HARM^3-Fusion: Hierarchical Attentional Representation Learning of Multi-Modal, Multi-Temporal, and Multi-Sequence Fusion for Pathological Complete Response Prediction of Head and Neck Squamous Cell Carcinoma

Jianye Wang[1*], Xinyue Liu[1*], Zhiying Gong[2], Lingjie Yang[3], Hanwen Zhang[2,1], Yu Long[2,1], Yimeng Fan[2,1], Yuncheng Jiang[4], Xiaohui Duan[3]✉, Weibing Zhao[1]✉

[1] Guangdong Laboratory of Machine Perception and Intelligent Computing, Shenzhen MSU-BIT University, China
[2] Beijing Key Laboratory of Intelligent Information Technology, School of Computer Science Technology, Beijing Institute of Technology, China
[3] Department of Radiology, Sun Yat-Sen Memorial Hospital, Sun Yat-Sen University
[4] Shenzhen Future Network of Intelligence Institute, School of Science and Engineering, The Chinese University of Hong Kong (Shenzhen), China
weibingzhao@smbu.edu.cn, duanxh5@mail.sysu.edu.cn

Abstract. The precise prediction of Pathological Complete Response (pCR) following Neoadjuvant Chemo-ImmunoTherapy (NCIT) in Head and Neck Squamous Cell Carcinoma (HNSCC) is crucial for optimizing therapeutic strategies and prognostic evaluation. Current methods exhibit limitations in simultaneously modeling multi-temporal treatment dynamics, multi-sequence magnetic resonance imaging (MRI) correlations, and multi-modal feature interactions. To address this challenge, we present a novel multi-modal representation and fusion framework, **HARM^3-Fusion**, which innovatively processes multi-temporal, multi-sequence MRI data and hierarchically fuses it with whole slide image (WSI) to enhance the accuracy of pCR prediction. Specifically, our method comprises three key modules: a multi-temporal MRI fusion module based on Loss-enhanced Dual-stream Convolutional Variational Auto-Encoder (LD-VAE), designed to decouple features from pre-treatment and post-treatment MRI scans; a multi-sequence MRI fusion module based on self-attention for integrating MRI features from T1 and T2 weighted sequences; and a multi-modal MRI-WSI fusion module based on cross-attention to fuse complementary information between MRI and WSI.To evaluate the efficacy of HARM^3-Fusion, we establish **HNSCC-pCR**, the first multi-modal dataset for HNSCC. HNSCC-pCR dataset comprises 407 patients, with each case including pre-treatment and post-treatment T1-weighted and T2-weighted MRI scans, WSI of pre- biopsy specimens,

* Equal contribution. ✉ Corresponding author.
Code available at: https://github.com/Jianye-Wang-WJY/HARM3-Fusion

and pathologically confirmed surgical pCR. Based on this dataset, experimental results demonstrate that HARM^3-Fusion achieves superior performance for pCR prediction compared to other single-modal and multi-modal approaches.

Keywords: Multi-modal Learning · Pathological Complete Response Prediction · Head and Neck Squamous Cell Carcinoma

1 Introduction

HNSCC, the sixth most prevalent malignant tumor globally [1], presents significant clinical challenges in therapeutic decision making. While landmark clinical trials (e.g., KEYNOTE-048 [2]) have demonstrated that NCIT can significantly improve patient survival, clinical data reveal that only ∼60.3% of patients achieve pCR due to substantial inter-individual heterogeneity in NCIT sensitivity [3]. Current pCR assessment relies on WSI of invasive biopsy specimens, which is not only intrinsically traumatic but also prone to sampling bias. Critically, pCR status directly guides subsequent surgical planning and organ-preservation strategies [4]. Thus, developing a non-invasive, reliable, and clinically deployable pCR prediction method is imperative for optimizing personalized HNSCC treatment paradigms.

