M2Fusion: Multi-time Multimodal Fusion for Prediction of Pathological Complete Response in Breast Cancer

Song Zhang[1,2], Siyao Du[3], Caixia Sun[4], Bao Li[5], Lizhi Shao[1,2], Lina Zhang[6], Kun Wang[7], Zhenyu Liu[1,2], and Jie Tian[1,4]

[1] CAS Key Laboratory of Molecular Imaging, Institute of Automation, Chinese Academy of Sciences, Beijing, China zhenyu.liu@ia.ac.cn, jie.tian@ia.ac.cn
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3] Department of Radiology, The First Hospital of China Medical University, Shenyang, China
[4] School of Engineering Medicine, Beihang University, Beijing, China
[5] Center for Biomedical Imaging, University of Science and Technology of China, Hefei, China
[6] Department of Radiology, The Fourth Affiliated Hospital of China Medical University, Shenyang, China
[7] Department of Breast Cancer, Guangdong Provincial People's Hospital, Guangzhou, China

Abstract. Accurate identification of patients who achieve pathological complete response (pCR) after neoadjuvant chemotherapy (NAC) is critical before surgery for guiding customized treatment regimens and assessing prognosis in breast cancer. However, current methods for predicting pCR primarily rely on single modality data or single time-point images, which fail to capture tumor changes and comprehensively represent tumor heterogeneity at both macro and micro levels. Additionally, complementary information between modalities is not fully interacted. In this paper, we present **M2Fusion**, pioneering the fusion of multi-time multimodal data for treatment response prediction, with two key components: the multi-time magnetic resonance imagings (MRIs) contrastive learning loss that learns representations reflecting NAC-induced tumor changes; the orthogonal multimodal fusion module that integrates orthogonal information from MRIs and whole slide images (WSIs). To evaluate the proposed M2Fusion, we collect pre-treatment MRI, post-treatment MRI, and WSIs of biopsy from patients with breast cancer at two different collaborating hospitals, each with the pCR assessed by the standard pathological procedure. Experimental results quantitatively reveal that the proposed M2Fusion improves treatment response prediction and outperforms other multimodal fusion methods and single-modality approaches. Validation on external test sets further demonstrates the generalization and validity of the model. Our code is available at https://github.com/SongZHS/M2Fusion.

Keywords: Multi-time imaging · Multimodal fusion · Pathological complete response · Breast cancer.

1 Introduction

Patients with breast cancer achieving pCR could benefit from breast-conserving surgery, even omitting surgery instead of breast mastectomy [1–3]. Accurate assessment of pCR before surgery is essential for tailoring surgery plans and could select patients with good prognosis in advance, which is an urgent need. However, the gold standard of pCR depends on the pathological results of surgical specimens. Various explorations have been attempted to develop biomarkers for predicting the response to NAC in patients with breast cancer, including imaging techniques like pre-treatment Magnetic Resonance Imaging (MRI), post-treatment MRI, and Whole Slide Imaging (WSI) [4–7]. While these studies have produced encouraging results, the majority are restricted to a single modality of data, neglecting complementary information between modalities [8]. Fusing modalities containing fully orthogonal information would dramatically promote predictive power and beyond single modality [9, 10]. For example, MRI provides insight into the anatomical structure and morphological characteristics of tumors, while WSIs contain information about the tumor microenvironment, complementing the information missed by MRI from a microscopic perspective. Therefore, integrating orthogonally derived data such as MRI and WSI presents an opportunity to discover and develop novel multimodal biomarkers for predicting treatment response.

Recently, most efforts have been concentrated on analyzing images collected from a single time point. Some studies considered intratumoral heterogeneity or introduced topological representations to utilize pre or post-treatment images for pCR prediction [11, 12]. These possess inherent limitations for the purpose of response prediction, as the impact of NAC is not taken into account. Importantly, imaging before and after treatment can dynamically reflect the regression pattern and changes of the tumor from a macroscopic perspective [13, 14]. Several studies have demonstrated that utilizing MRI at multiple time points could enhance predictive power [15, 16]. However, simple concatenation and convolution operations failed to fully capture the differences in MRI scans before and after treatment, as well as extract features associated with treatment response from them.

