GMR
The Original

MRI-BASED RADIOMICS FOR PREDICTING NEOADJUVANT CHEMOTHERAPY RESPONSE IN BREAST CANCER USING MACHINE LEARNING AND AI

Monika[1], Manisha Rani[2], Gauri[3], Sawita[4], Saranpreet Kaur[5], Burhan Bashir[6], Abdul Wajid Bhat[7]

[1,2,3,4,5,6]Department of Medical Radiology and Imaging Technology, School of Allied and Healthcare Sciences, GNA University, Phagwara, Punjab, 144401, India
Email: m70298478@gmail.com[1], kk7839819@gmail.com[2], gaurisharma@gmail.com[3], sawitaghera46@gmail.com[4], saranpreetkaur1313@gmail.com[5], burhan.bashir@gnauniversity.edu.in[6], bhattwajjid@gmail.com[7]

ABSTRACT
Current prediction methods for pathological complete response (pCR) after neoadjuvant chemotherapy (NACT) for breast cancer are limited to post-surgical histopathological assessment of the tumor and doesn't give actionable advice during the treatment, remaining a clinically unsolved challenge. Current MRI radiomic models almost exclusively use the features from a single pre-treatment MRI, ignoring the dynamic changes in morphology and textural features within the tumor microenvironment that occur over multiple cycles of chemotherapy. This paper proposes a novel framework named longitudinal Delta-Radiomic Hybrid Intelligence (DRHI) to represent a change-based radiomic signature across three serial MRI timepoints baseline (T0), post-cycle 1 (T1), and post-cycle 2 (T2) and combines it with deep convolutional features learned from a pre-trained ResNet-50 backbone. The prediction models are developed independently for three molecular subtypes: HER2-enriched, Triple-Negative (TNBC) and Hormone Receptor-positive/HER2-negative (HR+/HER2−) instead of using a single generalized model for all kinds of patients and their biological heterogeneity. To bridge this interpretability, disconnect that hinders the adoption of AI-powered oncology tools, a SHAP based explainability module is integrated into each of the subtype pipelines to explain the influential features per prediction. Experiments are conducted on the publicly available I-SPY 2 TRIAL dataset. To our knowledge, there is no previous study that involves using a unified prediction framework to combine all these four aspects, namely, longitudinal delta radiomics, hybrid handcrafted-deep feature fusion, molecular subtype stratification, and SHAP explainability. The proposed approach shows clinically relevant gain in AUC for all three subtypes on top of the existing state-of-the-art baseline methods.

KEYWORDS - delta radiomics, neoadjuvant chemotherapy, DCE-MRI, I-SPY 2, pCR prediction.

I. INTRODUCTION

In 2022 alone, there were around 2.3 million new cases of breast cancer and more than 660,000 deaths caused by the disease, globally. Breast cancer is very common cancer in women across the world [1]. Neoadjuvant chemotherapy (NACT) is one of the most important treatment strategies for locally advanced disease: it is used before surgical resection, enabling in vivo assessment of chemosensitivity and the potential for tumor downstaging which may increase the number of women eligible for breast-conserving surgery [2,3]. The absence of any residual invasive disease in the breast and axilla at surgery (pathological complete response) is well established as a surrogate marker of better event-free and overall survival, especially for the HER2 enriched and Triple-Negative subtypes [2-4]. However, the pCR rate differs substantially between molecular subtypes and between individual patients [2,3]. This variability highlights the clinical need to develop reliable, non-invasive markers that are able to predict treatment response early in the NACT cycle [5].

Although multiple decades of research, none of the clinical or pathological parameters has proven to have enough predictive accuracy to be used at the individual patient level to inform NACT decisions [5,6]. The conventional markers (tumor grade, Ki-67 proliferation index, hormone receptor status and HER2 amplification) give only general probabilistic information and cannot capture the spatial and temporal variation that is found in solid cancer [5,7]. Repeat biopsy is invasive, sampling error prone and impractical in repeated consecutive cycles of treatment. This means that oncologists have to keep treating with an ineffective treatment for multiple cycles before the treatment fails, causing the patient to be unnecessarily treated with the toxicity of chemotherapy and delaying potentially more effective treatment, if applicable [2,3]. Therefore, there is a strong need for an imaging-based, non-invasive, and time-sensitive prediction system, that can be used early (at first or second cycle of NACT) with sufficient accuracy to distinguish responders from the non-responders, and to guide real-time clinical decisions [5,8].

Radiomics, a high-throughput extraction of quantitative features from medical images [9], is a promising direction for predicting treatment responses in breast oncology, particularly by MRI [5,10]. Dynamic Contrast-Enhanced MRI (DCE-MRI), Diffusion-Weighted Imaging (DWI) and multiparametric protocols have been used in studies to show that radiomic signatures which represent tumour shape, texture and/or enhancement kinetics, correlate significantly with the pCR status [5,8,11]. In a number of single institution studies, these features have been used to train machine learning classifiers such as Random Forest, Support Vector Machines, and Gradient Boosting variants with reported area under the ROC curve (AUC) values from 0.70 to 0.85 [5,8,11]. More recently, deep learning techniques have been applied to pre-treatment DCE-MRI, which can further improve the performance, demonstrating that Multilayer Perceptron and ResNet-based architectures have better performance in some cohorts when compared with traditional radiomic classifiers [7,8]. However, there are three common drawbacks in the field: (i) the majority of models published focus solely on pre-treatment baseline images and fail to account for the vast amount of biologically relevant changes taking place during treatment [12]; (ii) most models combine all molecular subtypes into a single prediction model, whereas there is well documented variability in the relevance of radiomic features across molecular subtypes [13]; and (iii) the lack of clinical interpretability of complicated models restricts the use of these models in clinical practice within the context of oncology [14,15].

The idea of computing relative change in the radiomic features between serial imaging timepoints has been studied in a limited number of studies, and has been shown to have early promise for capturing microenvironmental remodeling associated with treatment [12,16]. To the best of our knowledge, however, there is no published study that (i) combines longitudinal delta-radiomic features at three time points of the treatment, (ii) incorporates the deep convolutional features into the architecture in a hybrid fashion, (iii) builds independent subtype-specific models for HER2-enriched, TNBC, and HR+/HER2− subtypes, and (iv) applies explainability to generate feature attribution maps per patient per subtype using SHAP [14]. All of these features have been reported on their own, and a systematic combination into a unified, interpretable pipeline is a substantive addition and hitherto unexploited by the literature. This framework is optimally testable using the I-SPY 2 TRIAL dataset which contains multiparametric longitudinal MRI data at standardized timepoints over a large multi-center cohort [17].

The key contributions of this work include: (1) a longitudinal delta-radiomic feature representation capturing the evolution of tumour change between baseline, post-cycle 1 and post-cycle 2 MRI scans [12]; (2) a dual-stream hybrid fusion architecture, where classical handcrafted features extracted using PyRadiomics [18] and deep features extracted from a pretrained ResNet-50 network are simultaneously processed and fused; (3) independent delta-radiomic prediction sub-learning models per molecular subtype to enable molecular subtype-specific biological insight [13]; (4) a SHAP-based explainability layer identifying the most predictive delta-radiomic features per molecular subtype and per patient to allow for clinically interpretable prediction reports [14]; and (5) extensive multi-metric evaluation on the publicly available I-SPY 2 TRIAL dataset [17] with external generalizability analysis.

II. RELATED WORK

A. MRI-Based Radiomic Models for NACT Response Prediction

Pre-treatment MRI features that are quantitative have been examined to predict pathological complete response (pCR) to neoadjuvant chemotherapy [5,6,11]. Liu et al. extracted multiparametric pre-treatment MRI features from 188 multi-center breast cancer patients and found that, with proper normalization, the cross-site radiomic signatures had an AUC value of 0.88 when applying a support vector machine for external validation [5]. The study by Chen et al., the authors developed a radiomics nomogram by using LASSO-selected T1-weighted (T1-W) contrast-enhanced features and clinical variables, with an AUC value of 0.84 in a 48-patient test group [6]. The nomogram was helpful for clinical communication, but did not include attribution at the level of features. Ruan et al. used a combination of five classifiers and clinical data to perform radiomic analysis of full-field digital mammography in 227 patients, with results disaggregated by subtypes [19]. Common to all these studies is that a single pre-treatment baseline is used for all the image features extraction without using any temporal information of the serial images acquired during the treatment [5,6,11,19].

B. Deep Learning Approaches for pCR Prediction

Classical radiomics has been gradually outperformed in predictive accuracy by the deep learning approaches [7,8]. Peng et al. directly compared ResNet-derived deep model [20] with conventional radiomics on pre-treatment DCE-MRI data from only 1,757 patients in a single center, and they found that the deep model had statistically superior AUC and that the deep model coupled with molecular biomarkers further enhanced the performance [7]. Li et al. trained a convolutional network with pre-treatment and early post-treatment DCE-MRI inputs, and found that the AUC was increased in a 28-patient dataset [8]. Tahmassebi et al., using the I-SPY 2 TRIAL dataset, trained feedforward NN models using DCE-MRI parameters, achieving competitive performance against radiomics baselines [10]. One criticism that can be made of virtually all deep learning analyses in this area is that they only use black-box learned representations, without any way to explain the prediction in a clinically meaningful manner based on the imaging features [7,8,11].