Currently, many studies are dedicated to accurately predicting pCR. Studies indicate that patients achieving pCR exhibit distinct tumor lesions on pre-imaging, with near-complete elimination of invasive lesions on post-treatment scans. In contrast, non-pCR patients exhibit persistent residual tumors, leading to a high structural similarity between pre-/post- treatment MRI scans. Therefore, dynamic information from pre-/post- treatment images was captured using multi-task learning to predict treatment response [5]. However, there are still theoretical limitations in distinguishing treatment response-related changes from the inherent heterogeneity of the tumor [6]. Another approach predicted pCR by learning stromal histology features from pretreatment WSI using a deep learning model [7]. However, the WSI-dependent microscopic characterization lacks the support of macroscopic dynamic information, making it difficult to independently support the multi-dimensional prediction task of pCR. To address these issues, a multi-modal Transformer architecture was designed for breast cancer [8], and an orthogonal fusion strategy was proposed [9], utilizing contrastive learning to optimize MRI features before and after treatment. However, most methods fail to systematically investigated the synergy between intra-modal optimization and inter-modal complementarity, particularly lacking a hierarchical fusion strategy for MRI temporal dynamic features and WSI spatial heterogeneity features. To address current challenges in predicting NCIT response for HNSCC, we propose HARM^3-Fusion, which is the first hierarchical attention-based multimodal fusion framework for NCIT response prediction in HNSCC. Our approach innovatively integrates multi-temporal, multi-sequence MRI and WSI data through a novel temporal feature decoupling and multi-modal fusion strategy, aiming at overcoming key limitations of existing methods in modeling tumor dynamic evolution,

capturing heterogeneity, and optimizing cross-modal interactions. Specifically, for multi-temporal modeling, we designed a LD-VAE to decouple treatment-induced tumor dynamic evolution features from tumor heterogeneity features through joint optimization of cross-reconstruction and contrastive losses. Besides, we designed a self-attention fusion module to correlate multi-sequence MRI features (T1 and T2), and a multi-modal cross-attention fusion module to model pathological-radiological interactions between WSI and MRI. These components collectively establish a hierarchical attention fusion mechanism. To support this research, we have established the largest multimodal dataset for pCR prediction in HNSCC, named HNSCC-pCR, comprising 407 fully-annotated patient cases. Each case includes multi-sequence MRI (T1/T2) acquired pre-/post- treatment, WSI prior to NCIT, and definitive pCR status labels. This comprehensive dataset enables robust investigation of treatment response dynamics for HNSCC.

The main contributions are as follows:

- We designed a LD-VAE to effectively disentangle latent variations in MRI before and after treatment.
- We propose a hierarchical multi-modal fusion mechanism that leverages both self-attention and cross-attention strategies to achieve complementary multi-sequence information integration and cross-modal feature interaction.
- We establish the largest multi-modal dataset for HNSCC pCR prediction to date, comprising 407 complete patient cases with paired T1/T2, pre-/post-treatment MRI and pre-treatment WSI, which will serve as a benchmark resource for advancing research for HNSCC.

2 Methods

As illustrated in Fig. 1, our framework comprises three core components: (1) a multi-temporal MRI fusion module, (2) a multi-sequence MRI fusion module, and (3) a multi-modal MRI-WSI fusion module. The multi-temporal MRI fusion module processes paired pre-/post- treatment MRI scans to generate temporally-disentangled feature representations. These features are then sent into the multi-sequence MRI fusion module, which employs self-attention mechanisms to integrate complementary information from T1 and T2 sequences, producing a unified MRI feature embedding. Finally, the multi-modal MRI-WSI fusion module establishes diagnostic-relevant interactions between the MRI embedding and WSI features through cross-attention, ultimately generating the pCR prediction.

2.1 Multi-Temporal MRI Fusion

Comparing two temporal images before and after treatment, tumor lesions in patients with pCR nearly disappear after treatment. Therefore, we expect significant differences in the features of the image pairs from pCR patients before and after treatment, while the opposite is true for non-pCR patients. Therefore, we designed a LD-VAE for decoupling the features of multi-temporal MRIs.