To address the existing challenges, this paper proposes **M**ulti-time **M**ultimodal **Fusion** (**M2Fusion**) to interact multi-time MRIs and WSIs for treatment response prediction. The contributions of this work are as follows:

- We propose multi-time MRIs contrastive learning loss to extract representations reflecting treatment-induced tumor change by optimizing the distance between pre- and post-treatment imaging features.
- We present a novel module for incorporating multimodal data like MRI and WSI, where orthogonal information is integrated and features from different modalities can be better exploited.

- To the best of our knowledge, this work is the first to use multi-time multimodal data simultaneously for treatment response prediction. The proposed model yields enhanced predictive performance and maintains stable generalization to the external dataset.

2 Related Work

2.1 Multi-time Imaging Prediction

Analyzing radiology images collected from a single time point can only provide the tumor's static information and does not adequately capture the tumor changes induced by NAC during treatment. To allow for a comprehensive analysis, Huang *et al.* utilized radiomic features and deep learning features from pre-NAC and post-NAC MRI to predict pCR [14].To better fuse radiomic features at two time points, a disentangled representation learning was proposed [15]. These two studies are mainly developed based on designed hand-crafted features. Jin *et al.* extracted multi-scale features from UNet based on MRI before and after treatment, leveraging tumor segmentation, and concatenated them to conduct treatment response prediction [13]. Liu *et al.* integrated multi-time multi-scale features through subtraction and global average pooling for pCR prediction [16].

2.2 Multimodal Fusion Prediction

Emerging evidence suggests that incorporating multimodal data can fully enrich the description of tumor heterogeneity across multiple views, thereby boosting the prediction performance [17–21]. Shah *et al.* concatenated radiomic features and pathological features to improve risk stratification [22]. They also proposed a model that takes the designed features of each modality as input to obtain weighted scores and then integrated them to predict immunotherapy response [23]. Additionally, inspired by the concept of cross-modal information interaction in the visual question answering (VQA) system, a hierarchical multimodal co-attention transformer was developed to progressively integrate radiology images and WSIs to obtain aggregation embedding [24]. A similar idea was also applied to integrate preoperative ultrasound images and WSIs [25].

3 Methods

The overview of the proposed network M2Fusion is depicted in Fig. 1, comprising three main modules: the multi-time MRIs contrastive learning module, the WSIs feature extraction module, and the multimodal fusion module. Given the i-th patient in the dataset, the observation is denoted by $(X^i_{pre}, X^i_{post}, X^i_{wsi}, Y^i)$, where X^i_{pre},X^i_{post},X^i_{wsi} and Y^i represent pre-treatment MRI, post-treatment MRI, WSIs before NAC treatment and treatment response classification label. $Y^i = 1$ indicates that the patient has achieved pathological complete response.

WSIs feature extraction module takes X^i_{wsi} as the input and outputs WSI embeddings of different magnifications. The multi-time MRI contrastive learning module takes multi-time MRI embeddings as the input. The multimodal fusion module aggregates WSI embeddings and MRI embeddings to integrated features for final pCR prediction. Now we delve into different modules in the following subsections.

