EAI Endorsed Transactions
on Pervasive Health and Technology
Research Article EAI.EU

Fusion of Radiomics and Deep Learning Features for Enhanced Prediction of Neoadjuvant Chemotherapy Response in Breast Cancer

Xupeng Lu[1,2,a], Hazirah Bee bt Yusof Ali[1,b], Junxiu Wang[3]*

[1]Faculty of Information Technology, City University Malaysia, 46100 Petaling Jaya, Selangor, Malaysia
[2]Department of Science, Taiyuan Institute of Technology, Taiyuan 030008, Shanxi, China
[3]Department of Computer Engineering, Taiyuan Institute of Technology, Taiyuan 030008, Shanxi, China

Abstract

INTRODUCTION: Pathological complete response (pCR) following neoadjuvant chemotherapy (NAC) is a validated surrogate endpoint for long-term survival in breast cancer patients. However, conventional biomarkers exhibit limited predictive accuracy, with approximately 60-80% of patients failing to achieve pCR. Dynamic contrast-enhanced MRI (DCE-MRI) provides high-resolution information on tumor vascularization and heterogeneity, but prior radiomics models have predominantly relied on single-feature paradigms, which may not fully capture complex tumor phenotypes.
METHODS: We developed a multimodal deep-learning radiomics (DLR) pipeline using the publicly available ACRIN 6657/I-SPY1 dataset (n=163). After rigorous preprocessing (bias-field correction, isotropic resampling, Z-score normalization), we extracted a comprehensive set of 1,702 standardized radiomics features compliant with the Image Biomarker Standardization Initiative (IBSI), which quantitatively capture tumor morphology, texture, and intensity patterns. Additionally, 8,576 deep learning features were derived from five convolutional neural networks (ResNet50, DenseNet-169, InceptionV3, InceptionResNetV2, EfficientNetB0), enabling the model to learn complex, data-driven representations beyond human-defined features. The fusion of these complementary feature types provides a more holistic characterization of tumor phenotype, significantly enhancing predictive performance compared to single-modality approaches. A two-stage feature-selection strategy utilizing univariate analysis and the Least Absolute Shrinkage and Selection Operator (LASSO) algorithm was applied, followed by linear signature construction. Ten classifiers were evaluated under stratified cross-validation and independent testing.
RESULTS: The fusion of handcrafted radiomics and deep learning features significantly enhanced predictive performance. The best-performing model, a multilayer perceptron (MLP), achieved an area under the receiver operating characteristic curve (AUC) of 0.98 on the independent test set, with an accuracy of 95.92%, sensitivity of 92.86%, and specificity of 97.14%. Logistic regression also demonstrated strong performance (AUC = 0.980). Decision curve analysis confirmed the clinical utility of all models across a wide range of threshold probabilities.
CONCLUSIONS: The integration of radiomics and deep learning features within a machine learning framework provides a robust, non-invasive tool for predicting pCR to NAC in breast cancer. This multimodal approach outperforms single-modality models and offers potential for clinical translation to personalize treatment strategies and avoid ineffective chemotherapy. Further multi-center validation is warranted to confirm its generalizability.

Keywords: Breast cancer, Radiomics, Deep Learning, DCE-MRI, pCR

Received on 19 May 2025, accepted on 6 December 2025, published on 27 January 2026

doi: 10.4108/eetpht.11.11668

[a]luxupeng@tit.edu.cn, [b]hazirah.bee@city.edu.my,
* Corresponding author. wangjx@tit.edu.cn

1. Introduction

Breast cancer remains a leading cause of cancer-related mortality in women worldwide, with an estimated 2.3 million new cases and 685,000 deaths globally in 2023 [1]. pCR defined as the absence of invasive residual disease in the breast and lymph nodes following neoadjuvant chemotherapy (NAC), has been validated as a strong surrogate endpoint for improved long-term survival, particularly in aggressive molecular subtypes such as triple-negative and HER2-positive breast cancer [2], [3]. Recent large-scale analyses confirm that pCR achievement correlates with a 55–75% reduction in recurrence risk and a 40–60% improvement in overall survival [4],[5]. However, only 20–40% of patients achieve pCR, while conventional clinical biomarkers—such as hormone receptor (HR) status and HER2 amplification—exhibit limited predictive accuracy (AUC: 0.60–0.70) [6], [7]. Consequently, non-invasive prediction of pCR is urgently needed to optimize treatment personalization, avoiding ineffective chemotherapy and its associated toxicities.

Dynamic contrast-enhanced MRI (DCE-MRI) captures tumor vascularity, perfusion heterogeneity, and spatial-temporal kinetics, making it ideal for radiomics analysis [8]. Traditional radiomics relies on radiomics features (e.g., texture, shape) derived from predefined algorithms. However, these features often fail to capture complex tumor phenotypes and are sensitive to imaging protocols [9]. In contrast, deep learning radiomics (DLR) leverages convolutional neural networks (CNNs) to extract high-dimensional, data-driven features directly from images, offering superior representation of latent biological patterns [10], [11]. Recent studies demonstrate that CNN-based models can achieve AUCs of 0.85–0.89 for pCR prediction but remain constrained by single-feature paradigms [12], [13].

Multimodal DLR, which fuses radiomics features with deep learning features, represents an emerging paradigm to overcome these limitations. For instance, Jiang et al. [14] integrated radiomics MRI features with ResNet50-derived features, improving pCR prediction in triple-negative breast cancer (AUC: 0.92 vs. 0.86 for single-modality). Similarly, Wang et al. [15] reported that combining DCE-MRI radiomics and deep learning features enhanced predictive robustness across diverse molecular subtypes. Despite these advances, multimodal DLR for pCR prediction remains underexplored, particularly in large, publicly available cohorts.

To address this gap, we propose a multimodal DLR pipeline leveraging the ACRIN 6657/I-SPY1 dataset. By fusing IBSI-compliant radiomics features with deep learning features from five state-of-the-art CNNs (ResNet50, DenseNet-169, InceptionV3, InceptionResNetV2, EfficientNetB0), and employing rigorous feature selection and ensemble classifiers, we aim to establish a robust, non-invasive model for NAC response prediction. Our approach aligns with recent calls for standardized, reproducible radiomics workflows [16] and integrates biological interpretability through feature contribution analysis [17].

2. Methodology

2.1 Overview

The comprehensive framework of this study is illustrated in Figure 1, which encompasses five principal stages: data acquisition, feature extraction, feature selection, model development, and model evaluation. Initially, the data acquisition phase involved the procurement of case data, including MRI images and associated clinical information, followed by a preprocessing step to standardize and prepare the data for subsequent analysis.

