

DEEP LEARNING METHODS FOR BREAST CANCER DIAGNOSIS: A STRUCTURED CRITICAL REVIEW OF ARCHITECTURES

¹Mustafa Raad Ali^{1,2}Normy Norfiza Abdul Razak¹, Rashid Jan³

^{1 1} Institute of Energy Infrastructure (IEI), Department of Civil Engineering, College of Engineering, Universiti Tenaga Nasional (UNITEN), Putrajaya Campus, Jalan IKRAM-UNITEN, 43000 Kajang, Selangor, Malaysia

^{1 2} AL-TURATH UNIVERSITY, 10001, BAGHBAD, IRAQ

^{1 3} Department of Mathematics, College of Science, Qassim University, Buraydah 51452, Saudi Arabia.

^{1 *} Corresponding author: m99448910@gmail.com

Abstract

This structured critical review evaluates the application of deep learning techniques in breast cancer diagnosis, emphasizing a rigorous qualitative and methodological appraisal of validation practices. While Convolutional Neural Networks (CNNs), Attention Mechanisms, and Vision Transformers have demonstrated strong performance in retrospective studies processing complex imaging modalities (mammography, ultrasound, MRI, and histopathology), their transition to clinical deployment remains severely hindered. This review critically discusses recurring methodological vulnerabilities prevalent in contemporary literature, most notably data leakage via image-level dataset splitting, the absence of independent external validation, and the consequent inflation of diagnostic accuracy. By highlighting these systemic bottlenecks, we propose a paradigm shift toward inherent model interpretability, federated learning, and strict patient-level dataset partitioning to ensure the safe clinical translation of artificial intelligence in oncology.

Index Terms - Breast Cancer, Deep Learning, CNN, Attention Mechanisms, Data Leakage, Model Validation, Medical Imaging, External Validation, Patient-Level Split, Vision Transformers.

Introduction

Deep neural networks have catalyzed a paradigm shift in medical image analysis, providing sophisticated capabilities for extracting hierarchical features from complex pathological data. In breast cancer diagnosis, artificial intelligence (AI) has rapidly evolved from traditional, hand-crafted feature extraction methods to end-to-end deep learning architectures. Specifically, Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and recently, Vision Transformers (ViTs), have shown exceptional theoretical performance across diverse imaging modalities, including full-field digital mammography, ultrasound, magnetic resonance imaging (MRI), and histopathological whole-slide imaging [1], [2], [3], [4], [5].

Despite the proliferation of highly parameterized models reporting diagnostic accuracies frequently exceeding 95% [6], [7], [8], the integration of these systems into active clinical workflows remains negligible. This discrepancy between experimental success and clinical reality is driven by profound methodological flaws that are often overlooked in traditional reviews. High classification metrics are frequently artifacts of improperly partitioned datasets, homogeneous training cohorts, and an over-reliance on limited public datasets, rather than indicators of true biological feature learning.

Contribution of this review: To bridge the gap between experimental algorithms and clinical utility,

this review contributes the following: 1) An organized appraisal of deep learning architectures categorized by modality and validation risk. 2) A critical methodological critique of prevalent vulnerabilities, specifically data leakage and the absence of external validation. 3) Editorial and methodological recommendations for the minimum acceptable reporting standards in breast cancer AI studies.

Review Scope and Literature Selection

The scope of this structured critical review is specifically aimed at evaluating deep learning architectures and their corresponding validation protocols in breast cancer imaging. The literature selection covers representative studies published in peer-reviewed journals and computational conferences between 2018 and 2024, with targeted update checks up to 2026 for emerging architectures. Certain foundational studies prior to 2018 are included strictly to provide historical context for architectural evolution.

This review utilizes a structured purposive selection strategy to isolate informative and methodologically distinct papers across major modalities. Studies were retained when providing explicit data on imaging modalities, architectures, and validation partitioning. A representative body of primary papers was evaluated to support our structural and qualitative arguments regarding algorithmic deployment risks.

