Physiology-Guided Cross-View Spatial-Temporal Aligning for Early Prediction of pCR in Breast Cancer from Longitudinal Multimodal MRI

Xin Li[1], Mingrui Zhuang[1,2], Mingke Cui[3], Zhixuan Liao[3], Qiang Zhang[3], and Hongkai Wang[1,2,4(✉)]

[1] School of Biomedical Engineering, Faculty of Medicine, Dalian University of Technology, Dalian 116024, China
wang.hongkai@dlut.edu.cn
[2] Central Hospital of Dalian University of Technology (Dalian Municipal Central Hospital), Dalian 116033, China
[3] Department of Breast Surgery, Cancer Hospital of Dalian University of Technology (Liaoning Cancer Hospital and institute), Shenyang 110242, China
[4] Dalian Key Laboratory of Intelligent Health and Digital Twin, Dalian, China

Abstract. Early prediction of Pathological Complete Response (pCR) in breast cancer patients undergoing Neoadjuvant Chemotherapy (NAC) is critical but challenging as early-stage morphological changes in tumors are subtle. Existing methods relying on long-term longitudinal images fail to extract lesion development trends in limited early-stage data. To address this challenge, we propose CALM (**C**ross-View Spatial-Temporal **A**ligning for **L**ongitudinal **M**ultimodal data), a framework designed to accurately predict pCR at an early treatment stage by effectively fusing subtle lesion features from multimodal longitudinal images, thereby compensating for the absence of later-stage data. To enable effective multimodal information integration and mitigate inter-modal discrepancies in acquisition views and spatial resolutions, we introduce a Physical-distance-weighted Cross-view Inter-modal Attention(PCIA) mechanism within the multimodal transformer block. Furthermore, to exploit complementary physiological information across modalities, we develop a Physiology-Gated Guidance Module(PGM), which leverages the high tumor-signal saliency of the DWI modality to guide cross-modal attention fusion among multimodal images. In addition, we propose a Trajectory Margin Contrastive Loss(TMCL) to explicitly model longitudinal pathological evolution trajectories on the latent feature manifold, encouraging the capture of both the directionality and the potential progression trend of therapeutic response. Experimental results on a collected longitudinal multimodal MRI dataset of 146 patients demonstrate that CALM achieves superior performance using only early-stage data, outperforming state-of-the-art (SOTA) methods, showing the potential of our heterogeneous multimodal fusion approach for early medical prediction tasks. Our code is available at https://github.com/Anchor-Li/CALM.

Keywords: Longitudinal Imaging Analysis · Early Prediction · Neoadjuvant Chemotherapy · Multimodal Learning.

1 Introduction

Neoadjuvant Chemotherapy (NAC) is a standard treatment for locally advanced breast cancer [1], with Pathological Complete Response (pCR) serving as a robust surrogate endpoint for Disease-Free Survival and Overall Survival [14]. Accurate early prediction of pCR is pivotal for adaptive therapy. However, conventional clinical assessment protocols typically rely on post-surgical histopathology after prolonged treatment, comprising 8 cycles of three-week therapeutic courses. The Non-pCR patients risk missing the optimal therapeutic window for switching to alternative therapies. Consequently, identifying potential responders and non-responders at early stage when tumor morphology has not undergone significant alteration is critical but remains a significant challenge [8].

Longitudinal medical imaging has demonstrated significant potential in capturing multidimensional feature evolution across various clinical tasks [13]. Research on NAC response prediction has transitioned from static single-timepoint analysis to dynamic longitudinal sequence modeling [10, 15, 16]. An example of tumor evolution during NAC is illustrated in Fig.1. The treatment phases are defined as follows: pre-NAC (T_0), early-NAC (T_1, after 2 cycles), mid-NAC (T_2, after 4 cycles), late-NAC (T_3, after 6 cycles), and post-NAC (T_4). At the end of the second cycle, tumors often haven't yet undergone significant volumetric reduction, imposing stringent requirements on algorithmic sensitivity. Existing deep learning methods frequently struggle in this context as they predominantly rely on morphological changes. For instance, while the DTFDF framework proposed by Hao et al. [6] excels in long-sequence prediction, ablation studies reveal a significant performance drop when using only early-stage data (T_0 and T_1).

