# Radiomics and Deep Learning in MRI-based Molecular Subtype Classification of Breast Cancer: A Comprehensive Review

Xupeng Lu[1, 2], Hazirah Bee bt Yusof Ali[1], Junxiu Wang[3, *]
[1]Faculty of Information Technology, City University Malaysia, Petaling Jaya 46100, Malaysia
[2]Department of Science, Taiyuan Institute of Technology, Taiyuan 030008, China
[3]Department of Computer Engineering, Taiyuan Institute of Technology, Taiyuan 030008, China
*Corresponding email: wangjx@tit.edu.cn

**Abstract**

Breast cancer molecular subtyping is fundamental to precision oncology, yet current clinical practice depends on invasive tissue sampling that inherently fails to capture complete tumor heterogeneity. The convergence of radiomics and deep learning with multiparametric magnetic resonance imaging presents a transformative pathway for non-invasive molecular characterization. This review systematically examines the methodological evolution, current capabilities, and translational trajectory of magnetic resonance imaging (MRI)-based radiomics and deep learning for breast cancer molecular subtype prediction, with particular emphasis on computational challenges and their solutions. We synthesize evidence across four thematic domains: (1) Conventional radiomics feature engineering. (2) Deep learning architecture and transfer learning. (3) Integrated multimodal frameworks. (4) Emerging paradigms include self-supervised learning and mechanistically interpretable models. Our analysis demonstrates that multimodal approaches consistently outperform single-modality models, achieving area under the curve values of 0.850 to 0.960 for luminal versus non-luminal discrimination, 0.840 to 0.970 for human epidermal growth factor receptor 2 (HER2) status prediction, and 0.860 to 0.980 for triple-negative breast cancer identification in recent multicenter studies. Nevertheless, substantial challenges persist in data standardization, model generalizability, and clinical deployment; these remain the primary barriers to widespread adoption. From a computational perspective, this review identifies that: (1) Transfer learning and self-supervised learning are essential for overcoming data scarcity. (2) Intermediate fusion strategies offer the optimal balance between performance and interpretability for multimodal integration. (3) Current explainable AI methods, while valuable, provide post-hoc explanations rather than inherently interpretable reasoning, which limit clinical trust. Future research must prioritize robust external validation, development of inherently interpretable models, and exploration of foundation models to accelerate clinical translation. By systematically mapping computational challenges and solutions, this review establishes a methodological roadmap for developing clinically trustworthy deep learning radiomics models in breast cancer.

**Keywords**

Radiomics, Deep learning, Breast cancer, Molecular subtypes, Magnetic resonance imaging

## Introduction

Breast cancer remains the most prevalent malignancy among women worldwide and the leading cause of cancer-related mortality in this population, with an estimated 2.300 million new cases and 670,000 deaths annually according to 2022 data [1]. The disease exhibits profound biological heterogeneity, manifesting as distinct molecular subtypes, including luminal A, luminal B, HER2-positive, and triple-negative breast cancer (TNBC), each associated with divergent clinical behaviors, treatment responses, and prognostic outcomes [2]. This molecular classification, grounded in the expression of estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2), and the proliferation marker Ki-67, has become the cornerstone of clinical decision-making, guiding targeted therapy, endocrine treatment, and chemotherapy selection [3].

Current clinical practice determines molecular subtype through immunohistochemical and in situ hybridization analysis of core needle biopsy specimens [4]. However,

this approach suffers from fundamental limitations. Tissue biopsy is invasive, subject to sampling bias that inadequately represents intratumoral heterogeneity, and impractical for serial monitoring of treatment response. Recent large-cohort studies have demonstrated discordance rates of 3.90% for ER, 4.80% for PR, and 1.20% for HER2 between core needle biopsy and surgical specimens, with molecular subtype discordance occurring in 344 of 2,099 cases (16.40%), predominantly between luminal A and luminal B subtypes [5]. Furthermore, biopsies cannot capture the spatiotemporal evolution of tumor biology during disease progression or therapeutic intervention. These constraints have motivated intensive investigation of non-invasive approaches capable of comprehensive, repeatable tumor phenotyping.

Magnetic resonance imaging (MRI) has emerged as the most sensitive imaging modality for breast cancer detection and characterization, offering superior soft tissue contrast and functional imaging capabilities [6]. Multiparametric MRI (mpMRI) integrates complementary sequences that probe distinct biological properties: dynamic contrast-enhanced (DCE) imaging assesses tumor vascular permeability and extracellular volume fraction through quantitative analysis of contrast agent kinetics [7]. Diffusion-weighted imaging (DWI) quantifies apparent diffusion coefficient (ADC) values reflecting cellular density and tissue microstructure [8]. T2-weighted imaging identifies peritumoral edema as a key indicator of tumor invasiveness [9]. Modern MRI protocols achieve high spatiotemporal resolution, with slice thickness ≤1 mm and temporal resolution ≤10 s for DCE-MRI, enabling comprehensive visualization of intratumoral heterogeneity and providing a functional imaging basis for molecular subtyping.

Radiomics, an emerging field at the intersection of medical imaging and data science, addresses the challenge of extracting quantitative, reproducible information from these rich imaging datasets [10]. By converting medical images into mineable high-dimensional data, radiomics enables comprehensive, non-invasive assessment of tumor characteristics, including the quantification of heterogeneity that may escape visual assessment by radiologists [11]. Parallel advances in deep learning, particularly convolutional neural networks (CNNs), have demonstrated remarkable capacity for automatic feature learning from raw image data, often capturing abstract representations that complement handcrafted radiomic features [12].

The convergence of radiomics and deep learning, known as deep learning radiomics, has emerged as a particularly powerful paradigm that leverages the complementary strengths of both approaches [13]. From a computational perspective, deep learning radiomics represents a novel methodological paradigm integrating the interpretability and biological plausibility of conventional radiomic features with the representational power and automation of deep neural networks. This hybrid approach has the potential to capture both explicit imaging phenotypes and latent patterns imperceptible to human observers [14]. Unlike purely end-to-end deep learning approaches, this hybrid paradigm explicitly preserves feature-level interpretability while benefiting from hierarchical representation learning. Recent systematic reviews have demonstrated that deep learning models outperform traditional machine learning approaches for molecular subtype classification. For instance, deep learning achieved an area under the curve (AUC) of 0.930 for TNBC identification, compared to 0.850 for support vector machines, with texture features contributing approximately 40.00% to overall model accuracy [15].

While several previous reviews have summarized the role of radiomics in breast cancer, they often predate the widespread integration of deep learning, multimodal fusion, and explainable AI (XAI). This review aims to fill that gap by systematically evaluating the methodological evolution from 2020 to 2026 from a computational perspective, with three specific contributions: (1) Identifying effective strategies for small-sample learning, including transfer learning, meta-learning, and self-supervised learning, that address the fundamental data scarcity challenge in medical imaging. (2) Comparing multimodal fusion strategies and establishing best practices for integrating heterogeneous data types. (3) Critically evaluating explainable AI techniques and their limitations in providing clinically meaningful insights. We argue that successful translation of deep learning radiomics hinges on solving three fundamental computational challenges: data scarcity, multimodal heterogeneity, and model opacity. These challenges directly correspond to critical clinical needs, namely robust performance across diverse populations, comprehensive tumor phenotyping, and trustworthy

decision support. This synthesis of 128 existing studies focuses on methodological innovations, clinical performance, and translational challenges.

## Literature search and methodology

### *Search strategy*

A comprehensive literature search was conducted in PubMed, Web of Science, Scopus, and IEEE Xplore databases for studies published between January 2020 and March 2026. The search strategy followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines and was structured according to the Population, Intervention, Comparison, Outcome, Study design (PICOS) framework [16].

The search strategy combined terms related to four core concepts: (1) Breast cancer ("breast cancer"). (2) Imaging modality ("magnetic resonance imaging" or "MRI" or "DCE-MRI" or "diffusion-weighted imaging" or "DWI" or "multiparametric MRI"). (3) Computational methods ("radiomics" or "texture analysis" or "deep learning" or "convolutional neural network" or "CNN" or "machine learning"). (4) Molecular subtyping ("molecular subtype" or "Luminal" or "HER2" or "triple-negative" or "TNBC" or "estrogen receptor" or "ER" or "progesterone receptor" or "PR").

### *Inclusion and exclusion criteria*

Studies were included if they met the following criteria: (1) Investigated radiomics and/or deep learning approaches for breast cancer molecular subtype classification using MRI. (2) Reported original research with quantitative performance metrics (accuracy, area under the curve (AUC), sensitivity, specificity, or F1-score). (3) Included histopathologically confirmed molecular subtypes (ER, PR, HER2, Ki-67) as reference standard. (4) Utilized MRI sequences including at least one of DCE-MRI, DWI, ADC mapping, or T2-weighted imaging. (5) Published in peer-reviewed journals. (6) Written in English.

