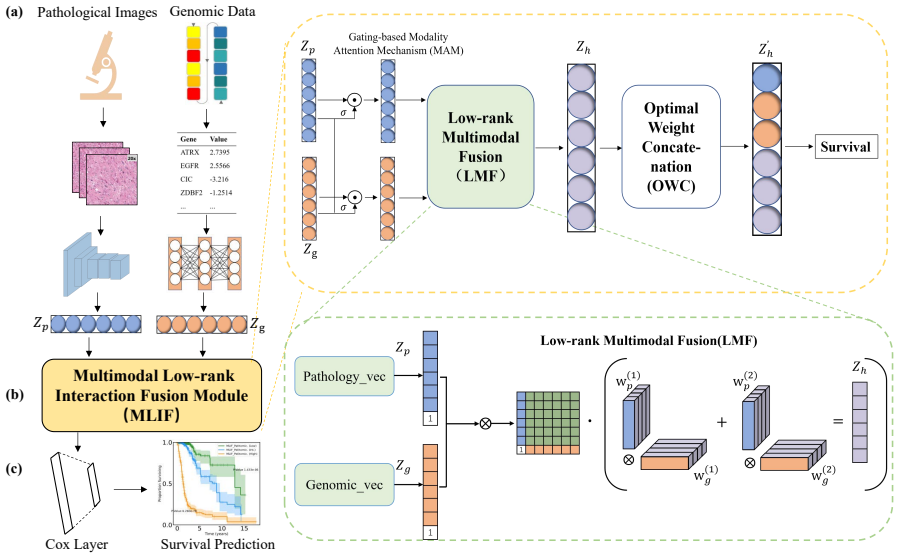

Fig. 1. The proposed PG-MLIF framework. (a) Training single-modal networks for pathological images and genomic data separately. (b) Utilization of the outcomes from stage (a) as inputs to the MLIF model for low-rank multimodal fusion. (c) Applying the Cox model for survival prediction.

sion (LMF), and optimal weight concatenation (OWC). The framework enables effective fusion at the feature level and model level of different data. (ii) Parallel decomposition is introduced for the first time to efficiently achieve low-rank fusion of medical multimodal data, which reduces the number of parameters and computational costs. (iii) The proposed method outperforms current SOTA ones in the glioma and clear cell renal cell carcinoma data. A comprehensive visualizations have comfirmed its robust predictive capabilities.

2 Methodology

An overview of the PG-MLIF framework is illustrated in Fig. 1. PG-MLIF is co-designed by two parts: unimodal extraction of features and fusion network architectures. Sections 2.1 to 2.2 will present these two parts separately. Additionally, section 2.3 elaborates on our pivotal fusion method, denoted as Low-rank Multimodal Fusion (LMF). This approach leverages a parallel decomposition strategy for efficient low-rank cross-modal interaction fusion.

2.1 Unimodal Extraction of Feature

Among them, pathological features have a profound impact on the prognostic assessment and treatment planning of tumors. In order to capture these fea-

tures, ResNet-50 is employed as the backbone for feature extraction, and the model is fine-tuned using pre-trained ImageNet weights. Multimodal genomic data has complex properties such as nonlinearity, high dimensionality, sparsity, and strong correlation, and in this study, self-normalizing network (SNN) was employed to achieve globally optimal end-to-end training, which contributes to better subsequent survival prediction. Finally, we extracted the embedding vectors $Z_p \in R^{32\times1}$, $Z_g \in R^{32\times1}$ in pathology as well as genomics, respectively, and used them as inputs to the fusion network MLIF.

2.2 Multimodal Low-rank Interaction Fusion Framework

Our multimodal network architecture, known as MLIF consists of three key components: gating-based modality attention mechanism (MAM), low-rank multimodal fusion (LMF), and optimal weight concatenation (OWC).

To harness the full potential of these disparate modalities, we present a gating-based attention mechanism [1] that dynamically determines the contribution of each modality to the feature expression, thereby adapting to the intrinsic importance of each modality. Subsequently, we present LMF to transform the feature matrices of both modalities into low-rank counterparts. This approach engenders an intricate interaction between the features, capturing the full spectrum of potential cross-modal interactions. The comprehensive methodological details of this feature fusion approach are elaborated upon in section 2.3. With this approach, we can fully exchange the information of the two modalities, mine more correlations. It is also possible to reduce a large number of parameters in order to better train the model. Finally, for optimal feature representation, we propose a method known as OWC. This method works by adaptively assigning different weights to different modalities in order to better utilize the respective information. The details of the proposed OWC approach are elaborated upon in section 2.4. In conclusion, the framework effectively addresses the challenges of incomplete as well as inefficient cross-modal fusion and provides diverse features for improving the performance of prognostic prediction.

