Predicting Radiation Therapy Response based on Dynamic Temporal Feature Difference Fusion from Longitudinal MRI

Xinyu Hao[1,2], Hongming Xu[1]✉, Qibin Zhang[1], Qi Xu[3], Xiaofeng Wang[4], Ilkka Polonen[2], and Fengyu Cong[1,2]

[1] School of Biomedical Engineering, Faculty of Medicine, Dalian University of Technology, Dalian 116024, China
[2] Faculty of Information Technology, University of Jyvaskyla, Jyvaskyla 40014, Finland
[3] School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China
[4] Department of Quantitative Health Sciences, Cleveland Clinic, Cleveland 44195, USA
mxu@dlut.edu.cn

Abstract. Significant progress has been made in AI-based prediction of therapeutic response to neoadjuvant chemotherapy (NAC) in breast cancer. However, current studies primarily rely on data from a single time point, neglecting the dynamic changes in tumor characteristics during treatment. To address this limitation, we propose a novel Dynamic Temporal Feature Difference Fusion (DTFDF) framework, which integrates image features from multiple time points throughout the treatment process to predict therapy response more precisely. Based on tumor spatial features, we design an innovative DTFDF strategy and introduce a treatment response-based triplet contrastive loss function to facilitate the learning of longitudinal tumor changes and enhance feature representation. Additionally, we incorporate biomarker prediction as an auxiliary task and introduce a feature decoupling-based multi-task learning module. This module generates feature representations for different tasks by accounting for both shared and task-specific information, improving response prediction. Experiments with data from 786 patients in the I-SPY 2 trial dataset demonstrate that our method achieves the highest AUC of 0.835 in predicting radiation therapy response, outperforming state-of-the-art (SOTA) approaches on longitudinal dynamic contrast-enhanced MRI data. Our source code is available at `https://github.com/AlexNmSED/DTFDF`.

Keywords: Longitudinal Medical Image Analysis · Temporal Feature Difference Fusion · Contrastive Loss.

1 Introduction

Unlike single time-point medical image analysis, longitudinal medical imaging captures changes over time, which is crucial for various medical applications, such as disease prediction, treatment response assessment, and prognosis modeling [5,13]. The trajectory of biomarker changes during treatment reflects tumor temporal heterogeneity, a key factor in predicting the neoadjuvant chemotherapy (NAC) response in breast cancer patients. Achieving pathological complete response (pCR) after NAC is the most desirable outcome and serves as a strong predictor of long-term survival for patients [2]. Additionally, accurate preoperative assessment of pCR can guide clinical decisions regarding surgical operations. The high heterogeneity of breast cancer exists throughout the various stages of NAC. The structural and functional changes in the tumor microenvironment during NAC can reflect the patient's treatment response and contain potential prognostic factors [8]. Although data from a single time-point are easier to acquire, they fail to capture dynamic changes during NAC and cannot fully characterize tumor heterogeneity. Therefore, integrating longitudinal imaging information to fully assess biological changes in breast cancer patients is essential.

Previous artificial intelligence (AI)-based methods for predicting the response to NAC in breast cancer have primarily focused on constructing models using single time-point imaging features, overlooking the dynamic monitoring of tumor temporal heterogeneity during the treatment process [4]. Recently, some studies have shown that incorporating longitudinal imaging data offers a more comprehensive understanding of disease progression, thereby enhancing predictive performance [15,17]. For example, Zhou et al. [19] employed contrastive disentangled representation learning to classify longitudinal CT radiomics features into shared and stage-specific categories. They utilized cycle cross-entropy loss to ensure feature consistency and contrastive disentanglement loss to separate shared and unique features. Additionally, several studies have explored deep feature fusion, highlighting the importance of integrating longitudinal imaging data. By combining multi-time-point information into a unified representation, these approaches effectively capture disease progression dynamics. Feature fusion strategies usually include sequential models and cross-attention mechanisms. In particular, time-series models like LSTM have been applied to predict treatment response and perform survival analysis using evolving clinical and imaging features [15,12,3]. Hu et al. [6] and Holste et al. [5] proposed VGG-TSwinformer and the longitudinal survival analysis transformer, respectively. Both studies leveraged the attention mechanism of Transformers to integrate longitudinal imaging data and improve disease prediction. Although the aforementioned studies demonstrate the potential of longitudinal medical imaging in evaluating the efficacy of breast cancer treatment, they merely fuse longitudinal image features along the time dimension, either through time-series models or cross-attention mechanisms. They neglected dynamic changes at different time points and failed to characterize the role of these changes in assessing the efficacy of therapy.