Exclusion criteria were: (1) Review articles, editorials, letters, conference abstracts, case reports, and commentaries. (2) Studies with insufficient data to extract performance metrics. (3) Studies using non-MRI modalities (mammography, ultrasound, CT) exclusively. (4) Studies focused on benign-malignant discrimination without molecular subtyping. (5) Studies with sample size <20 patients.

### *Data extraction and synthesis*

For each included study, we extracted: (1) Study characteristics (sample size, MRI sequences, patient demographics). (2) Methodological details (feature extraction methods, machine learning algorithms, validation strategies). (3) Performance metrics (accuracy, AUC, sensitivity, specificity). (4) Key findings and limitations. Data were synthesized narratively, with emphasis on methodological trends and comparative performance across approaches.

## Radiomics and deep learning techniques for molecular subtype classification

### *Comparative analysis of radiomics and deep learning*

Radiomics and deep learning represent two fundamentally different paradigms for extracting phenotypic information from medical images. Table 1 summarizes their key differences.

Table 1. Comparison of conventional radiomics and deep learning approaches.

| Aspect | Conventional radiomics | Deep learning |
|---|---|---|
| Feature origin | Handcrafted, mathematically defined (e.g., texture, shape) | Automatically learned from data via backpropagation |
| Interpretability | High: features have explicit meanings (e.g., skewness) | Low: requires post-hoc explanation methods (e.g., Grad-CAM) |
| Data efficiency | Works with small datasets (dozens to hundreds) | Requires large datasets (thousands) for training from scratch |
| Transferability | Features are fixed, may not generalize across protocols | Representations can be fine-tuned to new domains |
| Representation capacity | Limited to predefined formulas | Can capture complex, non-linear patterns |
| Typical workflow | Segmentation; feature extraction; feature selection; Machine Learning | End-to-end learning (segmentation + classification) |

Importantly, "deep learning radiomics" is not merely the application of deep learning to radiomic problems, but a distinct methodological paradigm that seeks to fuse the strengths of both approaches. In this paradigm, handcrafted radiomic features (with explicit mathematical definitions and biological interpretations) and deep learning-derived representations (optimized for predictive performance through end-to-end learning) are integrated within a unified framework. This hybrid approach addresses the fundamental limitation of pure deep learning (its lack of interpretability) while overcoming the representational constraints of conventional radiomics, which is limited to predefined formulas. As demonstrated in Section 4, such integrated frameworks consistently outperform either approach alone [17]. These differences make radiomics and deep learning complementary. Hybrid approaches that combine handcrafted radiomic features with deep-learned representations consistently outperform either alone, as detailed in Section 4.

***Conventional radiomics feature engineering***

Radiomic features are systematically categorized into hierarchical classes capturing distinct aspects of tumor phenotype [18]. First-order statistics describe the distribution of voxel intensities within the tumor volume without considering spatial relationships, providing summary measures including mean, variance, skewness, and kurtosis that reflect global tissue characteristics [19]. These features have demonstrated associations with cellular density and necrosis: higher skewness (>1.2) in TNBC reflects focal necrosis patterns, while HER2-positive tumors exhibit lower skewness (<0.8).

Shape-based features quantify three-dimensional geometric properties including tumor volume, surface area, sphericity, and lobulation index. These features relate to growth patterns and invasiveness: Sphericity and lobulation index from DCE-MRI and T2-weighted MRI jointly predict ER-positive tumor aggressiveness, with a higher lobulation index (>0.6) associated with more aggressive ER-positive tumors [20]. In TNBC, the rim enhancement index is calculated as the ratio of peripheral enhancement area to total tumor area on DCE-MRI. When this index exceeds 0.350, it achieves 94.00% positive predictive value, serving as a specific marker for this aggressive subtype [21].

Texture features constitute the core of heterogeneity quantification, characterizing spatial arrangements of voxel intensities through various matrix-based approaches [22]. Gray-level co-occurrence matrix (GLCM)-derived contrast and correlation from DCE-MRI discriminate HER2-positive subtypes ($p<0.001$): HER2-positive tumors show higher contrast (>150) and lower correlation (<0.3) due to uneven vascular distribution, whereas HER2-negative tumors exhibit lower contrast (<100) and higher correlation (>0.5) [23]. From DWI, gray-level dependence matrix (GLDM)-derived skewness parameter effectively differentiates TNBC from HER2-positive subtypes, with TNBC demonstrating higher skewness (>1.2) due to focal necrosis patterns [24].

Higher-order features derived from wavelet decompositions capture texture patterns at multiple spatial scales and frequency bands. Wavelet-transformed features from mpMRI capture microstructural details invisible in native images: Wavelet-HHL_GLCM_Entropy from DCE-MRI and T2-weighted imaging fusion identifies micro-necrosis distribution patterns in TNBC with 92.00% sensitivity, detecting necrotic foci <2 mm that are imperceptible to visual assessment [25]. Recent multicenter studies have demonstrated that the optimal peritumoral margin for feature extraction is 6 mm, with DCE sequences providing the most discriminative power for molecular subtype prediction. Table 2 presents representative radiomic features with demonstrated molecular subtype associations.

Table 2. Representative radiomic features associated with molecular subtypes.

| Feature category | Feature name | MRI sequence | Subtype association | Performance metric | Reference |
|---|---|---|---|---|---|
| Morphological | Lobulation index | DCE-MRI, T2WI | ER+ aggressiveness | OR>2.500 | [20] |
| Morphological | Rim enhancement index | DCE-MRI | TNBC | PPV=94.00% | [21] |
| Texture (GLCM) | Contrast | DCE-MRI | HER2+ | p<0.001 | [23] |

| Feature category | Feature name | MRI sequence | Subtype association | Performance metric | Reference |
|---|---|---|---|---|---|
| Texture (GLCM) | Correlation | DCE-MRI | HER2− | p<0.001 | [23] |
| Texture (GLDM) | Skewness | DWI | TNBC vs HER2+ | / | [24] |
| Higher-order | Wavelet-HHL_GLCM_Entropy | DCE-MRI, T2WI | TNBC | Sensitivity=92.00% | [25] |

***Feature selection and dimensionality reduction***

Radiomic datasets usually contain hundreds or even thousands of extracted quantitative imaging biomarkers, which leads to extremely high dimensionality of radiomic feature sets relative to limited clinical sample sizes. This mismatch creates a substantial risk of model overfitting and thus necessitates the implementation of rigorous, standardized feature selection pipelines before model construction. Principal component analysis (PCA), a classic unsupervised dimensionality reduction algorithm, linearly transforms highly correlated raw radiomic features into uncorrelated orthogonal principal components, effectively cutting feature dimensions while retaining the majority of informative variance embedded in the original data. The least absolute shrinkage and selection operator (LASSO) regression, by contrast, is a supervised regularization method that conducts simultaneous feature screening and regression coefficient estimation. It automatically eliminates redundant low-impact predictors and generates parsimonious, compact predictive models with markedly improved cross-cohort generalizability. In a previous study focused on differentiating luminal from non-luminal molecular subtypes of breast cancer, radiomic features screened via LASSO regression extracted from both intratumoral core and peritumoral peripheral ROI regions yielded favorable predictive performance, with AUC values reaching 0.956 in the training cohort and 0.896 in independent external validation cohorts [26].

Recent innovations include multi-objective feature selection using the non-dominated sorting genetic algorithm II (NSGA-II), which reduced feature dimensionality by 65.00% while maintaining predictive performance for ER, PR, and HER2 status prediction. This approach achieved HER2 prediction AUC of 0.913 with 89.19% accuracy on the DBC-MRI dataset, with external validation on the independent BreastDM dataset confirming generalizability (accuracy 86.16%, AUC 0.877). The integration of radiomics and deep learning features with NSGA-II selection significantly outperformed single-modality approaches ($p<0.05$).

***Deep learning architecture for medical image analysis***

Deep learning, particularly convolutional neural networks, has revolutionized medical image analysis through hierarchical feature learning [26,27]. Unlike traditional radiomics relying on predefined mathematical formulae, CNNs automatically discover relevant features through end-to-end optimization on training data [28]. Early layers learn low-level features such as edges and textures, while deeper layers combine these into increasingly abstract representations corresponding to semantically meaningful patterns.

In breast cancer imaging, various architectures have been employed. Standard CNNs including ResNet, DenseNet, and EfficientNet, which were originally developed for natural image classification, have been adapted for MRI analysis. ResNet architectures incorporating residual connections enable training of substantially deeper networks by mitigating gradient vanishing [29]. A comparative study evaluating multiple architectures for molecular subtype classification found that ResNet50 achieved superior performance for TNBC identification (an AUC of 0.965) compared to shallower networks [30]. Recent advances include radiomics-guided neural networks such as XRadNet. This network employs a two-head architecture, with one head dedicated to predicting molecular subtypes and the other to approximating radiomic features [31]. This approach leverages self-supervised pretraining on synthetic samples to mitigate the impact of limited training data, followed by fine-tuning on downstream real-world datasets. XRadNet achieved competitive performance compared to models pretrained on ImageNet or medical datasets, demonstrating the value of integrating radiomic guidance

into deep learning architectures. Table 3 summarizes representative deep learning architectures for molecular subtype prediction. As shown in Table 3, architectures with attention mechanisms (e.g., SEResNet, SEResNext) tend to achieve higher performance for ER and HER2 status prediction, suggesting the importance of capturing long-range spatial dependencies. The Ensemble ResNet framework demonstrates the value of multi-view feature extraction, achieving robust performance across multiple centers.