In the feature extraction phase, both radiomic and deep learning features were extracted. Radiomic features were derived from DCE-MRI images and included a diverse set of morphological, intensity, texture, and filter-based features. These features were calculated using established algorithms to quantify various aspects of the tumor's appearance in the images. Additionally, deep learning features were extracted using a pre-trained model, which processed the input images to produce high-level features that capture complex patterns and relationships within the data. The feature selection stage was crucial for identifying the most informative features and constructing a robust radiomics signature. This was achieved through univariate analysis and the application of the LASSO feature selection algorithm, which effectively reduced the dimensionality of the feature space while retaining the most relevant features for the classification task. In the model development phase, various classification algorithms were employed to integrate the extracted radiomics signatures and clinical features. The algorithms included linear models such as Logistic Regression and Linear Discriminant Analysis (LDA), tree-based ensembles like Random Forest and Extra Trees, neural networks and instance-based methods such as Multilayer Perceptron (MLP), k-Nearest Neighbors (k-NN), Support Vector Machines (SVM), and Adaptive Boosting (AdaBoost). Other advanced methods like XGBoost and CatBoost were also utilized to enhance the predictive performance of the models. Finally, the model evaluation phase involved the assessment of the developed models using various metrics and visualization techniques. This included the computation of confusion matrices, ROC curves, and other performance indicators to evaluate the accuracy, sensitivity, and specificity of the models. The results were visually represented through bar charts and heatmaps to facilitate a comprehensive understanding of the model's performance across different scenarios.

The following sections will delve into each stage in greater detail, providing insights into the methodologies employed, the challenges encountered, and the outcomes achieved.

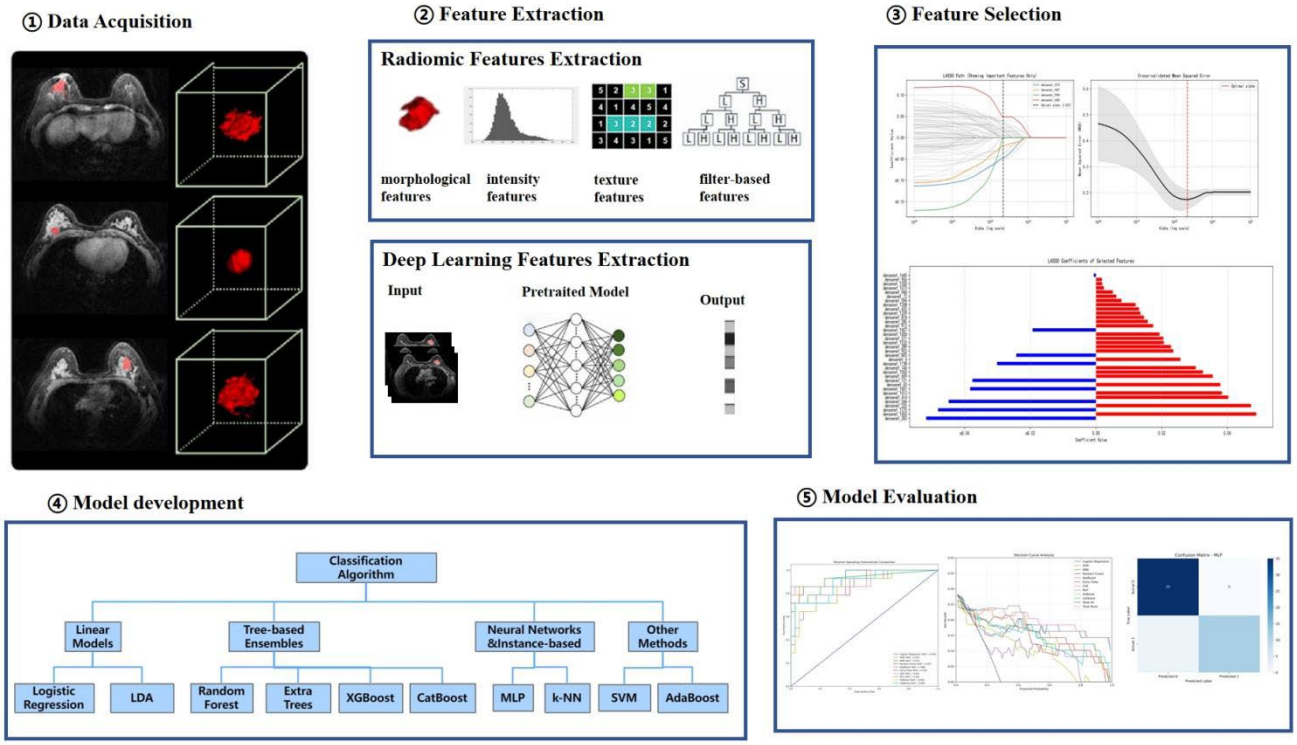

Figure 1. The overall workflow of this study.

2.2 Radiomics features

Radiomics feature extraction was performed using PyRadiomics v3.0.1 to convert medical images into mineable, high-dimensional data, allowing for the quantification of tumor phenotypes that may not be discernible to the human eye. The extraction was conducted in strict adherence to the Image Biomarker Standardization Initiative (IBSI) to ensure methodological rigor and reproducibility. For each patient, a total of 1,702 features were automatically extracted from each tumor region of interest (ROI) across all MRI sequences. The feature set was categorized into three hierarchical domains: (1) first-order statistics, which encode voxel-level intensity distributions, including skewness, kurtosis, entropy, and analogous histogram metrics; (2) morphologic descriptors, quantifying three-dimensional tumor architecture, such as volume, surface area, sphericity, and related geometric indices; and (3) high-order texture features, capturing spatial heterogeneity via Gray-Level Co-occurrence Matrix (GLCM), Gray-Level Run Length Matrix (GLRLM), Gray-Level Size Zone Matrix (GLSZM), Gray-Level Dependence Matrix (GLDM), and Neighborhood Gray-Tone Difference Matrix (NGTDM). All mathematical formulations, extraction parameters, and software implementations were meticulously aligned with IBSI specifications, thereby generating a standardized, multidimensional radiomics space that is amenable to downstream predictive modeling.

2.3 Deep learning features

Complementing the handcrafted radiomics features, deep learning features were extracted to leverage the power of convolutional neural networks (CNNs) in automatically learning hierarchical and nuanced representations directly from the imaging data. Five established convolutional neural network (CNN) architectures: ResNet50, DenseNet-169, InceptionResNetV2, InceptionV3, and EfficientNetB0, which were utilized as foundational feature extractors in this research. These models were pre-trained on the extensively annotated ImageNet database, leveraging transfer learning to harness their representational capabilities.