Motivated by the aforementioned challenges, we propose a novel DTFDF framework that integrates longitudinal imaging data via a treatment response-based triple contrastive loss (TR-TCL) to predict the neoadjuvant treatment response for breast cancer patients. Our main contributions include: (1) Built upon multi-stage spatial feature fusion, we develop an innovative longitudinal

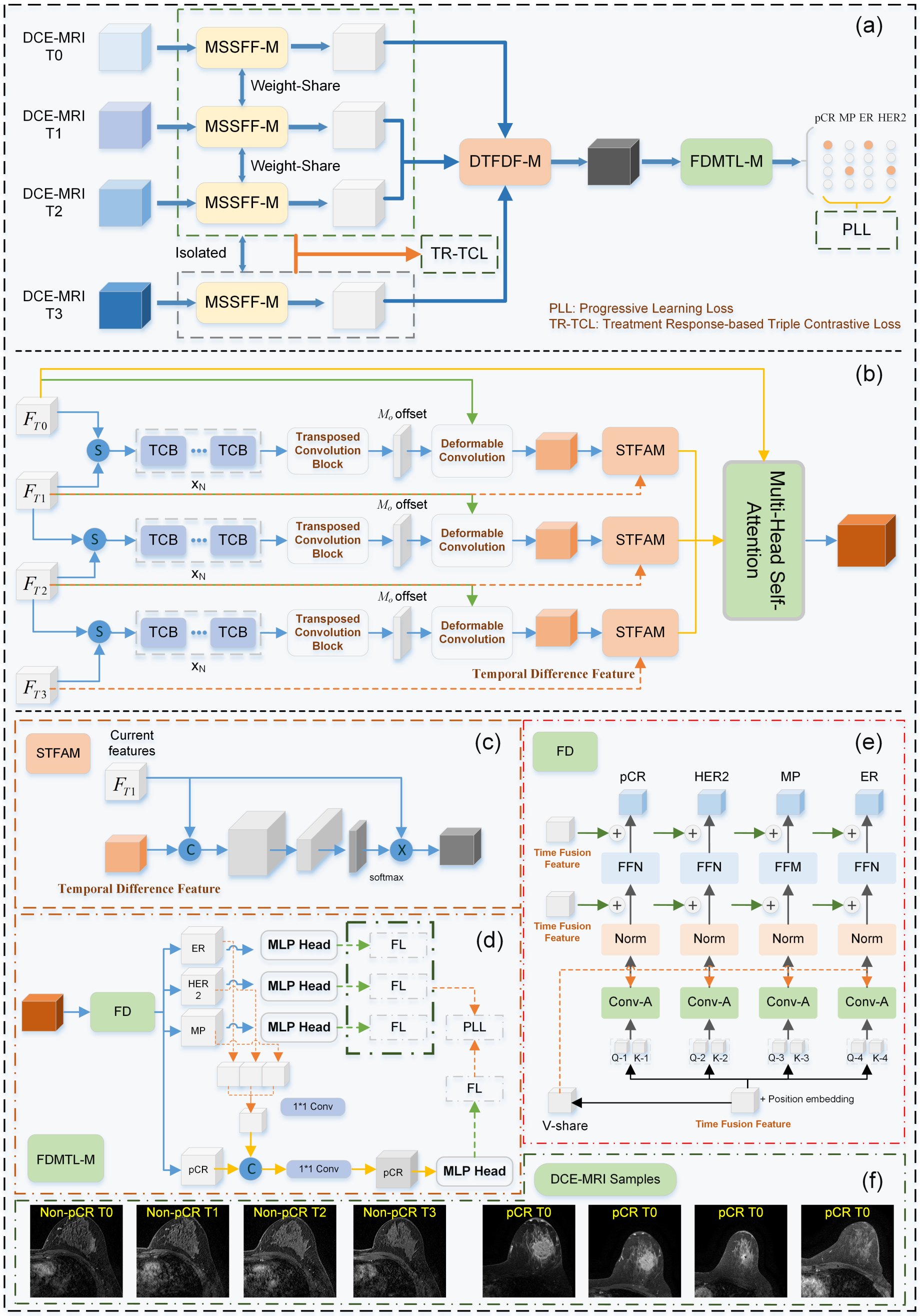

Fig. 1. Illustration of our proposed DTFDF framework. (a) Architecture of our DTFDF model. (b) DTFDF module. (c) STFAM module. (d) FDMTL module. (e) Feature decoupling module. (f) Examples of DCE-MRI samples.