2.3 Low-rank Multimodal Fusion

The primary objective of our study is to integrate single-modal features into compact ones which are suitable for downstream tasks. A pivotal component of the comprehensive MLIF fusion module is the LMF (Fig. 1). The technique captures important interactions by introducing low-rank factors and utilizing parallel decomposition of the low-rank power tensor and the input tensor for fusion. In contrast to traditional tensor fusion networks (TFN) techniques, the distinctive feature of LMF fusion lies in its ability to avoid explicit weight creation for capturing interactions. Additionally, LMF fusion exhibits linear scalability in modalities, reducing model parameters.

First, before fusion, in order to capture all the important interactions between unimodal and bimodal data, we add another dimension to each feature vector before computing the Kronecker product. This ensures that the unimodal

features remain unaffected, and simultaneously facilitates a more comprehensive fusion of information between the two modalities while preserving the original features. This operation, as proposed by Zadeh et al. [19], involves adding a dimension to the unimodal features before taking the outer product. Specifically, the input tensor Z, formed by computing the unimodal features, is given by:

$$Z_h = \begin{bmatrix} Z_p \\ 1 \end{bmatrix} \otimes \begin{bmatrix} Z_g \\ 1 \end{bmatrix}, Z_{p,g} \in \mathbb{R}^{d_m}. \tag{1}$$

In this computation, $\otimes$ is the outer product, i.e., Kronecker product, that is applied to a tensor with m modalities, each i-modality having a different feature dimension d_i, and these features form a differential multimodal tensor Z_h in a two-dimensional Cartesian space. To subsequently feed its results into the predictive model, reducing the feature tensor Z_h to h dimensions is usually necessary, so we use a linear layer $g(\cdot)$ to generate a vector representation as the output of tensor $Z \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_M}$:

$$h = g(Z; W, b) = W \cdot Z + b, h, b \in \mathbb{R}^{d_y}, \tag{2}$$

where W is the weight of this layer and b is the bias. In performing the tensor dot product $W \cdot Z$, we consider W to be composed of d_h M-order tensors. Since this operation involves full connectivity, it is necessary to explicitly create a high-dimensional tensor Z, whose dimensionality grows exponentially with the number of modalities, and the weight tensor W to be learned grows exponentially accordingly. This increases the computational complexity and the risk of the model being exposed to overfitting. The LMF method introduced in this experiment[11] is an improvement built upon TFN. In the tensor dot product $W \cdot Z$, we still use the above procedure and consider W as d_h matrices, and for each matrix $\tilde{W}_k \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_M}, k = 1, ..., d_h$, there exists an exact decomposition vector of the form.

$$\tilde{W}_k = \sum_{i=1}^{R} \otimes_{m=1}^{M} w_{m,k}^{(i)}, w_{m,k}^{(i)} \in \mathbb{R}_m^d, \tag{3}$$

where the minimum R that makes the decomposition valid is called the rank of the tensor, in this experiment, the rank is set to a fixed value r. Therefore, based on the decomposition of W and then on $Z = \otimes_{m=1}^{M} z_m$, we can extrapolate the original equation for calculating h as follows (in the case of a bimodal state):

$$\begin{aligned} h &= \left(\sum_{i=1}^{r} \otimes_{m=1}^{M} w_m^{(i)} \right) \cdot Z = \sum_{i=1}^{r} \left(\otimes_{m=1}^{M} w_m^{(i)} \cdot \otimes_{m=1}^{M} z_m \right) \\ &= \left(\sum_{i=1}^{r} w_p^{(i)} \cdot Z_p \right) \circ \left(\sum_{i=1}^{r} w_g^{(i)} \cdot Z_g \right), \end{aligned} \tag{4}$$

where $\circ$ denotes the element level product, i.e., Hadamard product, an essential aspect of this simplification involves leveraging the parallel decomposition from

Z and W so that we can calculate h without actually creating the tensor Z from the input representation Z_h.

For the specific implementation of the above LMF method, the core fusion process involves constructing an interaction matrix using distinct matrices for pathology and genetics, applying tensor decomposition to partition the feature tensor into low-rank components, extracting and combining shared data from all modalities, and forming a final output vector by linearly combining the low-rank and weight matrices. This approach effectively mitigates dimensional explosion and facilitates model training by transforming multidimensional information into low-dimensional feature interactions.