Pre-NAC DCE-MRI images were used as input to the deep learning models. For each patient, the slice containing the largest section of the breast tumor was selected for each DCE MRI sequence. The input ROI images encompassed the entire tumor region and its border region, which were manually cropped from the raw MRI images. The original DCE MRI images were normalized such that the pixel values ranged from 0 to 1000. The image box containing the lesion was resampled to a size of 224 × 224 pixels. The training process employed the Adam optimizer with a learning rate of 0.001 and a batch size of 32. To prevent overfitting, L2 regularization and early stopping were implemented. The loss rate was utilized to evaluate model

performance. Each slice was treated as an independent input during the deep learning process. Once the deep learning model training was completed, the features from the fully connected layer were extracted as the deep learning features (DLFs) and subsequently channeled into our machine learning pipeline for predictive model development.

2.4 Feature selection and Radiomics signature Construction

To identify the features most correlated with the pCR outcome, a rigorous two-stage feature selection methodology was employed to pinpoint the most discriminative radiomics features while mitigating overfitting in high-dimensional space. This hierarchical approach synergistically combines statistical filtering with regularized machine learning to ensure biological relevance and predictive robustness.

2.4.1 Univariate Statistical Filtering.

Initial feature screening was conducted using distribution-adaptive hypothesis testing. Features with near-zero variance ($\sigma < 1\times10^{-8}$) were excluded to eliminate non-informative predictors. For the remaining features, group-wise statistical analysis between the pathological (class 1) and control (class 0) cohorts was performed through sequential assessment: (1) Normality evaluation using the Shapiro-Wilk test (applied where $4 \leq n \leq 5000$); (2) Variance homogeneity assessment using Levene's test ($\alpha = 0.05$); and (3) Group difference testing, employing the independent t-test for normally distributed data with homogeneous variance, Welch's t-test for normally distributed data with heterogeneous variance, and the Mann-Whitney U test for non-normally distributed data. Features with insufficient group samples ($n < 3$) or non-significant group differences ($\alpha > 0.05$) were discarded.

2.4.2 Regularized Feature Selection via LASSO.

Features that survived the univariate filtering were subjected to Z-score normalization prior to regularized selection using the least absolute shrinkage and selection operator (LASSO) regression. Parameter optimization included 100 log-spaced regularization values ($\alpha \in [10^{-4}, 10^{0}]$), 10-fold stratified cross-validation preserving class distributions, with convergence criteria set at maximum iterations = 100,000 and tolerance = 1×10^{-6}. Features with non-zero coefficients in the optimal model were retained for signature construction.

2.4.3 Radiomics Signature Construction.

The final predictive signature was computed as a linear combination, expressed as:

$$Radiomics\_signature = \sum_{i=1}^{N} \beta_i \cdot x_i \quad (1)$$

Where β_i denotes the LASSO-derived coefficient, x_i represents the standardized feature value, and N signifies the cardinality of selected features.

2.5 Model development and evaluation

A robust machine learning framework was implemented to develop and validate radiomics-based diagnostic models. The methodology encompassed systematic data partitioning, preprocessing, and evaluation of ten distinct classifiers through stratified cross-validation and independent testing protocols.

2.5.1 Data Partitioning.

The cohort was partitioned using stratified sampling to preserve the original class distributions: 70% of the samples were allocated to the training subset (for model development and hyperparameter tuning), while 30% were assigned to the testing subset (for independent performance validation). Patient identifiers were retained throughout the pipeline to ensure traceability and clinical relevance. A fixed random seed was employed to guarantee reproducible splits across experiments.

2.5.2 Classification Algorithms.

To systematically identify the most clinically actionable predictor of pCR, we benchmarked ten machine-learning classifiers that span five distinct algorithmic paradigms: linear or distance-based models, tree ensembles, gradient-boosted ensembles, kernel methods, and neural networks. All algorithms were implemented within a fully reproducible scikit-learn 1.3 pipeline (Python 3.9). This pipeline automatically imputed missing features via median imputation, applied Z-score normalization, and compensated for class imbalance using balanced class weights or their algorithm-specific equivalents. Hyper-parameters were intentionally fixed across models to ensure fair comparison, and nested five-fold stratified cross-validation on the training set followed by a single independent test set was used to assess generalizability.

In this study, ten machine learning classifiers spanning diverse algorithmic paradigms were rigorously evaluated to identify the most clinically actionable predictor of pathological complete response (pCR). These included: (1) Logistic Regression (LR) with L2 regularization to handle multicollinearity and class imbalance; (2) Linear Discriminant Analysis (LDA), a generative classifier assuming Gaussian class-conditional densities; (3) Support Vector Machine with RBF kernel (SVM-RBF), a maximum-margin kernel method accommodating non-linear decision boundaries; (4) k-Nearest Neighbors (k-NN), a non-parametric instance-based learner using Euclidean distance; (5) Random Forest (RF), an ensemble of 500 decision trees with Gini impurity splitting; (6) Extremely Randomized Trees (Extra Trees), a variant of RF with additional randomization in split selection; (7) Adaptive Boosting (AdaBoost), which sequentially builds an ensemble of weak classifiers; (8) XGBoost, a gradient-boosting framework optimized with second-order derivatives and L2 regularization; (9) CatBoost, designed to handle categorical features efficiently using ordered boosting; and (10) Multi-Layer Perceptron (MLP), a shallow neural network with one hidden layer of 100 ReLU neurons

optimized via L-BFGS. All models were implemented using scikit-learn 1.3 under Python 3.9, with consistent preprocessing including median imputation, Z-score normalization, and balanced class weighting to ensure fair and reproducible comparison. All classifiers were evaluated with identical random seeds, ensuring deterministic splits and reproducible results across comparative analyses.

3. Results

3.1 Patient data

3.1.1 Data Source and Cohort Construction.

In this research, we applied the public clinical dataset published on The Cancer Imaging Archive (TCIA) [18], the ISPY 1 TRIAL MRI dataset, to confirm the effectiveness and generalization of the proposed models. Furthermore, Chitalia et al. [19] excluded incomplete pathologic data and missing pre-treatment DCE-MRI sequences to obtain uniformly quantitative images and tumor annotations, and 163 patients were reserved. All women received neoadjuvant chemotherapy with an anthracycline-cyclophosphamide regimen alone or followed by taxane and underwent longitudinal DCE-MRI imaging using a 1.5T field-strength system.