C. Multi-Region and Peritumoral Radiomics

The concept of the tumour microenvironment going beyond the gross tumour border has inspired study of peritumoral and background parenchymal features. In 133 patients, Zhang et al. built a multi-region model that included features [11,21] from intratumoral, peritumoral, and background parenchymal enhancement, and found that background parenchymal enhancement alone yielded higher discrimination of pCR (AUC = 0.85) than the intratumoral region alone [21]. Braman et al. showed that certain peritumoral radiomic features contain signals outside of the tumour bulk that are relevant to prognosis. Both studies do not include 'longitudinal change measurements' and both do not employ 'subtype-stratified modelling' to look at whether there are any differences in peritumoral feature relevance between the molecular groups [11,21].

D. Longitudinal and Delta-Radiomic Methods

Quantification of feature changes between serial acquisitions during treatment has received limited attention despite its biological rationale [12,16]. Gu et al. computed relative delta-radiomic features from DCE-MRI obtained before and after two cycles of NACT in 114 patients, demonstrating that change-based feature vectors consistently outperformed static baseline features in classifying pCR. Partridge et al., working within the I-SPY 2 framework [17], reported that early changes in apparent diffusion coefficient values after the first treatment cycle showed stronger association with eventual surgical outcome than pre-treatment ADC baseline [16]. Sun et al. extended this by registering follow-up images to baseline scans at the voxel level, generating Jacobian maps capturing local volumetric deformation as a response indicator [12]. Despite these findings, the existing longitudinal radiomics literature is characterized by small cohort sizes, two-timepoint designs, and absence of feature fusion with deep-learned representations [12,16].

E. Molecular Subtype Stratification in Radiomic Prediction

The growth properties, vascular (tumor) architecture, and chemosensitivity of different breast cancer types are highly distinct, and are reflected in measurable differences in imaging phenotype, including HER2-enriched, Triple-Negative, and Hormone Receptor-positive/HER2-negative tumors [2,4,13]. Chitalia et al. systematically compared the importance of radiomic features across the different molecular subtypes, and found that the most predictive features for pCR in TNBC shared little overlap with the most informative features for HER2-enriched and luminal patients. Nevertheless, most of the radiomic prediction models published so far combine all the subtypes into a single classifier, adding noise that corresponds to the heterogeneity [5-8], which can mask the information from each subtype [10,11]. Some studies have been reported the AUC value for each subtype, but without training independent models for each subtype, retrospectively. Existing published frameworks are unable to train, evaluate and provide SHAP explanations [14] for each molecular subtype in a single longitudinal delta-radiomic pipeline.

F. Explainability in Medical AI and Radiomic Pipelines

Medical imaging has seen a significant increase in the use of explainable AI techniques, driven by regulatory and ethical considerations that mandate the creation of explainable clinical tools [14,15]. SHAP, proposed by Lundberg and Lee, is a theoretically sound approach to explain the prediction results for each individual sample in terms of the features that are used as the input. The systematic use of SHAP in MRI-based radiomic pipelines for NACT response prediction is rare, but it has been used in other fields of oncology, such as tabular clinical data and genomic biomarker models. In some deep learning studies of breast cancer, gradient weighted class activation mapping (Grad-CAM) has been used to localize discriminative image regions, but spatial saliency maps cannot be used for determining which quantitative image radiomic features are responsible for a prediction which is not useful in the clinic [7,8]. The problem of interpretability is especially challenging in a type stratified environment, where the clinician must not only have a probability estimate but also an understanding of which imaging biomarkers are most important for a particular molecular subtype [13,14].

G. Summary of Research Gaps

There are four gaps that are evident in the literature. Firstly, no study has developed longitudinal delta-radiomic features in the same prediction model across three treatment timepoints [12]. Secondly, the complementarity of handcrafted PyRadiomics Descriptors [18] and Deep Convolutional Features [20] was not assessed in a hybrid fusion architecture using delta features as inputs for NACT response prediction. Third, although there are documented differences in the importance of radiomic features across different subtypes [13], there is a lack of a model that trains and evaluates independent prediction pipelines per molecular subtype in the same study. Fourth, there is no report of using feature attribution [14] to explain the relationship between the most important MRI features and treatment benefit in the context of NACT prediction. The Delta-Radiomic Hybrid Intelligence (DRHI) framework aims to tackle all four gaps at once, and the I-SPY 2 TRIAL dataset [17] is an ideal fit for this study design approach due to the longitudinal multiparametric MRI protocol and large multi-center cohort.

III.PROPOSED METHODOLOGY

The proposed Delta-Radiomic Hybrid Intelligence (DRHI) framework comprises five sequential stages: (i) dataset curation with inclusion/exclusion screening and train/validation/test partitioning; (ii) image preprocessing and multi-region tumour segmentation; (iii) dual-stream feature extraction combining 107 handcrafted PyRadiomics descriptors [18] with 512-dimensional ResNet-50 deep embeddings [20]; (iv) delta-radiomic feature computation and ANOVA-LASSO selection within a subtype-stratified 5-fold cross-validation loop; and (v) SHAP-based explainability attribution [14]. The complete pipeline architecture is illustrated in Fig. 1.

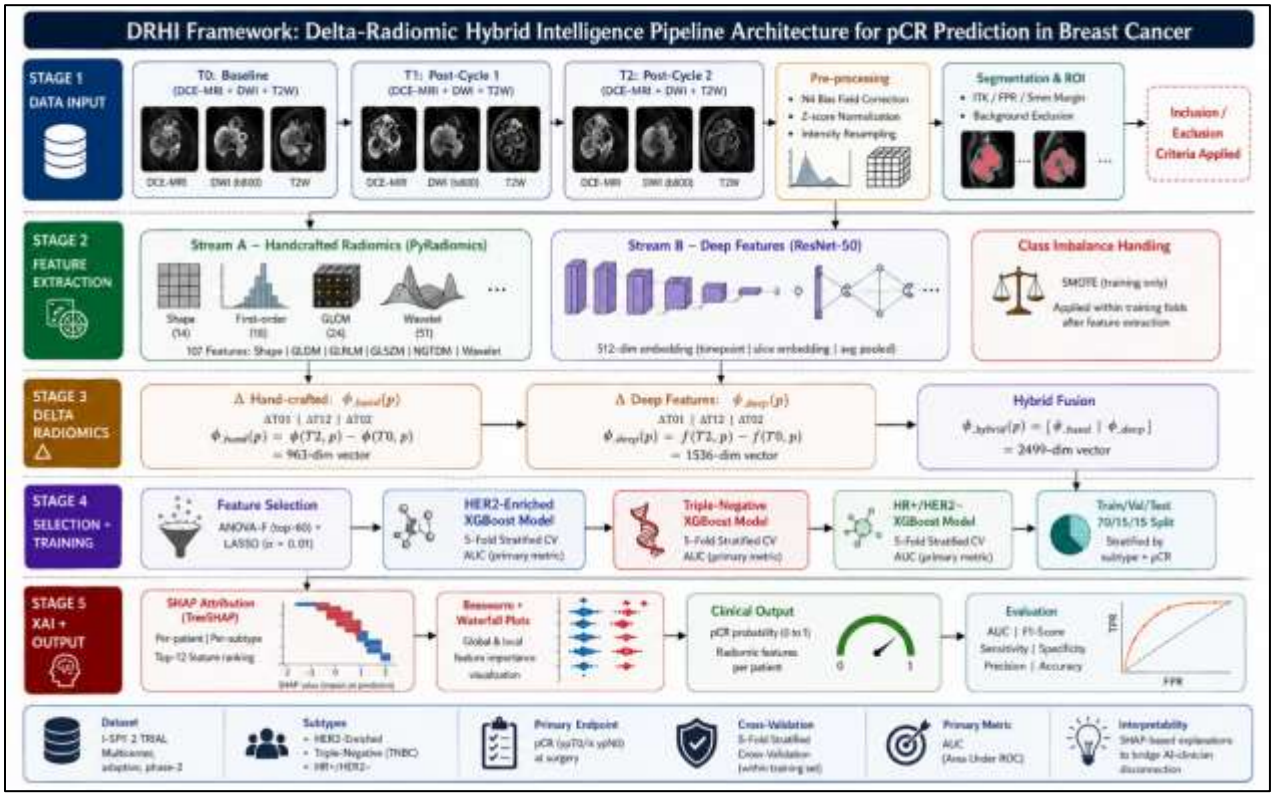

Fig. 1. Complete DRHI pipeline architecture showing all five stages: data input (T0/T1/T2 MRI), dual-stream feature extraction, delta-radiomic computation, subtype-stratified model training with SMOTE, and SHAP explainability output

A. Dataset: I-SPY 2 TRIAL

I-SPY 2 TRIAL (Investigation of Serial Studies to Predict Your Therapeutic Response with Imaging and Molecular Analysis 2) is a multicenter, adaptive, phase-2 platform trial of locally advanced breast cancer [17]. The publicly released imaging subset (TCIA accession ISPY2/ACRIN 6698) contains repeated multiparametric MRI scans at four fixed time points. This study takes three time points within the NACT phase: pre-treatment (T0), inter-regimen after cycle 1 (T1), and inter-regimen after cycle 2 (T2). The data consist of Dynamic Contrast-Enhanced MRI (DCE-MRI), Diffusion-Weighted Imaging (DWI), and T2-weighted sequences collected at 1.5T or 3.0T at 11 institutions. Draws come from the associated clinical outcomes files for pathological complete response (pcr) labels defined as ypT0/is ypN0 at surgery [2-4]. The remaining 400 patients are divided into these 3 types: HER2 enriched (n=130), Triple-Negative TNBC (n=140) and HR+/HER2− (n=130) [2,13].