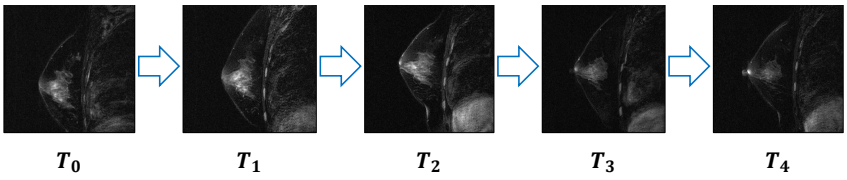

Fig. 1. The dynamic changes of tumor during NAC of a pCR patient, of whom significant volumetric reduction of tumor is only visually apparent at late treatment phases.

Integrating multimodal data is an imperative strategy to capture microscopic variations [3, 9, 11, 19]. Zhang et al. [18] proposed a framework fusing Dynamic Contrast Enhanced (DCE) MRI with Whole Slide Images (WSI), enhancing prediction via multi-temporal contrastive learning. Ma et al. [10] introduced a molecular subtype-aware mechanism to embed demographic data into MRI feature extraction. Similarly, the MAP framework proposed by Zhang et al. [17] utilized HoVer-Trans [12] to integrate multimodal imaging data. Although these studies validate the potential of multimodal data in NAC response prediction, they primarily employ conventional cross-attention mechanisms, relying on sig-

nificant tumor changes in late treatment phases but are not sensitive to early stage features.

To fully leverage the rich lesion features derived from multimodal anisotropic imaging data and achieve precise NAC response prediction at the early treatment stage, we propose a novel Cross-View Spatial-Temporal Aligning for Longitudinal Multimodal data (CALM) framework, which focus on mining early-stage data to capture faint response signals rather than relying on additional time points to boost performance. The main contributions of this study include: (1) Proposing a novel Physical-distance-weighted Cross-view Inter-modal Attention(PCIA), which establishes correlations between axial and sagittal features without resampling that blurs the texture of tumors to achieve feature complementarity; (2) Designing a lightweight Physiology-gated Guidance Module (PGM), which generates soft attention gates to focus the network on metabolically active regions and suppress Background Parenchymal Enhancement (BPE) noise without explicit segmentation; and (3) Introducing a strategy of distance and direction constraints on feature variations in the latent space called Trajectory Margin Contrastive Loss (TMCL) to enhance the sensitivity of model to dynamic pathological evolutions.

2 Method

This study aims to accurately predict pCR status early in the treatment course utilizing multimodal MRI images from T_0 and T_1. Multi-modality image inputs include DCE-MRI, DWI (Diffusion Weighted Imaging), and T2WI (T2-weighted Imaging) sequences. Notably, DCE-MRIs and DWIs are axial scans with high in-plane resolution but larger slice thickness, whereas T2WIs are sagittal scans offering higher Z-axis resolution. The overall architecture of CALM is illustrated in Fig.2(a), comprising three key modules: Cross-view Spatial Enhancement Module for heterogeneous multimodal fusion, Feature Extraction Positioning Guidance Module, and Temporal Fusion Module for constraining treatment response trajectories in the latent space. The model employs axial DCE-MRIs as the primary backbone for hemodynamic feature extraction, supplemented by pseudo-masks generated from axial DWIs for lesion localization guidance, and leverages sagittal T2WIs to complete high-resolution morphological details.

Physical-distance-weighted Cross-view Inter-modal Attention(PCIA). Although axial DCE-MRI provides rich hemodynamic information, it is limited by imaging protocols that typically mandate thicker slices along the Z-axis, leading to partial volume effects and edge blurring. In contrast, sagittal T2WI possesses higher spatial resolution in the Z-axis dimension. As illustrated in Fig.2(b), our proposed PCIA establishes correlations between heterogeneous image data. Let $F_{ax} \in \mathbb{R}^{C\times H_{ax}\times W_{ax}\times D_{ax}}$ and $F_{sag} \in \mathbb{R}^{C\times H_{sag}\times W_{sag}\times D_{sag}}$ represent the feature maps from axial and sagittal planes. Instead of performing volumetric resampling, we treat the sagittal features as a continuous function $F_{sag} : \mathbb{R}^3 \to \mathbb{R}^C$. For each feature point i in F_{ax},we calculate its center position