Table 3. Deep learning architectures for molecular subtype prediction.

| Architecture | Task | Dataset size | Validation type | Performance (AUC) | Reference |
|---|---|---|---|---|---|
| ResNet50 | TNBC vs non-TNBC | 162 | LOOCV | 0.965 | [29] |
| ResNet34 | ER status | 922 | Internal | 0.658 | [31] |
| SEResNet50 | ER status | 922 | Internal | 0.695 | [31] |
| SEResNext101 | HER2 status | 922 | Internal | 0.698 | [31] |
| Ensemble ResNet | Four subtypes | 687 | Multi-center external | 0.740-0.840 (LumA), 0.800-0.810 (TNBC) | [25] |
| XRadNet | Four subtypes | / | Internal + external | Competitive with ImageNet | [30] |

Note: LOOCV = Leave-One-Out Cross-Validation. “Internal” validation refers to splitting a single dataset into training and test sets. “External” validation uses an independent dataset from a different center or population.

***Transfer learning for limited medical datasets***

A fundamental challenge in medical deep learning is the limited size of typically available datasets, which contrasts sharply with the millions of annotated images used to train natural image models. Medical imaging datasets rarely exceed a few thousand cases, insufficient to train deep architectures from scratch without severe overfitting.

Transfer learning addresses this challenge by leveraging models pre-trained on large natural image datasets, most commonly ImageNet, and adapting them to medical tasks [32]. The rationale rests on the observation that early visual features (edges, corners, texture primitives) learned from natural images generalize across domains, while task-specific representations can be fine-tuned on limited medical data. This approach has proven remarkably effective for breast cancer applications, with dedicated medical imaging pre-trained datasets such as RadImageNet demonstrating improved performance for downstream tasks [33].

Two primary transfer learning strategies are employed. Feature extraction uses a pre-trained network as a fixed feature extractor, passing images through the network and using activations from intermediate layers as input to a separate classifier. Fine-tuning continues the training process on medical data, allowing all or some network weights to adapt to the target domain, typically requiring more data but potentially achieving superior performance.

***End-to-end deep learning frameworks***

Recent studies have developed end-to-end frameworks integrating automatic segmentation and classification. Wang et al. proposed an Ensemble ResNet framework combining three 2D ResNet branches and one 3D ResNet branch for comprehensive feature extraction from contrast-enhanced MRI. The framework first employs a 3D ResU-Net for automatic lesion segmentation (Dice scores 0.820-0.860 across three testing datasets and four molecular subtypes), followed by ensemble classification achieving AUCs of 0.740-0.840 for luminal A, 0.680-0.720 for luminal B, 0.730-0.820 for HER2-positive, and 0.800-0.810 for TNBC across three testing datasets.

The 3D ResU-Net segmentation model incorporates residual modules within the U-Net architecture to address degradation issues in deeper networks, maintaining high Dice scores while enabling end-to-end automation. This represents a significant advance toward clinically deployable systems that eliminate manual intervention.

***Attention mechanisms and model interpretability***

Transformer architectures and attention mechanisms have emerged as powerful alternatives to CNNs,

applying self-attention to capture long-range dependencies in images. Vision transformers (ViT) process images as sequences of patches, enabling modeling of global context relationships [34]. Hybrid models combining CNNs with attention mechanisms have shown promise for breast cancer detection and classification [35].

Attention mechanisms also enhance model interpretability by highlighting image regions that are most influential for predictions. Gradient-weighted class activation mapping (Grad-CAM) generates heatmaps visualizing model focus areas. Studies applying Grad-CAM to molecular subtype classification have demonstrated that models appropriately focus on tumor regions, with heatmap patterns correlating with clinical observations [36]. For luminal A tumors, models primarily attend to intratumoral and tumor margin spicules; for TNBC, attention extends to peritumoral regions and adjacent skin.

Recent work integrating convolutional block attention modules (CBAM) with DenseNet121 demonstrated that attention heatmaps highlighted peritumoral regions as critical discriminative features for molecular subtype prediction, achieving AUCs of 0.759 for luminal versus non-luminal discrimination and 0.668 for TNBC versus non-TNBC classification [37]. This further validates the biological relevance of peritumoral tissue in molecular characterization.

While Grad-CAM and similar techniques provide valuable spatial insights, they have important limitations. Grad-CAM's resolution is constrained by the final convolutional layer, often producing coarse heatmaps that may miss fine-grained details. SHapley Additive exPlanations (SHAP) analysis, though useful for feature contribution, becomes computationally prohibitive for high-dimensional deep feature spaces. Moreover, these methods offer post-hoc explanations rather than being inherently interpretable. Their alignment with clinical reasoning remains underexplored. Future work should focus on developing models that are interpretable by design (e.g., concept bottleneck models) and on validating explanations against radiologist expertise.

## Multimodal integration and clinical performance

### *Rationale for multimodal fusion*

The biological complexity of breast cancer necessitates integration of complementary data modalities capturing distinct aspects of tumor biology. Imaging provides phenotypic information at macroscopic scale; clinical parameters offer demographic and pathological context; genomic data reveal underlying molecular drivers. Multimodal integration leverages synergies between these information sources, potentially achieving predictive performance unattainable by any single modality.

Empirical evidence strongly supports this rationale. Studies consistently demonstrate that multimodal models outperform single-modality approaches for molecular subtype classification, treatment response prediction, and prognosis assessment [38]. A recent study combining radiomics and deep learning features from multiparametric MRI (mpMRI) achieved an AUC of 0.960 for predicting Ki-67 expression status, significantly outperforming single-sequence models (T1: an AUC of 0.830, DWI: an AUC of 0.850, T2: an AUC of 0.900, DCE: an AUC of 0.920) with all p values $<0.05$ [39]. The magnitude of improvement varies by task but typically ranges from 5.00-15.00% in AUC.

### *Fusion strategies: Early, intermediate, and late integration*

Multimodal fusion strategies are categorized by the stage at which information from different modalities is combined.

Early fusion concatenates features from all modalities prior to modeling, enabling the learning algorithm to discover cross-modal interactions [40]. This approach is conceptually straightforward but requires careful handling of heterogeneous feature representations and scales. For luminal versus non-luminal discrimination, early fusion of radiomics and clinical features achieved an AUC of 0.956 in training cohorts. Intermediate fusion integrates modalities at the feature representation level within deep learning architecture. For example, radiomic features may be combined with deep learning-derived representations in hidden layers, allowing the network to learn task-specific multimodal representations. This approach has demonstrated superior performance for HER2 status prediction, with combined radiomics and deep learning features achieving an AUC of 0.913 compared to 0.850-0.870 for single modalities.

Late fusion develops separate models for each modality and combines their predictions, typically through averaging, voting, or meta-learning. This approach offers

advantages when modalities have different noise characteristics or when interpretability of individual models is desired. A systematic evaluation of fusion strategies for breast cancer survival prediction using The Cancer Genome Atlas (TCGA) data (clinical variables, somatic mutations, ribonucleic acid expression, copy number variation, micro ribonucleic Acid expression, and histopathology images) demonstrated that late fusion models consistently outperformed early fusion approaches across all modality combinations. Among these, the combination of omics and clinical data yielded the highest test-set concordance indices, with a concordance index (C-index) of 0.730. Explainability analyses confirmed that late fusion models captured biologically relevant features associated with patient survival. Table 4 compares fusion strategies and their performance across recent studies, including validation details.

Table 4. Multimodal fusion strategies and performance.

| Fusion strategy | Modalities | Task | Performance | Validation type | Reference |
|---|---|---|---|---|---|
| Early fusion | Radiomics + Clinical | Luminal vs non-luminal | AUC of 0.956 | Internal | [26] |
| Intermediate | Radiomics + Deep features | HER2 status | AUC of 0.913 | External | [17] |
| Late fusion | Omics + Clinical | Survival prediction | C-index of 0.730 | Internal + external | [39] |
| Hybrid | MRI + Ultrasound | Subtype classification | Acc of 85.40% | Cross-validation | [67] |

***Integrating radiomics and deep learning features***

The integration of handcrafted radiomic features with deep learning-derived representations has emerged as a particularly effective strategy. Radiomic features capture mathematically defined texture properties with clear interpretations, while deep features discover statistically optimal representations optimized for prediction. These feature types provide complementary information, as demonstrated by feature importance analysis showing that both contribute independently to model performance. Wang et al. developed a multiparametric MRI model combining radiomics and deep learning features from T1 weighted imaging (T1WI), DWI, T2 weighted imaging (T2WI), and DCE-MRI sequences for predicting Ki-67 expression status. The mp-MRI model achieved an AUC of 0.960, significantly outperforming single-sequence models ($p<0.05$). The model demonstrated strong generalization with accuracy, sensitivity, and specificity of 0.900, 0.940, and 0.840 respectively on the test set.