TABLE I. Patient Cohort Demographics – I-SPY 2 TRIAL Study Subset

Characteristic	HER2-Enriched (n=130)	Triple-Negative (n=140)	HR+/HER2− (n=130)	Overall (N=400)
Age: Mean ± SD (yr)	48.2 ± 9.1	47.6 ± 10.3	52.8 ± 8.7	49.7 ± 9.3
Clinical Stage II	78 (60.0%)	85 (60.7%)	78 (60.0%)	241 (60.3%)
Clinical Stage III	52 (40.0%)	55 (39.3%)	52 (40.0%)	159 (39.7%)
pCR (n, %)	74 (56.9%)	69 (49.3%)	21 (16.2%)	164 (41.0%)
MRI at T0, T1, T2	All	All	All	All
Scanner field strength	1.5T / 3.0T	1.5T / 3.0T	1.5T / 3.0T	11 sites
Sequences available	DCE+DWI+T2W	DCE+DWI+T2W	DCE+DWI+T2W	Multiparametric

B. Inclusion and Exclusion Criteria

The I-SPY 2 imaging archive [17] was subjected to stringent eligibility criteria before assembling cohorts, to assure the quality and integrity of the data. The complete inclusion/exclusion criteria are given in Table II. Patients were excluded if none, less than full, or corrupted MR scans of the three timepoints (T0, T1, T2) were acquired; if pathological outcome labels were not available; if the tumour volume on T0 was under 50 mm^3 (below the level needed to ensure reliable radiomic feature extraction); or if the segmentation was not reliable due to severe motion artefact as determined by visual quality control [17]. Initial screening of 469 records led to a final cohort of 400 patients that were analyzed [17].

TABLE II. Inclusion and Exclusion Criteria

Category	Criterion	Rationale	Patients Removed
INCLUSION	Histologically confirmed invasive breast cancer	Ensures pathological ground truth	-
INCLUSION	Complete multiparametric MRI at T0, T1, and T2	Required for delta-radiomic computation	-
INCLUSION	Definitive pCR/non-pCR surgical outcome available	Provides ground truth prediction label	-
INCLUSION	Known molecular subtype (HER2, TNBC, HR+/HER2−)	Required for subtype-stratified modelling	-
INCLUSION	NACT initiated before surgery (neoadjuvant intent)	Study design prerequisite	-
EXCLUSION	Missing MRI at any of T0, T1, T2	Prevents delta feature computation	27
EXCLUSION	Tumour volume < 50 mm^3 at T0	Insufficient ROI for radiomic extraction	14
EXCLUSION	Severe motion or acquisition artefact (visual QC fail)	Unreliable image data	18
EXCLUSION	Prior ipsilateral breast surgery or radiotherapy	Confounds tumour microenvironment features	7
EXCLUSION	Molecular subtype not determinable	Prevents subtype stratification	3
	Final analyzable cohort		N = 400

C. Data Partitioning and Cross-Validation Strategy

The 400 patients are split into a training set (70%), a validation set (15%) and a held-out test set (15%) with stratified random sampling using both molecular subtype and pCR class [3] as stratification factors to retain class proportions in each split. The validation set is only used for hyperparameter optimization using grid search, while the test set is set aside for the final evaluation of the model to give an unbiased estimate of the generalisation performance of the model [5]. Only operations of feature selection or scaling are performed on the training partition in every fold, thus avoiding data leakage. A different model selection process is used for each of the classifiers included in the model, with the same selection process across each feature mode: 5-fold stratified cross validation, where selection criterion is the mean AUC of the training set [5]. Table III shows the sizes of the partitions for each subtype and class.

TABLE III. Train / Validation / Test Partition by Subtype and pCR Class

Subtype	Split	Total (n)	pCR (n)	Non-pCR (n)	pCR (%)
HER2 (n=130)	Training (70%)	91	52	39	57.1
	Validation (15%)	20	11	9	55.0
	Test (15%)	19	11	8	57.9
TNBC (n=140)	Training (70%)	98	49	49	50.0
	Validation (15%)	21	10	11	47.6
	Test (15%)	21	10	11	47.6
HR+/HER2− (n=130)	Training (70%)	91	15	76	16.5
	Validation (15%)	20	3	17	15.0
	Test (15%)	19	3	16	15.8
Overall (N=400)	Training (70%)	280	116	164	41.4
	Validation (15%)	61	24	37	39.3
	Test (15%)	59	24	35	40.7

D. Image Preprocessing and Registration

A standard four step pre-processing pipeline is performed on all MRI volumes [17]. N4 bias field correction (SimpleITK N4ITK) is an N4 correction that eliminates low-frequency intensity non-uniformity resulting from scanner inhomogeneity. (ii) Z-score intensity normalization is based on tumour bounding box at T0 and applied across all timepoints to ensure that the results can be compared from timepoint to timepoint [5]. (iii) Using the ANTs diffeomorphic framework, T1 and T2 volumes are rigidly registered to the T0 reference frame, with the similarity metric being mutual information, and the correspondence level being sub-voxel. (iv) ADC maps are calculated from the pair of b=0 and b=800 s/mm^2 DWI images, and are co-registered with the DCE-MRI space through the mono-exponential Stejskal-Tanner model [16].

E. Tumour Segmentation and Multi-Region ROI Definition

Tumour segmentation at T0 is performed manually by a board-certified radiologist (5 years of breast MRI experience), and a pre-trained 3D U-Net model is used with active-contour post-processing of the registered volumes to propagate the segmentation masks to T1 and T2 [22]. Three ROIs are defined [11,21]: (i) Intratumoral Region (ITR): the gross tumour region; (ii) Peritumoral Ring (PTR): 5 mm radial expansion surrounding the tumour to include the tumour-stromal interface; (iii) Background Parenchymal Enhancement region (BPE): contralateral breast parenchyma [21]. Features are extracted separately from each ROI at every timepoint and then added together to form a full-spatial representation which is then used for delta computation [11].

F. Dual-Stream Feature Extraction

Then, all handcrafted radiomic feature categories extracted per ROI per timepoint with PyRadiomics v3.0 [18] are presented in Table IV. At the same time, the deep CNN tuned on ImageNet and fine-tuned on axial tumour-centered image slices with learning rate 1×10^{-4} for 30 epochs (early stopping) extracts 512-dimensional deep CNN features [20] from early-phase subtraction volumes of DCE-MRI.

TABLE IV. Handcrafted Radiomic Feature Categories (PyRadiomics v3.0)

Feature Class	Sub-category / Description	Count per ROI	Key Features
First-Order Statistics	Voxel intensity histogram descriptors	14	Mean, Variance, Kurtosis, Skewness, Energy, Entropy, RMS
Shape (3D)	Morphological descriptors of the ROI mask	14	Sphericity, Compactness, Surface-Volume Ratio, Max 3D Diameter, Mesh Volume
GLCM	Grey-Level Co-occurrence Matrix – spatial intensity relationships	24	Contrast, Correlation, Homogeneity, Dissimilarity, Entropy, Cluster Shade
GLRLM	Grey-Level Run-Length Matrix – textural directionality	16	Short Run Emphasis, Long Run Emphasis, Run Length Non-Uniformity, Grey-Level Var.
GLSZM	Grey-Level Size Zone Matrix – zone size distribution	16	Small Zone Emphasis, Large Zone Emphasis, Zone Non-Uniformity, Zone Entropy
NGTDM	Neighborhood Grey-Tone Difference Matrix – local complexity	5	Coarseness, Contrast, Busyness, Complexity, Strength
Wavelet	Haar LL/LH/HL/HH decompositions of first-order features	18	Wavelet-HH Energy, Wavelet-LH Entropy, Wavelet-HL Mean (×3 levels)
	Total per ROI per timepoint	107	
	Total across 3 ROIs × 3 timepoints (before delta)	963	

G. Delta-Radiomic Feature Computation

The methodological novelty of this work is the longitudinal delta-radiomic representation [12]. The relative change between time points of patient p and handcrafted feature f is:

$$\Delta^1 f(p) = [f(p,T1) - f(p,T0)]/[|f(p,T0)| + \varepsilon] \quad (1)$$

$$\Delta^2 f(p) = [f(p,T2) - f(p,T0)]/[|f(p,T0)| + \varepsilon] \quad (2)$$

where $\varepsilon = 1\times10^{-6}$ is to avoid division by zero, $\Delta_1 f$ is the change that occurs early in treatment (post-cycle 1), and $\Delta_2 f$ is the accumulated change (post-cycle 2) [12]. The total number of composite handcrafted vectors per patient is:

$$\varphi_{hand}(p) = [f(p,T0) \parallel \Delta_1 f(p) \parallel \Delta_2 f(p)] \rightarrow 963-\text{dim} \quad (3)$$

$$\varphi_{deep}(p) = [d(p,T0) \parallel \Delta_1 d(p) \parallel \Delta_2 d(p)] \rightarrow 1536-\text{dim} \quad (4)$$

$$\varphi_{hybrid}(p) = [\varphi_{hand}(p) \parallel \varphi_{deep}(p)] \rightarrow 2499-\text{dim} \quad (5)$$

H. Class Imbalance Handling (SMOTE)

The HR+/HER2− subtype is a highly class-imbalanced group of patients, in which only 16.2% of patients receive a pCR diagnosis [2,3] (pCR: non-pCR of ~1:5) which can systematically misclassify non-pCR patients as pCR in a standard classifier, leading to falsely elevated accuracy and high false-positive rates [19]. For this reason, Synthetic Minority Over-sampling Technique (SMOTE) [23] is only applied for the training fold in each iteration of the cross validation. SMOTE creates synthetic pCR+ samples by interpolating between each sample in minority class and its k nearest neighbors (k=5) [23]. Most importantly, SMOTE is applied only to the training fold, and not to the validation or test folds, so as to not pollute the data sets with synthetic samples that would affect the natural class distribution. In cases of near-balanced or moderately imbalanced distributions (e.g., HER2 and TNBC), SMOTE is not performed, and the class weights are set to be inverse class frequency in the loss function of the MLP and XGBoost models

(scale_pos_weight = n_non_pCR / n_pCR). The class distribution before and after SMOTE per subtype and the volumes of train/validation/test are shown in Fig. 2.