medical image feature fusion strategy; (2) We introduce the TR-TCL to capture the subtle changes in tumor features along the timeline during treatment; (3) We propose a feature decoupling-based multi-task learning (FDMTL) strategy to further improve the prediction ability. Experimental results on the breast dynamic contrast-enhanced (DCE) MRI Trial (I-SPY 2) demonstrate that our method outperforms SOTA longitudinal image analysis methods. Our DTFDF model can be easily extended to other longitudinal medical image analysis frameworks for the personalized diagnosis of other diseases.

2 Method

Fig. 1(a) illustrates the architecture of our proposed DTFDF, which consists of several different modules: the multi-stage spatial feature fusion module (MSSFF-M), the dynamic temporal feature difference fusion module (DTFDF-M), and the feature decoupling-based multi-task learning module (FDMTL-M). Additionally, we introduce the TR-TCL to enhance the capability of MSSFF-M to recognize spatial feature disparities between during-treatment and post-treatment DCE-MRI of breast cancers. Details of different modules are provided below.

Multi-Stage Spatial Feature Fusion Module (MSSFF-M). As illustrated in Fig. 1(a), MSSFF-M extracts features from multi-stage DCE-MRI, including pre-treatment (T0), inter-treatment (T1, T2), and post-treatment (T3). It consists of structurally identical yet independent components, sharing parameters between the pre-treatment and inter-treatment stages. MSSFF-M incorporates the HiFuse block [7], which employs a hierarchical parallel fusion structure to integrate local features and global representations at different scales, leveraging the strengths of both CNNs and Transformers. Global feature extraction utilizes Windows-based Multi-head Self-Attention (W-MSA & SW-MSA) modules [11], while local feature extraction employs depthwise separable convolutions with skip connections [1], effectively reducing computational demands.

Treatment Response-based Triple Contrastive Loss (TR-TCL). Triple Contrastive Loss (TCL) [18] optimizes the embedding space by minimizing the distance between anchor and positive samples while maximizing the distance between anchor and negative samples. In the NAC process for breast cancer patients, achieving pCR signifies complete tumor regression after treatment, whereas non-pCR patients exhibit limited or even progressive tumor regression. Therefore, we categorize pre-treatment and post-treatment DCE-MRI images of pCR patients as "with changes", while those of non-pCR patients as "no changes". Let $X_{Batch} = \{X_1, X_2, ..., X_n\}$ represent a mini-batch of patients. For each patient X_i, we extract the spatial feature $F_{i,pre}$ from the pre-treatment DCE-MRI and the spatial features at three treatment time points using the MSSFF-M. To optimize the MSSFF-M, we introduce the TR-TCL as follows:

$$\mathcal{L}(f_A, f_P, f_N) = \max\left(\|f_A - f_P\|^2 - \|f_A - f_N\|^2 + \alpha, 0\right), \tag{1}$$

where f_A denotes the post-treatment spatial feature $F_{i,post}$, serving as the anchor; f_P denotes the feature of a positive sample, corresponding to DCE-MRI representations classified as "no changes", while f_N denotes the feature of a negative sample, whose representations are classified as "with changes". The Euclidean distance is computed between the anchor and positive/negative samples. The α in Eq. (1) defines the boundary threshold for separating positive and negative samples during the comparison process, which is set as 0.2 following [21]. In this study, if a patient achieves pCR, the pre-treatment spatial feature $F_{i,pre}$ is assigned as the negative sample feature. To select the positive sample for this patient, we choose the sample that has the maximum distance between $F_{i,post}$ and post-treatment features of all positive samples. When a patient does not achieve pCR, the positive sample selection follows the same strategy as the patient achieving pCR. To select the negative sample for this patient, we select the sample that has the minimum distance between $F_{i,post}$ and post-treatment features of all negative samples.