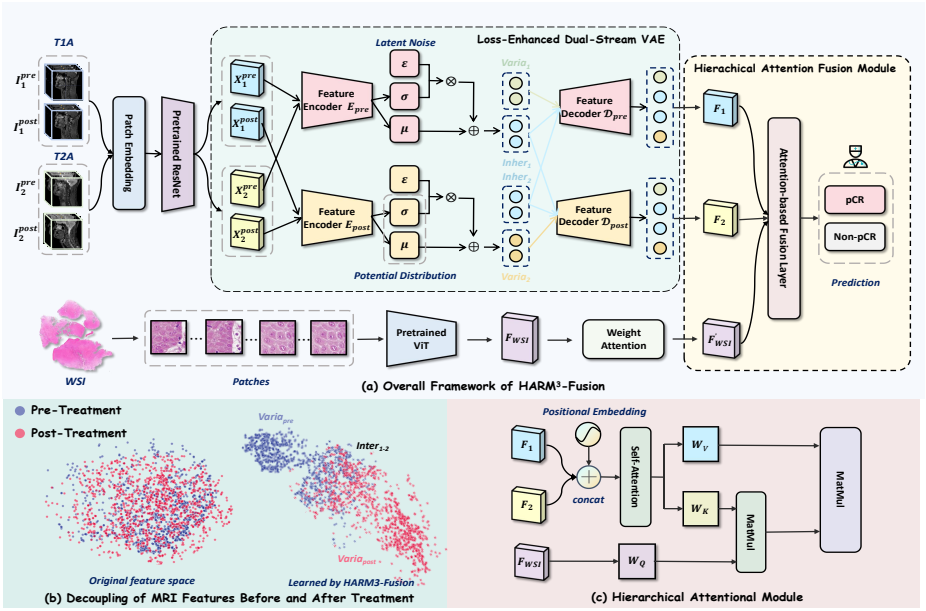

Fig. 1. Overview architecture of our proposed HARM^3-Fusion model: (a) HARM^3-Fusion contains two key components: temporal decoupling learning and hierarchical attention fusion module. (b) Illustration of the temporal feature decoupling. By designing a LD-VAE architecture, the inherent component Inher(Z_i) and the variational component Varia(Z_i) are extracted, making the decoupling effect more apparent. (c) Schematic diagram of the hierarchical attention fusion module.

LD-VAE contains two VAE [10] branches that process pre-/post- treatment images respectively. In each VAE branch, the input features $\{X_{\text{pre}}, X_{\text{post}}\}$ which are extracted from the original MRIs $\{I_{\text{pre}}, I_{\text{post}}\}$ are processed through the encoder $\{E_{\text{pre}}, E_{\text{post}}\}$ to obtain the mean and variance of the input data, followed by the addition of noise sampled from a standard normal distribution to obtain the latent variable $\{Z_{\text{pre}}, Z_{\text{post}}\}$,

$$Z_{\text{pre}} = E_{\text{pre}}(X_{\text{pre}}),\ Z_{\text{post}} = E_{\text{post}}(X_{\text{post}}). \tag{1}$$

Subsequently, the latent variable is decomposed using fully connected layers to extract the inherent $Z_{i\text{-inher}}$ and variant components $Z_{i\text{-varia}}$,

$$Z_i = [Z_{i\text{-inher}}, Z_{i\text{-varia}}], i \in \{\text{pre}, \text{post}\}. \tag{2}$$

The decoder $D_{\text{pre}}, D_{\text{post}}$ then fuses the inherent variables of the pre-/post-treatment branches with their respective variant components to obtain the feature sequence for each branch. Finally, the sequences are concatenated to form the temporal fusion sequence F_{temporal},

$$F_{\text{temporal}} = \sum_i D_i(Z_{i\text{-inher}}, Z_{i\text{-varia}}),\ i \in \{\text{pre}, \text{post}\}. \tag{3}$$

To enhance the decoupling effect, the overall loss function is the weighted sum of the reconstruction loss $\mathcal{L}_{\text{recon}}$ and the contrastive loss $\mathcal{L}_{\text{contra}}$.