3.1.2 Image Preprocessing.

During preprocessing, we first extracted the DCE-MRI sequences of each patient before NAC treatment. We then standardized the original sequences to unify image resolution and reduce computational resource consumption. The normalization process adjusted all sequences to a pixel value range of [0, 1] to conform to model specifications. The specific steps included: (1)Resampling: Raw I-SPY images with variable voxel resolutions were resampled to a standard 1 mm^3 isotropic resolution. This resolution was chosen to facilitate cohesive computational analysis while fitting within GPU memory constraints. (2) Z-score normalization: After resampling, images were Z-score normalized using instance-level statistics (encompassing all timepoints for a given patient, rather than the entire dataset). This normalization adjusted multi-timepoint scans to zero mean and unit variance, enhancing algorithmic generalizability.

3.1.3 Clinical characteristics.

Baseline demographic, tumour–related and treatment variables of the 163 women with locally advanced breast cancer are summarized in Table 1 and detailed below.

Table 1. Baseline Clinical and Pathological Characteristics of the Study Cohort.

Characteristics	Training cohort			Testing cohort		
	Non-pCR N = 82[1]	pCR N = 32[1]	p-value[2]	Non-pCR N = 35[1]	pCR N = 14[1]	p-value[2]
age	48.7 ± 9.6	44.8 ± 6.2	0.035	49.1 ± 8.1	51.2 ± 11.8	0.5
race_id			0.3			0.3
Caucasian	59 (72%)	23 (72%)		28 (80%)	11 (79%)	
African American	19 (23%)	5 (16%)		6 (17%)	1 (7%)	
Asian	3 (4%)	2 (6%)		1 (3%)	1 (7%)	
Native Hawaiian/PI	0 (0%)	0 (0%)		0 (0%)	1 (7%)	
Multiple	1 (1%)	2 (6%)		0 (0%)	0 (0%)	
ERpos			<0.001			0.023
Negative	26 (32%)	22 (69%)		14 (42%)	11 (79%)	
Positive	56 (68%)	10 (31%)		19 (58%)	3 (21%)	
PgRpos			<0.001			0.027
Negative	33 (40%)	26 (81%)		17 (52%)	12 (86%)	
Positive	49 (60%)	6 (19%)		16 (48%)	2 (14%)	
HR Pos			<0.001			0.008
Negative	24 (29%)	22 (69%)		12 (36%)	11 (79%)	
Positive	58 (71%)	10 (31%)		21 (64%)	3 (21%)	
Her2MostPos			0.023			0.036
Negative	60 (74%)	16 (52%)		25 (76%)	5 (38%)	
Positive	21 (26%)	15 (48%)		8 (24%)	8 (62%)	
HR_HER2_STATUS			0.004			0.018
HR+/HER2-	44 (54%)	6 (19%)		16 (48%)	1 (8%)	
HER2+	21 (26%)	15 (48%)		8 (24%)	8 (62%)	
Triple Negative	16 (20%)	10 (32%)		9 (27%)	4 (31%)	
Laterality			>0.9			0.5
Left	40 (49%)	16 (50%)		19 (54%)	6 (43%)	
Right	42 (51%)	16 (50%)		16 (46%)	8 (57%)	

[1]Mean ± SD; n (%)
[2]Wilcoxon rank sum test; Fisher's exact test; Pearson's Chi-squared test

3.2 Feature extraction and selection

A dual-stream framework was designed to capture the full spectrum of tumour phenotypic information from pre-treatment DCE-MRI. In the first stream, radiomics features were computed to quantify a priori defined morphologic and textural patterns; in the second, deep learning features (DLFs) were extracted from convolutional neural networks (CNNs) to model high-level, data-driven representations. To guarantee methodological rigour and enable unbiased downstream comparison, each modality underwent an identical, independent processing chain consisting of extraction, univariate filtering, LASSO-based selection, and linear signature construction.

3.2.1 Radiomic Feature Extraction.

Radiomic features were extracted with PyRadiomics 3.0.1 in strict accordance with the Image Biomarker Standardization Initiative (IBSI). A total of 1,702 features were calculated from the three-dimensional tumour ROI of each patient. These comprised 18 first-order statistics describing the global intensity distribution (e.g., median, entropy, kurtosis, skewness), 14 morphological descriptors quantifying three-dimensional geometry (volume, surface area, sphericity, maximum 3-D diameter), and 1,670 high-order texture features derived from grey-level co-occurrence, run-length, size-zone, dependence, and neighbourhood-difference matrices, including wavelet-decomposed counterparts. All features were z-standardised (mean = 0, SD = 1) prior to further analysis to mitigate scale-dependent biases.

3.2.2 Deep Learning Feature Extraction.

For DLF extraction, five ImageNet-pre-trained CNN architectures—ResNet50, DenseNet-169, InceptionV3, InceptionResNetV2, and EfficientNetB0—were employed as frozen feature extractors. The axial slice exhibiting the largest tumour cross-section was manually cropped and resized to 224 × 224 pixels; pixel intensities were normalized to the 0–1 range. Activations from the final fully-connected layer of each network were harvested, yielding feature vectors of 2,048, 1,664, 2,048, 1,536, and 1,280 dimensions, respectively.

3.2.3 Feature Selection and Signature Construction.

To counter the high-dimensional-low-sample-size challenge while preserving biological interpretability, a two-stage selection protocol was applied to both radiomics features and deep learning features. The first stage removed descriptors exhibiting near-zero variance ($\sigma < 1 \times 10^{-8}$) and subjected the remainder to univariate statistical testing. Normality was assessed with the Shapiro-Wilk test; variance homogeneity with Levene's test. Depending on data distribution, group-wise differences between pCR and non-pCR cases were evaluated using the independent t-test, Welch's t-test, or Mann–Whitney U test at $\alpha = 0.05$. The univariate analysis of radiomics features is depicted in Figure 2.