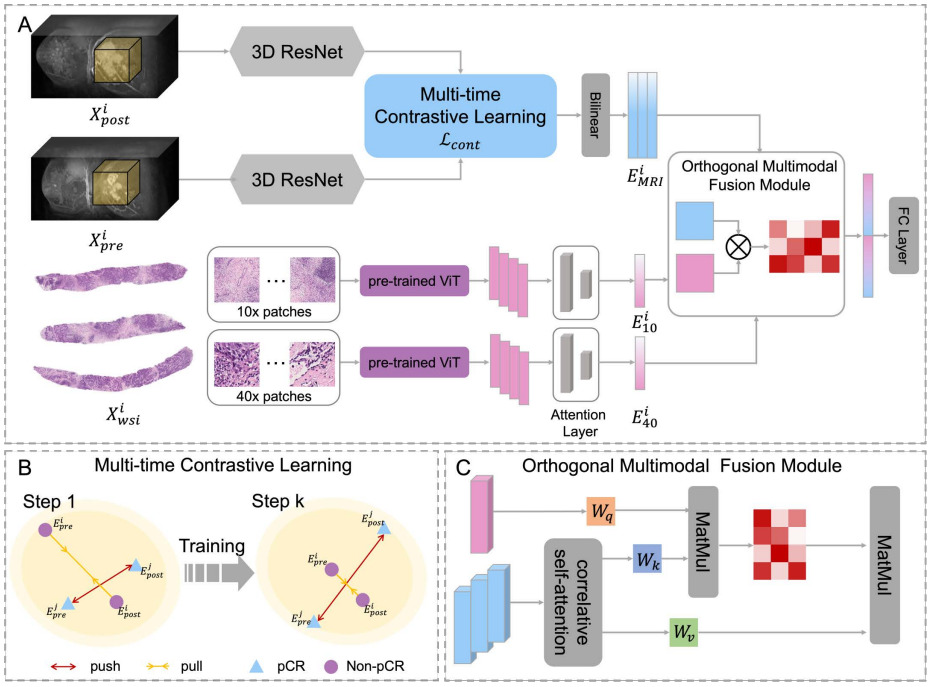

Fig. 1. Overview architecture of our proposed M2Fusion model. **A.** M2Fusion with two key components: the multi-time MRIs contrastive learning loss and the orthogonal multimodal fusion module. **B.**Graphical demonstration of the multi-time MRIs contrastive learning training process. After training, from step 1 to step k,E^i_{pre}, E^i_{post} of pCR patients are pushed away, and E^i_{pre}, E^i_{post} of non-pCR patients are pulled close. **C.** The illustration of orthogonal multimodal fusion module.

3.1 Multi-time MRIs Contrastive Learning

For the MRI scans taken before and after treatment, we observe that after neoadjuvant chemotherapy, there are no invasive residual tumors visible on the MRI scans of patients who achieve pCR. In some pCR cases, there might be cancer in situ on the post-treatment MRI. However, for patients who do not achieve pCR, tumors do not completely regress, and tiny tumors are still present on the

MRI scans after treatment. That's to say, in the case of pCR patients, the complete regression of the tumor results in substantial changes between the MRI scans before and after treatment, leading to a lower similarity between them. Conversely, for non-pCR patients, the presence of residual tumors leads to a relatively higher similarity between the MRI scans before and after treatment. Inspired by contrastive learning, therefore, we hypothesize that the image features of pCR patients before and after treatment should ideally exhibit significant dissimilarity. Conversely, for non-pCR patients, the image features before and after treatment should ideally be as similar as possible.

Based on the observation and the hypothesis, we propose a multi-time MRI contrastive learning loss function $\mathcal{L}_{cont}$, aimed at learning representations by minimizing the distance between pre-treatment MRI and post-treatment MRI for non-pCR patients, while simultaneously maximizing this distance for pCR patients. In detail, 3D ResNet [26] is employed to obtain $E^i_{pre} = f_{pre}(X^i_{pre}) = ResNet(X^i_{pre}), E^i_{post} = f_{post}(X^i_{post}) = ResNet(X^i_{post})$ respectively for a given patient, where f_{pre}, f_{post} share similar network architecture but have different network weights. Then, $\mathcal{L}_{cont}$ is utilized for pushing E^i_{pre}, E^i_{post} of pCR patients away and pulling E^i_{pre}, E^i_{post} of non-pCR patients close. It can be formulated as

$$\mathcal{L}_{cont} = \sum_i (1 - Y^i) MSE(E^i_{pre}, E^i_{post}) + Y^i COS(E^i_{pre}, E^i_{post}) \tag{1}$$

where $MSE(\cdot)$ represents mean-squared loss and $COS(\cdot)$ represents cosine similarity. Finally, we integrate E^i_{pre}, E^i_{post} by bilinear pooling to obtain multi-time MRI representations E^i_{MRI} for multimodal fusion later.

$$E^i_{MRI} = Bilinear(E^i_{pre}, E^i_{post}) = E^i_{pre} A E^i_{post} + b \tag{2}$$

where A is a trainable parameter and b is a bias term.