Dynamic Temporal Feature Difference Fusion Module (DTFDF-M). As illustrated in Fig. 1(b), we propose the DTFDF-M to integrate tumor spatial representations from longitudinal DCE-MRI. Given representations from two consecutive time points, such as F_{T0} and F_{T1}, let their shapes be $[C, H, W]$, where C denotes the number of channels, H represents the tensor height, and W represents its width. To fuse these spatial representations, F_{T0} and F_{T1} are stacked together to form $F_s \in \mathbb{R}^{2\times C\times H\times W}$. Subsequently, F_s is fed into a temporal convolutional block (TCB) with a residual structure. The TCB comprises two dilated causal convolutional layers, normalization layers, non-linear activation functions, and dropout layers, ensuring effective feature fusion and temporal modeling. In this study, we stack four TCB layers to capture richer feature representations. The output from the last TCB layer is fed into a transposed convolution block to restore the feature dimensions, resulting in ${F_s}'$. A 3×3 convolution operation is then employed to generate the offset M_o from ${F_s}'$. Both M_o and the tumor representation from the previous time point, e.g., F_{T0}, are fed into a deformable convolution layer to extract tumor dynamic features F_d over the temporal interval. To effectively integrate the temporal differences in tumor representations with current spatial features, e.g., F_{T1}, a Spatio-Temporal Feature Aggregation Module (STFAM) is developed for adaptive information fusion, as illustrated in Fig. 1(c). The STFAM generates the enhanced feature representation $\mathrm{f}_{F_{\mathrm{T1}}}$ as follows:

$$\mathrm{f}_{F_{\mathrm{T1}}} = \omega(F_{\mathrm{T1}}, F_d) \otimes F_{\mathrm{T1}}, \tag{2}$$

where $\omega(\cdot)$ denotes the operations to compute adaptive weights using a softmax function, and $\otimes$ denotes the element-wise multiplication. We integrate the temporal difference features from each time interval into the subsequent time point to predict pCR. To capture the temporal dependencies throughout the entire treatment period, a multi-head self-attention mechanism is employed to fuse the pre-treatment tumor representation and temporal difference features, resulting in the temporal fusion representation, denoted as F_{pCR}.

Feature Decoupling-based Multi-Task Learning Module (FDMTL-M). In clinical practice, the proportion of patients achieving pCR is generally low, and the performance of pCR prediction varies across molecular subtypes, reflecting their intrinsic biological heterogeneity. Relying solely on single-task learning methods for pCR prediction may be insufficient to achieve optimal results. Therefore, we propose incorporating the prediction of biomarkers closely associated with pCR as auxiliary tasks within the multi-task learning (MTL) framework. To achieve this, we introduce a feature decoupling-based MTL approach (see Fig. 1(d)), which generates feature representations for different tasks by accounting for both shared and task-specific information.

As shown in Fig. 1(e), we extend Conv-Former block [16] into a multi-task feature decoupling module (FD-M). In this module, the temporal fusion representation F_{pCR} contains shared information across different tasks, which is utilized as the value matrix V_{share}. The task-specific representation F_t is derived from task-related query-key feature representations, Q_t and K_t, where $t \in \{\text{pCR, ER, HER2, MP}\}$, which is calculated as:

$$f = \text{Norm}\big(V_{\text{share}} + \text{ConvAttention}(V_{\text{share}}, Q_t, K_t)\big), \tag{3}$$

$$F_t = \text{Norm}\big(V_{\text{share}} + \text{FFN}(f)\big). \tag{4}$$

F_t is then passed through its respective classification head to create a binary classifier. For the pCR prediction task, we incorporate the features of auxiliary tasks. Focal Loss (FL) $\mathcal{L}_{\text{Focal}}$ is employed to address class imbalance issues. In our experiments, $\mathcal{L}_{\text{Focal}}$ is utilized to evaluate each specific task, with α and γ set to [1,1] and 2, respectively. Additionally, we introduce a progressive learning loss strategy for the loss function of the MTL framework. Initially, equal weights are assigned to the primary and auxiliary task losses. As training progresses, the weight of the primary task loss gradually increases, while those of the auxiliary tasks decrease. This ensures that, towards the end of training, the primary task plays a more dominant role. The overall loss $\mathcal{L}$ is calculated as follows:

