Testable Concept Bottlenecks for Pretreatment pCR Prediction in Breast DCE-MRI: Faithfulness and Reliability

Jiyang Tan[1⋆], Xinglong Liang[2,3⋆], Xinyi Hu[1], Youran Zhao[1], Jun Xu[1⋆⋆], and Tianyu Zhang[2,3⋆⋆]

[1] Harbin Institute of Technology, Shenzhen, China
[2] Department of Radiology, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX, Amsterdam, The Netherlands
[3] Department of Radiology and Nuclear Medicine, Radboud University Medical Centre, Geert Grooteplein 10, 6525 GA, Nijmegen, The Netherlands
xujunqgy@hit.edu.cn, t.zhang@nki.nl

Abstract. Pretreatment prediction of pathological complete response (pCR) from breast dynamic contrast-enhanced MRI (DCE-MRI) is clinically meaningful, yet explanations are often hard to validate and may be unstable under perturbations. We propose a testable projection-based concept bottleneck for pretreatment pCR classification that produces radiologist-readable concept evidence and enables quantitative evaluation of explanation quality. A 3D ResNet is trained end-to-end and then frozen as a feature extractor; a lightweight bottleneck projects image representations into a biomedical text-embedding space and scores K=13 BI-RADS-style radiologic concepts via cosine similarity using a frozen PMC-CLIP text encoder. Faithfulness is assessed by controlled concept interventions (top-m zeroing versus random controls), while reliability is quantified by concept stability under test-time augmentation (TTA) using a case-level score and summarized through coverage–performance analysis under reliability-based subset selection. Experiments on pretreatment I-SPY2 DCE-MRI show performance comparable to the end-to-end baseline while providing controllable, testable, and stability-aware concept evidence. Code is available at `https://github.com/dairoTJY/breast-mri-cbm`.

Keywords: Pretreatment pCR Prediction · Breast DCE-MRI · Concept Bottleneck Models · Faithfulness · Reliability.

1 Introduction

In breast cancer, neoadjuvant therapy (NAT) is administered before surgery, and pathological complete response (pCR) assessed on post-NAT pathology is a key endpoint associated with favorable prognosis [2]. Predicting pCR before therapy

⋆ Jiyang Tan and Xinglong Liang contributed equally to this work.
⋆⋆ Corresponding author

initiation is therefore clinically meaningful. Breast dynamic contrast-enhanced MRI (DCE-MRI) captures enhancement kinetics and morphological phenotypes related to tumor vascularity and heterogeneity, making it an important imaging source for response prediction. We focus on pretreatment-only DCE-MRI (T0), which aligns with the therapy-start decision point and avoids reliance on on-treatment or longitudinal scans that encode early response changes [15].

Under this pretreatment-only constraint, models must infer outcome from baseline phenotypes alone. While recent methods improve pCR prediction by leveraging multi-time imaging, longitudinal trajectories, and multimodal variables [22,13,9], these settings address early response monitoring where treatment-course signals are available. This leaves a gap: how to achieve competitive pretreatment pCR prediction with explanations that are both testable and reliable.

Explainability remains a barrier to clinical translation. Clinically useful explanations should support testable faithfulness (whether the evidence truly drives the prediction under controlled changes) and case-level reliability (whether the evidence is stable enough for deployment-oriented quality control). Post-hoc attribution heatmaps (e.g., Grad-CAM [17]) are intuitive but provide continuous visualizations without a discrete intervention interface, making controlled counterfactual tests ambiguous; they can also be sensitive to perturbations. In contrast, concept-based evidence introduces explicit concept variables that can be intervened on, and reliability can be quantified via stability under test-time augmentation (TTA) [19], enabling coverage–performance analysis via reliability-based subset selection [4].

Concept Bottleneck Models (CBMs) enable testable explanations by explicitly modeling concepts and supporting concept-level interventions [10,21,14], but instance-level concept labels and expert corrections are often unavailable at scale in medical imaging. We therefore adopt a projection-based concept bottleneck predictor for pretreatment breast DCE-MRI pCR prediction under minimal concept supervision: we use a 3D ResNet backbone as the encoder [5], compare it with other 3D CNN backbones in the experiments, freeze the backbone, and learn only a lightweight bottleneck and linear head that map image representations into a biomedical text-embedding space to produce radiologist-readable concept scores. We evaluate explanations with a faithfulness–reliability scheme, using controlled concept interventions and TTA-based stability summarized through coverage–performance analysis.