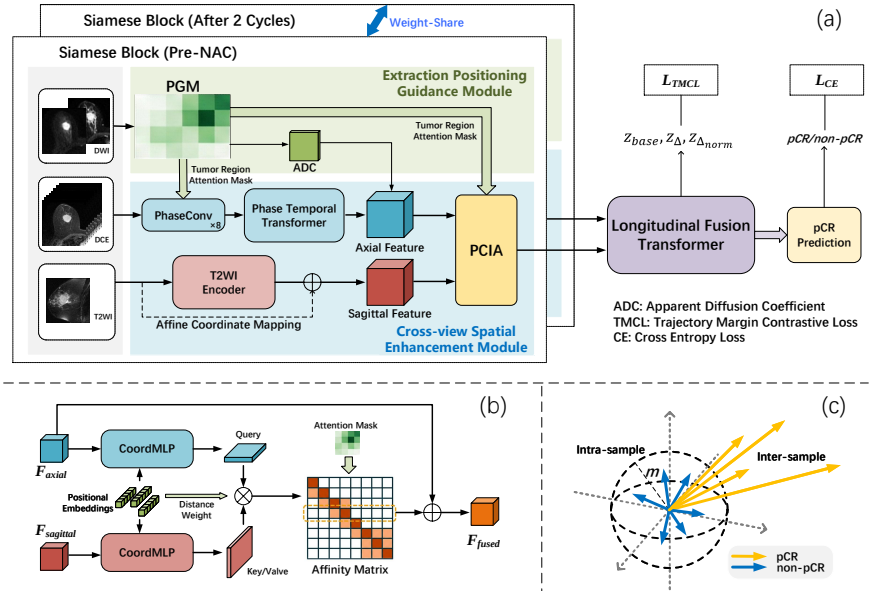

Fig. 2. Illustration of our proposed model CALM. (a)Architecture of our CALM model. (b)Physical-distance-weighted Cross-view Inter-modal Attention(PCIA) mechanism. (c)A schematic diagram of Trajectory Margin Contrastive Loss(TMCL).

$P_{ax}(i) = (x_i, y_i, z_i)$ in the physical imaging space using its affine matrix M_{ax}. Similarly, we calculate the position $P_{sag}(j)$ for each point j in F_{sag}.

To reduce the computational complexity of 3D global attention and eliminate interference from irrelevant backgrounds, we adopt a distance-based sparse search strategy with a threshold δ. For a query point $i \in F_{ax}$, we construct its neighborhood set in F_{sag} within a bounded physical radius as $\mathcal{N}(i) = \{j \in F_{sag} \mid \|P_{ax}(i) - P_{sag}(j)\|_2 < \delta\}$, where δ was set to 5 mm in this study. The attention weights α_{ij} are modulated by the relative physical offset:

$$\alpha_{ij} = \text{softmax}_{j \in \mathcal{N}(i)} \left(\frac{Q_i K_j^T + \psi(P_{ax}(i) - P_{sag}(j))}{\sqrt{d}} \right), \tag{1}$$

where $\psi(\cdot)$ is a Fourier Feature Mapping that encodes high-frequency geometric relationships. The final fused feature is obtained by aggregating spatially relevant information from the sagittal view:

$$F_{fused}(i) = F_{ax}(i) + \sum_{j \in \mathcal{N}(i)} \alpha_{ij} V_j. \tag{2}$$

Through PCIA, the model performs an implicit neural representation-style feature aggregation. It allows axial feature stream to aggregate high-frequency structural information directly from sagittal stream whenever Z-axis details are

required without explicit voxel alignment, thereby preserving the anisotropic complementary advantages of the multimodal data.

Physiology-gated Guidance Module (PGM). The alterations in the tumor microenvironment induced by early-stage NAC are faint and susceptible to interference from Background Parenchymal Enhancement (BPE), a common physiological noise in DCE-MRIs dependent on hormonal status [4]. Conversely, high-b-value DWIs reflect cellular density and is immune to BPE. Tumor regions in DWIs typically appear as distinct hyperintensities due to restricted water diffusion caused by high cellular density. To incorporate the positional prior without compromising the textures of DCE-MRI, we constructed a lightweight PGM, which generates a single-channel soft spatial attention map $M_{dwi} = \sigma(F_{dwi})$ to indicate high-probability tumor regions. We align M_{dwi} to the DCE-MRI feature space to obtain M'_{dwi} using affine matrices to compute the alignment transformation grid. Subsequently, a residual attention mechanism $F'_{dce} = F_{dce} \odot (1 + \sigma(M'_{dwi}))$ is employed to enhance the DCE-MRI features. We also treated the Apparent Diffusion Coefficient(ADC), computed from two distinct b-value images, as an individual channel within the cross-sectional features for subsequent fusion. These allow the network to focus on hyper-cellular regions while suppressing BPE without requiring explicit voxel-level tumor annotations.