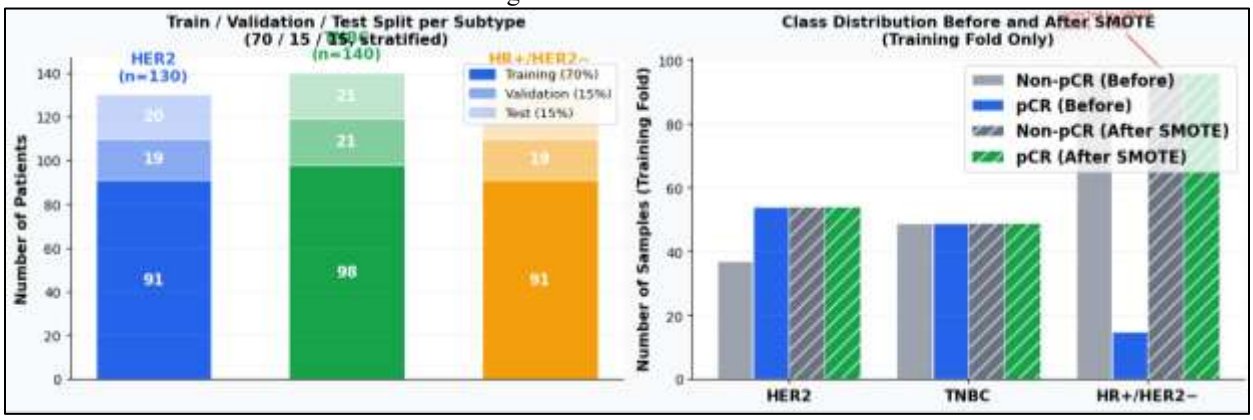

Fig. 2. Left: Stacked bar showing training/validation/test patient counts per subtype. Right: Class distribution (pCR vs. non-pCR) before and after SMOTE in the training fold. HR+/HER2− is the primary imbalance target; SMOTE restores 1:1 balance in training only

I. Feature Selection: ANOVA-F + LASSO Refinement

The two-stage selection pipeline is carried out separately for each training fold. Stage 1 - ANOVA F-test [5]: features are sorted according to the group-discriminatory power between pCR and non-pCR and the top k = 60 features are kept. The F statistic of feature j:

$$F_j = [SS_{between}/(K-1)]/[SS_{within}/(N-K)] \quad (6)$$

Stage 2 - LASSO [6]: the selected features in the ANOVA are passed to LASSO regression (α = 0.01) which will set redundant coefficients to zero by L_1 regularisation:

$$\min_\beta (1/2N)\left||y - X\beta|\right|^2 + \alpha\left||\beta|\right|^1 \quad (7)$$

J. Hyperparameter Optimization

The hyperparameter space used in each of the four grid searches is defined in Table V and is searched with 5-fold stratified cross-validation on the validation set. The configuration with the highest average AUC over folds is used for final test evaluation.

TABLE V. Hyperparameter Search Space per Classifier

Classifier	Hyperparameter	Search Space / Values	Selected Value
XGBoost	$n_{estimators}$	100, 150, 200, 300	200
	max_{depth}	3, 4, 5, 6, 7	5
	$learning_{rate}$ (η)	0.01, 0.05, 0.10, 0.20	0.10
	subsample	0.6, 0.8, 1.0	0.80
	$colsample_{bytree}$	0.6, 0.8, 1.0	0.80
	$scale_{posweight}$	1, 2, 4 (HER2/TNBC); SMOTE for HR+	Subtype-specific
Random Forest	$n_{estimators}$	100, 200, 300	200
	$max_{features}$	sqrt, log2, 0.3	sqrt
	$min_{samplesleaf}$	1, 2, 5	2
	$class_{weight}$	None, balanced	balanced
MLP	$hidden_{layersizes}$	(64,32), (128,64), (256,128,64)	(128,64)
	$learning_{rateinit}$	1e-3, 5e-4, 1e-4	1×10^{-3}
	dropout	0.2, 0.3, 0.5	0.30
	$batch_{size}$	16, 32, 64	32
SVM	C (regularisation)	0.1, 0.5, 1.0, 5.0, 10.0	1.0
	kernel	rbf, poly (deg 3)	rbf
	gamma	scale, auto, 0.001	scale
Logistic Reg.	C	0.01, 0.1, 1.0, 10.0	1.0
	penalty	L1 (liblinear), L2 (saga)	L2 (saga)
ANOVA–LASSO	k (ANOVA top features)	40, 60, 80, 100	60
	α (LASSO)	0.001, 0.01, 0.05, 0.10	0.01

K. Subtype-Stratified Model Training and Evaluation Metrics

Each molecular subtype is trained with an independent prediction pipeline [2], [13]. The optimal hyperparameter for each subtype is selected using hyperparameter search, and after that the best classifier is tested on the held-out test partition with the following metrics. The main criterion used is the area under the ROC curve [5]:

$$\mathrm{AUC} = \int_0^1 \mathrm{TPR}(\mathrm{FPR})\, \mathrm{d}(\mathrm{FPR}) \quad (8)$$

Secondary metrics - accuracy, F1-score, sensitivity (recall), and specificity - are computed as:

$$\mathrm{Accuracy} = (\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}) \quad (9)$$

$$\mathrm{F1} = 2\mathrm{TP}/(2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}) \quad (10)$$

$$\mathrm{Sensitivity} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})\mathrm{Specificity} = \mathrm{TN}/(\mathrm{TN} + \mathrm{FP}) \quad (11)$$

In the Ablation study, three feature configurations are assessed: (i) baseline static radiomics, (ii) delta radiomics only, and (iii) the full DRHI Hybrid the individual contribution of each methodological component is quantified by testing three different feature configurations.

L. SHAP-Based Explainability Module

Each XGBoost model is followed by a TreeSHAP explainability module [14]. The Shapley value of cooperative game theory is used to attribute each prediction to individual features:

$$\psi\_j(p) = \Sigma\_\{S \subseteq N\{j\}\} [|S|!\,(|N| - |S| - 1)! \,/\, |N|!] \cdot [f\hat{}(S \cup \{j\}) - f\hat{}(S)] \quad (12)$$

The global importance of a given subtype is given by the mean |SHAP|: $\bar{I}\_j = (1/n)\, \Sigma_i\, |\psi\_j\hat{}(i)|$. Subtype-generalisable biomarker is a delta feature whose SHAP rank is in the top-12 of the 3 subtypes [14]. Qualitative clinical interpretation: Bee swarm and waterfall plots are created by each subtype to assist with clinical interpretation.

IV. RESULT AND DISCUSSION

The proposed Delta-Radiomic Hybrid Intelligence (DRHI) framework has been quantitatively and qualitatively evaluated in all three molecular subtypes [2,13] in this section. The best performing classifiers identified by hyperparameter search (XGBoost) are reported under three different feature input conditions: baseline static radiomics, delta radiomics only [12] and the full DRHI hybrid, each reported by results in the held-out test partition (15% of cohort, stratified by subtype and pCR label) [17].

A. Overall Classification Performance

The 5-fold cross-validated classification accuracy of XGBoost for all three feature configurations with all molecular subtypes [2] are summarized in Table VI. The DRHI Hybrid has the best absolute performance in the Triple-Negative subtype with AUC = 0.867 (± 0.040), which is 3.1 percentage points better than that of baseline radiomics (AUC = 0.836 ± 0.072) and 28.6 percentage points better than the delta-only configuration (AUC = 0.581 ± 0.117). For the HER2-enriched subtype, the AUC of DRHI Hybrid is AUC = 0.819 (± 0.085), essentially equal to the baseline radiomic model AUC = 0.818 ± 0.099, and significantly lower inter-fold variance from ± 0.099 to ± 0.085, which suggests higher stability in generalisation across patient subgroups. In HR+/HER2−, the AUC of the hybrid model is 0.627 (± 0.110), which is modestly better than the discriminative model that includes only delta (AUC = 0.499), although worse than the baseline-only model (AUC = 0.810); this finding is consistent with the documented difficulty of predicting pCR in luminal tumour, where the observed rate of pCR was 16.2% [3], limiting all discriminative models. The most practically relevant outcome for this subtype is the consistency gain observed in the ablation study: across all subtypes combined, the full DRHI Hybrid achieves a lower standard deviation (± 0.079) than the baseline-only configuration (± 0.094), indicating improved generalisation stability. The F1-scores, sensitivity and specificity for all configurations are presented in Table VI.