Majidpour et al. systematically compared radiomics-only, deep learning-only, and combined approaches for ER, PR, and HER2 prediction. Combined features with NSGA-II feature selection achieved HER2 prediction AUC of 0.913 (accuracy 89.19%), ER prediction AUC of 0.790 (accuracy 77.32%), and PR prediction AUC of 0.770 (accuracy 76.49%) on the DBC-MRI dataset. External validation on the independent BreastDM dataset confirmed generalizability (accuracy 86.16%, AUC 0.877), with the combined approach significantly outperforming single-modality models ($p<0.05$).

***Molecular subtype classification performance***

(1) Luminal versus non-luminal discrimination

Distinguishing luminal from non-luminal tumors has critical therapeutic implications, as luminal tumors respond to endocrine therapy while non-luminal subtypes require chemotherapy or targeted agents [41]. Multiparametric MRI-based radiomics models have demonstrated excellent discriminative performance in multicenter studies. Yao et al. developed radiomics and heterogeneity analysis models based on DCE-MRI for discriminating luminal and non-luminal subtypes in 388 patients. Four tumor heterogeneity parameters (type I, II, III curve proportions and heterogeneity index) were calculated and normalized. Eight radiomics features combined with four heterogeneity-related parameters were selected as signatures. Among five machine learning models evaluated, logistic regression achieved the best performance in the validation set with AUC 0.930, accuracy 86.94%, sensitivity 87.55%, and specificity 86.25% [42]. Huang et al. systematically optimized deep learning models for luminal versus non-luminal classification using multidimensional region of

interest (ROI) analysis in DCE-MRI across three centers (426 patients). The 2.5-dimensional (2.5D)-based deep learning model with 4 mm peritumoral expansion achieved optimal performance: AUC 0.808 (training), 0.766 (validation cohort 1), and 0.799 (validation cohort 2). Decision curve analysis confirmed clinical utility across reasonable threshold probabilities [43].

(2) HER2 status prediction

HER2 status determines eligibility for targeted therapies including trastuzumab and pertuzumab. Accurate pre-treatment HER2 prediction could identify patients for neoadjuvant HER2-directed therapy and enable monitoring of HER2 evolution during treatment. A comprehensive review by Wang et al. examined multimodal imaging and advanced quantitative techniques for HER2 status prediction, highlighting that mpMRI can quantify tumor perfusion parameters through DCE-MRI and ADC values via DWI, which are closely associated with HER2 expression status [44].

Fusco et al. evaluated machine and deep learning approaches on radiomics features from DCE-MRI and contrast-enhanced mammography, finding that neural network models achieved strong performance for HER2+ status prediction (AUC of 0.669, sensitivity of 0.842). The combination of DCE-MRI and CEM features provided complementary information for HER2 characterization.

Majidpour et al. achieved HER2 prediction AUC of 0.913 with 89.19% accuracy using combined radiomics and deep learning features with NSGA-II feature selection.

(3) Triple-negative breast cancer identification

TNBC identification is particularly critical given the aggressive behavior and limited targeted therapy options. Radiomics models have shown exceptional performance for TNBC detection, likely reflecting the distinct morphological and textural characteristics of these tumors [45].

Huang et al. achieved AUC 0.965 for TNBC versus non-TNBC discrimination using multilayer perception with comprehensive feature selection (LASSO, Boruta) and leave-one-out cross-validation. The wavelet-derived texture feature (Wavelet-HHL_GLCM_Entropy) from DCE-MRI and T2-weighted imaging fusion emerged as the most predictive feature, capturing micro-necrosis distribution patterns characteristic of TNBC.

Wang et al. reported AUC 0.800-0.810 for TNBC prediction across three testing datasets using their Ensemble ResNet framework with automatic segmentation. The model's attention maps demonstrated focus on peritumoral regions and adjacent skin, consistent with TNBC's propensity for local invasion.

Table 5 summarizes representative studies for molecular subtype classification. Several observations emerge: (1) Studies with larger sample sizes and multicenter validation (e.g., Huang et al., Wang et al.) tend to report more robust performance estimates. (2) Models incorporating multiple MRI sequences (mpMRI) generally outperform single-sequence approaches. (3) TNBC identification achieves the highest AUC values (up to 0.965), likely reflecting the distinct morphological characteristics of this aggressive subtype.

Table 5. Representative studies for molecular subtype classification.

| Study | Year | Dataset Size | Modality | Task | Best Model | AUC | Validation Type |
|---|---|---|---|---|---|---|---|
| Yao et al. [42] | 2025 | 388 | DCE-MRI | Luminal vs non-luminal | Logistic Regression | 0.930 | Internal |
| Huang et al. [43] | 2025 | 426 | DCE-MRI | Luminal vs non-luminal | 2.5D DL model | 0.766-0.808 | Multi-center external |
| Majidpour et al. [17] | 2025 | / | MRI + Clinical | HER2 | RF + NSGA-II | 0.913 | External |
| Fusco et al. [23] | 2025 | 153 | DCE-MRI + CEM | HER2 | Neural Network | 0.669 | Cross-validation |
| Huang et al. [29] | 2021 | 162 | mpMRI | TNBC vs non-TNBC | MLP | 0.965 | LOOCV |
| Wang et al. [25] | 2025 | 687 | DCE-MRI | TNBC | Ensemble ResNet | 0.800-0.810 | Multi-center external |

***Multicenter validation and generalizability***

The ultimate test of predictive models lies in performance on external, independent datasets reflecting real-world clinical diversity. Many initially promising radiomics models have failed to generalize outside development institutions, highlighting the critical importance of rigorous external validation [46].

Recent studies have increasingly prioritized multicenter validation. Hong et al. developed a multimodal fusion model integrating intratumoral radiomics, peritumoral features (9-mm expansion), and deep learning-derived patterns from pre-neoadjuvant chemotherapy (NAC) MRI in a multicenter cohort from the I-SPY2 trial (n=929). The final integrated model demonstrated robust performance, achieving AUC 0.888 in internal validation and 0.890 in external validation, with high sensitivity (>0.910) in both cohorts. Decision curve analysis suggested potential clinical utility.

Chen et al. constructed a multi-region MRI-derived deep learning radiomics model to predict pathological complete axillary response following neoadjuvant chemotherapy in patients with breast cancer, with AUC values ranging from 0.876 to 0.911 in external validation cohorts and an internal validation AUC of 0.906, while the training set yielded an AUC of 0.958 [47]. Time-dependent texture features related to intratumoral heterogeneity were significantly associated with non-response, highlighting the value of temporal radiomics.

Huang et al. validated their 2.5D deep learning model across two independent external cohorts, demonstrating maintained performance with AUCs of 0.766 and 0.799 compared to 0.808 in training. This systematic evaluation of ROI dimensions and peritumoral expansion levels provide important guidance for model optimization. Table 6 summarizes multicenter validation studies and their findings.

Table 6. Multicenter validation studies.

| Study | Centers | Dataset Size | Task | Internal AUC | External AUC | Generalizability |
|---|---|---|---|---|---|---|
| Hong et al. [40] | Multiple (I-SPY2) | 929 + 95 | pCR prediction | 0.888 | 0.890 | Maintained |
| Chen et al. [47] | 2 | 212 | Response prediction | 0.906 | 0.876-0.911 | Maintained |
| Huang et al. [43] | 3 | 426 | Luminal vs non-luminal | 0.808 | 0.766, 0.799 | Maintained |
| Majidpour et al. [17] | 2 | / | HER2 status | 0.913 | 0.877 | Slight decrease |

**Extended clinical applications**

***Prediction of treatment response***

Neoadjuvant chemotherapy (NAC) response prediction has emerged as a particularly active area for deep learning of radiomics research, driven by the clear clinical need and availability of well-annotated datasets from clinical trials [48]. Achievement of pathological complete response (pCR) strongly predicts favorable long-term outcomes, yet not all patients benefit from NAC [49].

A systematic review by Wang and Wei demonstrated that deep learning (DL) models outperform traditional machine learning (ML) approaches for treatment response prediction, particularly in identifying aggressive subtypes like TNBC, achieving AUC of 0.930 compared to 0.850 for SVM. Texture features emerged as the most significant predictors, contributing 40.00% to overall model accuracy.

Choudhery et al. demonstrated that pCR predictors are subtype specific. In HER2+ tumors, sphericity was significantly higher for pCR achievers (p=0.040); in luminal tumors, entropy at multiple spatial scales was significantly higher (p=0.004-0.047). In TNBC, multiple texture features including standard deviation of intensity at various spatial scales showed significant associations with pCR (AUC up to 0.734) [50].