Additionally, we utilize M'_{dwi} as a Region of Interest (ROI) selector for PCIA to select only the Top-K voxels with the highest response values as Queries for cross-view attention. This strategy drastically reduces computational overhead and further mitigates interference from normal tissues.

Trajectory Margin Contrastive Loss (TMCL). To address the issue of minimal feature differences between T_0 and T_1 in early prediction, we explicitly model the treatment trajectory within the latent space. Let the features extracted at T_0 and T_1 be v_{t0} and v_{t1} respectively. To mitigate overfitting caused by small sample sizes and force the network to learn the most discriminative dynamic features, we propose the TMCL (L_{traj}) comprising intra-sample and inter-sample components.

The Intra-sample Loss aims to constrain the magnitude of feature displacement based on the clinical label. For pCR patients ($y = 1$), the tumor microenvironment undergoes substantial changes compared to non-pCR patients ($y = 0$), whose tumor status remains relatively stable. We encourage a larger cosine distance between v_{t0} and v_{t1} in the latent space. The loss function is defined as:

$$\mathcal{L}_{intra} = (1 - y) \cdot d(v_{t0}, v_{t1})^2 + y \cdot max(0, m - d(v_{t0}, v_{t1}))^2, \tag{3}$$

where $d(\cdot)$ denotes cosine distance and m is a preset margin threshold set to 0.6. This explicitly amplifies the distinctiveness of subtle early changes.

The Inter-sample Loss further constrains the direction and variation of trajectory features. We construct a composite trajectory vector z_{traj} incorporating

the baseline state z_{base}, change quantity z_{Δ}, and normalized change $z_{\Delta norm}$:

$$z_{base} = \frac{v_{t0} + v_{t1}}{2}, z_{\Delta} = v_{t1} - v_{t0}, z_{\Delta_{norm}} = \frac{v_{t1} - v_{t0}}{|v_{t0}| + |v_{t1}| + \epsilon}, \tag{4}$$

$$z_{traj} = Proj(concat(z_{base}, z_{\Delta}, z_{\Delta_{norm}})). \tag{5}$$

Here, z_{base} reflects the centroid of the tumor, preserving the anatomical background and intrinsic texture features. z_{Δ} captures the absolute changes in tumor, representing the magnitude and direction of feature vector changes. $z_{\Delta norm}$ eliminates the influence of feature magnitude via normalization, compelling the model to focus on the proportion of change rather than absolute size. Using a supervised contrastive formulation [7], we cluster the trajectory representations of patients with the same lable while separating those with different outcomes to form a clear decision boundary for treatment response in the latent space. For a sample i in batch B, we define the inter-sample loss function as:

$$\mathcal{L}_{inter} = \sum_{i \in B} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\text{sim}(z_i, z_p)/\tau)}{\sum_{a \in A(i)} \exp(\text{sim}(z_i, z_a)/\tau)}, \tag{6}$$

where $P(i)$ is the set of samples with the same label as i, $A(i)$ refers to all samples in the batch, $\text{sim}(\cdot)$ is cosine similarity, and τ is the temperature coefficient. The overall loss of the proposed CALM is calculated as:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda \mathcal{L}_{traj} = \mathcal{L}_{CE} + \lambda_1 \mathcal{L}_{intra} + \lambda_2 \mathcal{L}_{inter}, \tag{7}$$

where λ_1 is set to 0.1 and λ_2 is set to 0.2. Through TMCL, the model is compelled to shift focus from static appearances to evolutionary patterns, significantly enhancing sensitivity to subtle early treatment-induced changes.