Unlike recent umbrella reviews that primarily aggregate and celebrate high diagnostic accuracies [9], this qualitative synthesis re-evaluates the foundational methodologies of the extracted studies. We critically analyze how the high accuracies reported in contemporary literature intersect with the pervasive issue of data leakage.

Breast Cancer Imaging Datasets

Robust deep learning necessitates large, meticulously annotated datasets. The performance of deep neural networks in breast cancer diagnosis is intrinsically tied to the characteristics, volume, and modality of the training data. This review identifies four primary imaging modalities utilized across the literature: mammography, ultrasound, MRI, and histopathology [9], [10], [11].

While public datasets have catalyzed early architectural innovations and provided necessary benchmarks, their homogenous nature presents a formidable barrier to clinical generalization. Models trained exclusively on specific public cohorts frequently fail to maintain their reported accuracy when applied to private, multi-institutional clinical data—a phenomenon known as domain shift [12], [13], [14]. Table I summarizes the most utilized public benchmark datasets and their typical validation risks.

TABLE I: PUBLIC BREAST CANCER IMAGING DATASETS USED IN DEEP LEARNING STUDIES

Dataset	Modality	Approx. Size	Validation Risk / Limitations
CBIS-DDSM	Mammography	2,620 images	Older digitized films, high domain shift
INbreast	Mammography	410 images	Small sample size, severe overfitting risk
MIAS	Mammography	322 images	Extremely small, outdated resolution

Dataset	Modality	Approx. Size	Validation Risk / Limitations
BUSI	Ultrasound	780 images	Limited institutional diversity
BreakHis	Histopathology	7,909 images	Patient-level split often reported unclearly
BACH	Histopathology	400 images	High imbalance in microscopic classes
TCGA- BRCA	Multi-omics / WSI	Variable	High genomic heterogeneity, non-standard MRI cohort

Deep Learning Architectures in Breast Cancer Diagnosis

The evolution of diagnostic AI has seen a rapid succession of network architectures. This section details the primary architectural families applied to breast imaging.

CNN and Transfer Learning

CNNs remain the foundational backbone for extracting localized spatial features in breast imaging. Due to the limited size of medical datasets, transfer learning from ImageNet-pretrained models (e.g., ResNet, DenseNet, EfficientNet) is the ubiquitous standard. For instance, in mammography and histopathology, standardizing architectures like ResNet-50 provides a highly effective baseline for mass classification without requiring massive data collection [10].

Hybrid CNN-RNN Models

While standard CNNs excel at static 2D images, they struggle with sequential or 3D contexts. Hybrid architectures leverage CNNs for spatial feature extraction and RNNs (LSTMs/GRUs) for temporal or volumetric context modeling. This is particularly crucial for Dynamic Contrast-Enhanced MRI (DCE-MRI), where analyzing the temporal enhancement kinetics across sequential frames is essential for differentiating benign from malignant lesions [15], [16].

Attention Mechanisms and Vision Transformers

To address the limitations of localized receptive fields in CNNs, Attention Mechanisms and Vision Transformers (ViTs) have emerged as powerful tools capable of modeling long-range global dependencies. However, because ViTs lack the inductive biases (like translation invariance) inherent to CNNs, they require immense pretraining data. ViTs may perform well in histopathology when sufficient pretraining or large-scale datasets are available, though their clinical utility remains constrained by data volume requirements [17], [18], [19].

Lightweight and Edge-Oriented Networks

For point-of-care ultrasound devices or resource-constrained clinical environments, lightweight models such as MobileNet, ShuffleNet, and SqueezeNet have been extensively explored. These networks employ depthwise separable convolutions to radically reduce parameter counts and computational overhead. Studies demonstrate that while they maintain acceptable sensitivity for initial screening, they may experience slight diagnostic degradation in highly complex cases compared to deep DenseNet counterparts [20], [21].