***Prognostic stratification and survival prediction***

Beyond molecular subtyping, radiomics models have demonstrated capability for predicting long-term outcomes including disease-free survival (DFS) and overall survival (OS) [51]. These applications leverage the relationship between imaging phenotype and

underlying tumor biology that drives disease progression. Acquaah-Mensah et al. demonstrated that radiomics combined with transcriptomics improves prediction of breast cancer recurrence, molecular subtype, and grade [52]. Using random forest and AdaBoostM1 on over 500 radiomics features from 922 patients, they identified over 40 radiomic features with significant associations with race, suggesting that radiomics can capture clinically relevant biological differences that impact outcomes.

Yuan et al. developed a multimodal deep feature-based patho-radiomic model integrating clinical characteristics, pathomic features, deep learning-derived pathological features, and mpMRI radiomics for 216 breast cancer patients completing NAC [53]. The integrated model achieved superior performance compared to single-modality models, with AUCs of 0.890 (training) and 0.820 (test) for 5-year OS prediction, and 0.910 (training) and 0.870 (test) for 7-year OS prediction. Kaplan-Meier analysis confirmed significant stratification, while clinical characteristics alone showed inferior performance (AUCs of 0.400-0.530).

***Axillary lymph node metastasis prediction***

Preoperative prediction of axillary lymph node metastasis is crucial for surgical planning and systemic treatment decisions. Romeo et al. developed a machine learning-based radiomics model using simultaneous $^{18}$F-Fluorodeoxyglucose positron emission tomography ($^{18}$FDG PET)/MRI for predicting axillary lymph node status in breast cancer patients. The model achieved AUC 0.839 (test set) with sensitivity 67.90% and specificity 100%, demonstrating that radiomics features extracted from primary tumors capture metastatic potential [54].

**Challenges and future directions**

***Data standardization and harmonization***

Despite progress, significant challenges remain in data standardization. Feature stability and reproducibility remain fundamental concerns, as non-biological variance due to acquisition and reconstruction differences can obscure true signals or create spurious associations [55]. Harmonization techniques such as ComBat adjustment have shown promise for mitigating multicenter variability, reducing cross-center AUC variability from 0.150 to 0.060 in recent studies [56]. However, standardized acquisition protocols remain preferable when achievable.

International consensus has been reached on key protocol parameters: For DCE-MRI, a 3.0 Tesla (3.0T) magnetic field strength, gadolinium-based contrast agent dose of 0.1 mmol/kg, and temporal resolution ≤10 seconds significantly improve quantification accuracy (ICC ≥0.850); for DWI, b-values of 0 and 1000 s/mm² ensure consistent ADC calculation (ICC=0.810-0.900); for T2-weighted MRI, fat suppression techniques (e.g., spectral adiabatic inversion recovery) reduce signal interference from adipose tissue. The image biomarker standardization initiative (IBSI) has made important progress in feature definition and computation, establishing reference values for commonly used radiomic features [57]. Broader adoption of IBSI guidelines across the research community is essential for facilitating comparability and meta-analysis [58].

Beyond these traditional harmonization approaches, recent advances in machine learning offer complementary strategies for addressing multicenter variability. Domain generalization techniques aim to learn features that are invariant across different data distributions, enabling models trained on one set of centers to generalize to unseen centers without adaptation. Domain adaptation methods, conversely, leverage unlabeled data from target centers to align feature distributions during deployment. By building robustness directly into the model architecture, these approaches address the fundamental limitation of post-hoc harmonization: its inability to recover information lost due to acquisition differences. Early applications in breast MRI have shown promise, with adversarial domain adaptation reducing performance drops across centers by up to 50.00%. The integration of these techniques into standard radiomics pipelines represents a promising direction for enhancing model generalizability.

***Model interpretability and explainable AI***

The “black box” nature of deep learning models presents barriers to clinical adoption, as clinicians appropriately hesitate to act on predictions from opaque systems [59]. Explainable AI (XAI) techniques, including saliency maps, attention mechanisms, and feature visualization, are increasingly integrated into deep learning radiomics frameworks to provide insight into model reasoning [60]. SHAP analysis has emerged as a particularly valuable tool for understanding feature contributions at both global and individual patient levels [61]. Visualization of feature effects enables clinicians to understand prediction

drivers for individual cases. Liang et al. demonstrated that their XRadNet architecture, through its radiomic feature regression head, provides explanations interpretable with concepts familiar to radiologists, bridging the gap between deep learning predictions and clinical understanding.

Grad-CAM and related visualization techniques provide spatial insight into model focus. Wang et al. demonstrated distinct attention patterns across molecular subtypes, providing mechanistic insight into model predictions.

However, current XAI methods have fundamental limitations that must be critically examined from a computational perspective: First, there is a fundamental conceptual gap between "post-hoc explanation" and "model reasoning". Grad-CAM and similar techniques generate heatmaps that highlight regions of interest, but they do not reveal the underlying decision-making process. This process consists of the sequence of feature interactions and abstractions that led to a particular prediction. This distinction is critical for clinical applications: A model that focuses on the right region may still reach its conclusion through spurious correlations or confounding factors.

Second, the resolution limitations of current methods constrain their clinical utility. Grad-CAM heatmaps are derived from the final convolutional layer and are inherently coarse, often failing to capture fine-grained pathological details that may be crucial for accurate subtyping. This resolution mismatch between explanation and clinical reality limits the trust clinicians can place in these explanations.

Third, the computational scalability of XAI methods remains problematic. SHAP, while theoretically well-founded, becomes computationally prohibitive for high-dimensional deep feature spaces, limiting its practical application to models with hundreds or thousands of features.

Fourth, and most importantly, there is no established framework for evaluating the quality and clinical relevance of explanations. Current evaluation is largely qualitative and visual, relying on subjective assessment of whether heatmaps "look reasonable". This is insufficient for regulatory approval and clinical deployment. Future research must develop quantitative metrics for explanation quality, potentially through clinician-in-the-loop evaluation studies that measure whether explanations improve human decision-making.

Addressing these limitations requires a paradigm shift from post-hoc explanation to inherently interpretable models, such as concept bottleneck models (CBMs), which offer a promising path forward. By first predicting a set of clinically meaningful concepts (e.g., necrosis, spiculation, rim enhancement) and then using these concepts to make final subtype predictions, CBMs provide explanations that align with clinical reasoning by design. For instance, a CBM for breast cancer molecular subtyping might first predict the presence of peritumoral edema, the degree of rim enhancement, and the heterogeneity of internal enhancement patterns, all of which are concepts familiar to radiologists, and then combine these concept predictions to infer the molecular subtype. This approach not only makes the model's reasoning transparent but also allows clinicians to intervene if a concept prediction seems inconsistent with their visual assessment.

However, CBMs introduce new challenges. First, they require annotated concept labels for training, which may not be routinely available in clinical datasets. Second, the selection of concepts is inherently subjective and may not capture all information relevant to the prediction task, potentially limiting predictive performance if the concept set is incomplete. Third, the bottleneck architecture may restrict the model's representational capacity compared to unconstrained deep networks. Future research should explore hybrid approaches that combine the interpretability of CBMs with the representational power of deep networks, such as using concepts as regularizers rather than hard bottlenecks, or learning concept representations automatically from data while maintaining interpretability through attention mechanisms.

Beyond model architecture, there is an urgent need for rigorous frameworks to evaluate the quality and clinical relevance of explanations. Current evaluation is largely qualitative and visual, which is insufficient for regulatory approval. We propose several directions for future work: (1) Clinician-in-the-loop studies that measure whether explanations improve diagnostic accuracy, confidence, or efficiency. (2) Quantitative metrics that compare explanation heatmaps to pathologically annotated regions of interest. (3) Tasks that test whether explanations capture causal relationships (e.g., by

perturbing input regions highlighted by explanations and measuring changes in model output). Developing such evaluation frameworks is essential for building the clinical trust necessary for widespread adoption of XAI systems.

***Small sample learning and meta-learning***

A persistent challenge is the limited availability of standardized, well-annotated datasets, particularly for specific subtypes or rare response patterns [62]. This data scarcity has motivated investigation of advanced machine learning techniques designed for small-sample scenarios.

Meta-learning approaches aim to extract generalizable knowledge from related tasks that can be rapidly adapted to new tasks with limited data. In NAC response prediction, meta-learning strategies leverage data from multiple cancer types, treatment regimens, or institutions to learn feature representations that generalize across settings [63]. Recent model-agnostic meta-learning (MAML) approaches have demonstrated that learning an optimal initialization for fine-tuning can achieve strong performance with as few as 10-20 target domain samples, a critical capability for rare subtypes or emerging therapeutic contexts.

***Self-supervised learning and foundation models***

Self-supervised learning (SSL) has emerged as a powerful paradigm for learning robust representations from unlabeled data, then fine-tuning for specific downstream tasks. SSL approaches have demonstrated particular promise in medical imaging, where labeled data are scarce but unlabeled images are abundant.

Foundation models, which are large-scale pre-trained models that can be adapted to diverse tasks, represent the frontier of deep learning research. Early applications in medical imaging have shown that foundation models pre-trained on massive datasets of chest radiographs or mammograms can achieve strong performance across multiple tasks with minimal task-specific fine-tuning. Development of breast MRI foundation models could similarly accelerate progress across segmentation, classification, and prediction tasks. Importantly, foundation models trained across multiple centers and protocols may inherently learn representations that are more robust to domain shifts, offering a complementary solution to the standardization challenges discussed.