The reconstruction loss $\mathcal{L}_{\text{recon}}$ includes the mean square error loss and the KL divergence. The mean square error is adopted to minimize the difference between the reconstructed features and the input data, where the reconstructed features include the own intrinsic component and the cross intrinsic component reconstruction. KL divergence is employed to minimize the latent variables and the prior distribution. $\mathcal{L}_{\text{recon}}$ is shown as follows,

$$\mathcal{L}_{\text{recon}} = \sum_{i} \sum_{j} ||X_i - D_i(Z_{j\text{-Inher}}, Z_{i\text{-varia}})||_2^2 + \text{KL}(q(Z_i|X_i)||p(Z_i)), \quad (4)$$

where $i, j \in \{\text{pre}, \text{post}\}$.

The contrastive loss $\mathcal{L}_{\text{contra}}$ encourages the model to uncover subtle features associated with pCR, which employs cosine similarity (COS) loss and mean squared error (MSE) loss to retain inherent features while distinguishing variant features. $\mathcal{L}_{\text{contra}}$ is designed as follows,

$$\mathcal{L}_{\text{contra}} = \lambda \,\text{COS}(Z_{\text{pre-Varia}}, Z_{\text{post-Varia}}) + (1-\lambda)\,\text{MSE}(Z_{\text{pre-Inher}}, Z_{\text{post-Inher}}), \quad (5)$$

where λ is a hyper-parameter that controls the balance between these feature components.

2.2 Multi-Sequence MRI Fusion

For MRI multi-sequence fusion, we take contrast-enhanced T1-weighted images (T1) and T2-weighted images (T2) as inputs. T1 sequences highlight lesion enhancement features by injecting contrast agents into the blood, while T2 sequences reveal specific lesion features with high signal intensity. Therefore, in clinical practice, comparing images from these two sequences is essential for further assessing the condition of the patient. Leveraging this feature complementarity, we designed a self-attention [11] fusion module to capture the internal relationships between two sequences. First, we concatenate the T1 and T2 sequences, then incorporate relative positional encoding to enhance the self-attention module's ability to differentiate two sequences. Finally, the fused image, which contains complementary information, is generated as the output.

2.3 Multi-Modal MRI-WSI Fusion

Due to the significant differences between MRI and WSI modalities, we choose cross-attention to flexibly capture the dynamic relationships between modalities for deep fusion of heterogeneous modalities.

To overcome the challenge posed by the extremely high resolution of WSI images, which complicates processing, we first carry out feature extraction and then proceed with fusion. We extract the WSI feature sequence in three steps: segmentation, patching, and feature extraction [12]. First, the WSI image is converted from the RGB color space to the HSV color space, where a saturation

threshold is applied to distinguish the foreground from the background and delineate the tissue boundaries. Next, the image is divided into 224×224 patches, where a patch is considered valid if one of its four corner points lies within the tissue region, while patches outside the tissue region are excluded from feature extraction. Subsequently, we use a vision transformer-based feature extractor [13] to extract a feature sequence from each patch, and these sequences are concatenated to form the feature map for the WSI of each patient. Prior to multi-modal fusion, multi-head attention [14] is applied to pool the sequence length into a fixed value, ensuring alignment of the WSI features.

We used cross-attention [15] to fuse the MRI feature F_{MRI} with the WSI feature F_{WSI}. W_Q, W_K, W_V are learnable parameters. Q_{WSI} is the dot product of F_{WSI} and W_Q, and $K_{\text{MRI}}, V_{\text{MRI}}$ are the dot products between F_{MRI} and W_K, W_V. d_k is the dimension of F_{WSI} and F_{MRI}, which is 512 in our experiment. Cross-attention weight matrix A can be calculated as,

$$A = \text{Softmax}\left(\frac{Q_{\text{WSI}} K_{\text{MRI}}^T}{\sqrt{d_k}}\right). \tag{6}$$

The fused feature F_{fused} is the dot product of the weight matrix A and V_{MRI},

$$F_{\text{fused}} = A \cdot V_{\text{MRI}}. \tag{7}$$

After fusion, F_{fused} is processed by a convolutional neural network (CNN) followed by a multilayer perceptron (MLP) to perform pCR prediction task.