3.2 WSIs Feature Extraction

To fully leverage WSI-level labels and gigapixel WSIs, we adopt an attention-based multiple-instance learning approach [27]. First, tissue regions are automatically segmented. Then, 512×512 patches are cut from the segmented foreground contour at the magnification of $40\times$ and $10\times$. We use pre-trained vision transformers [28] to extract patch features at different magnifications, denoted as $P^i_{40} = \{p^{i_j}_{40}\} \in \mathbb{R}^{i_J \times d}$ and $P^i_{10} = \{p^{i_k}_{10}\} \in \mathbb{R}^{i_K \times d}$, where i_J and i_K are the total number of patches at $40\times$ and $10\times$ for a patient. Finally, patch features are integrated by learned attention score $a^{i_j}_{40}$ and $a^{i_k}_{10}$ to obtain WSIs embeddings at $40\times$ and $10\times$ for a patient.

$$E^i_{40} = \sum_j a^{i_j}_{40} p^{i_j}_{40} \qquad E^i_{10} = \sum_k a^{i_k}_{10} p^{i_k}_{10} \tag{3}$$

3.3 Orthogonal Multimodal Fusion

Currently, the fusion of MRIs and WSIs primarily employs early fusion and late fusion strategies. Both strategies typically overlook the interactions across modalities. Attention-based fusion modules take this into consideration [24, 25]. Inspired by that, we argue that incorporating orthogonal information will enhance the representation ability and facilitate downstream tasks [9, 29]. Based on that, we propose an orthogonal multimodal fusion module (abbreviated as OMF), which is illustrated in Fig.1.C.

During the attention, each element in QK^T measures the orthogonality between q_i and k_j, with larger values indicating greater similarity. $q_i k_j^T = 0$ means that q_i and k_j are orthogonal, indicating that they are perpendicular in feature space with the lowest linear correlation. A high weight score should be assigned to them, and $1 - \frac{q_i k_j^T}{\|q_i\|\|k_j\|}$ could satisfy this. For the fusion of E_{WSI} and E_{MRI}, the multiplication of a query Q from the WSIs embedding E_{WSI} and a key K from the MRIs embedding E_{MRI} is used to find k_j orthogonal to q_i and assign high weight score. Then, the weight score is utilized for the feature aggregation of E_{MRI}. This operator can be formulated as:

$$\begin{aligned} H_f &= \text{OrthAttention}(E_{WSI}W_q, E_{MRI}W_k, E_{MRI}W_v) \\ &= \left(J - \left|E_{WSI}W_q(E_{MRI}W_k)^T\right|\right) E_{MRI}W_v \\ &= \left(J - \left|QK^T\right|\right) V \\ &= \left(J - \left|\left(\frac{q_1}{\|q_1\|}, \cdots, \frac{q_n}{\|q_n\|}\right)\left(\frac{k_1}{\|k_1\|}, \cdots, \frac{k_m}{\|k_m\|}\right)^T\right|\right) V \end{aligned} \tag{4}$$

Where J is a matrix with all elements equal to 1, W_q, W_k,andW_v are trained weight matrices, Q is the query generated from E_{WSI}, and K and V are the key and value generated from E_{MRI}. For a given patient, we have E^i_{MRI}, E^i_{40} and E^i_{10}. Integrating WSI embeddings and MRI embeddings at different magnifications through the orthogonal multimodal fusion module, we obtain $H^i_{f_{40}}$ and $H^i_{f_{10}}$.

$$\begin{aligned} H^i_{f_{40}} &= \text{OrthAttention}(E^i_{40}W_q, E_{MRI}W_k, E_{MRI}W_v) \\ H^i_{f_{10}} &= \text{OrthAttention}(E^i_{10}W'_q, E_{MRI}W'_k, E_{MRI}W'_v) \end{aligned} \tag{5}$$

It should be noted that before E_{MRI} is input into the orthogonal multimodal fusion module [30], it is passed into the correlative self-attention mechanism. Finally, we concatenate $H^i_{f_{40}}$, $H^i_{f_{10}}$,E^i_{40} and E^i_{10} to obtain H^i_f for classification.

$$H^i_f = Concat(H^i_{f_{40}}, H^i_{f_{10}}, E^i_{40}, E^i_{10}) \tag{6}$$

4 Experiments

4.1 Datasets and Implementation Details

Data Collection In this study, to evaluate our proposed approach, patients with breast cancer from our two collaborating hospitals were collected between