$$\mathcal{L} = w_1\mathcal{L}_{\text{TR-TCL}} + (1 - w_1)\mathcal{L}_{\text{PLL}}, \tag{5}$$

$$\mathcal{L}_{\text{PLL}} = w_2\mathcal{L}_{\text{Focal}_{\text{pCR}}} + (1 - w_2)\left(\frac{\sum \mathcal{L}_{\text{Focal}_{a_t}}}{N}\right), \tag{6}$$

where w_1 is fixed at 0.01, $a_t \in \{\text{ER, HER2, MP}\}$, and N is the number of auxiliary tasks. During training, w_2 is initially set to 0.5, which progressively increases such that $1 - w_2$ decreases to 0.01 by the end of training.

3 Experiments and Results

3.1 Dataset and Experimental Settings

The dataset used in this study is obtained from the I-SPY 2 trial [10]. The I-SPY 2 imaging cohort comprises 985 patients, of whom 199 were excluded due to missing multi-time point MRI scans, poor image quality, or misaligned sequences.

The remaining 786 patients, each undergoing four MRI scans before and during NAC (see Fig. 1(f)), were randomly divided into training and testing sets at a 4:1 ratio. The training set was further divided into five folds for cross-validation. The time points are defined as follows: pre-treatment (T0, pre-NAC); after 3 cycles (T1, early NAC); after 12 cycles and between drug regimens (T2, mid-NAC); and post-treatment (T3, post-NAC, before surgery). To ensure consistent spatial resolution, all MRI images are resampled to a voxel size of $1 \times 1 \times 1mm^3$. To reduce noise disturbance, the first and last five slices of each MRI scan are discarded, and the remaining slices are resampled to $64 \times 256 \times 256$ pixels. Our DTFDF model was trained on four NVIDIA P100 GPUs with a batch size of 32 for 50 epochs, employing an early stopping strategy. The initial learning rate was set to 0.002, with a weight decay of 0.05. AUC, accuracy (ACC), sensitivity (SEN), and specificity (SPE) are used as evaluation metrics.

3.2 Comparisons with Existing Methods

We compare the proposed DTFDF model with three SOTA methods for longitudinal medical image analysis: LSTM [12], Transformer [9], DiT [14], and LOMIA-T [20]. All methods utilize MRI data from all time points of each patient, with HiFuse employed as the feature extractor. Table 1 presents the comparative results. It is observed that DTFDF achieves the best performance in predicting pCR, attaining the highest AUC value of 0.835, significantly outperforming DiT (0.785), LOMIA-T (0.769), and Transformer (0.771). This superiority is mainly attributed to the proposed dynamic temporal feature difference fusion strategy. Unlike transformer-based methods (e.g., DiT & LOMIA-T), which treat feature maps as tokens to compute self-attention for modeling temporal relationships, our fusion strategy effectively captures longitudinal tumor changes, which are critical for pCR prediction accuracy. In addition, the TR-TCL module constrains the model to learn discriminative features between pre- and post-treatment, further enhancing prediction performance. Notably, the LSTM method [12] does not perform well, suggesting that the self-attention mechanism in Transformers offers superior temporal modeling capabilities for pCR prediction.

Table 1. Comparison of our model with existing methods.

Methods	AUC	ACC	SEN	SPE
LSTM [12]	$0.514_{\pm 0.018}$	$0.509_{\pm 0.013}$	$0.559_{\pm 0.118}$	$0.451_{\pm 0.125}$
Transformer [9]	$0.771_{\pm 0.036}$	$0.699_{\pm 0.013}$	$0.676_{\pm 0.022}$	$0.710_{\pm 0.020}$
DiT [14]	$0.785_{\pm 0.016}$	$0.746_{\pm 0.026}$	$0.618_{\pm 0.024}$	$0.824_{\pm 0.054}$
LOMIA-T [20]	$0.769_{\pm 0.022}$	$0.766_{\pm 0.009}$	$0.596_{\pm 0.042}$	$\mathbf{0.850_{\pm 0.001}}$
DTFDF (Ours)	$\mathbf{0.835_{\pm 0.043}}$	$\mathbf{0.764_{\pm 0.059}}$	$\mathbf{0.745_{\pm 0.035}}$	$0.775_{\pm 0.088}$