2.4 Optimal Weight Concatenation

Our proposed OWC method uses an adaptive dynamic allocation strategy. This approach assigns weights to each modality based on its prognostic performance, adjusting these weights at the model level according to the contribution of the input data. Specifically, we combine the feature vectors of all models into a composite vector Z_i from all models into a combined vector Z=[Z_p,Z_g,Z_h] and then use a small neural network (combining a feedforward network and a Sigmoid activation) to learn optimal weights are finally determined to maximize the performance of the model. Once the optimal weights are determined, apply these weights to the corresponding feature vectors. The final combined feature vector can be expressed as $Z_h^{'} = [w_p Z_p, w_g Z_g, w_h Z_h]$. By this method, the information within and between different modalities is fully utilized, which makes the whole fusion process correlate with the original data and get the best feature expression, thus improving the generalization performance of the feature expression.

3 Experiments and Results

3.1 Dataset

To validate our MLIF model, we gathered glioma and clear cell renal cell carcinoma datasets from TCGA [6]. We selected 1505 diagnostic tissue images with survival outcomes and grading labels, representing 769 patients. For genomic data, we downloaded RNA-seq data from TCGA-GBMLGG and TCGA-KIRC via cBioPortal [8,4]. After DESeq2 analysis, we chose the top 240 prognostic genes, combining copy number variations (79), mutations (1), and expression levels. Each patient's genomic data comprises 320 elements.

3.2 Implementation Details

In practice, we use the Monte Carlo 15-fold cross-validation (MCCV) method and randomly partition the data into the training and test set (according to the ID number) with a ratio of 8:2. The MLIF model is trained with Adam optimizer (lr=2e-4), batch sizes of 8 for pathology and 64 for genes. Implemented in PyTorch, training was conducted on NVIDIA GeForce RTX 3080Ti GPU.

Table 1. Comparison of PG-MLIF with other multimodal methods in GBMLGG.

Model	Pathology	Genomic	Multi-modal	Fusion Method
MultiCoFusion[16]	0.783±0.016	0.838±0.018	0.857±0.015	Concatenation
MCAT[7]	0.787±0.028	0.598±0.054	0.817±0.021	Co-Attention Transformer
DOF[2]	0.715±0.054	0.716±0.063	0.788±0.067	TFN+MMO Loss
M^2F [13]	0.739	0.798	0.827	Transformer Encoder
Pathomic Fusion [6]	0.814±0.023	0.866±0.014	0.878±0.009	TFN
GPDBN[17]	NA	NA	0.812±0.015	Inter + intra TFN
Ours	**0.821±0.020**	**0.873±0.011**	**0.895±0.007**	LMF + OWC

Table 2. The C-index value of PG-MLIF survival prognosis and ablation experiments in KIRC.

Model	Feature Type	C-index	$P-Value^*$
VGG19	Pathology	0.671±0.023	4.1943e-04
ResNet-50	Pathology	0.676±0.024	2.1542e-04
SNN(ReLU)	Genomic	0.684±0.025	3.5955e-11
SNN(SiLU)	Genomic	0.703±0.028	1.7532e-11
GPDBN[17]	Pathology + Genomic	0.712±0.038	1.4327e-14
Pathomic Fusion [6]	Pathology + Genomic	0.719±0.031	1.1537e-11
MLIF(OWC Only)	Pathology + Genomic	0.718±0.019	4.0134e-13
MLIF(LMF Only)	Pathology + Genomic	0.725±0.030	5.1309e-12
MLIF(LMF + OWC)	Pathology + Genomic	**0.728±0.028**	5.7684e-12

* The P-value is obtained by using each feature type as a classifier and then calculating the Log-rank test based on the risk values obtained from the testing.

3.3 Evaluation metrics

To assess the performance of our method, we utilize the C-index as an evaluation metric. The C-index assesses the accuracy of survival prediction models by comparing predicted and valid survival time orders. The formula is:

$$C-index = \frac{\sum_{i,j} 1_{t_j<1_i} \cdot 1_{\hat{Y}_j^{(1)}>\hat{Y}_i^{(1)}} \cdot \sigma_j}{\sum_{i,j} 1_{t_j<t_i} \cdot \sigma_j}, \tag{5}$$

where the term $1_{t_j<1_i}$ indicates that $1_{t_j<1_i} = 1$ when $t_j < 1_i$, and otherwise it is equal to 0. The C-index ranges from 0 to 1, with 0.5 indicating that the model's predictive performance is no better than random chance. A C-index above 0.5 signifies improved predictive accuracy.