The main contributions are: (1) 13 BI-RADS-style radiologic concepts for pretreatment two-phase DCE-MRI, with an LLM used only for text-level vocabulary curation (summarization/normalization/de-duplication); (2) a pretreatment-only projection-based concept bottleneck predictor that preserves competitive predictive performance while producing concept evidence; and (3) a quantitative evaluation protocol for faithfulness (interventions) and reliability (TTA stability) demonstrating a practical quality-control signal via coverage–performance analysis.

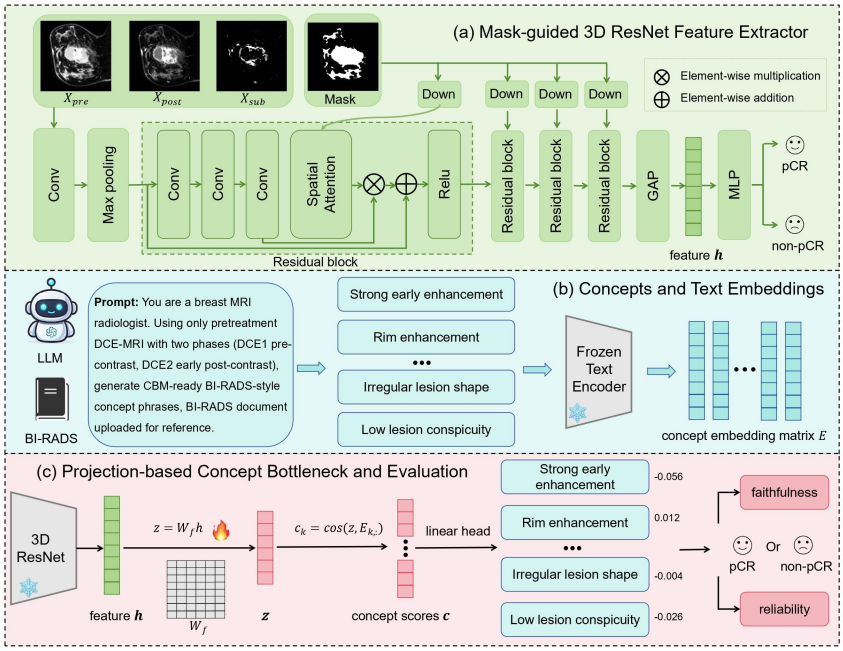

Fig. 1. Overview of our framework for pretreatment pCR prediction from two-phase breast DCE-MRI. (a) A mask-guided 3D ResNet-50 extracts a case-level feature h from X. (b) BI-RADS-style concept prompts are encoded by a frozen text encoder to form a fixed concept embedding matrix E. (c) A trainable projection maps h to z. Concept scores c are computed by comparing z with the concept embeddings E, and a linear concept head predicts pCR. Faithfulness and reliability are then evaluated.

2 Methods

Given pretreatment (T0) breast DCE-MRI, our goal is to predict pCR $y \in \{0, 1\}$ (y=1 indicates pCR) and to produce intervenable concept evidence for evaluating explanation quality. An overview of the proposed pipeline is shown in Fig. 1. Let $X \in \mathbb{R}^{3\times D\times H\times W}$ denote the three-channel T0 volume and $M \in \mathbb{R}^{1\times D\times H\times W}$ the aligned tumor ROI mask. We first train an end-to-end 3D ResNet and then freeze its backbone to obtain a stable representation h; we project h into a text-embedding space and compute cosine similarities to K=13 predefined radiologic concept vectors, yielding a concept score vector $c(X, M) \in \mathbb{R}^K$ for final prediction and for faithfulness/reliability assessment via direct edits on c.

2.1 3D ResNet Backbone

We construct $X \in \mathbb{R}^{3\times D\times H\times W}$ from the pre-contrast image X_{pre}, the early post-contrast image X_{post}, and the subtraction image $X_{\text{sub}} = X_{\text{post}} - X_{\text{pre}}$:

$$X = [X_{\text{pre}},\, X_{\text{post}},\, X_{\text{sub}}]. \tag{1}$$

The tumor mask M is incorporated via mask-guided spatial attention to emphasize tumor-relevant regions. Let $\psi(\cdot)$ denote the ResNet stem. We inject M after the stem by resizing it to the current feature resolution and concatenating it with the feature tensor:

$$Z_0 = \mathrm{Concat}\big(\mathrm{Down}(M),\, \psi(X)\big), \tag{2}$$

where $\mathrm{Down}(\cdot)$ denotes interpolation/downsampling to match the spatial size of $\psi(X)$. The subsequent 3D ResNet $f_\theta(\cdot)$ applies a mask-guided spatial attention gate in each residual block and yields a case-level representation via global average pooling:

$$h = f_\theta(Z_0) \in \mathbb{R}^{D_f}. \tag{3}$$

An end-to-end baseline uses a two-layer MLP $g(\cdot)$ to produce logit $s = g(h)$ and probability $p = \sigma(s)$. After training, we freeze f_θ as a feature extractor for the concept-bottleneck stage.