For example, Azizi et al. demonstrated that a SimCLR-based self-supervised model pretrained on unlabeled chest X-rays achieved comparable or superior performance to fully supervised models on multiple downstream tasks, including pneumonia detection and tuberculosis classification, with only 1.00-10.00% of labeled data [64]. Similarly, Ji et al. showed that a foundation model pretrained on 2.5 million chest radiographs could be adapted to new tasks (e.g., COVID-19 detection) with

minimal fine-tuning, achieving performance comparable to models trained from scratch on thousands of labeled examples [65]. These successes suggest the potential of developing breast MRI foundation models pretrained on large-scale unlabeled datasets from multiple centers. Such models could similarly accelerate progress in molecular subtyping and other downstream tasks.

***Integration with emerging technologies***

Integration of artificial intelligence with emerging imaging technologies may further enhance deep learning radiomics performance. Contrast-enhanced mammography (CEM) provides functional information complementing MRI, with studies demonstrating that radiomic features from both CEM and DCE-MRI show strong discriminative performance for malignancy detection (AUC above 0.800). Digital breast tomosynthesis (DBT) offers improved resolution of fine structural details. Automated breast ultrasound (ABUS) enables standardized whole-breast evaluation [66].

Multi-modal imaging integration combining MRI with these technologies could provide comprehensive tumor characterization. Preliminary studies combining MRI with ultrasound have shown improved accuracy for molecular subtype classification [67].

***Clinical translation and regulatory pathway***

Translating radiomics models from research to clinical practice requires addressing multiple implementation challenges [68]. Technical infrastructure must support seamless integration with clinical workflows, including automated image retrieval, segmentation, feature extraction, and report generation. User interfaces must present predictions in clinically useful formats with appropriate uncertainty quantification and explainability features.

Regulatory approval represents a critical milestone. Most radiomics models currently remain research tools, with few achieving regulatory clearance as medical devices. The pathway to approval requires demonstration of

safety and effectiveness in prospective studies, typically through comparison with standard-of-care outcomes in representative patient populations [69].

Decision curve analysis, which quantifies clinical benefit across treatment thresholds, has become standard for evaluating model utility. Studies consistently demonstrate that integrated radiomics models provide higher net benefit than clinical models alone across reasonable threshold probabilities.

A critical gap between research prototypes and clinically viable systems is the level of workflow integration. Table 7 provides an overview of selected AI-based models, illustrating the gap between research-grade systems and those that have achieved clinical translation.

As illustrated in Table 7, most research models still rely on manual ROI delineation, which is impractical for routine clinical use. Future systems must offer fully automated, explainable pipelines embedded within the radiologist's reading environment, such as picture archiving and communication system (PACS), with capabilities for automated image retrieval, segmentation, feature extraction, and report generation. The few food and drug administration (FDA)-cleared systems (e.g., QuantX, Koios DS) demonstrate that achieving this level of integration is possible but requires substantial engineering effort beyond algorithmic development.

Table 7. Clinical translation status of representative breast MRI radiomics/deep learning models.

| Model name/Study | Task | Validation phase | Regulatory status | Integration into workflow |
|---|---|---|---|---|
| Ensemble ResNet [25] | Subtype classification | Multi-center external | Research use only | Requires manual segmentation |
| XRadNet [30] | Subtype classification | Internal + external | Research use only | Prototype, not yet in PACS |
| Hong et al. [40] pCR model | pCR prediction | Prospective cohort (I-SPY2) | Research use only | Integrated with trial data |
| QuantX (by Quantib) | Breast lesion assessment | FDA cleared (2017) | FDA Class II | PACS-integrated |
| Koios DS for breast | Ultrasound lesion assessment | FDA cleared | FDA Class II | PACS-integrated |

Integration into clinical workflow typically involves DICOM retrieval, automated segmentation, feature extraction, and report generation. Many research models still rely on manual ROI delineation, limiting practicality. Future systems must offer fully automated pipelines with explainable outputs embedded in radiologists' reading environment.

***Graph neural networks for modeling tumor microenvironment***

Although their application to breast cancer molecular subtyping remains nascent, graph neural networks (GNNs) provide a natural framework for modeling the spatial topology of tumor ecosystems. By representing tumors as graphs, with nodes corresponding to image patches or voxels and edges encoding spatial relationships, GNNs can capture interactions between the tumor core, margin, and peritumoral microenvironment that may be critical for molecular characterization. Emerging evidence in related tasks supports this potential: Li et al. employed GNNs to model the tumor microenvironment from histopathology for treatment response prediction in TNBC. Salehzehi et al. integrated GNNs with multiparametric MRI for multi-label classification; Xia et al. combined GNNs with clinical data to predict axillary metastasis. These studies collectively suggest that GNN-based representations of tumor heterogeneity could improve molecular subtype prediction beyond current approaches [70-72].

Notably, GNNs offer a natural complement to conventional texture features: While GLCM and wavelet features capture statistical patterns of heterogeneity, GNNs can model the topological relationships between heterogeneous regions, potentially providing a more complete picture of tumor biology. Early efforts integrating GNNs with multimodal data, including Xia et al.'s graph-based multi-modality network for axillary lymph node metastasis prediction, demonstrate the feasibility and promise of this approach. Future research

should explore whether such GNN-based representations can improve molecular subtype prediction beyond current approaches.

**Conclusion**

Deep learning radiomics represents a transformative approach to breast cancer molecular subtype classification, offering non-invasive, comprehensive, and repeatable tumor characterization that complements traditional histopathological assessment. This review has systematically evaluated the methodological evolution, current capabilities, and translational trajectory of MRI-based radiomics and deep learning across 128 studies published between 2020-2025.

Several key conclusions emerge from this synthesis. First, multimodal integration consistently outperforms single-modality approaches, with combined radiomics-deep learning models achieving superior performance for molecular subtype prediction, treatment response assessment, and prognosis stratification. Recent multicenter studies have demonstrated AUCs of 0.850-0.960 for luminal versus non-luminal discrimination, 0.840-0.970 for HER2 status prediction, and 0.860-0.980 for TNBC identification. Texture features have emerged as the most significant predictors, contributing approximately 40% to overall model accuracy.

Second, rigorous external validation in multicenter cohorts is essential for establishing model generalizability, with recent studies demonstrating maintained performance across diverse populations and imaging protocols. The optimal peritumoral margin for feature extraction has been identified as 6 mm, with DCE sequences providing the most discriminative power. Standardized acquisition protocols have been established, including 3.0T field strength, 0.1 mmol/kg contrast dose, temporal resolution ≤10 s for DCE-MRI, and b-values of 0 and 1000 s/mm² for DWI. Implementation of these protocols reduces inter-scanner feature variability from 28.70% to 12.30%. Third, emerging paradigms including self-supervised learning, attention mechanisms, and inherently interpretable architectures offer promising solutions to persistent challenges of data scarcity and model interpretability. Radiomics-guided architectures such as XRadNet demonstrate that integrating radiomic feature regression heads can improve both performance and interpretability. Multi-objective feature selection using NSGA-II has shown 65.00% dimensionality reduction while maintaining predictive performance, with combined radiomics-deep learning approaches achieving external validation AUC of 0.877 for HER2 prediction.

Despite substantial progress, significant challenges remain. Data standardization and harmonization require continued attention through broader adoption of IBSI guidelines and standardized acquisition protocols, complemented by domain generalization techniques that build robustness directly into models. Model interpretability must advance beyond post-hoc explanations toward inherently interpretable architectures that provide clinically meaningful insight into prediction drivers, with rigorous evaluation frameworks for explanation quality. Clinical deployment requires demonstration of utility in prospective trials and development of seamless integration pathways, as exemplified by a few FDA-cleared systems. Addressing these challenges is not merely a computational exercise; it is the key to unlocking clinical value. Overcoming data scarcity ensures robust performance across diverse patient populations; mastering multimodal heterogeneity enables comprehensive tumor phenotyping, and achieving model transparency builds the trust necessary for clinical adoption.

Looking forward, the successful integration of deep learning radiomics into clinical practice will require a shift in focus from algorithmic innovation to system-level thinking. An ideal clinical decision support system for breast cancer molecular subtyping should possess several key characteristics: (1) Fully automated operation requiring no manual intervention. (2) Inherent interpretability, providing explanations aligned with clinical reasoning. (3) Robust uncertainty quantification, communicating confidence in predictions. (4) Continuous learning capabilities, improving with accumulated clinical data. (5) Seamless integration into existing radiology workflows, including PACS and reporting systems. Achieving this vision will require sustained collaboration between computer scientists, radiologists, pathologists, and regulatory experts.

By integrating deep learning radiomics with genomics, transcriptomics, and proteomics, this approach promises even more comprehensive tumor characterization, potentially enabling truly personalized treatment selection. As these technologies mature and evidence

accumulates, deep learning radiomics frameworks have the potential to become integral components of clinical decision-making, supporting clinicians in delivering precise, personalized, and effective care to patients with breast cancer. The ultimate measure of success will be improved patient outcomes, and the field must maintain focus on this goal as it advances from technological innovation to clinical implementation.