2019 and 2023, denoted as **the in-house cohort A** and **the in-house cohort B**. All patients underwent neoadjuvant chemotherapy according to the guidelines. Each patient includes pre-treatment dynamic contrast-enhanced MRI(DCE-MRI), post-treatment DCE-MRI, pre-treatment WSIs of core biopsies, and pCR labels. The phase with peak tumor enhancement at DCE-MRI was used. pCR is defined as no residual invasive cancer in both the breast and axillary lymph nodes, while cancer in situ was allowed in some cases. **The in-house cohort A** consists of 375 patients, including 134 patients achieving pCR, and 241 non-pCR. The in-house cohort A is randomly split into internal training and validation sets at a 4:1 ratio. **The in-house cohort B** consists of 204 patients from another different hospital, including 74 patients achieving pCR, and 130 non-pCR. The in-house cohort B acts as an external set to further demonstrate the performance of the proposed method.

Implementation and Evaluation We use the cross entropy loss and $\mathcal{L}_{cont}$ to train our model. To avoid over-fitting, we also impose an l_1-norm and l_2-norm to the learnable network weight Θ. Finally, the loss function can be formulated as follows:

$$\mathcal{L} = \mathcal{L}_{cont} + \mathcal{L}_{ce} + \lambda_1 \|\Theta\|_1 + \lambda_2 \|\Theta\|_2 \tag{7}$$

$\mathcal{L}_{cont}$ is only used during training. The input size of MRI is $48 \times 96 \times 96$ and the patch size of WSIs is 512×512. To train M2Fusion, we use SGD optimization with a learning rate of 0.01 and employ a poly learning rate policy. All code was implemented in PyTorch and executed on a NVIDIA RTX 4090 GPU. To enable end-to-end training of our model, a batch size of 1 with 32 steps for gradient accumulation is utilized. The area under the receiver operating characteristic curve (AUC) is employed to evaluate the performance of treatment response prediction.

4.2 Results

Ablation Study: Multi-time MRIs contrastive learning loss ($\mathcal{L}_{cont}$) and OMF are two key components in M2Fusion. In Table 1, we investigate the performance of M2Fusion without $\mathcal{L}_{cont}$ and OMF respectively. Following results are observed: (1) Incorporating orthogonal information from MRIs and WSIs enhances the performance of M2Fusion, (2)$\mathcal{L}_{cont}$ enables M2Fusion to extract better representations of multi-time MRIs, which is beneficial to the classification ability of M2Fusion, (3) both $\mathcal{L}_{cont}$ and OMF contribute to the improved predictive power of M2Fusion.

Comparison to other methods: To evaluate the performance of our model, we compare M2Fusion with current multimodal fusion models, including Concat [22], SFusion [31], HMCAT [24]. Additionally, we compare M2Fusion with models that utilize single modality data, where M_{MRI} is developed based on multi-time MRIs with $\mathcal{L}_{cont}$ and M_{WSI} is developed based on WSIs. All of the

Table 1. Ablation Study on model components with the AUC metric on internal validation set and external test set

Baseline	OMF	$\mathcal{L}_{cont}$	AUC	
			Internal Validation Set	External Test Set
✓			0.6975	0.7129
✓	✓		0.7052	0.7403
✓		✓	0.7299	0.7415
✓	✓	✓	0.7346	0.7992

experiments are conducted on the same training/validation split. As shown in Table 2, M2Fusion achieves an AUC of 0.7346 in the internal validation set, superior to other multimodal fusion methods and single modality methods. Compared with SFusion, which aggregates each modality with learnable assigned weights, M2Fusion stands out by considering modal interaction. M2Fusion outperforms attention-based methods that consider modal interaction, such as HMCAT, demonstrating the validity of the assumption that extracting orthogonal representations can improve model performance. The findings demonstrate that: (1) our method enhances modal interaction, particularly through the extraction of orthogonal representations; (2) we extract more informative representations of tumor changes from multi-time MRI, which is beneficial for predicting treatment response; and (3) leveraging more complementary features from multimodal data can further improve performance.