2.2 Radiologic Concepts and Text Embeddings

BI-RADS (Breast Imaging Reporting and Data System) is a standardized lexicon for breast imaging interpretation and reporting, including MRI descriptors of lesion morphology, enhancement, and distribution. We define K=13 radiologist-readable concept nodes following the BI-RADS MRI lexicon, restricted to attributes observable from pretreatment two-phase DCE (pre and early post) and derived images (e.g., early subtraction) [1,18]. An LLM (GPT-5.2 in our implementation) is used only to curate the BI-RADS–style concept vocabulary (summarization/normalization/de-duplication under the two-phase constraint). It does not generate instance-level concept labels or any learning signal, and is not used in training or inference.

We use the text encoder from PMC-CLIP as the frozen encoder $\phi(\cdot)$ to map each prompt t_k into an ℓ_2-normalized concept vector [12]:

$$v_k = \frac{\phi(t_k)}{\|\phi(t_k)\|_2} \in \mathbb{R}^{D_t}, \quad k = 1, \ldots, K. \tag{4}$$

PMC-CLIP is pretrained on biomedical figure–caption pairs, yielding text embeddings that better align with biomedical terminology; this makes it well-suited for encoding BI-RADS-style radiologic concept prompts used in our pCR prediction setting. When multiple templates are used per concept, we average embeddings and re-normalize. Stacking $\{v_k\}$ yields

$$E \in \mathbb{R}^{K \times D_t}, \quad E_{k,:} = v_k^\top, \tag{5}$$

which serves as the fixed reference for similarity-based concept scoring.

2.3 Projection-based Concept Bottleneck Predictor

Let $h = f_\theta(Z_0)$ be extracted by the frozen backbone, where Z_0 is constructed from (X, M) as in Eq. 2. We learn a linear projection $W_f \in \mathbb{R}^{D_t \times D_f}$ and normalize the projected feature:

$$z = W_f h, \;\; \bar{z} = \frac{z}{\|z\|_2}. \tag{6}$$

We define the concept score vector as cosine similarities between $\bar{z}$ and concept embeddings:

$$c(X, M) = \bar{z} E^\top \in \mathbb{R}^K, \;\; c_k \in [-1, 1]. \tag{7}$$

A linear concept head produces the logit and probability

$$\ell = w^\top c + b, \;\; p(X, M) = \sigma(\ell), \tag{8}$$

where $w \in \mathbb{R}^K$ and $b \in \mathbb{R}$. For instance i, we define concept contribution

$$a_{i,k} = w_k \, c_{i,k}, \tag{9}$$

used to rank concepts (e.g., top-m by $|a_{i,k}|$). All faithfulness and reliability evaluations are performed directly on c.

2.4 Faithfulness via Concept Interventions

We assess faithfulness by controlled interventions on c: if a concept dimension is ranked as influential, editing it should induce a larger change in model output. For any index set $S \subseteq \{1, \ldots, K\}$, we apply a zeroing intervention

$$\tilde{c}_k = \begin{cases} 0, & k \in S, \\ c_k, & \text{otherwise}, \end{cases} \tag{10}$$

and obtain post-intervention probability $\tilde{p}$ by feeding $\tilde{c}$ into Eq. (8). For each instance, we select S_m as the top-m concepts by $|a_{i,k}|$ and compare against a random-m control (uniform sampling; repeated 5 times and averaged). We used m=3 and m=5 to represent small- and medium-sized concept interventions, respectively, while keeping the number of edited concepts interpretable. We quantified intervention effects from complementary perspectives. $\mathbb{E}[|\Delta p|]$ captures the average magnitude of prediction change, while PCR_ϵ measures how often interventions induce a non-trivial change beyond a tolerance ϵ. To assess whether interventions alter overall discrimination rather than only per-sample probabilities, we also reported ΔAUC:

$$\mathbb{E}[|\Delta p|] = \frac{1}{N} \sum_{i=1}^{N} |\tilde{p}_i - p_i| \,. \tag{11}$$

$$\Delta\mathrm{AUC} = \mathrm{AUC}(\tilde{p}) - \mathrm{AUC}(p). \tag{12}$$

$$\mathrm{PCR}_\epsilon = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(|\tilde{p}_i - p_i| > \epsilon) \,. \tag{13}$$