Comparative Analysis of Architectures

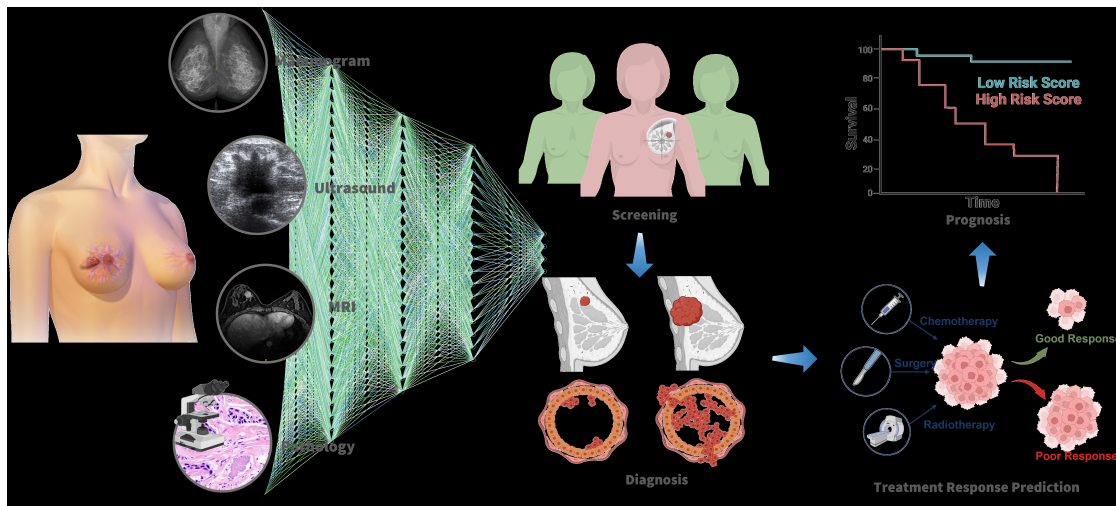


Fig. 1: Author-generated schematic summarizing common feature-extraction pipelines.

The evolutionary shift from localized spatial feature extraction (CNNs) to global contextual modeling (ViTs) presents a complex trade-off between diagnostic accuracy and data efficiency. CNNs inherently possess inductive biases—such as translation invariance and local connectivity—which allow them to generalize effectively even when trained on the relatively small datasets typical of medical imaging.

Conversely, Vision Transformers lack these inductive biases, rendering them highly data-hungry. Thus, the selection of architecture must be strictly dictated by dataset volume rather than theoretical complexity. Table II provides a comprehensive overview of various modeling paradigms.

TABLE II: COMPARATIVE OVERVIEW OF ARCHITECTURE FAMILIES

Family		Primary Strength	Primary Weakness		Validation Risk	
CNNs		Local features, data efficient	Limited context	global	Overfitting cohorts	small
ViTs		Global context modeling	Highly data hungry		Requires pretraining	large
Hybrid	CNN-LSTM	Volumetric/sequence modeling	High cost	computational	Slice-level leakage	
YOLO / Det.	Object	Real-time mass localization	High annotation cost		Bounding box bias	
Lightweight CNNs		Low computational overhead	Slight accuracy drop		Underfitting	
Self-Supervised		Unlabeled data utilization	Complex pipeline		Pretraining instability	

Critical Methodological Challenges

Despite models reporting accuracies that often surpass human radiologist baselines, the clinical translation of deep learning systems in breast oncology remains stymied. A structured qualitative synthesis of the literature reveals systemic methodological flaws that artificially inflate performance metrics, limiting their reliability in real-world clinical settings.

Small Datasets and Class Imbalance

Deep neural networks require massive, balanced datasets to learn generalizable features. In breast cancer imaging, pathological findings are inherently less frequent than benign tissue. Many reviewed studies attempt to resolve this via aggressive synthetic data generation or simplistic oversampling. However, applying these augmentations before partitioning the dataset frequently results in synthetic data leakage [22], [23], [24].