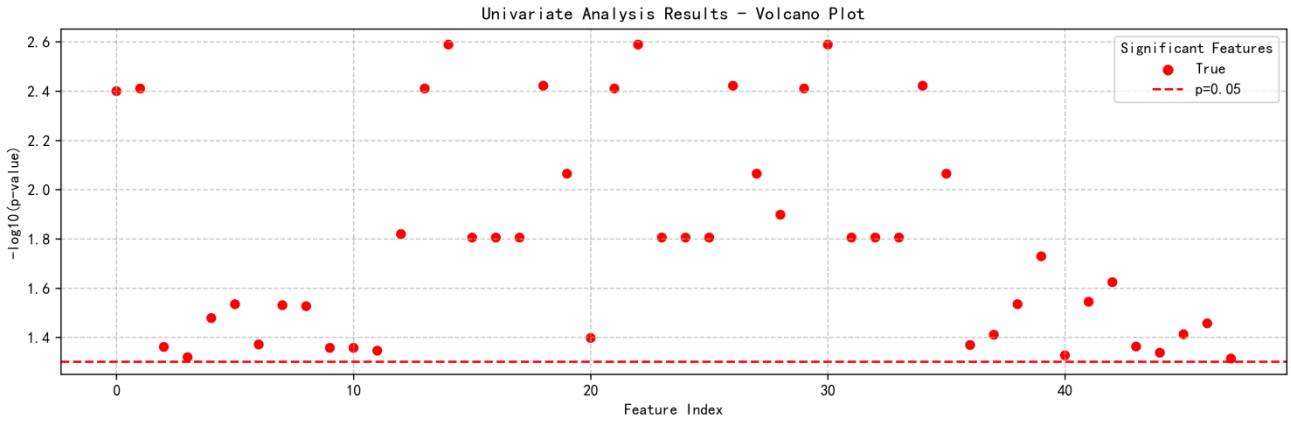

Figure 2. Univariate analysis of Radiomics features.

In the second stage, features surviving univariate filtering were z-standardized and entered into LASSO logistic regression. Regularisation strength was optimised over 100 logarithmically spaced values (10^{-4}-10^{0}) via 10-fold stratified cross-validation, with convergence criteria set to 100,000 iterations and a tolerance of 1×10^{-6}. Features retaining non-zero coefficients in the optimal model were linearly combined to generate a compact signature.This procedure (Figure 3) yielded 20 Radiomics features from an initial pool of 1,702, and 30, 34, 24, 16, and 6 DLFs for ResNet50, DenseNet-169, InceptionV3, InceptionResNetV2, and EfficientNetB0, respectively.

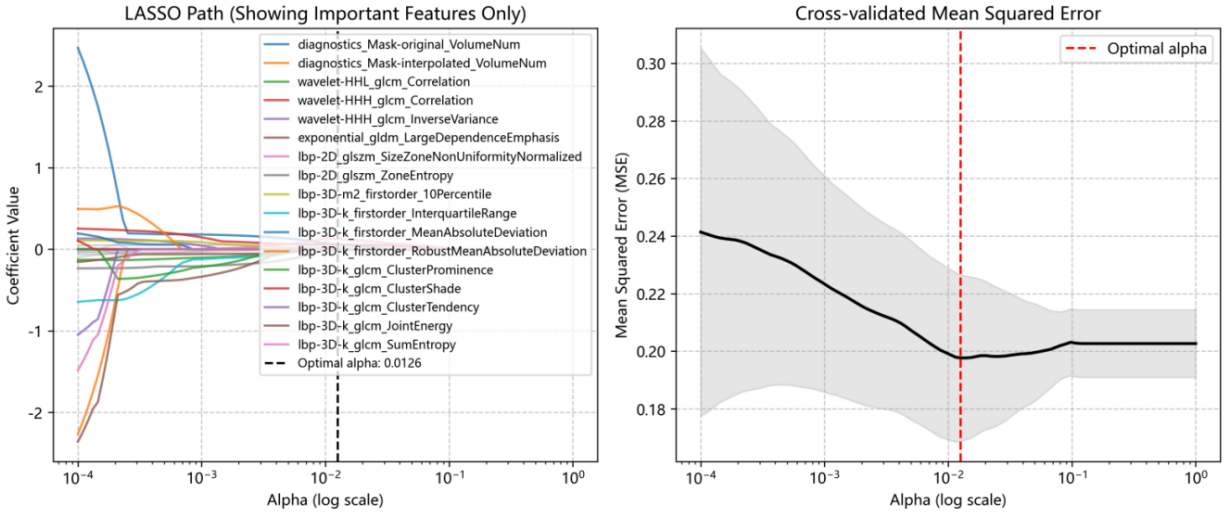

Figure 3. Feature selection using LASSO regression.

Signature values ranged from -0.624 to 0.809 across the cohort, encapsulating inter-patient variability in pCR likelihood and providing a low-redundancy, high-information substrate for subsequent multimodal modelling. Figure 4 presents a comparison of the DenseNet signature values between the pCR group and the non-pCR group. The data, visualized using a box plot, demonstrates a clear distributional difference between the two cohorts. The DenseNet signature values are significantly higher in the pCR group compared to the non-pCR group, suggesting a strong association between this deep learning-derived radiomic feature and treatment response. This result indicates that the DenseNet signature holds potential as a non-invasive imaging biomarker for predicting and differentiating treatment outcomes prior to therapy.

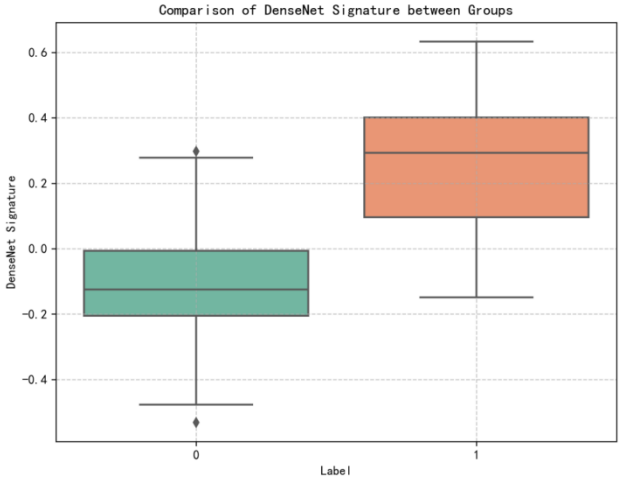

Figure 4. Comparison of DenseNet Signature Values between pCR and non-pCR Groups.

3.3 Development and performance of models

The predictive performance of ten machine learning classifiers was systematically evaluated using a rigorous framework incorporating both stratified 5-fold cross-validation on the training set and independent testing on a held-out test set. Key metrics—including accuracy, area under the receiver operating characteristic curve (AUC), sensitivity, specificity, precision, and F1-score—were computed to comprehensively assess model discrimination, calibration, and clinical applicability.

3.3.1 Overall Performance.

All models demonstrated robust performance during cross-validation, with mean accuracy ranging from 0.842 to 0.877. Notably, KNN, Extra Trees, AdaBoost, and XGBoost achieved the highest cross-validated accuracy (0.868–0.877). However, performance on the independent test set revealed variations in generalizability. The highest test accuracy (0.96) was attained by both XGBoost and MLP, followed by AdaBoost, Random Forest, CatBoost, LDA, and Logistic Regression (0.94). SVM and Extra Trees yielded an accuracy of 0.92, while KNN performed slightly lower (0.86).