2.5 Concept Reliability under Test-Time Augmentation

We quantify reliability as stability of concept evidence under mild perturbations. For each case, we run T=20 test-time augmentation (TTA) trials to obtain concept outputs $\{c^{(t)}\}_{t=1}^{T} \subset \mathbb{R}^K$ and compute the mean per-concept variability:

$$U_{\text{mean}} = \frac{1}{K} \sum_{k=1}^{K} \text{Std}_t\left(c_k^{(t)}\right). \tag{14}$$

Smaller U_{mean} indicates more stable concept evidence. To evaluate practical utility, we rank cases by increasing U_{mean} and compute predictive metrics on retained subsets at preset coverages $p \in \{50, 70, 90, 100\}\%$, where coverage denotes the fraction of test cases retained after sorting by increasing U_{mean}. This yields a coverage–performance relationship used to evaluate the utility of U_{mean} as a quality-control score.

3 Experiments

3.1 Dataset and Implementation Details

We used the eligible pretreatment (T0) I-SPY2 cohort with available pre- and early post-contrast DCE-MRI, pCR labels, and tumor masks [16]. We used patient-level train/val/test splits of 8:1:1 (949 cases; 309 pCR-positive). All experiments were run on a single NVIDIA GeForce RTX 4090 GPU (24 GB).

For each case, we extracted the two pretreatment phases from the stacked DCE series. Volumes were resampled to $2\times2\times2\,\text{mm}^3$. Tumor masks were generated by a pretrained 3D nnU-Net [8,3] and used as ROI mask M. Each channel was z-score normalized and padded/cropped to $160 \times 160 \times 160$. Training used random flip/affine and additive Gaussian noise; no augmentation was applied at validation/testing.

The end-to-end baseline was a 3D ResNet-50 trained with a weighted random sampler and the mixed loss:

$$\mathcal{L} = 0.5\,\mathcal{L}_{\text{focal}} + 0.5\,\mathcal{L}_{\text{bce}} \tag{15}$$

where $\mathcal{L}_{\text{focal}}$ is focal loss [11] and $\mathcal{L}_{\text{bce}}$ is BCEWithLogitsLoss. We used AdamW for optimization with a batch size of 16. Training was run for 50 epochs, and the best checkpoint was selected by validation AUC. For the concept-bottleneck stage, the backbone was frozen and only the projection and linear concept head were trained.

Performance was evaluated using area under the ROC curve (AUC), accuracy (ACC), F1 score, sensitivity (SEN), and specificity (SPE). For threshold-dependent metrics, the operating threshold was selected on the validation set by maximizing the Youden index and applied to the test set [20]. Results are reported as mean ± 95% CI.

Table 1. Test-set performance (mean ± 95% CI).

Model	AUC	F1	SEN	SPE	ACC
ResNet-18	0.68 ± 0.11	0.52 ± 0.15	0.64 ± 0.19	0.71 ± 0.11	0.69 ± 0.09
ResNet-34	0.70 ± 0.11	0.52 ± 0.13	0.80 ± 0.16	0.54 ± 0.11	0.61 ± 0.09
ResNet-50	0.71 ± 0.12	0.52 ± 0.15	0.64 ± 0.19	0.71 ± 0.10	0.69 ± 0.09
SE-ResNet50	0.66 ± 0.13	0.45 ± 0.14	0.68 ± 0.19	0.51 ± 0.12	0.56 ± 0.09
SE-ResNet101	0.69 ± 0.12	0.47 ± 0.15	0.56 ± 0.19	0.71 ± 0.10	0.67 ± 0.09
DenseNet121	0.69 ± 0.12	0.49 ± 0.16	0.56 ± 0.19	0.74 ± 0.10	0.69 ± 0.09
DenseNet169	0.68 ± 0.13	0.50 ± 0.16	0.60 ± 0.19	0.71 ± 0.10	0.68 ± 0.09
CBM(Ours)	0.69 ± 0.12	0.50 ± 0.16	0.55 ± 0.20	0.78 ± 0.09	0.73 ± 0.08

Table 2. Faithfulness evaluation via concept zeroing. m denotes the number of concept dimensions zeroed per instance (top-m vs. random-m).