3.4 Results

Comparison Study. We compared our fusion module with six previous multimodal fusion methods on the TCGA-GBMLGG dataset. The results in Table 1 indicate that MLIF outperforms other methods in predicting glioma survival prognosis, with improved C-index. Both our approach and Pathomic Fusion performed well, prompting further validation using the TCGA-KIRC dataset. Using MCCV, our model and Pathomic Fusion achieved C-index values of 0.728 ± 0.028 and 0.719 ± 0.031, respectively (Table 2).

Table 3. The C-index value of PG-MLIF survival prognosis and ablation experiments in GBMLGG.

Model	Feature Type	C-index	$P-Value^*$
Cox	Age + Gender	0.740±0.019	5.6146e-29
	Grade	0.751±0.013	2.2572e-59
	Subtype	0.773±0.011	3.5269e-49
	Subtype + Gender	0.789±0.013	4.3432e-59
ResNet-50	Pathology	0.821±0.020	4.0153e-88
SNN(ReLU)	Genomic	0.866±0.014	2.9415e-66
SNN(SiLU)	Genomic	0.873±0.011	7.0080e-41
MLIF(OWC Only)	Pathology + Genomic	0.881±0.014	9.6306e-47
MLIF(LMF Only)	Pathology + Genomic	0.891±0.013	2.0130e-83
MLIF(LMF + OWC)	Pathology + Genomic	**0.895±0.007**	6.2605e-78

Table 4. Comparison of the training and testing speeds between LMF and TFN.

Model	Parameters	Training Speed(IPS)	Testing Speed(IPS)
Pathomic Fusion(TFN)	773,219	636.94	1424.20
MLIF(LMF)	186,115	1210.65	2250.54

Ablation Study. In comparing survival prognosis tasks across two distinct datasets, MLIF demonstrates superior performance compared to its foundational experimentation outcomes. The success of MLIF is attributed to integrating LMF's parallel decomposition strategy and OWC's optimal weight allocation, effectively combining diverse multimodal features. To evaluate the pivotal roles of LMF and OWC, we conduct a comprehensive ablation study with two modes: (1) LMF only. (2) OWC only. Using the second approach exclusively, the C-index achieves noteworthy values of 0.881 and 0.718 for the respective TCGA-GBMLGG and TCGA-KIRC datasets, showing an improvement of approximately 5% over the average baseline performance (Table 3). Employing only the first method results in improved C-index values for both datasets. Integrating both modes within MLIF model further elevates performance, with statistically significant results thoroughly validated through multiple experimental runs.

Complexity Analysis. In practice, our model has significantly fewer parameters compared to Pathomic Fusion. Additionally, we evaluated MLIF's computational complexity by comparing training and testing speeds with Pathomic Fusion. Using the Time package, we measured 1000 inferences in both models. The results in Table 4 demonstrate the superiority of the LMF method in medical image survival analysis, showing improved training and testing speed.

Visualization and Analysis. To delve deeper into improving patient stratification through multimodal interactive low-rank fusion, we constructed Kaplan-Meyer (K-M) curves (see Supplementary Materials). All models demonstrated statistically significant differences, closely matching actual grade stratification. This highlights the clinical importance of PG-MLIF in predicting patient strati-

fication and improving the prognosis of glioma patients.In addition, scrutinizing predicted patient risk scores, we compared actual vs. anticipated risk within K-M curves and showed risk distribution plots generated by different networks.

4 Conclusions

This study proposes a novel PG-MLIF framework, that employs deep learning, for integrating pathological images and genomic data. To effectively integrate multimodal data, we use MAM for meticulous feature selection, introduce the LMF technique to improve operational efficiency as well as capture comprehensive features, and propose OWC to further enrich the representational depth of our features. Extensive experimental results on two datasets demonstrate that our survival prognostic results outperform both unimodal and existing multimodal fusion methods. In addition, the visualization results indicate the framework's strong ability to distinguish short-term survivors from long-term ones.

Acknowledgments. This work was supported in part by Guangxi Natural Science Foundation (No. 2024GXNSFFA010014 and 2019GXNSFFA245014) and National Natural Science Foundation of China (No. 82360356, 62172120, and 62376038).