Table 2. Comparison with the AUC metric on internal validation set and external test set

model	AUC	
	Internal Validation Set	External Test Set
M_{MRI}	0.6968	0.6950
M_{WSI}	0.6605	0.6917
Concat [22]	0.6975	0.7129
Concat w/ $\mathcal{L}_{cont}$	0.7299	0.7415
HMCAT [24]	0.7168	0.7023
SFusion [31]	0.6906	0.7093
M2Fusion	0.7346	0.7992

Generalization on external set: To further validate the effectiveness and generalization of M2Fusion, we conduct experiments on the external test set without any additional training. As shown in Table 2, M2Fusion achieves an AUC of up to 0.7992, which is 8.0% higher than the second-best model. This suggests that our model maintains satisfactory predictive performance when generalizing to datasets from other hospitals, with no significant decline. Results in Table 1 also reveal the validity of two key components in M2Fusion.

5 Conclusion

In this work, we propose M2Fusion model to fuse multi-time multimodal imaging data for treatment response prediction in patients with breast cancer. M2Fusion utilizes multi-time MRIs contrastive learning to extract MRI features reflecting NAC-induced tumor change and applies orthogonal multimodal fusion to incorporate orthogonal information from multimodal features. We collected multi-time multimodal data consisting of 579 patients from different hospitals to demonstrate the performance of our model. Experiments reveal that M2Fusion outperforms state-of-the-art fusion methods as well as single-modality models. Besides, M2Fusion still maintains strong generalization ability and maintains satisfactory predictive performance on the external validation.

Acknowledgments. This work was supported by the National Key R&D Program of China, China (2021YFF1201003), the National Natural Science Foundation of China, China (62333022, 92259301, 81930053, 62027901, 82371947, 82302165), Beijing Natural Science Foundation, China(JQ23034).

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

References

1. Cortazar, P., et al. :Pathological complete response and long-term clinical benefit in breast cancer: the CTNeoBC pooled analysis. The Lancet **384**(9938),164–172 (2014)
2. Spring L.M., et al. :Pathologic complete response after neoadjuvant chemotherapy and impact on breast cancer recurrence and survival: a comprehensive meta-analysis. Clinical cancer research **26**(12),2838–2848(2020)
3. Yee D., et al. :Association of event-free and distant recurrence–free survival with individual-level pathologic complete response in neoadjuvant treatment of stages 2 and 3 breast cancer: three-year follow-up analysis for the I-SPY2 adaptively randomized clinical trial. JAMA oncology **6**(9),1355–1362 (2020).
4. Liu, Z., et al. :Radiomics of multiparametric MRI for pretreatment prediction of pathologic complete response to neoadjuvant chemotherapy in breast cancer: a multicenter study. Clinical Cancer Research **25**(12): 3538-3547(2019)
5. Ogier, du T.J., et al.:Federated learning for predicting histological response to neoadjuvant chemotherapy in triple-negative breast cancer. Nature medicine **29**(1), 135–146 (2023)
6. Huang, Z., et al.: Artificial intelligence reveals features associated with breast cancer neoadjuvant chemotherapy responses from multi-stain histopathologic images. NPJ Precision Oncology **7**(1): 14 (2023)
7. Li, B., et al.:Deep learning with biopsy whole slide images for pretreatment prediction of pathological complete response to neoadjuvant chemotherapy in breast cancer: A multicenter study. The Breast **66**,183–190(2022)
8. Lipkova, J., et al. :Artificial intelligence for multimodal data integration in oncology. Cancer cell **40**(10): 1095–1110(2022)
9. Boehm, K.M., et al. :Harnessing multimodal data integration to advance precision oncology. Nature Reviews Cancer **22**(2): 114–126(2022)
10. Steyaert, S., et al. :Multimodal data fusion for cancer biomarker discovery with deep learning. Nature Machine Intelligence **5**(4): 351–362(2023)
11. Shi, Z., et al. :MRI-based quantification of intratumoral heterogeneity for predicting treatment response to neoadjuvant chemotherapy in breast cancer. Radiology **308**(1): e222830(2023)
12. Du, S., et al. :Distilling Knowledge from Topological Representations for Pathological Complete Response Prediction. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.). International Conference on Medical Image Computing and Computer-Assisted Intervention– MICCAI 2022. LCNS,vol.13432,pp.56–65 Springer, Cham.(2022).https://doi.org/10.1007/978-3-031-16434-7_6
13. Jin, C., et al. :Predicting treatment response from longitudinal images using multitask deep learning. Nature communications **12**(1): 1851(2021)
14. Huang, Y.H., et al. :Longitudinal MRI-based fusion novel model predicts pathological complete response in breast cancer treated with neoadjuvant chemotherapy: a multicenter, retrospective study. EClinicalMedicine **58**(2023)
15. Yue, H., et al: MLDRL: Multi-loss disentangled representation learning for predicting esophageal cancer response to neoadjuvant chemoradiotherapy using longitudinal CT images. Medical image analysis **79**: 102423 (2022)
16. Liu, Y., et al. :Early prediction of treatment response to neoadjuvant chemotherapy based on longitudinal ultrasound images of HER2-positive breast cancer patients by Siamese multi-task network: A multicentre, retrospective cohort study. EClinicalMedicine **52**(2022)