	top-m			random-m		
m	$\mathbb{E}[\lvert\Delta p\rvert]$	$\mathrm{PCR}_{0.05}$	ΔAUC	$\mathbb{E}[\lvert\Delta p\rvert]$	$\mathrm{PCR}_{0.05}$	ΔAUC
3	0.046	0.021	−0.375	0.013	0.002	−0.194
5	0.052	0.695	−0.378	0.019	0.006	−0.166

3.2 Results

Comparison with the End-to-End Baseline Table 1 summarizes the test results across several 3D backbones, including ResNet, DenseNet [7], and SE-ResNet [6]. We select 3D ResNet-50 as the baseline due to its strongest overall discrimination. Relative to this baseline, the concept bottleneck preserves performance at a comparable level (AUC 0.71 ± 0.12 vs. 0.69 ± 0.12; F1 0.52 ± 0.15 vs. 0.50 ± 0.16). CBM attains higher ACC/SPE point estimates (0.73 ± 0.08/0.78 ± 0.09), while the baseline is more sensitive (0.64 ± 0.19 vs. 0.55 ± 0.20), reflecting different operating trade-offs.

Faithfulness Results As shown in Table 2, top-m interventions consistently outperformed random-m controls; for example, at m=5 we observed a much higher $\mathrm{PCR}_{0.05}$ (0.695 vs. 0.006) and a larger drop in AUC (−0.378 vs. −0.166).

Reliability Results We examined whether concept evidence was stable at the case level and whether stability could inform quality control. Using $T = 20$ TTA trials, we computed U_{mean} as a stability score for each case. On the test set, $U_{\mathrm{mean}} = 0.992 \pm 0.494 \times 10^{-3}$ (P10/P50/P90: $(0.505, 0.844, 1.607) \times 10^{-3}$), indicating generally stable concept evidence with a tail of higher-variability cases. Table 3 reports coverage–performance under reliability-based subset selection (Sec. 2.5). Higher-stability subsets achieved better AUC/F1 overall, supporting U_{mean} as a practical quality-control signal: low-U_{mean} cases support more reliable concept evidence, whereas high-U_{mean} cases can be flagged for review.

Table 3. Coverage–performance under reliability-based subset selection.

Coverage (%)	AUC	ACC	F1	SEN	SPE
50	**0.743**	**0.729**	**0.581**	**0.600**	**0.788**
70	0.696	0.716	0.558	0.600	0.766
90	0.684	0.686	0.471	0.522	0.746
100	0.687	0.684	0.464	0.520	0.743

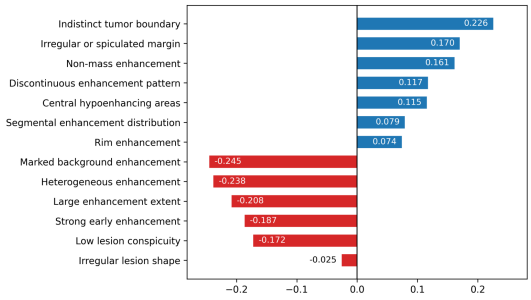

Fig. 2. Global concept weights of the linear head.

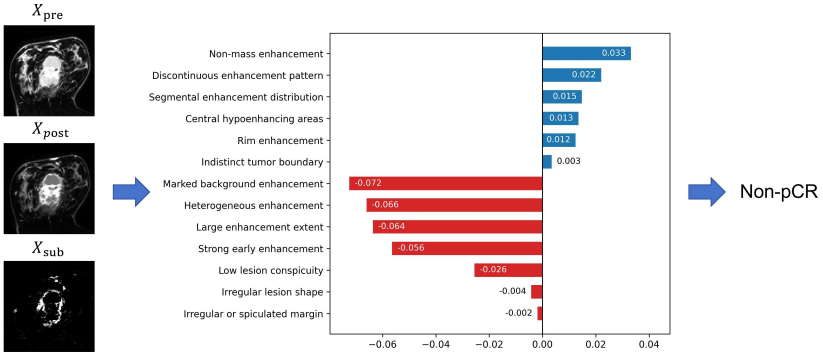

Fig. 3. Case-level concept evidence.