Data Leakage via Image-Level Splitting

One of the most consequential flaws identified in contemporary literature is data leakage resulting from improper dataset partitioning. Deep learning models must be validated on data entirely unseen during training. In clinical datasets, a single patient often contributes multiple images. When datasets are randomly divided into training and testing sets at the *image level*, images from the same patient inevitably appear in both sets. Consequently, the neural network learns to recognize patient-specific anatomical landmarks rather than generalized pathological features. True validation requires strict *patient-level splitting*.

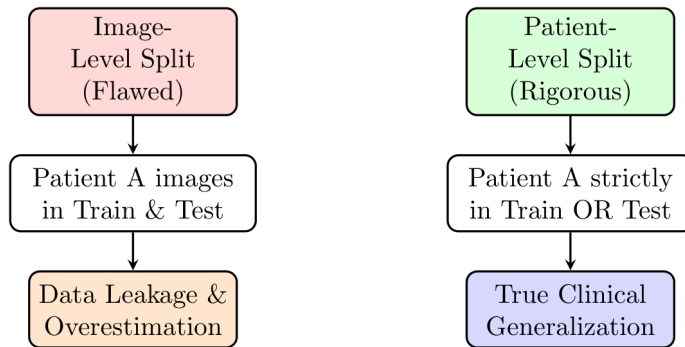


Fig. 2: The critical distinction between image-level and patient-level dataset partitioning.

Lack of Independent External Validation

Many reviewed studies achieving high accuracy are validated exclusively on retrospective, internal datasets. While cross-validation provides an estimate of performance, it does not account for domain shifts caused by different imaging hardware or hospital protocols. Without independent external validation, high internal accuracy is frequently an indicator of domain overfitting [25], [26], [27].

Accuracy Inflation and Misleading Metrics

The pervasive reliance on raw 'Accuracy' as the primary evaluation metric is deeply flawed given natural class imbalances. Rigorous evaluation demands reporting Sensitivity, Specificity, AUC, and F1-score to genuinely assess predictive capability on minority classes.

The Black Box Dilemma and Explainability

The opaque nature of complex architectures limits clinical trust. While methods like Gradient-weighted Class Activation Mapping (Grad-CAM) provide rudimentary heatmaps and serve as helpful auxiliary tools, they often highlight spurious background correlations (e.g., surgical markers) instead of true pathology, meaning they cannot substitute for causal clinical reasoning.

TABLE III: COMMON METHODOLOGICAL FLAWS AFFECTING CLINICAL RELIABILITY

Methodological Flaw	Clinical Consequence
Image-level Splitting	Induces data leakage; memorizes patient anatomy.

Methodological Flaw	Clinical Consequence
Internal Validation Only	Model overfits to single hospital protocol.
Metric Inflation	Overestimates performance by relying on raw accuracy.
No Subgroup Analysis	Fails to verify fairness across diverse demographics.
No Calibration Analysis	Output probabilities do not reflect true clinical risk.
No Confidence Intervals	Masks the statistical variance of the model.
No Code/Data Availability	Prevents independent reproducibility and auditing.
Evidence-Based Appraisal of Validation Methodologies	
To evaluate the structural integrity of contemporary deep learning deployment in breast oncology, a critical synthesis of validation configurations was conducted across the selected primary literature. The analysis indicates that many reported high-performing models remain dependent on retrospective, internally validated datasets, which limits their clinical generalizability. A primary technical vulnerability resides in image-level or patch-level dataset partitioning. In clinical scenarios where a single patient contributes multiple views, bilateral mammograms, histopathology patches, or multi-slice DCE-MRI sequences, random shuffling at the image or patch level can route highly correlated samples from the same subject into both training and testing subsets. Consequently, the network may learn patient-specific anatomical landmarks, scanner signatures, or acquisition artifacts rather than generalized pathophysiological features. This validation weakness is particularly important in breast cancer imaging because public datasets are often limited in size, institutional diversity, and demographic variability. Therefore, high retrospective accuracy should not be interpreted as clinical readiness unless supported by patient-level partitioning, external validation, transparent reporting of evaluation metrics, and prospective or multi-institutional testing. True clinical translation requires a transition from retrospective benchmark optimization toward patient-stratified validation and cross-institutional auditing protocols.	
Clinical Translation and Future Directions	
To transition deep learning from an experimental computational exercise to an active clinical diagnostic tool, future research must fundamentally pivot from chasing raw accuracy to ensuring algorithmic robustness. The following directions are imperative:	
Federated Learning	
To combat the limitations of single-institution datasets without violating patient privacy laws, Federated learning frameworks may provide a practical route forward. FL allows multiple hospitals to train a shared global model collaboratively while keeping raw imaging data strictly localized, drastically improving domain generalization.	
Mandatory Patient-Level Validation	
Journal editorial boards should increasingly require strict reporting standards. At minimum, studies should clearly report patient-level data partitioning; where feasible, external validation should be prioritized over internal cross-validation alone to provide genuine proof of generalizability.	