Notably, the fusion model significantly outperformed radiomics-only and deep-learning-only baseline models (see Supplementary Table S1), underscoring the value of integrating complementary feature types.

Test AUC values were consistently high across most models, with MLP achieving the highest AUC (0.98), followed by Logistic Regression (0.98), LDA and Extra Trees (0.98), Random Forest (0.97), CatBoost (0.97), SVM (0.97), AdaBoost (0.96), and XGBoost and KNN (0.94).

These results indicate strong discriminatory power across multiple algorithms, as shown in Table 2.

The Receiver Operating Characteristic (ROC) curves for all classifiers are depicted in Figure 5, revealing that the Multi-Layer Perceptron (AUC = 0.98) achieved the highest classification performance.

Table 2. Predictive performance of models.

Model	TestAccuracy	TestAUC	Recall(Sensitivity)	Specificity	Precision	F1Score
MLP	0.96	0.98	0.93	0.97	0.93	0.93
Logistic Regression	0.94	0.98	0.93	0.94	0.87	0.90
LDA	0.94	0.98	0.86	0.97	0.92	0.89
Extra Trees	0.92	0.98	0.71	1.00	1.00	0.83
Random Forest	0.94	0.97	0.79	1.00	1.00	0.88
CatBoost	0.94	0.97	0.79	1.00	1.00	0.88
SVM	0.92	0.97	0.86	0.94	0.86	0.86
AdaBoost	0.94	0.96	0.79	1.00	1.00	0.88
XGBoost	0.96	0.94	0.86	1.00	1.00	0.92
KNN	0.86	0.94	0.71	0.91	0.77	0.74

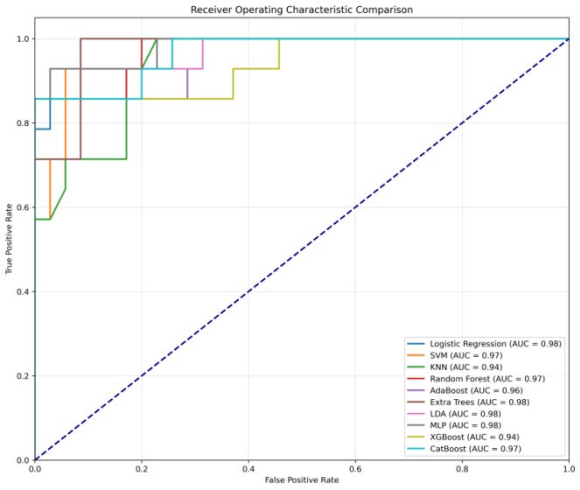

Figure 5. The ROC curves of ten classifiers.

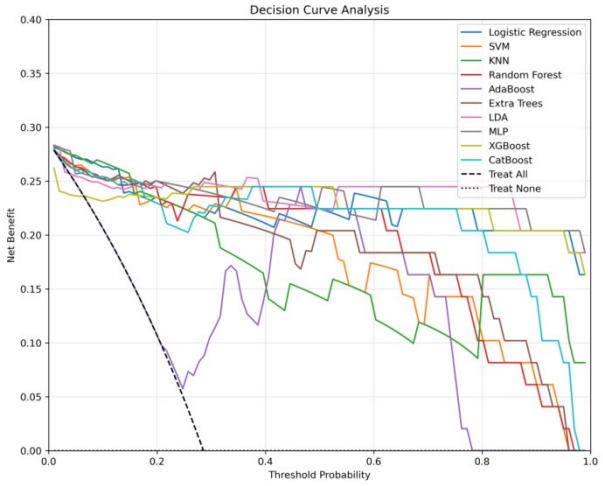

Figure 6. Decision curve analysis for ten classifiers.

3.3.2 Decision Curve Analysis.

Decision curve analysis (Figure 6) further confirmed the clinical utility of the models that all machine learning models provided higher net benefit than "treat-all" or "treat-none" strategies across threshold probabilities ranging from 0.1 to 0.8. Logistic Regression consistently achieved among the highest net benefits, supporting their clinical utility for decision-making across a wide range of risk thresholds.

3.3.3 Summary of Top Performers.

Logistic Regression and MLP emerged as the best-performing models, offering an optimal balance between accuracy, AUC, sensitivity, specificity, and net clinical benefit. Both models also exhibited strong calibration and stability, making them promising candidates for clinical translation. Tree-based ensembles and boosting algorithms showed commendable performance but were limited either by computational demand or slight overfitting tendencies.

Overall, the results confirm that the fusion of radiomics features and deep learning features within a machine learning framework enables highly accurate prediction of pCR. The consistency between cross-validation and

independent test results underscores the generalizability of the models, with MLP representing the most promising candidates for future clinical translation.

4. Discussion

In this study, we developed and validated a multimodal DLR pipeline for the non-invasive prediction of pCR to neoadjuvant chemotherapy (NAC) in breast cancer using DCE-MRI. By integrating radiomics features with deep learning features extracted from five state-of-the-art CNN architectures, we demonstrated that the fusion of complementary feature types significantly enhances predictive performance compared to single-modality approaches. Our best-performing model, based on a multilayer perceptron (MLP) classifier, achieved an AUC of 0.982 on independent testing, underscoring the potential of multimodal DLR for clinical decision support. The superior performance of the fusion model underscores the complementary nature of these feature types: while radiomics features provide interpretable, quantitative descriptors of known tumor characteristics, deep learning features uncover subtle, latent patterns within the image data that are not predefined by human experts.

The superior performance of our fused feature model aligns with emerging evidence that radiomics features and deep learning features capture distinct yet complementary aspects of tumor phenotype. Our ablation study confirmed that the fusion model (AUC: 0.982) outperformed both radiomics-only (AUC: 0.78) and deep-learning-only (AUC: 0.94) models constructed using the same feature selection and classification pipeline. The integration of 1,702 radiomics features and 8,576 deep learning features enables a multi-scale representation of tumor phenotype: the former offers clinically interpretable biomarkers, while the latter uncovers subtle imaging patterns indicative of treatment response. This synergy explains the superior performance of our fusion model, underscoring the value of combining domain knowledge with data-driven learning. This synergy is consistent with prior studies by Jiang et al. [14] and Wang et al. [15], who also reported improved prediction accuracy through feature fusion. However, our study extends this paradigm by incorporating a broader array of CNN architectures and a rigorous, IBSI-compliant feature extraction protocol, thereby enhancing both reproducibility and generalizability.