3 Experiments and Results

3.1 Dataset and Implementation Details

In this study, we collected a longitudinal multimodal breast MRI dataset from 146 patients at our affiliated hospital, including DCE-MRI, DWI and T2WI covering two treatment phases T_0 and T_1, for method evaluation due to the scarcity of public datasets containing longitudinal data with the three modalities. The 3D MRI scans of three modalities aligned in physical imaging space for each patient at each phase. All images were cropped to the breast region using morphological erosion algorithm [5]. The details of the three modality images are as follows: (1) DCE-MRI: axial scans with 8 dynamic phases ($32 \times 200 \times 200$); (2) DWI: axial scans with $b = 0$ and 800 s/mm^2 ($16 \times 106 \times 106$); (3) T2WI: sagittal scans that provide high-resolution information along the Z-axis ($10 \times 326 \times 326$). Our CALM model was trained and validated on two NVIDIA RTX 4090 GPUs with a batch size of 4 for 100 epochs, employing a five-fold cross-validation strategy. The initial learning rate is set to 0.004. Area Under the Curve (AUC), Accuracy (ACC), Sensitivity (SEN), and Specificity (SPE) were used as evaluation metrics.

3.2 Comparisons with Existing Methods

To validate the effectiveness of our proposed method, we compared the CALM model with three SOTA methods tailored for longitudinal or multimodal medical image analysis, including standard vision transformer ViT [2], longitudinal sequence modeling network LOMIA-T [15] and multimodal fusion framework MAP [17]. For a fair comparison, all methods utilized MRI data from three modalities at two time points (T_0 and T_1) for each patient. The results are presented in Table 1. In this challenging early prediction task, our CALM model demonstrated superior performance. Specifically, CALM achieved the highest AUC of 0.804 and ACC of 0.774, outperforming all comparison Transformer-based methods. The performance gain of CALM suggests that an architecture specifically designed for the geometric structure of medical imaging is more effective even when compared to the specialized multimodal method MAP, which employs fusion strategies but primarily relies on conventional cross-attention mechanisms that neglect the intrinsic spatial heterogeneity between different views. This is primarily attributed to our model's ability to not only integrate multimodal information but also preserve high-frequency details through cross-view alignment and capture subtle early pathological changes via explicit trajectory modeling. Additionally, CALM demonstrated exceptional clinical robustness by achieving a balanced performance between SEN (0.791) and SPE (0.759), while competing methods exhibited severe class imbalance. This indicates that existing methods struggle to identify responders amidst background noise. The superior Sensitivity of CALM validates its clinical utility in correctly identifying potential pCR patients at an early stage, thereby minimizing the risk of missing optimal therapeutic windows.

Table 1. Comparison of our model with existing methods.

Methods	AUC	ACC	SEN	SPE
ViT [2]	$0.739_{\pm 0.073}$	$0.717_{\pm 0.081}$	$0.577_{\pm 0.237}$	$0.833_{\pm 0.259}$
LOMIA-T [15]	$0.749_{\pm 0.058}$	$0.719_{\pm 0.063}$	$0.689_{\pm 0.154}$	$0.737_{\pm 0.103}$
MAP [17]	$0.771_{\pm 0.019}$	$0.737_{\pm 0.034}$	$0.576_{\pm 0.073}$	$\mathbf{0.875_{\pm 0.068}}$
CALM(Ours)	$\mathbf{0.804_{\pm 0.034}}$	$\mathbf{0.774_{\pm 0.027}}$	$\mathbf{0.791_{\pm 0.038}}$	$0.759_{\pm 0.073}$

3.3 Ablation Study

To strictly dissect the contribution of each core component within the CALM framework, we conducted extensive ablation studies under identical training hyperparameters, as detailed in Table 2. The removal of the PCIA module resulted in the most precipitous decline in overall performance, with the AUC plummeting by approximately 6.3% from 0.804 to 0.741. This sharp drop confirms that relying solely on axial DCE-MRI features is insufficient due to their thick-slice