Regulatory and Clinical Workflow Integration

For models to be adopted, developers must consider the realities of the clinical environment. Future algorithms should be designed with Picture Archiving and Communication Systems (PACS) integration in mind, targeting FDA or CE-cleared regulatory pathways. Studies must gradually shift from retrospective evaluations to prospective, radiologist-in-the-loop workflow analyses.

Limitations of the Current Review

While this review strictly follows a methodological appraisal framework, several limitations must be acknowledged. First, the literature search was restricted to articles published in English, which may exclude relevant methodologies reported in other languages. Second, the rapid and continuous evolution of foundational models (Foundation Models) and multimodal architectures up to mid-2026 means that any purely quantitative synthesis is inherently subject to rapid obsolescence. Consequently, this review prioritizes conceptual methodological flaws—such as data leakage—over a transient ranking of state-of-the-art accuracies or a rigid quantitative indexing of surveyed papers.

Conclusion

Deep learning has shown promising performance in retrospective breast cancer imaging studies. However, its clinical reliability remains constrained by limited external validation, inconsistent reporting of patient-level data partitioning, and overreliance on accuracy-based evaluation. Future studies should prioritize robust validation, transparent reporting, clinically meaningful metrics, and interpretable model behavior. This review argues that methodological reliability, rather than marginal improvements in retrospective accuracy, should become the primary criterion for evaluating breast cancer AI systems [28], [29].

References

- [1] T. Navghare, A. Muley, and V. Jadhav, "Classification of breast cancer patients using deep learning techniques," *Indian Journal Of Science And Technology*, 2024.
- [2] Y. Alashban, "Breast cancer detection and classification with digital breast tomosynthesis: A two-stage deep learning approach," *Diagnostic and Interventional Radiology*, 2024.
- [3] S. Alkentar and A. Assalem, "A hybrid multistage deep learning system for breast cancer classification," *Sylwan*, 2024.
- [4] M. Gagliardi, T. Ruga, E. Zumpano, and E. Vocaturo, "Breast cancer classification via deep learning approaches," *SPAST Reports*, 2024.
- [5] L. Sun and S. Li, "A study of breast cancer classification algorithms by fusing machine learning and deep learning," *Applied Sciences*, 2023.
- [6] T. Swapna, "Enhanced image segmentation in breast cancer classification leveraging deep learning," *Communications on Applied Nonlinear Analysis*, 2024.
- [7] J. R. Fernandis, "ALOA: Ant lion optimization algorithm-based deep learning for breast cancer classification," *Multimedia Research*, 2021.
- [8] K. S. Manojee, "Patho-net: Enhancing breast cancer classification using deep learning and explainable artificial intelligence," *American Journal of Cancer Research*, 2025.
- [9] Z. Liu, "Improving breast cancer classification using histopathology images through deep learning," *Applied and Computational Engineering*, 2024.