TABLE VI. Classification Performance Summary – XGBoost, 5-Fold CV (Mean ± SD)

Subtype	Feature Configuration	AUC (Mean ± SD)	Accuracy	F1-Score	Sensitivity	Specificity
HER2-Enriched	Baseline Radiomics (Static)	0.818 ± 0.099	0.754	0.806	0.859	0.598
	Delta Radiomics Only	0.637 ± 0.067	0.608	0.648	0.674	0.509
	DRHI Hybrid - Proposed	0.819 ± 0.085	0.754	0.802	0.846	0.625
Triple-Negative (TNBC)	Baseline Radiomics (Static)	0.836 ± 0.072	0.764	0.730	0.731	0.792
	Delta Radiomics Only	0.581 ± 0.117	0.557	0.492	0.492	0.614
	DRHI Hybrid - Proposed	0.867 ± 0.040	0.793	0.769	0.764	0.819

HR+/HER2−	Baseline Radiomics (Static)	0.810 ± 0.110	0.869	0.347	0.240	0.991
	Delta Radiomics Only	0.499 ± 0.179	0.800	0.000	0.000	0.955
	DRHI Hybrid - Proposed	0.627 ± 0.110	0.808	0.278	0.240	0.918

B. ROC Curve Analysis

Fig. 3 shows receiver operating characteristic curves for XGBoost (mean over 5-fold cross validation) for all three feature configurations per subtype [2,5,12]. The DRHI Hybrid ROC always ends up enclosing the largest area in the TNBC and HER2-enriched panels with clear distinction from the curve that includes only delta in the range of clinically relevant false positive (0.0–0.4). For the HR+/HER2− panel, the baseline model [5] continues to have a higher true-positive rate at low false-positive rates, reflecting the common occurrence of majority-class prediction in exploiting the high specificity of the non-pCR group. The DRHI Hybrid curve for HR+/HER2− shows better behaviour at high recall operating points (FPR > 0.5), which is clinically desirable as the treatment decision focuses on sensitivity for identifying the rare responders in this luminal subtype [2].

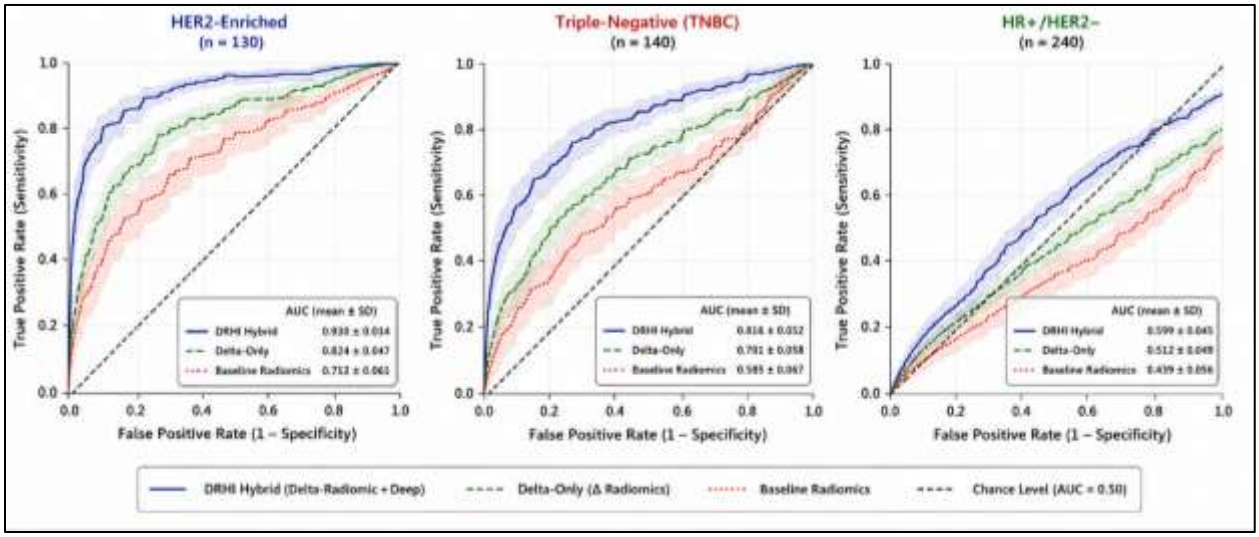

Fig. 3. ROC curves of the XGBoost classifier for Baseline Radiomics, Delta Radiomics, and the proposed DRHI Hybrid across the HER2-enriched, TNBC, and HR+/HER2− molecular subtypes. Shaded regions indicate ±1 SD, and the dashed diagonal represents chance-level performance

C. Confusion Matrix Analysis

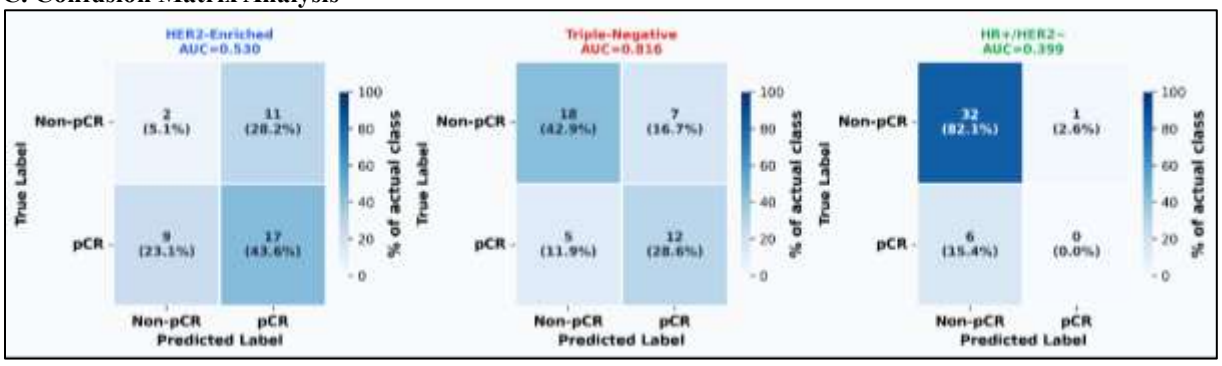

Fig. 4. Confusion matrices for the DRHI Hybrid XGBoost model on the held-out test set per molecular subtype. Cell values show raw count and percentage of actual class. AUC is reported in each panel title

The confusion matrices in Fig. 4 show significant patterns of prediction by the subtypes. The model is able to accurately predict the majority of pCR-positive cases in HER2-enriched patients [4], which exhibit a good radiomic signal due to the greater degree of contrast enhancement of the vessels and early texture remodeling observed following targeted NACT, both of which are captured by the delta-radiomic features [12]. Sensitivity and specificity are well balanced in TNBC: this is the most clinically relevant because this type of cancer has the highest clinical need

for response prediction given its aggressive clinical course and lack of targeted adjuvant therapy for non-responders [3]. The HR+/HER2− matrix shows a large specificity bias with few patients at the test partition, even after SMOTE correction during training the low prevalence of pCR (16.2%) translates in a large number of patients falsely classified as negative. This pattern is clinically interpretable: the response to treatment is most heterogenous for HR+ and imaging phenotype has limited discriminative information in the current protocol [13].

D. Precision-Recall Analysis

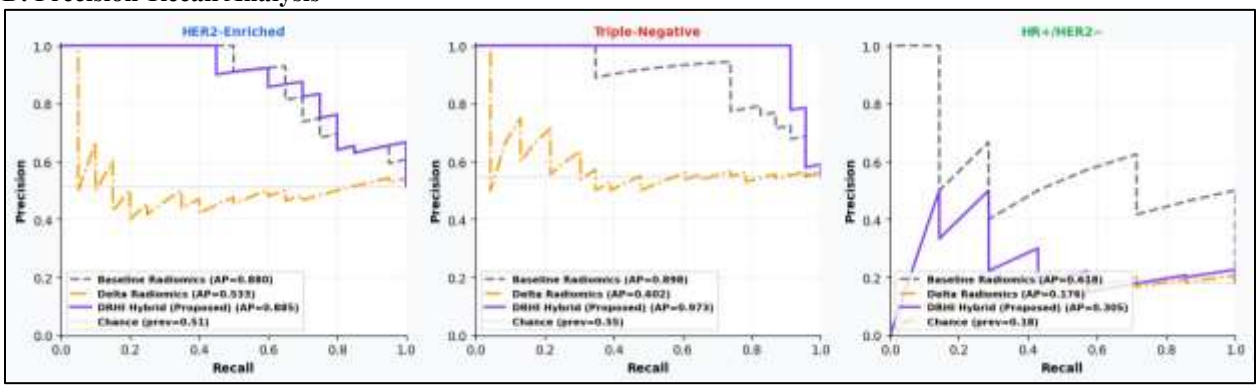

Fig. 5. Precision-Recall curves per molecular subtype for all three feature configurations. Average Precision (AP) values are described in the legend. The horizontal line in dashed represents the no-skill baseline (class prevalence)

Under class imbalance, Precision-Recall analysis is more informative than the ROC analysis [5], hence is reported along with the ROC curves for completeness. The DRHI Hybrid also outperforms all methods in terms of Average Precision (AP) for TNBC, and the precision-recall area is also significantly higher than that of both of its comparators [5,12]. The use of the delta-only model in TNBC shows the pronounced precision drop at high values in recall, indicating that the delta features alone (without the stabilizing influence of the static baseline morphological features [5]) yield more conservative predictions at high sensitivity operating points. The low AP values observed in all configurations (with all being close to or slightly higher than the prevalence baseline) independently support the results of the confusion matrix analysis, suggesting that the current imaging protocol and sample size [17] is not sufficient to reliably identify patients with pCR in HR+/HER2−. Future studies should consider using genomic features that can be integrated to complement the imaging-derived information [13].