nature. Without PCIA, the model loses the ability to supplement axial features with high-resolution Z-axis information from the sagittal T2WI stream, proving that the geometry-aware alignment is essential for preserving anisotropic data details without introducing resampling artifacts. Furthermore, eliminating the PGM led to a notable decrease in SEN, despite maintaining a relatively stable Specificity. This specific pattern of error indicates that without the physiological guidance from DWI, the model struggles to differentiate active tumor tissue from BPE noise in DCE-MRI. The network is easily distracted by physiological noise without it, resulting in a higher rate of missed diagnoses for subtle lesions. Finally, the exclusion of the TMCL caused the overall AUC to fall to 0.784. While this drop is smaller than that of PCIA, it is statistically significant for early prediction tasks where feature differences between T_0 and T_1 are minimal. This validates that standard classification losses are insufficient for capturing the magnitude of therapeutic response. TMCL forces the model to explicitly learn the vector evolution of the tumor in the latent space, thereby enhancing sensitivity to the minute dynamic changes that characterize early response. The superior performance of CALM is thus not due to a single module but the synergistic integration of geometric alignment, physiological guidance, and dynamic trajectory constraints.

Table 2. Ablation Studies on different strategies.

Components			Evaluation			
PCIA	PGM	TMCL	AUC	ACC	SEN	SPE
-	-	-	$0.701_{\pm 0.043}$	$0.683_{\pm 0.026}$	$0.612_{\pm 0.213}$	$0.745_{\pm 0.162}$
-	-	√	$0.724_{\pm 0.055}$	$0.714_{\pm 0.056}$	$0.609_{\pm 0.134}$	$\mathbf{0.806_{\pm 0.063}}$
-	√	√	$0.741_{\pm 0.029}$	$0.733_{\pm 0.041}$	$0.659_{\pm 0.109}$	$0.794_{\pm 0.050}$
√	-	√	$0.776_{\pm 0.042}$	$0.753_{\pm 0.026}$	$0.765_{\pm 0.069}$	$0.743_{\pm 0.058}$
√	√	-	$0.784_{\pm 0.028}$	$0.767_{\pm 0.038}$	$0.765_{\pm 0.115}$	$0.747_{\pm 0.116}$
√	√	√	$\mathbf{0.804_{\pm 0.034}}$	$\mathbf{0.774_{\pm 0.027}}$	$\mathbf{0.791_{\pm 0.038}}$	$0.759_{\pm 0.073}$

4 Conclusion

To address the challenge of subtle tumor morphological and physiological changes in the early prediction of breast cancer NAC response, we proposes a novel CALM framework in this study. By innovatively introducing Physical-distance-weighted Cross-view Inter-modal Attention, we successfully achieve feature fusion of anisotropy multimodal data without resampling, maximizing the complementary advantages of anatomical information. Leveraging DWI guidance, we realize automatic lesion focusing without manual annotation, effectively eliminating the interference of background parenchymal enhancement. Through Dynamics Modeling, we explicitly model the patient's treatment evolution trajectory in the latent space, significantly enhancing the model's sensitivity to minute

dynamic changes. Experimental results demonstrated the immense potential of multimodal cross-view deep learning in clinical precision medicine. In the future, we will focus on validating the generalization capability of this model on larger-scale multi-center cohorts to promote its application in clinical precision diagnosis and treatment.

Acknowledgments. This work was supported in part by the Joint Funds of the National Natural Science Fundation of China (No. U24A20753), the general program of National Natural Science Fund of China (No. 81971693), the China Postdoctoral Science Foundation (No. 2025M772947), the funding of Liaoning Key Lab of IC & BME System and Dalian Engineering Research Center for Artificial Intelligence in Medical Imaging and Liaoning technology innovation center of hyperpolarized MRI.

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

References

1. Asselain, B., Barlow, W., Bartlett, J., Bergh, J., Bergsten-Nordström, E., Bliss, J., Boccardo, F., Boddington, C., Bogaerts, J., Bonadonna, G., et al.: Long-term outcomes for neoadjuvant versus adjuvant chemotherapy in early breast cancer: meta-analysis of individual patient data from ten randomised trials. The Lancet Oncology **19**(1), 27–39 (2018)
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
3. Duanmu, H., Huang, P.B., Brahmavar, S., Lin, S., Ren, T., Kong, J., Wang, F., Duong, T.Q.: Prediction of pathological complete response to neoadjuvant chemotherapy in breast cancer using deep learning with integrative imaging, molecular and demographic data. In: International conference on medical image computing and computer-assisted intervention. pp. 242–252. Springer (2020)
4. Eskreis-Winkler, S., Sutton, E.J., D'Alessio, D., Gallagher, K., Saphier, N., Stember, J., Martinez, D.F., Morris, E.A., Pinker, K.: Breast mri background parenchymal enhancement categorization using deep learning: outperforming the radiologist. Journal of Magnetic Resonance Imaging **56**(4), 1068–1076 (2022)
5. Ghosh, S., Das, S.: Multi-scale morphology-aided deep medical image segmentation. Engineering Applications of Artificial Intelligence **137**, 109047 (2024)
6. Hao, X., Xu, H., Zhang, Q., Xu, Q., Wang, X., Polonen, I., Cong, F.: Predicting radiation therapy response based on dynamic temporal feature difference fusion from longitudinal mri. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 440–449. Springer (2025)
7. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in neural information processing systems **33**, 18661–18673 (2020)