- [10] K. J. Geras *et al.*, “High-resolution breast cancer screening with multi-view deep convolutional neural networks,” *IEEE Transactions on Medical Imaging*, 2018.
- [11] H. Mewada, “Extended deep-learning network for histopathological image-based multiclass breast cancer classification using residual features,” *Symmetry*, 2024.
- [12] H. Ding and Z. Dong, “A novel multimodal framework integrating pathomics, deep learning, and machine learning for breast cancer histological grades classification,” *Diagnostic Pathology*, 2026.
- [13] J. Abdollahi, N. Davari, Y. Panahi, and M. Gardaneh, “Detection of metastatic breast cancer from whole-slide pathology images using an ensemble deep-learning method,” *Archives of Breast Cancer*, 2022.
- [14] shak PACAL, “Deep learning approaches for classification of breast cancer in ultrasound (US) images,” *Idr niversitesi Fen Bilimleri Enstitüsü Dergisi*, 2022.
- [15] M. U. Dalmis *et al.*, “Using deep learning to segment breast and fibroglandular tissue in MRI volumes,” *Medical Image Analysis*, 2017.
- [16] Z. Zhu, E. Albadawy, A. Saha, J. Zhang, M. R. Harowicz, and M. A. Mazurowski, “Deep learning for identifying radiogenomic associations in breast cancer,” *IEEE Transactions on Medical Imaging*, 2020.
- [17] R. J. Chen *et al.*, “Scaling vision transformers to gigapixel images via hierarchical self-supervised learning,” *Medical Image Analysis*, 2022.
- [18] Y. Gao, M. Zheng, X. Wang, J. Shen, Y. Zheng, *et al.*, “Instance-based vision transformer for subtyping of breast cancer in whole slide images,” *IEEE Transactions on Medical Imaging*, 2021.
- [19] G. Ayana, J. Park, and S. Choe, “Abstract 5052: Patchless deep transfer learning for improved mammographic breast mass classification,” *Cancer Research*, 2022.
- [20] M. Byra, M. Galperin, H. Ojeda-Fournier, A. Tartar, *et al.*, “Breast mass classification in ultrasound with transfer learning and lightweight CNNs,” *Computerized Medical Imaging and Graphics*, 2019.
- [21] A. Gomez, M. Sanchez, P. Garcia, *et al.*, “Edge-oriented deep learning architectures for real-time breast ultrasound screening,” *Diagnostics*, 2021.
- [22] A. Yala, C. Lehman, T. Schuster, T. Portnoi, and R. Barzilay, “A deep learning mammography-based model for improved breast cancer risk prediction,” *Radiology*, 2019.
- [23] G. Campanella, M. G. Hanna, L. Geneslaw, A. Mirafior, *et al.*, “Clinical-grade computational pathology using weakly supervised deep learning on whole slide images,” *Nature Medicine*, 2019.
- [24] M. Bahl, R. Barzilay, A. B. Yedidia, N. J. Locascio, L. Yu, and C. D. Lehman, “High-risk breast lesions: A machine learning model to predict pathologic upgrade and reduce unnecessary surgical excision,” *Radiology*, 2018.
- [25] S. M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, *et al.*, “International evaluation of an AI system for breast cancer screening,” *Nature*, 2020.
- [26] W. Lotter, A. R. Diab, B. Haslam, J. G. Kim, G. Grisot, *et al.*, “Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach,” *Nature Medicine*, 2021.

- [27] M. J. Willemink, W. A. Koszek, C. Hardell, J. Wu, *et al.*, “Preparing medical imaging data for machine learning,” *Radiology*, 2020.
- [28] N. Kutluer, O. A. Solmaz, V. Yamacli, B. Eristi, and H. Eristi, “Classification of breast tumors by using a novel approach based on deep learning methods and feature selection,” *Breast Cancer Research and Treatment*, 2023.
- [29] A. Hosny, C. Parmar, J. Quackenbush, L. H. Schwartz, and H. J. W. L. Aerts, “Artificial intelligence in radiology,” *Nature Reviews Cancer*, 2018.