3.3 Ablation Study

We conduct extensive ablation studies to investigate the contribution of different time points and the effectiveness of our proposed components, as presented in Table 2. All experiments are performed under identical training hyperparameters. As shown in Table 2, using only pre-treatment images (T0) results in the poorest predictive performance. Combining T0 with early-treatment images (T1) does not lead to a significantly improvement in AUC value, which we attribute to the minimal tumor changes observed between T0 and T1. In contrast, using only post-treatment images (T3) results in a 2.3% improvement in AUC compared to T0, highlighting the higher predictive value of T3 images in the pCR prediction task. Furthermore, combining T0 and T3 leads to a further 2.9% improvement in AUC, reinforcing this finding. Notably, replacing the DTFDF-M with a Transformer-based fusion strategy, while keeping the other two proposed components (TR-TCL and FDMTL-M), also enhances performance, with the AUC increasing from 0.771 to 0.802. However, compared to our DTFDF, this approach leads to a 3.3% drop in AUC, emphasizing the crucial role of the DTFDF module in improving pCR prediction. Additionally, when compared to models lacking TR-TCL and FDMTL-M, our DTFDF achieves AUC improvements of 1.8% and 1.5%, respectively, further validating the effectiveness of the treatment-response-based triple contrastive loss and the multi-task learning strategy. Overall, models that exclude DTFDF-M or TR-TCL demonstrate inferior predictive performance, highlighting the importance of these strategies in enabling the model to learn treatment-related differential features crucial for pCR prediction, thus enhancing its accuracy.

Table 2. Ablation studies on different strategies.

Methods	M-1	M-2	M-3	AUC	ACC	SEN	SPE
Only T0	-	-	✓	$0.761_{\pm 0.005}$	$0.714_{\pm 0.023}$	$0.657_{\pm 0.016}$	$0.742_{\pm 0.039}$
Only T3	-	-	✓	$0.784_{\pm 0.056}$	$0.735_{\pm 0.055}$	$0.649_{\pm 0.033}$	$0.778_{\pm 0.098}$
T0+T1	✓	-	✓	$0.764_{\pm 0.033}$	$0.678_{\pm 0.026}$	$0.715_{\pm 0.058}$	$0.657_{\pm 0.053}$
T0+T3	✓	-	✓	$0.790_{\pm 0.038}$	$0.705_{\pm 0.040}$	$0.731_{\pm 0.058}$	$0.691_{\pm 0.059}$
w/o DTFDF-M	-	✓	✓	$0.802_{\pm 0.014}$	$0.758_{\pm 0.019}$	$0.626_{\pm 0.040}$	$\underline{\mathbf{0.822}_{\pm \mathbf{0.048}}}$
w/o TR-TCL	✓	-	✓	$0.817_{\pm 0.042}$	$0.754_{\pm 0.058}$	$0.718_{\pm 0.032}$	$0.774_{\pm 0.097}$
w/o FDMTL-M	✓	✓	-	$0.820_{\pm 0.030}$	$0.738_{\pm 0.040}$	$0.718_{\pm 0.057}$	$0.749_{\pm 0.060}$
Our DTFDF	✓	✓	✓	$\underline{\mathbf{0.835}_{\pm \mathbf{0.043}}}$	$\underline{\mathbf{0.764}_{\pm \mathbf{0.059}}}$	$\underline{\mathbf{0.745}_{\pm \mathbf{0.035}}}$	$0.775_{\pm 0.088}$

[1] M-1 for DTFDF-M, M-2 for TR-TCL, M-3 for FDMTL-M

4 Conclusion

In this study, we propose a novel longitudinal medical image analysis framework to predict the therapy response of breast cancer patients to NAC. Building on multi-stage tumor spatial feature fusion, we design a Dynamic Temporal Feature Difference Fusion (DTFDF) strategy, which demonstrated superior performance compared to existing fusion approaches based on LSTM and Transformer models. Additionally, to capture subtle temporal changes in tumor characteristics across patient MRIs, we introduce a treatment response-based triplet contrastive loss function to enhance feature learning and improve predictive accuracy. Evaluations on the publicly available I-SPY 2 trial dataset demonstrate the superiority of our method. An ablation study further validates the effectiveness of the included modules. However, the current evaluation relies on data from all four time points. In the future, we will explore how to incorporate the dynamic changes of the tumor into pre-treatment MRI features.