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

References

1. Arevalo, J., Solorio, T., Montes-y Gómez, M., González, F.A.: Gated multimodal units for information fusion. arXiv preprint arXiv:1702.01992 (2017)
2. Braman, N., Gordon, J.W., Goossens, E.T., Willis, C., Stumpe, M.C., Venkataraman, J.: Deep orthogonal fusion: Multimodal prognostic biomarker discovery integrating radiology, pathology, genomic, and clinical data. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24. pp. 667–677. Springer (2021)
3. Ramón y Cajal, S., Sesé, M., Capdevila, C., Aasen, T., De Mattos-Arruda, L., Diaz-Cano, S.J., Hernández-Losa, J., Castellví, J.: Clinical implications of intratumor heterogeneity: challenges and opportunities. Journal of Molecular Medicine **98**, 161–177 (2020)
4. 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.: The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discovery **2**(5), 401–404 (2012)
5. Chan, J.K.: The wonderful colors of the hematoxylin–eosin stain in diagnostic surgical pathology. International Journal of Surgical Pathology **22**(1), 12–32 (2014)
6. Chen, R.J., Lu, M.Y., Wang, J., Williamson, D.F., Rodig, S.J., Lindeman, N.I., Mahmood, F.: 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 (2020)

7. 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/CVF International Conference on Computer Vision. pp. 4015–4025 (2021)
8. Gao, J., Aksoy, B.A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S.O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., et al.: Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal. Science Signaling **6**(269), pl1–pl1 (2013)
9. Hankey, W., McIlhatton, M.A., Ebede, K., Kennedy, B., Hancioglu, B., Zhang, J., Brock, G.N., Huang, K., Groden, J.: Mutational mechanisms that activate wnt signaling and predict outcomes in colorectal cancer patients. Cancer Research **78**(3), 617–630 (2018)
10. Huang, Z., Zhan, X., Xiang, S., Johnson, T.S., Helm, B., Yu, C.Y., Zhang, J., Salama, P., Rizkalla, M., Han, Z., et al.: Salmon: survival analysis learning with multi-omics neural networks on breast cancer. Frontiers in Genetics **10**, 166 (2019)
11. Liu, Z., Shen, Y., Lakshminarasimhan, V.B., Liang, P.P., Zadeh, A., Morency, L.P.: Efficient low-rank multimodal fusion with modality-specific factors. arXiv preprint arXiv:1806.00064 (2018)
12. Lu, C., Shiradkar, R., Liu, Z.: Integrating pathomics with radiomics and genomics for cancer prognosis: A brief review. Chinese Journal of Cancer Research **33**(5), 563 (2021)
13. Lu, Z., Lu, M., Xia, Y.: M 2 f: A multi-modal and multi-task fusion network for glioma diagnosis and prognosis. In: International Workshop on Multiscale Multimodal Medical Imaging. pp. 1–10. Springer (2022)
14. Pan, X., Cheng, J., Hou, F., Lan, R., Lu, C., Li, L., Feng, Z., Wang, H., Liang, C., Liu, Z., et al.: Smile: Cost-sensitive multi-task learning for nuclear segmentation and classification with imbalanced annotations. Medical Image Analysis p. 102867 (2023)
15. Pan, X., Lin, H., Han, C., Feng, Z., Wang, Y., Lin, J., Qiu, B., Yan, L., Li, B., Xu, Z., et al.: Computerized tumor-infiltrating lymphocytes density score predicts survival of patients with resectable lung adenocarcinoma. iScience **25**(12) (2022)
16. Tan, K., Huang, W., Liu, X., Hu, J., Dong, S.: A multi-modal fusion framework based on multi-task correlation learning for cancer prognosis prediction. Artificial Intelligence in Medicine **126**, 102260 (2022)
17. Wang, Z., Li, R., Wang, M., Li, A.: Gpdbn: deep bilinear network integrating both genomic data and pathological images for breast cancer prognosis prediction. Bioinformatics **37**(18), 2963–2970 (2021)
18. Yang, J., Ju, J., Guo, L., Ji, B., Shi, S., Yang, Z., Gao, S., Yuan, X., Tian, G., Liang, Y., et al.: 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 (2022)
19. Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.P.: Tensor fusion network for multimodal sentiment analysis. arXiv preprint arXiv:1707.07250 (2017)
20. Zhou, H.Y., Yu, Y., Wang, C., Zhang, S., Gao, Y., Pan, J., Shao, J., Lu, G., Zhang, K., Li, W.: A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. Nature Biomedical Engineering pp. 1–13 (2023)