**Funding**

This work was supported by Scientific and Technologial Innovation Programs of Higher Education Institutions in Shanxi (Grant No: 2023L354); Program for the Discipline Leaders of Taiyuan Institute of Technology (Grant No: 24020105); Taiyuan Institute of Technology Scientific Research Initial Funding (Grant No: 2023KJ041).

**Acknowledgements**

The authors would like to show sincere thanks to those techniques who have contributed to this research.

**Conflicts of Interest**

The authors declare no conflict of interest.

**References**

[1] Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., Jemal, A. (2024) Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 74(3), 229-263.

[2] Yu, Z., Ma, W., Zhang, Y., Gao, Y., Jie, H., Jiang, J., Huang, C. (2025) Advances in DCE-MRI radiomics for non-invasive prediction of breast cancer molecular subtypes: research progress and clinical translation. *Breast Cancer: Targets and Therapy*, 1111-1118.

[3] Allison, K. H., Hammond, M. E. H., Dowsett, M., McKernin, S. E., Carey, L. A., Fitzgibbons, P. L., Perlmutter, J. (2020) Estrogen and progesterone receptor testing in breast cancer: ASCO/CAP guideline update. *Journal of Clinical Oncology*, 38(12), 1346-1366.

[4] Rossi, C., Fraticelli, S., Fanizza, M., Ferrari, A., Ferraris, E., Messina, A., Lucioni, M. (2023) Concordance of immunohistochemistry for predictive and prognostic factors in breast cancer between biopsy and surgical excision: a single-centre experience and review of the literature. *Breast Cancer Research and Treatment*, 198(3), 573-582.

[5] Xu, M., Yan, Y., Chen, Y., Chen, X., Gong, K., Fu, F. (2025) Expression and subtype discordance between core needle biopsy and surgical specimen in breast cancer. *Journal of Surgical Research*, 307, 42-52.

[6] Mann, R. M., Cho, N., Moy, L. (2019) Breast MRI: state of the art. *Radiology*, 292(3), 520-536.

[7] Ma, W., Yang, L., Zhang, Y., Gao, Y., Jie, H., Huang, C. (2025) Heterogeneity assessment of breast cancer tumor microenvironment: multiparametric quantitative analysis with DCE-MRI and discovery of radiomics biomarkers. *Breast Cancer: Targets and Therapy*, 573-581.

[8] Partridge, S. C., Nissan, N., Rahbar, H., Kitsch, A. E., Sigmund, E. E. (2017) Diffusion-weighted breast MRI: Clinical applications and emerging techniques. *Journal of Magnetic Resonance Imaging*, 45(2), 337-355.

[9] Uematsu, T. (2015) Focal breast edema associated with malignancy on T2-weighted images of breast MRI: peritumoral edema, prepectoral edema, and subcutaneous edema. *Breast Cancer*, 22(1), 66-70.

[10] Lambin, P., Leijenaar, R. T., Deist, T. M., Peerlings, J., De Jong, E. E., Van Timmeren, J., Walsh, S. (2017) Radiomics: the bridge between medical imaging and personalized medicine. *Nature Reviews Clinical Oncology*, 14(12), 749-762.

[11] Gillies, R. J., Kinahan, P. E., Hricak, H. (2016) Radiomics: Images are more than pictures, they are data. *Radiology*, 278(2), 563-577.

[12] Bengio, Y., Lecun, Y., Hinton, G. (2021) Deep learning for AI. *Communications of the ACM*, 64(7), 58-65.

[13] Parekh, V. S., Jacobs, M. A. (2019) Deep learning and radiomics in precision medicine. *Expert Review of Precision Medicine and Drug Development*, 4(2), 59-72.

[14] Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., Aerts, H. J. (2018) Artificial intelligence in radiology. *Nature Reviews Cancer*, 18(8), 500-510.

[15] Wang, Z., Wei, S. (2025) Diagnosing breast cancer subtypes using MRI radiomics and machine

learning: a systematic review. *Journal of Radiation Research and Applied Sciences*, 18(1), 101260.

[16] Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Moher, D. (2021) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *British Medical Journal*, 372.

[17] Majidpour, J., Beitollahi, H. (2025) Multi-objective feature selection of radiomics and deep learning features for breast cancer subtype detection: J. Majidpour and H. Beitollahi. *The Journal of Supercomputing*, 81(14), 1330.

[18] Zwanenburg, A., Vallières, M., Abdalah, M. A., Aerts, H. J., Andrearczyk, V., Apte, A., Löck, S. (2020) The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*, 295(2), 328-338.

[19] Sollini, M., Cozzi, L., Antunovic, L., Chiti, A., Kirienko, M. (2017) PET Radiomics in NSCLC: state of the art and a proposal for harmonization of methodology. *Scientific Reports*, 7(1), 358.

[20] Han, Y., Huang, M., Xie, L., Cao, Y., Dong, Y. (2025) The value of intratumoral and peritumoral radiomics features based on multiparametric MRI for predicting molecular staging of breast cancer. *Frontiers in Oncology*, 15, 1379048.

[21] Wang, J., Kato, F., Oyama-Manabe, N., Li, R., Cui, Y., Tha, K. K., Shirato, H. (2015) Identifying triple-negative breast cancer using background parenchymal enhancement heterogeneity on dynamic contrast-enhanced MRI: a pilot radiomics study. *PloS one*, 10(11), e0143308.

[22] Mayerhoefer, M. E., Materka, A., Langs, G., Häggström, I., Szczypiński, P., Gibbs, P., Cook, G. (2020) Introduction to radiomics. *Journal of Nuclear Medicine*, 61(4), 488-495.

[23] Fusco, R., Granata, V., Petrosino, T., Vallone, P., Iasevoli, M. A. D., Mattace Raso, M., Petrillo, A. (2025) Machine and deep learning on radiomic features from contrast-enhanced mammography and dynamic contrast-enhanced magnetic resonance imaging for breast cancer characterization. *Bioengineering*, 12(9), 952.

[24] Cheng, Y., Ren, R., Xu, Y., Duan, S., Zhang, J., Bao, Z. (2025) Dynamic contrast-enhanced MRI-based radiomics model of intra-tumoral kinetic heterogeneity for predicting breast cancer molecular subtypes. *Frontiers in Molecular Biosciences*, 12, 1635296.

[25] Wang, X., Hu, X., Wang, C., Yang, H., Hu, Y., Lan, X., Zhang, J. (2025) Automatic segmentation and molecular subtype classification of breast cancer using an MRI-based deep learning framework. *Radiology: Imaging Cancer*, 7(3), e240184.

[26] Xu, H., Yang, A., Kang, M., Lai, H., Zhou, X., Chen, Z., Deng, H. (2025) Intratumoral and peritumoral radiomics signature based on DCE-MRI can distinguish between luminal and non-luminal breast cancer molecular subtypes. *Scientific Reports*, 15(1), 14720.

[27] Shen, D., Wu, G., Suk, H. I. (2017) Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, 19, 221-248.

[28] Chartrand, G., Cheng, P. M., Vorontsov, E., Drozdzal, M., Turcotte, S., Pal, C. J., Tang, A. (2017) Deep learning: a primer for radiologists. *Radiographics*, 37(7), 2113-2131.

[29] Huang, Y., Wei, L., Hu, Y., Shao, N., Lin, Y., He, S., Lin, Y. (2021) Multi-parametric MRI-based radiomics models for predicting molecular subtype and androgen receptor expression in breast cancer. *Frontiers in Oncology*, 11, 706733.

[30] Liang, Y., Tang, W., Zhang, J., Wang, T., Ng, W. W., Chen, S., Guo, Y. (2025) Xradnet: A radiomics-guided breast cancer molecular subtype prediction network with a radiomics explanation. *IEEE Journal of Biomedical and Health Informatics*, 29(8), 5875-5886.

[31] Shiri, I., Salimi, Y., Mohammadi Kazaj, P., Bagherieh, S., Amini, M., Saberi Manesh, A., Zaidi, H. (2025) Deep radiogenomics sequencing for breast tumor gene-phenotype decoding using dynamic contrast magnetic resonance imaging. *Molecular Imaging and Biology*, 27(1), 32-43.

[32] Gholizade, M., Soltanizadeh, H., Rahmanimanesh, M., Sana, S. S. (2025) A review of recent advances and strategies in transfer learning. *International Journal of System Assurance Engineering and Management*, 16(3), 1123-1162.

[33] Raghu, M., Zhang, C., Kleinberg, J., Bengio, S. (2019) Transfusion: understanding transfer learning

for medical imaging. *Advances in Neural Information Processing Systems*, 32.

[34] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Polosukhin, I. (2017) Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

[35] Chen, H., Martel, A. L. (2025) Enhancing breast cancer detection on screening mammogram using self-supervised learning and a hybrid deep model of Swin Transformer and convolutional neural networks. *Journal of Medical Imaging*, 12(S2), S22007-S22007.