Acknowledgments. This work was supported in part by Liaoning Province Science and Technology Joint Program (2024-MSLH-065), the Fundamental Research Funds for Central Universities (DUT25Z2514, DUT24YG201), and scholarship from the China Scholarship Council (No. 202006060060). (Corresponding author: Hongming Xu)

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

References

1. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1251–1258 (2017)
2. Cortazar, P., Zhang, L., Untch, M., et al.: Pathological complete response and long-term clinical benefit in breast cancer: the ctneobc pooled analysis. The Lancet **384**(9938), 164–172 (2014)
3. Gao, Y., Ventura-Diaz, S., Wang, X., et al.: An explainable longitudinal multimodal fusion model for predicting neoadjuvant therapy response in women with breast cancer. Nature Communications **15**(1), 9613 (2024)
4. Hao, X., Xu, H., Zhao, N., et al.: Predicting pathological complete response based on weakly and semi-supervised joint learning in breast cancer multi-parametric mri. Biomedical Signal Processing and Control **93**, 106164 (2024)
5. Holste, G., Lin, M., Zhou, R., et al.: Harnessing the power of longitudinal medical imaging for eye disease prognosis using transformer-based sequence modeling. NPJ Digital Medicine **7**(1), 216 (2024)
6. Hu, Z., Wang, Z., Jin, Y., Hou, W.: Vgg-tswinformer: Transformer-based deep learning model for early alzheimer's disease prediction. Computer Methods and Programs in Biomedicine **229**, 107291 (2023)
7. Huo, X., Sun, G., Tian, S., et al.: Hifuse: Hierarchical multi-scale feature fusion network for medical image classification. Biomedical Signal Processing and Control **87**, 105534 (2024)
8. Kurtz, D.M., Esfahani, M.S., Scherer, F., et al.: Dynamic risk profiling using serial tumor biomarkers for personalized outcome prediction. Cell **178**(3), 699–713 (2019)
9. Li, T.Z., Xu, K., Gao, R., et al.: Time-distance vision transformers in lung cancer diagnosis from longitudinal computed tomography. In: Proceedings of SPIE–the International Society for Optical Engineering. vol. 12464. NIH Public Access (2023)

10. Li, W., Newitt, D.C., Gibbs, J., et al.: Predicting breast cancer response to neoadjuvant treatment using multi-feature mri: results from the i-spy 2 trial. NPJ breast cancer **6**(1), 63 (2020)
11. Liu, Z., Lin, Y., Cao, Y., et al.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
12. Mei, X., Liu, Z., Singh, A., et al.: Interstitial lung disease diagnosis and prognosis using an ai system integrating longitudinal data. Nature communications **14**(1), 2272 (2023)
13. van Timmeren, J., Bussink, J., Koopmans, P., et al.: Longitudinal image data for outcome modeling. Clinical Oncology (2024)
14. Tong, T., Li, D., Gu, J., et al.: Dual-input transformer: An end-to-end model for preoperative assessment of pathological complete response to neoadjuvant chemotherapy in breast cancer ultrasonography. IEEE Journal of Biomedical and Health Informatics **27**(1), 251–262 (2022)
15. Xu, Y., Hosny, A., Zeleznik, R., et al.: Deep learning predicts lung cancer treatment response from serial medical imaging. Clinical Cancer Research **25**(11), 3266–3275 (2019)
16. Xu, Z., Wu, D., Yu, C., et al.: Sctnet: Single-branch cnn with transformer semantic information for real-time segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 6378–6386 (2024)
17. Yamashita, R., Long, J., Longacre, T., et al.: Deep learning model for the prediction of microsatellite instability in colorectal cancer: a diagnostic study. The Lancet Oncology **22**(1), 132–141 (2021)
18. Yang, J., Duan, J., Tran, S., et al.: Vision-language pre-training with triple contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15671–15680 (2022)
19. Zhou, X., Yue, H., Zheng, Z., et al.: Predicting pathological response in esophageal squamous cell carcinoma with longitudinal ct radiomics and disentangled representation learning: a multicenter retrospective cohort study. International Journal of Surgery pp. 10–1097 (2024)
20. Sun, Y., Li, K., Chen, D., et al.: 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)
21. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823 (2015)