3 Experiments

3.1 Datasets and Evaluation Metrics

We established a dataset of HNSCC patients from a collaborating hospital, covering the period from 2020 to 2023. The dataset included T1/T2, pre-/post-treatment MRIs, WSI before NCIT, and the pCR label. The pCR label indicates the complete disappearance of malignant cells in the primary tumor and metastatic lymph nodes, as determined by pathology after NCIT. The dataset comprised 407 patients, of whom 122 achieved pCR and 285 did not. We split the dataset into training and validation sets with an 80/20 ratio and performed five-fold cross-validation to evaluate performance using five metrics: sensitivity, specificity, Area Under Curve (AUC), accuracy, and F1-score.

3.2 Implementation Details

The HARM^3-Fusion model was trained on an NVIDIA RTX 4090D 24GB GPU. To reduce overfitting, we employed L1 and L2 regularization, with the Adam optimizer. The initial learning rate was set to 10^{-4}, and the weight decay was 10^{-6}. The maximum number of epochs was set to 200, with a batch size of 8.

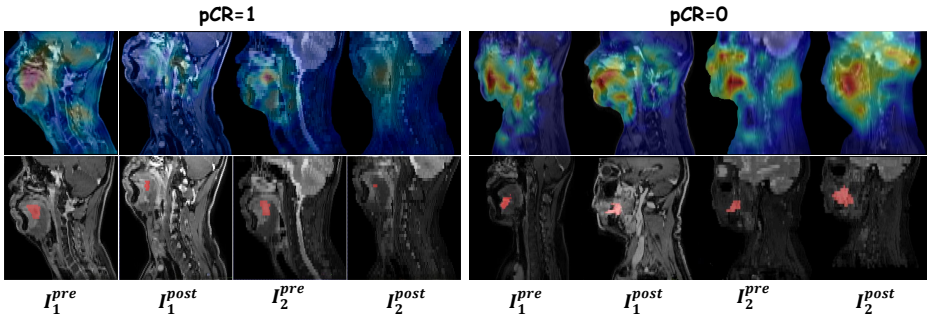

Fig. 2. Class Activation Maps (upper) and Voxel-wise annotations (lower) of HNSCC.

Table 1. Performance comparison between our proposed HARM^3-Fusion and state-of-the-art methods on HNSCC-pCR dataset. **Bold** denotes the best performance.

Model	MRI	WSI	AUC ↑	ACC ↑	F1-score ↑	SEN↑	SPE ↑
3D-CNN [16]	✓	–	0.5881	0.6854	0.5357	0.5212	0.614
ResNet [17]	✓	–	0.5218	0.6914	0.5879	0.5500	0.6705
3D-RPNET [5]	✓	–	0.7059	0.7365	0.6578	0.6337	0.7183
CLAM [12]	–	✓	0.6202	0.7020	0.6121	0.6278	0.6601
TransMIL [18]	–	✓	0.6132	0.6914	0.5745	0.6020	0.6191
HMCAT [19]	✓	✓	0.6968	0.7479	0.6741	0.6414	0.7350
M2Fusion [9]	✓	✓	0.6905	0.7256	0.6812	0.6455	0.7419
Ours	✓	✓	**0.7581**	**0.8272**	**0.7541**	**0.6578**	**0.8167**

3.3 Comparison to other methods

To evaluate the performance of our model, we compare HARM^3-Fusion with current mainstream or state-of-the-art unimodal and multi-modal models. The experimental results are shown in Tab. 1, where HARM^3-Fusion performs well on five core metrics, all of which are significantly better than the unimodal model. Compared with M2-fusion, an orthogonal multi-modal fusion model using a single MRI sequence, HARM^3-Fusion improves 6.76% on AUC, indicating that the hierarchical fusion mechanism effectively improves the ability of multi-sequence, multi-modal complementary information interaction. Meanwhile, compared with HMCAT, the F1-score is improved by 7.48%, indicating that HARM^3-Fusion's excellent decoupling ability for multi-sequence MRI can balance the impact of the long-tailed distribution of the data and accurately identify the positive samples, which has excellent clinical applicability.