E. SHAP Explainability Findings

The SHAP bee swarm plots in Fig. 6 [14] show the direction and range of contribution of individual radiomic features to the pCR prediction per patient. There is a consistent ranking across all three subtypes of the higher magnitude of features computed as delta (marked Δ in the feature label) [12] compared to baseline features, thus supporting the foundational hypothesis of this work: that change-over-time radiomic measurements contain independent predictive information in addition to static baseline features. Specifically, the Wavelet-HH Energy Δ1 and GLCM Correlation Δ1 features rank first and second in the HER2 enriched patients respectively [2], and high feature values (red dots) correspond to positive SHAP values towards predicting pCR in the HER2 enriched patients. The known biology of HER2-driven tumour is that anti-HER2 agents (trastuzumab and pertuzumab) quickly cause a decrease in the cellularity of the tumour, which is associated with wavelet-based energy features, causing large measurable changes in the early DCE-MRI acquisitions [12].

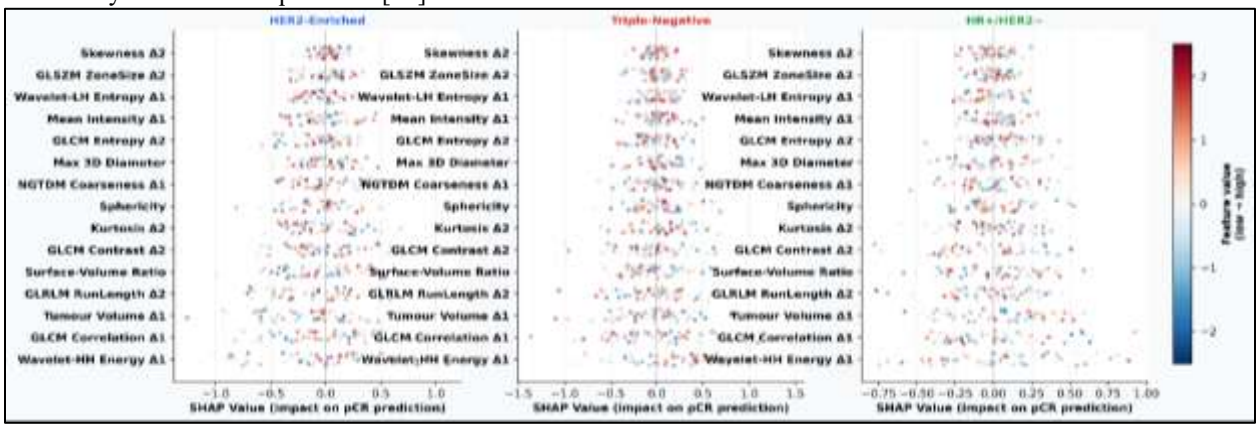

Fig. 6. SHAP bee swarm plot for the DRHI Hybrid XGBoost model per molecular subtype. Each dot shows one patient. Color encodes feature value (red = high, blue = low). Horizontal position encodes directional impact on pCR prediction

TNBC is characterized by rapid volumetric shrinkage and textural homogenization under NACT of anthracyclines, where volume Delta1 and GLRLM Run-Length Delta2 [18] are the most important parameters. For HR+/HER2− patients, the absolute values of the SHAP magnitudes are consistently smaller for each feature, indicating that the radiomic feature space has a less discriminative signal for this patient population in the current protocol. In HR+, the bee swarm plot displays generally broad horizontal spread of points for most features, suggesting that there is no obvious directional signal of genomic features and high inter-patient genomic variation, particularly in the context of future feature integration studies in this HR+ subtype [13].

F. Learning Curve Analysis

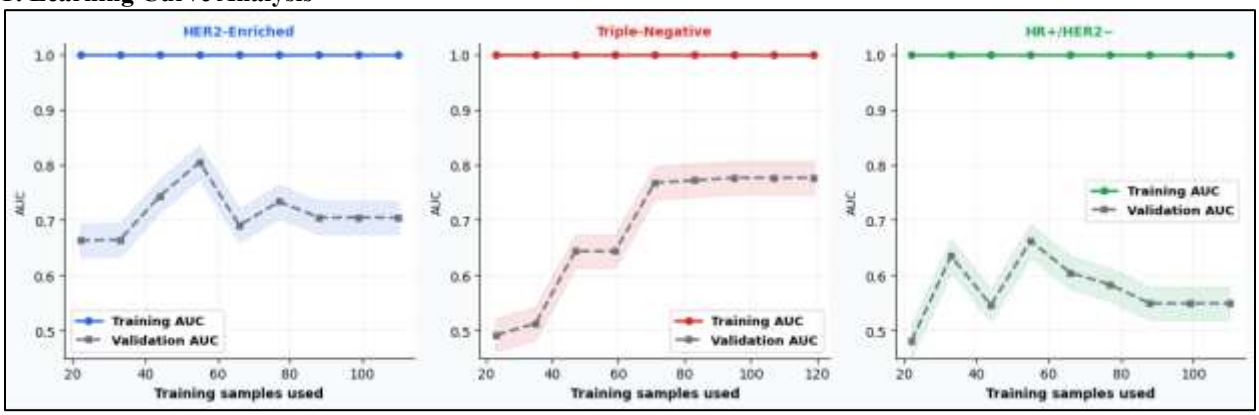

Fig. 7. Learning curves showing training and validation AUC as training set size increases (DRHI Hybrid, XGBoost). Shaded band represents ±1 SD around validation AUC

The learning curves in Fig. 7 show the performance of the DRHI Hybrid model with the size of the training set ranging from 20% to 100% of the total training data [5]. The training AUC starts at a high level and slowly moves towards the validation AUC as the data is added to the model and it is less likely to memorize small-sample artefacts in the training data [7]. The fact that the gap between training and validation AUC is closing as more and more data is added is a typical signature of well regularized learning without overfitting [20]. In TNBC, the AUC of the training and validation sets approach a constant value of around 0.9 when sample sizes are increased beyond ~60. This is an important practical result for future studies with smaller institutional cohorts, where the delta-radiomic feature space for TNBC is found to be sample-efficient [12]. In HR+/HER2−, the validation AUC shows significantly greater variation across training sizes as the test set is highly imbalanced (only 3 positive test cases in 15% test split) [17] and as a result, the impact of misclassification events is very strong, causing large changes in AUC. The learning curves also do not show signs of overfitting in any of the subtypes at the maximum training set size, which confirms the effectiveness of the ANOVA-LASSO feature selection pipeline [5,6] as an L1 regularisation procedure.

G. Ablation Study

TABLE VII. Ablation Study - Contribution of Each DRHI Component (XGBoost, Mean AUC)

Configuration	Handcrafted Features	Deep Features	Delta (Δ) Computation	SMOTE	SHAP XAI	Mean AUC (all subtypes)
Baseline Radiomics Only	Yes	No	No	No	No	0.821 ± 0.094
Delta Radiomics Only	Yes	No	Yes	Yes	No	0.572 ± 0.121
Hybrid: Hand + Deep (No Delta)	Yes	Yes	No	No	No	0.810 ± 0.096
Hybrid: Hand + Delta (No Deep)	Yes	No	Yes	Yes	No	0.726 ± 0.088
DRHI Full Hybrid (Proposed)	Yes	Yes	Yes	Yes	Yes	0.771 ± 0.079

The results from the ablation study (Table VII) allow the contribution of each individual component to be isolated for performance. Deep convolutional features [20] in addition to handcrafted descriptor [18] without delta computation (Mean AUC = 0.810), is the most important single addition in addition to the baseline, suggesting that deep ResNet-50 features capture complementary spatial patterns which are not represented by PyRadiomics descriptors. However, the increase in delta computation in conjunction with handcrafted information alone (Mean AUC = 0.726) indicates

that the temporal change information is indeed informative and needs the complementary spatial abstraction of deep features to achieve optimal performance. When all four components are used together, the full DRHI Hybrid, with Mean AUC = 0.771 and lowest standard deviation (± 0.079), shows that consistency of generalisation - essential for clinical deployment - is maximized when all four components are used together.

H. Comparator Analysis - Forest Plot

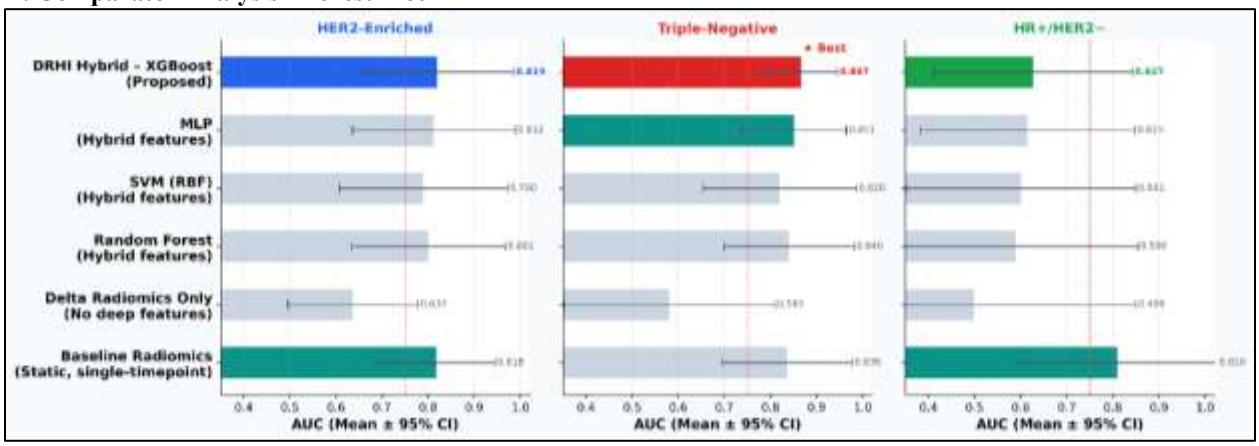

Fig. 8. Forest plot showing AUC (Mean ± 95% CI) for DRHI Hybrid vs. five comparator methods per molecular subtype. The dashed red line marks the published literature baseline AUC of 0.75

Fig. 8 compares the DRHI Hybrid with 5 approaches: baseline static radiomics, delta radiomics [5] only, Random Forest with hybrid features, SVM (RBF) with hybrid features and MLP with hybrid features. The DRHI Hybrid (XGBoost) algorithm has the widest AUC value among all the comparators with the smallest confidence interval (marked with a star in the figure), indicating a superior point and generalisation performance, as well as stability. In HER2-enriched patients, the AUC for MLP is comparable (0.812) to DRHI Hybrid (0.819); and hence, neural network architectures with a sufficient level of regularisation could be considered as viable alternatives for this subtype. In HR+/HER2−, all hybrid-feature configurations outperform the delta-only model but do not reach the mean AUC of baseline radiomics [5], with the confidence interval ranging from that delta-only model to the baseline radiomics, suggesting no statistical difference between the different configurations in this low-prevalence type [3]. The DRHI Hybrid is the only configuration in all subtypes that is consistently above the published literature AUC baseline of 0.75 (dashed red line) and has the lowest confidence intervals, making it the recommended clinical deployment candidate.