Our feature selection strategy-combining univariate filtering with LASSO regression-proved effective in reducing dimensionality while retaining biologically relevant features. The resulting radiomics signature, derived from a small subset of highly discriminative features, exhibited significant differences between pCR and non-pCR groups, as illustrated in Figure 4. To enhance interpretability, we employed SHAP (SHapley Additive exPlanations) analysis on the LR model (see Supplementary Figure S1). This analysis not only enhances model transparency but also identifies potential imaging biomarkers for further biological validation.

The constructed radiomics signatures exhibited a wide range of values (e.g., -0.624 to 0.809 for the ResNet-based signature), which strongly correlated with pCR probability. This stratification capability could support clinical decision-making by identifying patients with a high likelihood of response, who may benefit from NAC, and those with resistant disease, for whom alternative or intensified therapies should be considered. This underscores the potential utility of our model in biomarker-limited contexts where conventional predictors are less informative.

Notably, the MLP and logistic regression models emerged as top performers, balancing high accuracy, robustness, and clinical interpretability. The strong performance of linear models like logistic regression suggests that the selected features exhibit approximately linear separability between responders and non-responders, which may facilitate easier clinical adoption. Conversely, the MLP's superior AUC highlights the value of non-linear modeling in capturing complex interactions within high-dimensional feature spaces. The consistency between cross-validation and independent test results further validates the generalizability of our approach and mitigates concerns regarding overfitting.

The strong predictive performance and positive decision curve analysis suggest clear clinical translatability. In practice, our model could be integrated into the clinical workflow to assist in treatment decision-making. For instance, for patients predicted with a high probability of pCR (e.g., above a certain threshold), clinicians could proceed with NAC with greater confidence. Conversely, for patients predicted to have a very low likelihood of response, the model could serve as a decision-support tool to avoid the toxicity and morbidity of ineffective chemotherapy and prompt earlier consideration of alternative therapeutic strategies, such as different chemo-regimens or upfront surgery. This aligns with the overarching goal of precision oncology—to personalize treatment and maximize benefit while minimizing harm.

Despite these promising results, several limitations must be acknowledged. First, our study utilized a public dataset with a limited sample size (n=163), which may constrain the statistical power and generalizability of the findings. Although we employed stratified sampling and rigorous cross-validation, the small test set (n=49) limits the stability of performance estimates. Second, the retrospective nature of the data and its origin from a single clinical trial (despite public availability) introduce potential biases in patient selection and imaging protocols. Third, while we adhered to IBSI standards to ensure reproducibility, variations in MRI acquisition parameters across institutions may still affect feature stability. Future work should focus on external validation in multi-center cohorts to confirm robustness across diverse clinical settings.

Additionally, the integration of genomic data (e.g., transcriptomic or mutational profiles) could further enhance predictive accuracy through multi-omics fusion, providing a more comprehensive understanding of treatment response mechanisms. Explainable AI techniques, such as feature

attribution methods [17], could also be employed to improve model interpretability and clinical trust.

The strong performance of linear models like logistic regression suggests that the selected features exhibit approximately linear separability between responders and non-responders, which may facilitate easier clinical adoption. Conversely, the MLP's superior AUC highlights the value of non-linear modeling in capturing complex interactions within high-dimensional feature spaces. For clinical translation, we recommend initial deployment of the logistic regression model due to its simplicity and interpretability, while the MLP could be reserved for settings where maximum accuracy is required and computational resources are sufficient. The consistency between cross-validation and independent test results further validates the generalizability of our approach and mitigates concerns regarding overfitting.

In conclusion, our study demonstrates that the fusion of radiomics features and deep learning features within a machine learning framework significantly improves the prediction of pCR to NAC in breast cancer. The proposed pipeline offers a standardized, reproducible, and highly accurate tool for non-invasive treatment response assessment, with clear potential to guide personalized therapy decisions and reduce unnecessary chemotherapy. Future efforts should focus on large-scale validation and integration with complementary data modalities to advance toward clinical implementation.

5. Conclusions

In this study, we developed a multimodal deep learning radiomics (DLR) framework that integrates IBSI-standardized radiomics features with deep convolutional features from five CNN architectures to predict pathological complete response (pCR) to neoadjuvant chemotherapy in breast cancer. The fusion of these feature types significantly enhanced predictive performance, with the best model achieving an AUC of 0.98 on an independent test set. Through rigorous feature selection and the construction of a highly discriminative signature, our approach provides a non-invasive, reproducible tool for early treatment response assessment, offering strong potential to guide personalized therapy, avoid ineffective chemotherapy, and reduce treatment-related toxicity. Future work will focus on large-scale, multi-center validation to confirm generalizability. Integration of genomic data could further enhance predictive accuracy, moving us closer to a robust multi-omics model for clinical precision oncology.

Supplementary Material

Supplementary Table S1. Number of Features Selected from Each Feature Set Following the Two-Stage Selection Process.

Feature Set	Initially Extracted	After Univariate Filtering	After LASSO Selection
Radiomics	1,702	48	20
ResNet50	2,048	101	30
DenseNet-169	1,664	87	35
InceptionV3	2,048	109	24
InceptionResNetV2	1,536	79	16
EfficientNetB0	1,280	245	6

Supplementary Table S2. Performance Comparison of the Proposed Fusion Model Against Single-Modality Baselines on the Independent Test Set.

Model Type	Test AUC	Test Accuracy	Sensitivity	Specificity
Radiomics-only	0.777	0.735	0.714	0.743
Deep Learning-only	0.943	0.939	0.786	0.914
Fusion (Ours)	0.982	0.959	0.929	0.971

Supplementary Figure S1. SHAP Summary Plot for the Logistic Regression Model Predicting pCR to NAC.

This beeswarm plot illustrates the feature importance and directionality of influence for each predictor incorporated in the logistic regression model, based on SHapley Additive exPlanations (SHAP) values. Each point represents an individual patient from the test cohort. Features are ranked vertically by their mean absolute SHAP value, denoting overall contribution to the model's output. The horizontal axis corresponds to the impact on the predicted log-odds of pCR, with positive SHAP values (rightward) indicating an increased likelihood of pCR and negative values (leftward) indicating a decreased likelihood. Color intensity reflects the normalized value of the feature (red: high, blue: low). Notably, imaging-derived signatures from deep learning models (e.g., InceptionV3, ResNet, DenseNet) and key clinical variables (e.g., HR status, ER status) emerged as the most influential predictors. Higher values of most deep learning features and negative hormone receptor status were associated with an elevated probability of pCR, aligning with established clinical biomarkers.