We generated the decoupled feature heatmaps by extracting the weights and gradients of the last convolutional layer of the feature extraction network [20], as shown in Fig. 2, which provided the attention maps of HARM^3-Fusion across different regions of the MRI. The heatmaps include the feature extraction results for both T1 and T2 sequences, before and after treatment, along with the ground truth tumor location images manually annotated. By comparing the heatmaps

Table 2. Model performance with different attention configurations.

LD-VAE	Attention-1	Attention-2	AUC	ACC	F1-score	SEN	SPE
–	✓	✓	0.6258	0.6829	0.6586	0.5945	0.7367
✓	–	✓	0.6917	0.7784	0.6888	0.6210	0.7602
✓	✓	–	0.7195	0.7931	0.7206	0.6563	0.7862
✓	–	–	0.6810	0.7502	0.6812	0.6027	0.7586
✓	✓	✓	**0.7581**	**0.8272**	**0.7649**	**0.6978**	**0.8167**

with the ground truth, it is evident that the feature extraction network is able to focus on the tumor regions as well as other areas. This demonstrates that our MRI multi-temporal feature extraction model is capable of isolating the unique features associated with PCR and decoupling potential changes that cannot be captured by the ROI labels.

As a result, this study indicates that: (1) Our designed LD-VAE can effectively decouple the dynamic latent information of MRI before and after treatment. (2) Our multi-sequence fusion approach effectively utilizes the actual clinical information and improves the degree of interaction of T1-T2 information. (3) Our designed multi-modal fusion mechanism further enhances the classification prediction ability of HARM^3-Fusion.

3.4 Ablation Study

Multi-temporal MRI fusion module, multi-sequence MRI fusion module, and multi-modal MRI-WSI fusion modules are the three key components of HARM^3-Fusion. To evaluate the contribution of each component in our model to the pCR prediction performance, in Table 2, we progressively removed different modules in HARM^3-Fusion and compared the performance differences before and after the removal. The results are as follows: (1) When removing the multi-temporal MRI fusion module, we directly connected the features of MRI for subsequent prediction. The model's F1-score dropped by 10.6%, and sensitivity dropped by 10.33%, indicating that LD-VAE can effectively improve the decoupling ability of MRI temporal information and enhance the differentiation of positive samples. (2) When removing the multi-sequence MRI fusion module and multi-modal MRI-WSI fusion module and replacing it with feature linkage, the AUC decreased by 6.64% and 3.86%, respectively, indicating that the hierarchical attentional fusion module is critical for effectively enhancing the ability of multi-sequence and multi-modal information interaction. (3) Further, when only LD-VAE was retained for individual prediction, the AUC decreased by 7.71%, illustrating the critical role of fusing complementary anatomical and histopathological features for pCR prediction.

4 Conclusion

In this work, we propose HARM^3-Fusion, a framework that integrates clinical diagnostic data via three dedicated modules: a multi-temporal MRI fusion module, a multi-sequence MRI fusion module, and a multi-modal MRI-WSI fusion module to predict NCIT response for patients with HNSCC. We design a LD-VAE to fully decouple MRI temporal features, and subsequently integrate complementary multi-sequence MRI data via a self-attention mechanism, and finally employ a cross-attention mechanism to fuse WSI data. Ultimately, after validation on a comprehensive dataset of 407 patients, HARM^3-Fusion demonstrates superior robustness in pCR prediction.