Interpretability Visualization Fig. 2 visualizes the global concept-head weights w in $\ell = w^\top c + b$: the sign indicates whether a concept pushes the logit toward pCR or non-pCR, while the magnitude reflects its overall influence. The weights suggest the classifier relies mainly on enhancement-related BI-RADS phenotypes (e.g., early enhancement strength, heterogeneity, extent, and background enhancement), with morphology/distribution cues as complementary signals. Fig. 3 explains a prediction via signed concept evidence $a_{i,k} = w_k c_{i,k}$ (supporting vs. opposing), yielding a compact, radiologist-readable rationale. For the shown non-pCR case, evidence is dominated by enhancement-related concepts and can be qualitatively cross-checked on $X_{\text{post}}/X_{\text{sub}}$.

4 Conclusion

We proposed a projection-based, testable concept bottleneck for pretreatment pCR prediction from early-phase breast DCE-MRI. The approach preserves baseline discrimination while producing intervenable, radiologist-readable concept evidence. Faithfulness is supported by larger output shifts under top-concept interventions than random controls, and reliability is quantified via TTA-based concept stability for coverage–performance quality control. Current limitations include the lack of expert annotations, external validation, and interpretability baselines, which will be addressed through multi-center validation and clinician-guided refinement.

Acknowledgements T.Z. is Editorial Board member of *European Radiology Experimental*, Editorial Board member of *BMC Medicine*, and Trainee Editorial Board member of *Radiology: Artificial Intelligence*. This work was supported by the Shenzhen Science and Technology Program (Grant Nos. GXWD20231129101652001 and SYSPG20241211173609005), and the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2026A1515012130).

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

References

1. D'Orsi, C., Bassett, L., Feig, S.: Breast imaging reporting and data system (bi-rads). Oxford University Press, New York (2018)
2. Esserman, L.J., Berry, D.A., DeMichele, A., Carey, L., Davis, S.E., Buxton, M., Hudis, C., Gray, J.W., Perou, C., Yau, C., et al.: Pathologic complete response predicts recurrence-free survival more effectively by cancer subset: results from the i-spy 1 trial—calgb 150007/150012, acrin 6657. Journal of Clinical Oncology **30**(26), 3242–3249 (2012)
3. Garrucho, L., Kushibar, K., Reidel, C.A., Joshi, S., Osuala, R., Tsirikoglou, A., Bobowicz, M., Del Riego, J., Catanese, A., Gwoździewicz, K., et al.: A large-scale multicenter breast cancer dce-mri benchmark dataset with expert segmentations. Scientific data **12**(1), 453 (2025)
4. Geifman, Y., El-Yaniv, R.: Selective classification for deep neural networks. Advances in neural information processing systems **30** (2017)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
6. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
7. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)

8. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
9. Janíčková, I., Tan, Y.Y., Helbich, T.H., Miloserdov, K., Bago-Horvath, Z., Heber, U., Langs, G.: Temporal representation learning of phenotype trajectories for pcr prediction in breast cancer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 606–615. Springer (2025)
10. Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept bottleneck models. In: International conference on machine learning. pp. 5338–5348. PMLR (2020)
11. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
12. Lin, W., Zhao, Z., Zhang, X., Wu, C., Zhang, Y., Wang, Y., Xie, W.: Pmc-clip: Contrastive language-image pre-training using biomedical documents. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 525–536. Springer (2023)
13. 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)
14. Oikarinen, T., Das, S., Nguyen, L.M., Weng, T.W.: Label-free concept bottleneck models. arXiv preprint arXiv:2304.06129 (2023)
15. Reig, B., Lewin, A.A., Du, L., Heacock, L., Toth, H.K., Heller, S.L., Gao, Y., Moy, L.: Breast mri for evaluation of response to neoadjuvant therapy. Radiographics **41**(3), 665–679 (2021)
16. Rugo, H.S., Olopade, O.I., DeMichele, A., Yau, C., van't Veer, L.J., Buxton, M.B., Hogarth, M., Hylton, N.M., Paoloni, M., Perlmutter, J., et al.: Adaptive randomization of veliparib–carboplatin treatment in breast cancer. New England Journal of Medicine **375**(1), 23–34 (2016)
17. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
18. Spak, D.A., Plaxco, J., Santiago, L., Dryden, M., Dogan, B.: Bi-rads® fifth edition: A summary of changes. Diagnostic and interventional imaging **98**(3), 179–190 (2017)
19. Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T.: Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. Neurocomputing **338**, 34–45 (2019)
20. Youden, W.J.: Index for rating diagnostic tests. Cancer **3**(1), 32–35 (1950)
21. Yuksekgonul, M., Wang, M., Zou, J.: Post-hoc concept bottleneck models. arXiv preprint arXiv:2205.15480 (2022)
22. 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)