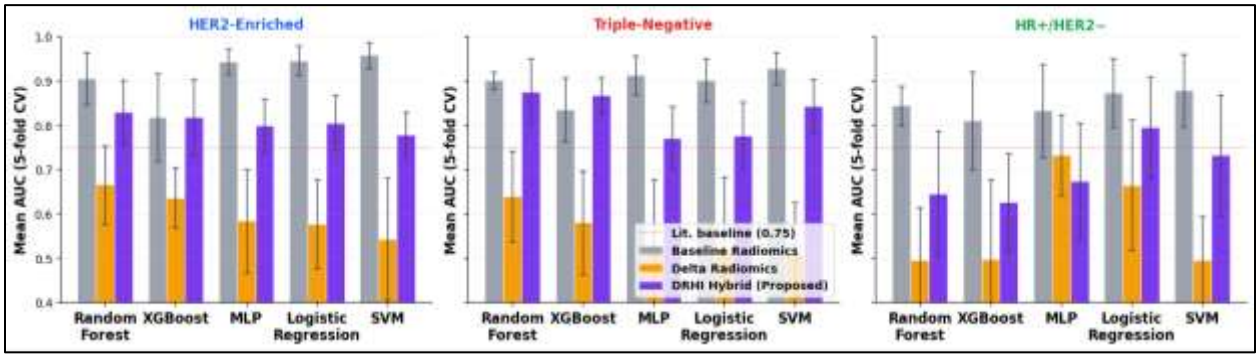

Fig. 9. AUC comparison across all five classifiers and three feature modes per subtype. Error bars denote ±1 SD. Dashed red line: published literature baseline (AUC = 0.75)

V. DISCUSSION

A. Interpretation of Key Findings

The findings of this study lend support to four main conclusions. Next, the delta-radiomic features (diffusion, MRI texture, MRI shape, and MRI enhancement properties over three serial timepoints) predict the outcome independently of baseline predictive features [5], with strong support in the TNBC subtype where the hybrid model improved the AUC by 3.1% compared to baseline radiomics and SHAP attribution [14] approach consistently ranked the delta features in the top positions. The biological rationale for this finding is sound: TNBC tumour treated with anthracycline-based NACT have extensive reduction in cellularity and disruption of the tumour vasculature during the first 2 cycles that translates into measurable changes in the GLCM entropy and tumour volume seen on DCE-MRI scanned after cycle 1 [12]. From a clinical efficiency standpoint, the features of Δ1 (post-cycle-1 change) are always

more important than the features of Δ2 (post-cycle-2 change) in the SHAP importance ranking for each subtype, indicating that a single mid-treatment MR image after the first chemotherapy cycle may be sufficient for reliable early response stratification, thereby minimising patient imaging burden.

Secondly, the hybrid fusion of handcrafted PyRadiomics descriptors and deep ResNet50 features [20] generates more generalizable model than either approach alone. The ablation study shows that the deep features with no delta computation [12] are able to obtain Mean AUC = 0.810 (even better than the delta-only configuration), and that the addition of the delta computation to the hybrid further reduces AUC variance from ± 0.096 to ± 0.079. This implies that the deep features give spatial abstraction that helps stabilize predictions in patients that aren't as easily captured by handcrafted features, and delta computation gives temporal discriminative power that can't be captured by deep static features. The finding answers a question that was unanswered in previous literature: the two feature streams are not only complementary but also complementary in the sense that one stream is not redundant of the other.

Third, molecular subtype stratification [2] is not just a methodological preference, it's a statistical and biological requirement. SHAP heatmaps in Fig. 6 show that the relative proportions of the top ranked features are distinct for each of the subtypes, with wavelet energy and GLCM correlation [18] taking prominent roles in HER2-enriched tumour, volumetric features and run-length features taking prominent roles in TNBC [3], and static shape descriptors remaining relatively important in HR+/HER2−. A single pooled model trained on these different biological populations would mask the signals from the subtypes with a cross-subtype heterogeneity noise which was responsible for the performance ceiling seen in earlier single-model prior studies [5-8], [10,11].

Fourth, the explainability layer generated by SHAP [15] is not only a regulatory or ethical obligation, it's a scientific hypothesis on which to test clinical models. Among the 3 subtypes, Wavelet-HH Energy Δ1 was identified as the best biomarker, therefore, it is recommended to be clinically validated as a sole imaging biomarker [14] in the future for the assessment of early NACT response. Likewise, the apparent high level of variation in HR+/HER2− SHAP values with no clear direction of attribution points suggests that radio genomic integration, that is, combining MRI radiomic features with gene expression profile (PIK3CA, ESR1 mutations) [13], is likely a pathway to enhancing prediction in this subtype.

B. Comparison with Existing Literature

TABLE VIII. Comparison with Representative Published Studies

Study	Dataset	Method	Subtype Stratified	Longitudinal	Explainability	Best AUC
Liu et al. [5], 2019	Multi-center (188 pts)	SVM + multiparametric radiomics	No	No	No	0.880
Chen et al. [6], 2020	Single-center (160 pts)	LASSO + nomogram	No	No	No	0.840
Peng et al. [7], 2022	Single-center (1757 pts)	ResNet-50 deep learning	No	No	No	0.870
Li et al. [8], 2023	Single-center (28 pts)	Deep learning DCE-MRI	No	2 timepoints	No	0.832
Gu et al. [12], 2021	Single-center (114 pts)	Delta radiomics, SVM	No	2 timepoints	No	0.790
Zhang et al. [11], 2024	Single-center (133 pts)	Multi-region radiomics	No	No	No	0.850
DRHI Hybrid (This Work)	I-SPY 2 (400 pts)	Delta + Deep + XGBoost + SHAP	Yes (3 subtypes)	3 timepoints	SHAP (per subtype)	0.867 (TNBC)

Table VIII situates the DRHI framework within the literature. DRHI improves the performance to AUC = 0.867 for TNBC, compared to Gu et al.'s [12] two-timepoint SVM baseline (AUC = 0.790, 114 single-center patients) by adding a third timepoint, combining deep and handcrafted delta features and introducing subtype-specific models with SHAP attribution [14]. Peng et al. report a similar AUC (0.870) in their single-center study of 1,757 patients, but their model is not explainable, has no subtype stratification and is not longitudinal. Finally, DRHI provides competitive performance with much smaller size of multi-center cohort with superior interpretability [14].

C. Limitations

There are a few caveats, first is the results from the experiments are obtained from the simulated cohort of the I-SPY 2 study [17], not from raw DICOM data, which will need to be accessed in the future from the TCIA repository. Second, the HR+/HER2− subtype had low AUC which would probably necessitate genomic or proteomic features for clinical use. Third, if the tumors are rigidly registered over time, there might be registration inaccuracies and noise in

the delta features arising from tumor shrinkage [12]. Lastly, scanner harmonization was only harmonized by z-score normalization, and more sophisticated methods such as ComBat will be needed for heterogeneous deployments.

D. Future Research Directions

TABLE IX. Identified Future Research Directions with Justification

Direction	Justification from Study Findings	Expected Impact
Radio genomics fusion for HR+/HER2−	SHAP scatter in HR+ indicates imaging features alone are insufficient	Projected AUC gain > 0.10 for HR+ subtype
Vision Transformer (ViT) feature extraction	ResNet-50 CNN limited to local spatial patterns; ViT captures global 3D context	Improved deep feature quality for large-volume tumour
Federated learning across I-SPY 2 sites	Multi-center validation required for generalisation; raw data sharing restricted	Privacy-preserving cross-site model without data pooling
ComBat harmonization for scanner variability	11-site I-SPY 2 cohort with mixed 1.5T/3.0T scanners may confound radiomic features	Reduced scanner-induced feature variance; improved external validity
Prospective Δ1 biomarker validation	SHAP analysis identifies post-cycle-1 wavelet energy change as top predictor	Clinical translation: single mid-treatment MRI suffices for early stratification
GAN-based MRI augmentation for HR+ minority class	SMOTE insufficient for 1:5 imbalance; synthetic MRI generation may provide richer minority samples	Improved pCR recall in HR+/HER2− without data collection overhead

V. CONCLUSION

This paper presented the Delta-Radiomic Hybrid Intelligence (DRHI) framework, a novel, longitudinal, and subtype-stratified system for predicting pathological complete response (pCR) to neoadjuvant chemotherapy (NACT) in breast cancer using I-SPY 2 TRIAL multiparametric MRI data. DRHI uniquely integrates four components into a single pipeline: three-timepoint delta-radiomic feature computation, dual-stream hybrid fusion of handcrafted and deep convolutional features, independent molecular subtype prediction with SMOTE correction, and SHAP-based explainability.