8. Krasniqi, E., Filomeno, L., Arcuri, T., Ferretti, G., Gasparro, S., Fulvi, A., Roselli, A., D'Onofrio, L., Pizzuti, L., Barba, M., et al.: Multimodal deep learning for predicting neoadjuvant treatment outcomes in breast cancer: a systematic review. Biology Direct **20**(1), 72 (2025)
9. Li, Q., Wang, S., Safari, M., Hu, X., Qiu, R., Roper, J., Lin, M., Lin, J., Torres, M.A., Yang, X.: Foundation model based multimodal transformer framework for survival analysis in her2 stratified breast cancer. Physics in Medicine & Biology **70**(20), 205009 (2025)
10. Ma, D., Cao, J., Cheng, H., Zhou, D., Liu, J., Zhang, X., Wu, K., Gu, C., Guan, X.: Longitudinal mri-clinical multimodal fusion for pcr prediction in breast cancer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 323–332. Springer (2025)
11. Mao, N., Dai, Y., Zhou, H., Lin, F., Zheng, T., Li, Z., Yang, P., Zhao, F., Li, Q., Wang, W., et al.: A multimodal and fully automated system for prediction of pathological complete response to neoadjuvant chemotherapy in breast cancer. Science advances **11**(18), eadr1576 (2025)
12. Mo, Y., Han, C., Liu, Y., Liu, M., Shi, Z., Lin, J., Zhao, B., Huang, C., Qiu, B., Cui, Y., et al.: Hover-trans: Anatomy-aware hover-transformer for roi-free breast cancer diagnosis in ultrasound images. IEEE Transactions on Medical Imaging **42**(6), 1696–1706 (2023)
13. Shen, C., Menten, M.J., Bogunović, H., Schmidt-Erfurth, U., Scholl, H.P., Sivaprasad, S., Lotery, A., Rueckert, D., Hager, P., Holland, R.: Spatiotemporal representation learning for short and long medical image time series. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 656–666. Springer (2024)
14. Spring, L.M., Fell, G., Arfe, A., Sharma, C., Greenup, R., Reynolds, K.L., Smith, B.L., Alexander, B., Moy, B., Isakoff, S.J., 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)
15. Sun, Y., Li, K., Chen, D., Hu, Y., Zhang, S.: Lomia-t: a transformer-based longitudinal medical image analysis framework for predicting treatment response of esophageal cancer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 426–436. Springer (2024)
16. Taylor, C.R., Monga, N., Johnson, C., Hawley, J.R., Patel, M.: Artificial intelligence applications in breast imaging: current status and future directions. Diagnostics **13**(12), 2041 (2023)
17. Zhang, J., Li, Y., Li, Z., Cui, Z., Li, P., Li, J., Li, Z., Xie, Y., Chan, K.W., Zhang, Q., et al.: Deep-learning-based her2 status assessment from multimodal breast cancer data predicts neoadjuvant therapy response. Nature Biomedical Engineering pp. 1–12 (2025)
18. Zhang, S., Du, S., Sun, C., Li, B., Shao, L., Zhang, L., Wang, K., Liu, Z., Tian, J.: M2fusion: multi-time multimodal fusion for prediction of pathological complete response in breast cancer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 458–468. Springer (2024)
19. Zhou, Y., Jin, F., Suo, G., Yang, J.: Resvit-gannet: a deep learning framework for classifying breast cancer histopathology images using multimodal attention and gan-based augmentation. BMC Medical Imaging **25**(1), 401 (2025)