17. Li B., et al.: Multi-omics fusion for prediction of response to neoadjuvant therapy in breast cancer with external validation. Proceedings of the 2021 San Antonio Breast Cancer Symposium-SABCS.AACR,vol. 82(4 Suppl),P2-12-11.(2022) https://doi.org/10.1158/1538-7445.SABCS21-P2-12-11
18. Sammut, S.J., et al. :Multi-omic machine learning predictor of breast cancer therapy response. Nature **601**(7894): 623–629(2022)
19. Feng, L., et al. :Development and validation of a radiopathomics model to predict pathological complete response to neoadjuvant chemoradiotherapy in locally advanced rectal cancer: a multicentre observational study. The Lancet Digital Health **4**(1): e8–e17(2022)
20. Acosta, J.N., et al. :Multimodal biomedical AI. Nature Medicine **28**(9): 1773–1784(2022)
21. Shao, L., et al. :Multiparametric MRI and whole slide image-based pretreatment prediction of pathological response to neoadjuvant chemoradiotherapy in rectal cancer: a multicenter radiopathomic study. Annals of surgical oncology **27**: 4296–4306(2020)
22. Boehm, K.M., et al. :Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer. Nature cancer **3**(6): 723–733 (2022)
23. Vanguri, R.S., et al. :Multimodal integration of radiology, pathology and genomics for prediction of response to PD-(L) 1 blockade in patients with non-small cell lung cancer. Nature cancer **3**(10): 1151–1164 (2022)
24. Li, Z., et al. :Survival prediction via hierarchical multimodal co-attention transformer: A computational histology-radiology solution. IEEE Transactions on Medical Imaging (2023)
25. Huang, Y., et al. :Deep learning radiopathomics based on preoperative US images and biopsy whole slide images can distinguish between luminal and non-luminal tumors in early-stage breast cancers. EBioMedicine **94** (2023)
26. Chen, S., Ma, K., Zheng, Y. :Med3d: Transfer learning for 3d medical image analysis. arXiv preprint arXiv:1904.00625(2019)
27. Lu, M.Y.,et al. :Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering **5**(6): 555–570(2021)
28. Wang, X., et al. :Transformer-based unsupervised contrastive learning for histopathological image classification. Medical image analysis **81**: 102559(2022)
29. Vaswani A.,et al. :Attention is all you need.In: Guyon, I., et al(eds.). Advances in neural information processing systems-NIPS.vol. 30, pp. 5998–6008. Curran Associates, Inc(2017).
30. Wang, F., Mei, J., Yuille, A. :SCLIP: :Rethinking Self-Attention for Dense Vision-Language Inference. arXiv preprint arXiv:2312.01597 (2023)
31. Liu, Z., et al. :SFusion: Self-attention Based N-to-One Multimodal Fusion Block. In: Greenspan, H., et al(eds.).International Conference on Medical Image Computing and Computer-Assisted Intervention-MICCAI 2023. LCNS,vol. 14221, pp. 159–169.Springer, Cham (2023). https://doi.org/10.1007/978-3-031-43895-0_15