Acknowledgments. This work was partially supported by the Shenzhen Science and Technology Program (Grant No. JCYJ20241202130548062), and the Guangdong Provincial Key Area Project of General Universities (Grant No. 2024ZDZX1017).

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

References

1. Daniel E Johnson, Barbara Burtness, C René Leemans, Vivian Wai Yan Lui, Julie E Bauman, and Jennifer R Grandis. Head and neck squamous cell carcinoma. *Nature reviews Disease primers*, 6(1):92, 2020.
2. Barbara Burtness, Kevin J Harrington, Richard Greil, Denis Soulières, Makoto Tahara, Gilberto de Castro, Amanda Psyrri, Neus Basté, Prakash Neupane, Åse Bratland, et al. Pembrolizumab alone or with chemotherapy versus cetuximab with chemotherapy for recurrent or metastatic squamous cell carcinoma of the head and neck (keynote-048): a randomised, open-label, phase 3 study. *The Lancet*, 394(10212):1915–1928, 2019.
3. Sabine Semrau, Antoniu-Oreste Gostian, Maximilian Traxdorf, Markus Eckstein, Sandra Rutzner, Jens von der Grün, Thomas Illmer, Matthias Hautmann, Gunther Klautke, Simon Laban, et al. Implementation of double immune checkpoint blockade increases response rate to induction chemotherapy in head and neck cancer. *Cancers*, 13(8):1959, 2021.
4. Davide Mattavelli, Gunnar Wichmann, Davide Smussi, Alberto Paderno, Maria Serrahima Plana, Ricard Nin Mesia, Micaela Compagnoni, Alessandro Medda, Susanna Chiocca, Stefano Calza, et al. Is precision medicine the solution to improve organ preservation in laryngeal/hypopharyngeal cancer? a position paper by the preserve research group. *Frontiers in Oncology*, 14:1433333, 2024.
5. Cheng Jin, Heng Yu, Jia Ke, Peirong Ding, Yongju Yi, Xiaofeng Jiang, Xin Duan, Jinghua Tang, Daniel T Chang, Xiaojian Wu, et al. Predicting treatment response from longitudinal images using multi-task deep learning. *Nature communications*, 12(1):1851, 2021.
6. Bao Li, Fengling Li, Zhenyu Liu, FangPing Xu, Guolin Ye, Wei Li, Yimin Zhang, Teng Zhu, Lizhi Shao, Chi Chen, 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.

7. Fengling Li, Yongquan Yang, Yani Wei, Yuanyuan Zhao, Jing Fu, Xiuli Xiao, Zhongxi Zheng, and Hong Bu. Predicting neoadjuvant chemotherapy benefit using deep learning from stromal histology in breast cancer. *NPJ Breast Cancer*, 8(1):124, 2022.
8. Kevin M Boehm, Omar SM El Nahhas, Antonio Marra, Pier Selenica, Hannah Y Wen, Britta Weigelt, Evan D Paul, Pavol Cekan, Ramona Erber, Chiara ML Loeffler, et al. Multimodal histopathologic models stratify hormone receptor-positive early breast cancer. *BioRxiv*, pages 2024–02, 2024.
9. Song Zhang, Siyao Du, Caixia Sun, Bao Li, Lizhi Shao, Lina Zhang, Kun Wang, Zhenyu Liu, and Jie Tian. 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*, pages 458–468. Springer, 2024.
10. Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013.
11. Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. Multi-head attention: Collaborate instead of concatenate. *arXiv preprint arXiv:2006.16362*, 2020.
12. Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021.
13. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
14. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
15. Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366, 2021.
16. Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.
17. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
18. Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34:2136–2147, 2021.
19. Zhe Li, Yuming Jiang, Mengkang Lu, Ruijiang Li, and Yong Xia. Survival prediction via hierarchical multimodal co-attention transformer: A computational histology-radiology solution. *IEEE Transactions on Medical Imaging*, 42(9):2678–2689, 2023.
20. Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: visual explanations from deep networks via gradient-based localization. *International journal of computer vision*, 128:336–359, 2020.