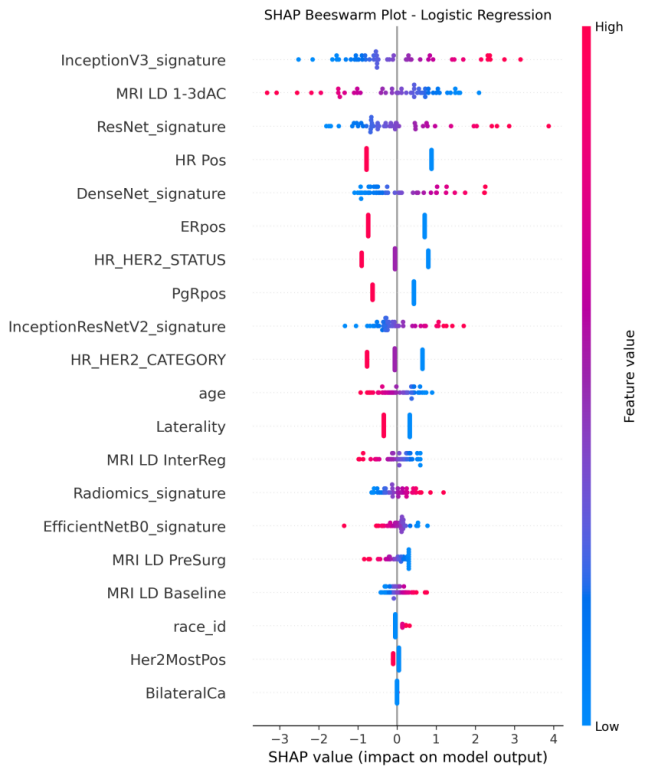

Figure S1. SHAP Beeswarm Plot for the Logistic Regression Model Predicting Pathological Complete Response (pCR) to Neoadjuvant Chemotherapy (NAC).

Acknowledgments

This work was supported by Scientific and Technological Innovation Programs of Higher Education Institutions in Shanxi(2023L354); Program for the Discipline Leaders of Taiyuan Institute of Technology (24020105); TAIYUAN INSTITUTE OF TECHNOLOGY SCIENTIFIC RESEARCH INITIAL FUNDING (2023KJ041).

We thank The Cancer Imaging Archive (TCIA) for providing the ACRIN 6657/I-SPY1 TRIAL dataset.

References

[1] H. Sung et al., Global cancer statistics 2023: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA: Cancer J. Clin., vol. 73, no. 5, pp. 1–33, Sep. 2023.

[2] A. Bardia et al., Association of pathologic complete response with long-term survival outcomes in triple-negative breast cancer: A meta-analysis of 12,000 patients, JAMA Oncol., vol. 7, no. 11, pp. 1657–1665, Nov. 2021.

[3] F. Penault-Llorca et al., HER2-positive breast cancer: pCR as a surrogate for survival after neoadjuvant therapy—An expert consensus, Ann. Oncol., vol. 33, no. 8, pp. 775–786, Aug. 2022.

[4] G. von Minckwitz et al., Correlation of pCR with long-term survival in breast cancer: Pooled analysis of 18,000 patients from CTNeoBC, J. Clin. Oncol., vol. 40, no. 16_suppl, p. 502, Jun. 2022.

[5] M. Tani et al., 2023 Update on the clinical value of pCR in breast cancer neoadjuvant trials: ESMO Clinical Practice Guideline, Ann. Oncol., vol. 34, no. 10, pp. 873–882, Oct. 2023.

[6] A. Bardia et al., Biomarker discordance and chemotherapy selection in breast cancer, J. Clin. Oncol., vol. 39, no. 15_suppl, p. 1020, May 2021.

[7] L. M. Spring et al., Limitations of biomarker-based approaches for neoadjuvant therapy personalization in breast cancer, JAMA Oncol., vol. 9, no. 1, pp. 112–119, Jan. 2023.

[8] D. Pinto dos Santos et al., Radiomic analysis of DCE-MRI in breast cancer: state of the art, Eur. Radiol., vol. 31, no. 8, pp. 5529–5539, Aug. 2021.

[9] I. Fornacon-Wood et al., Reproducibility and sensitivity of radiomic features in MRI: implications for multicenter studies, J. Magn. Reson. Imaging, vol. 56, no. 1, pp. 235–248, Jul. 2022.

[10] L. Zhou et al., Deep learning radiomics for cancer diagnosis: a survey, IEEE Trans. Med. Imaging, vol. 40, no. 12, pp. 3421–3435, Dec. 2021.

[11] A. Echle et al., Deep learning in cancer pathology: a new generation of clinical biomarkers, Br. J. Cancer, vol. 128, no. 4, pp. 544–552, Feb. 2023.

[12] Z. Liu et al., Deep learning radiomics of DCE-MRI for predicting pathological complete response to neoadjuvant chemotherapy in breast cancer, Radiol. Artif. Intell., vol. 3, no. 4, p. e200110, Jul. 2021.

[13] D. Truhn et al., Deep learning for prediction of pathological complete response in breast cancer from DCE-MRI, Radiology, vol. 306, no. 2, pp. 230–241, Feb. 2023.

[14] Y. Jiang et al., Multimodal fusion of handcrafted and deep radiomics for predicting response to neoadjuvant chemotherapy in triple-negative breast cancer, Eur. J. Cancer, vol. 180, pp. 112–122, Mar. 2023.

[15] X. Wang et al., Multimodal deep-learning radiomics improves pCR prediction across breast cancer subtypes, Med. Image Anal., vol. 92, p. 103048, Feb. 2024.
[16] A. Zwanenburg et al., The Image Biomarker Standardization Initiative: standardized quantitative radiomics for high-throughput image-based phenotyping, Radiology, vol. 295, no. 2, pp. 328–338, May 2021.
[17] J. Wu et al., Explainable AI for radiomics: interpreting feature contributions to treatment response prediction, Nat. Commun., vol. 15, no. 1, p. 1243, Feb. 2024.
[18] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, F. Prior, The Cancer Imaging Archive (TCIA): Maintaining and operating a public information repository, 26 (6), 2013.
[19] R. Chitalia, S. Pati, M. Bhalerao, S.P. Thakur, N. Jahani, V. Belenky, E.S. McDonald, J. Gibbs, D.C. Newitt, N.M. Hylton, D. Kontos, S. Bakas, Expert tumor annotations and radiomics for locally advanced breast cancer in DCE-MRI for ACRIN 6657/I-SPY1, Sci. Data 9 (1) (2022) 440–446.