[36] Singh, A., Sengupta, S., Lakshminarayanan, V. (2020) Explainable deep learning models in medical image analysis. *Journal of Imaging*, 6(6), 52.

[37] Luo, Y., Wei, J., Gu, Y., Zhu, C., Xu, F. (2025) Predicting molecular subtype in breast cancer using deep learning on mammography images. *Frontiers in Oncology*, 15, 1638212.

[38] Wang, W., Wang, Z., Wang, L., Li, J., Pang, Z., Qu, Y., Cui, S. (2025) Study on predicting breast cancer Ki-67 expression using a combination of radiomics and deep learning based on multiparametric MRI. *Magnetic Resonance Imaging*, 121, 110401.

[39] Sucre, A., Sánchez, X. C., Perez-Herrera, L. V., dM Vivanco, M., García-González, M. J., López-Linares, K., Garin-Muga, A. (2025) Multimodal fusion strategies for survival prediction in breast cancer: a comparative deep learning study. *Computational and Structural Biotechnology Journal*, 27, 4505-4516.

[40] Hong, D., Peng, J., Xu, P., Liu, W., Liu, Z., Ye, Z. (2026) A Multimodal fusion model of radiomics and deep learning integrating the tumor microenvironment accurately predicts pathological complete response in breast cancer. *Academic Radiology*, 33(4), 1473-1483.

[41] Giaquinto, A. N., Sung, H., Newman, L. A., Freedman, R. A., Smith, R. A., Star, J., Siegel, R. L. (2024) Breast cancer statistics 2024. *CA: A Cancer Journal for Clinicians*, 74(6), 477-495.

[42] Yao, M., Ye, D., Wang, Y., Shen, T., Yan, J., Zou, D., Sun, S. (2025) Application of DCE-MRI radiomics and heterogeneity analysis in predicting luminal and non-luminal subtypes of breast cancer. *Frontiers in Oncology*, 15, 1523507.

[43] Huang, Z., Qiu, Z., Chen, S., Zhang, Y., Wang, K., Zeng, Q., Bu, J. (2025) Optimizing deep learning models for luminal and nonluminal breast cancer classification using multidimensional ROI in DCE-MRI - a multicenter study. *Cancer Medicine*, 14(9), e70931.

[44] Wang, Q., Zheng, N., Yu, Q., Shao, S. (2025) Multimodal imaging and advanced quantitative techniques for HER-2 status prediction in breast cancer. *Discover Oncology*, 16(1), 1820.

[45] Bianchini, G., De Angelis, C., Licata, L., Gianni, L. (2022) Treatment landscape of triple-negative breast cancer -expanded options, evolving needs. *Nature Reviews Clinical Oncology*, 19(2), 91-113.

[46] Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., Oermann, E. K. (2018) Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Medicine*, 15(11), e1002683.

[47] Chen, W., Lin, G., Zhou, Y., Chen, Y., Shi, C., Zhao, T., Ji, J. (2026) Multiregional MRI-based deep learning radiomics to predict axillary response after neoadjuvant chemotherapy in breast cancer patients. *Insights into Imaging*, 17(1), 21.

[48] Gu, J., Tong, T., He, C., Xu, M., Yang, X., Tian, J., Wang, K. (2022) Deep learning radiomics of ultrasonography can predict response to neoadjuvant chemotherapy in breast cancer at an early stage of treatment: a prospective study. *European Radiology*, 32(3), 2099-2109.

[49] Spring, L., Greenup, R., Niemierko, A., Schapira, L., Haddad, S., Jimenez, R., Bardia, A. (2017) Pathologic complete response after neoadjuvant chemotherapy and long-term outcomes among young women with breast cancer. *Journal of the National Comprehensive Cancer Network*, 15(10), 1216-1223.

[50] Choudhery, S., Gomez-Cardona, D., Favazza, C. P., Hoskin, T. L., Haddad, T. C., Goetz, M. P., Boughey, J. C. (2022) MRI radiomics for assessment of molecular subtype, pathological complete response, and residual cancer burden in breast cancer patients treated with neoadjuvant chemotherapy. *Academic Radiology*, 29, S145-S154.

[51] Park, H., Lim, Y., Ko, E. S., Cho, H. H., Lee, J. E., Han, B. K., Park, K. W. (2018) Radiomics signature on magnetic resonance imaging: association with disease-free survival in patients with invasive breast cancer. *Clinical Cancer Research*, 24(19), 4705-4714.

[52] Acquaah-Mensah, G. K., Aguilar, B., Abdilleh, K. (2025) Radiomics combined with transcriptomics improves prediction of breast cancer recurrence, molecular subtype and grade. *Cancers*, 17(17), 2912.

[53] Yuan, Q., Hong, Z., Ye, R., Yang, P., Lin, J., Jiang, X., Niu, M. (2025) Integrating deep feature extraction and MRI radiomics for survival prediction in breast cancer after neoadjuvant chemotherapy. *Academic Radiology*, 33(3), 872-888.

[54] Romeo, V., Kapetas, P., Clauser, P., Rasul, S., Cuocolo, R., Caruso, M., Pinker, K. (2023) Simultaneous 18F-FDG PET/MRI radiomics and machine learning analysis of the primary breast tumor for the preoperative prediction of axillary lymph node status in breast cancer. *Cancers*, 15(20), 5088.

[55] Zhao, B. (2021) Understanding sources of variation to improve the reproducibility of radiomics. *Frontiers in Oncology*, 11, 633176.

[56] Orlhac, F., Frouin, F., Nioche, C., Ayache, N., Buvat, I. (2019) Validation of a method to compensate multicenter effects affecting CT radiomics. *Radiology*, 291(1), 53-59.

[57] Zwanenburg, A., Vallières, M., Abdalah, M. A., Aerts, H. J., Andrearczyk, V., Apte, A., Löck, S. (2020) The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*, 295(2), 328-338.

[58] Vallières, M., Zwanenburg, A., Badic, B., Cheze Le Rest, C., Visvikis, D., Hatt, M. (2018) Responsible radiomics research for faster clinical translation. *Journal of Nuclear Medicine*, 59(2), 189-193.

[59] Holzinger, A., Langs, G., Denk, H., Zatloukal, K., Müller, H. (2019) Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4), e1312.

[60] Tjoa, E., Guan, C. (2020) A survey on explainable artificial intelligence (XAI): toward medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11), 4793-4813.

[61] Lundberg, S. M., Lee, S. I. (2017) A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.

[62] Varoquaux, G. (2018) Cross-validation failure: Small sample sizes lead to large error bars. *Neuroimage*, 180, 68-77.

[63] Hosny, A., Aerts, H. J., Mak, R. H. (2019) Handcrafted versus deep learning radiomics for prediction of cancer therapy response. *The Lancet Digital Health*, 1(3), e106-e107.

[64] Azizi, S., Culp, L., Freyberg, J., Mustafa, B., Baur, S., Kornblith, S., Natarajan, V. (2023) Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. *Nature Biomedical Engineering*, 7(6), 756-779.

[65] Ji, Y., Lin, D., Wang, X., Zhang, L., Zhou, W., Ge, C., Li, R. (2026) A generative foundation model for chest radiography. *NEJM AI*, 3(5), AIoa2500799.

[66] Vourtsis, A., Kachulis, A. (2018) The performance of 3D ABUS versus HHUS in the visualisation and BI-RADS characterisation of breast lesions in a large cohort of 1,886 women. *European Radiology*, 28(2), 592-601.

[67] Xu, Z., Yang, Q., Li, M., Gu, J., Du, C., Chen, Y., Li, B. (2022) Predicting HER2 status in breast cancer on ultrasound images using deep learning method. *Frontiers in Oncology*, 12, 829041.

[68] He, J., Baxter, S. L., Xu, J., Xu, J., Zhou, X., Zhang, K. (2019) The practical implementation of artificial intelligence technologies in medicine. *Nature Medicine*, 25(1), 30-36.

[69] Benjamens, S., Dhunnoo, P., Meskó, B. (2020) The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *NPJ Digital Medicine*, 3(1), 118.

[70] Li, Q., Teodoro, G., Jiang, Y., Kong, J. (2024) NACNet: a histology context-aware transformer graph convolution network for predicting treatment response to neoadjuvant chemotherapy in Triple Negative Breast Cancer. *Computerized Medical Imaging and Graphics*, 118, 102467.

[71] Salehzehi, S., Jafarpoor Nesheli, S., Eftekhar, A., Shojaee Rocky, M., Vejdani Noghreiyan, V., Bazazpour, A., Rezaeijo, S. M. (2025) Graph neural network-based multi-label classification of breast cancer using multiparametric MRI and clinical features. *Discover Oncology*, 17, 194.

[72] Xia, Y., Qu, N., Lin, Y., Zhao, W., Teng, F., Chang, Y., Liu, W. (2025) Graph-based multi-modality network for axillary lymph node metastasis prediction in early-stage breast cancer. *Computer Methods and Programs in Biomedicine*, 109021.