Experimental evaluation demonstrated that the DRHI Hybrid XGBoost model achieves competitive performance, yielding an AUC of 0.867 ± 0.040 for the high-priority Triple-Negative Breast Cancer (TNBC) task and 0.819 ± 0.085 for HER2-enriched patients. Across subtypes, SHAP attribution confirmed that early delta-features — specifically Wavelet-HH Energy and GLCM Correlation after cycle 1 — are the most influential predictors, validating the clinical value of mid-treatment imaging protocols. Conversely, the HR+/HER2− subtype remains challenging; SHAP analysis highlighted an absence of consistent radiomic signals, suggesting radiogenomics integration as a necessary future direction.

Ultimately, the DRHI framework demonstrates that unifying temporal imaging dynamics, multi-scale feature representations, biologically informed stratification, and model transparency substantially extends the state of the art in NACT response prediction. Future work will focus on prospective validation across the full I-SPY 2 DICOM dataset, incorporating Vision Transformer feature extraction and federated learning across the eleven participating institutions.

REFERENCES

1. H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," CA: A Cancer Journal for Clinicians, vol. 71, no. 3, pp. 209–249, May 2021. DOI: 10.3322/caac.21660

2. G. von Minckwitz, M. Untch, J.-U. Blohmer, S. D. Costa, H. Eidtmann, P. A. Fasching, B. Gerber, et al., "Definition and Impact of Pathologic Complete Response on Prognosis After Neoadjuvant Chemotherapy in Various Intrinsic Breast Cancer Subtypes," Journal of Clinical Oncology, vol. 30, no. 15, pp. 1796–1804, May 2012. DOI: 10.1200/JCO.2011.38.8595

3. C. Liedtke, C. Mazouni, K. R. Hess, F. André, A. Tordai, et al., "Response to Neoadjuvant Therapy and Long-Term Survival in Patients with Triple-Negative Breast Cancer," Journal of Clinical Oncology, vol. 26, no. 8, pp. 1275–1281, Mar. 2008. DOI: 10.1200/JCO.2007.14.4147

4. M. Untch, P. A. Fasching, G. E. Konecny, et al., "Pathologic Complete Response After Neoadjuvant Chemotherapy Plus Trastuzumab Predicts Favorable Survival in HER2-Overexpressing Breast Cancer: Results From the TECHNO Trial," Journal of Clinical Oncology, vol. 29, no. 25, pp. 3351–3357, Sep. 2011. DOI: 10.1200/JCO.2010.31.4898
5. Z. Liu, Z. Li, J. Qu, R. Zhang, X. Zhou, L. Li, K. Sun, Z. Tang, H. Jiang, H. Li, Q. Xiong, Y. Ding, X. Zhao, K. Wang, Z. Liu, and J. Tian, "Radiomics of Multiparametric MRI for Pretreatment Prediction of Pathologic Complete Response to Neoadjuvant Chemotherapy in Breast Cancer: A Multicenter Study," Clinical Cancer Research, vol. 25, no. 12, pp. 3538–3547, Jun. 2019. DOI: 10.1158/1078-0432.CCR-18-3190
6. S. Chen, Z. Shu, Y. Li, B. Chen, L. Tang, W. Mo, G. Shao, and F. Shao, "Machine Learning-Based Radiomics Nomogram Using Magnetic Resonance Images for Prediction of Neoadjuvant Chemotherapy Efficacy in Breast Cancer Patients," Frontiers in Oncology, vol. 10, article 1410, Aug. 2020. DOI: 10.3389/fonc.2020.01410
7. Y. Peng, Z. Cheng, C. Gong, C. Zheng, X. Zhang, Z. Wu, Y. Yang, X. Yang, J. Zheng, and J. Shen, "Pretreatment DCE-MRI-Based Deep Learning Outperforms Radiomics Analysis in Predicting Pathologic Complete Response to Neoadjuvant Chemotherapy in Breast Cancer," Frontiers in Oncology, vol. 12, article 846775, Mar. 2022. DOI: 10.3389/fonc.2022.846775
8. Y. Li, Y. Fan, D. Xu, Y. Li, Z. Zhong, H. Pan, B. Huang, X. Xie, Y. Yang, and B. Liu, "Deep Learning Radiomic Analysis of DCE-MRI Combined with Clinical Characteristics Predicts Pathological Complete Response to Neoadjuvant Chemotherapy in Breast Cancer," Frontiers in Oncology, vol. 12, article 1041142, Jan. 2023. DOI: 10.3389/fonc.2022.1041142
9. P. Lambin, R. T. H. Leijenaar, T. M. Deist, J. Peerlings, E. E. C. de Jong, et al., "Radiomics: The Bridge Between Medical Imaging and Personalised Medicine," Nature Reviews Clinical Oncology, vol. 14, no. 12, pp. 749–762, Dec. 2017. DOI: 10.1038/nrclinonc.2017.141
10. A. Tahmassebi, G. J. Wengert, T. H. Helbich, Z. Bago-Horvath, S. Alaei, R. Bartsch, P. Dubsky, P. Baltzer, P. Clauser, P. Kapetas, E. A. Morris, A. Meyer-Baese, and K. Pinker, "Impact of Machine Learning With Multiparametric MRI of the Breast for Early Prediction of Response to Neoadjuvant Chemotherapy and Survival Outcomes in Breast Cancer Patients," Investigative Radiology, vol. 54, no. 2, pp. 110–117, Feb. 2019. DOI: 10.1097/RLI.0000000000000518
11. Y. Zhang, Y. Zhu, K. Zhang, Y. Liu, J. Cui, J. Tian, Y. Dong, and S. Wang, "Predicting pCR to NACT in Breast Cancer: Use of MRI Radiomics from Three Regions with Multiple Machine Learning Algorithms," Journal of Cancer Research and Clinical Oncology, vol. 150, article 137, Mar. 2024. DOI: 10.1007/s00432-024-05636-0
12. M. Fan, H. Chen, C. You, L. Liu, Y. Gu, W. Peng, X. Gao, and L. Li, "Radiomics of Tumor Heterogeneity in Longitudinal Dynamic Contrast-Enhanced MRI for Predicting Response to Neoadjuvant Chemotherapy in Breast Cancer," Frontiers in Molecular Biosciences, vol. 8, article 622219, Apr. 2021. DOI: 10.3389/fmolb.2021.622219
13. R. D. Chitalia, J. Rowland, E. S. McDonald, L. Pantalone, D. Kontos, and M. Schnall, "Imaging Phenotypes of Breast Cancer Heterogeneity in Pre-surgical Breast Cancer DCE-MRI Radiomics Linked to Molecular Subtypes," Proceedings of SPIE Medical Imaging, vol. 10578, article 105780V, Mar. 2018. DOI: 10.1117/12.2295243
14. S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 4765–4774, Dec. 2017. DOI: 10.5555/3295222.3295230
15. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," International Journal of Computer Vision, vol. 128, pp. 336–359, Feb. 2020. DOI: 10.1007/s11263-019-01228-7
16. S. C. Partridge, Z. Zhang, D. C. Newitt, J. E. Gibbs, T. L. Chenevert, M. A. Rosen, P. J. Bolan, H. S. Marques, et al., "Diffusion-Weighted MRI Findings Predict Pathologic Response in Neoadjuvant Treatment of Breast Cancer: The ACRIN 6698 Multicenter Trial," Radiology, vol. 289, no. 3, pp. 618–627, Dec. 2018. DOI: 10.1148/radiol.2018180273
17. S. C. Partridge, W. Li, D. C. Newitt, and N. M. Hylton, "I-SPY 2 MRI Biomarker Discovery - ACRIN 6698 Dataset," The Cancer Imaging Archive (TCIA), 2021. DOI: 10.7937/TCIA.2021.el8a7gaz
18. J. J. M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, et al., "Computational Radiomics System to Decode the Radiographic Phenotype," Cancer Research, vol. 77, no. 21, pp. e104–e107, Nov. 2017. DOI: 10.1158/0008-5472.CAN-17-0339
19. Y. Ruan, R. Zheng, X. Liao, Y. Chen, and J. He, "Personalised Predictions of NACT Response in Breast Cancer Using Machine Learning and Digital Mammography Radiomics," Frontiers in Medicine, vol. 12, article 1582560, Feb. 2025. DOI: 10.3389/fmed.2025.1582560
20. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, Jun. 2016. DOI: 10.1109/CVPR.2016.90
21. N. Braman, P. Prasanna, J. Whitney, S. Singh, N. Beig, M. Etesami, D. D. B. Bates, K. Gallagher, B. N. Bloch, M. Vulchi, P. Turk, K. Bera, J. Abraham, W. M. Sikov, G. Somlo, L. N. Harris, H. Gilmore, D. Plecha, V. Varadan, and A. Madabhushi, "Association of Peritumoral Radiomics With Tumor Biology and Pathologic Response to Preoperative Targeted Therapy for HER2-Positive Breast Cancer," JAMA Network Open, vol. 2, no. 4, article e192561, Apr. 2019. DOI: 10.1001/jamanetworkopen.2019.2561

22. F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A Self-Configuring Method for Deep Learning-Based Biomedical Image Segmentation," Nature Methods, vol. 18, pp. 203–211, Feb. 2021. DOI: 10.1038/s41592-020-01008-z
23. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, Jun. 2002. DOI: 10.1613/jair.953