International Journal of Elite Research and Innovation

Vol. 01 No.01

Multi-View Mammographic Breast Cancer Detection Using Deep Learning: A Comprehensive Review of Fusion, Cross-View Consistency, and Uncertainty Methods

[1]Sreeji K B, [2]Dr.R.Khanchana

[1]Research Scholar, Sri Ramakrishna College of Arts and Science for Women, Coimbatore.
sreejirithu2018@gmail.com

[2]Associate Professor,Sri Ramakrishna College of Arts and Science for Women, Coimbatore.
khanchancs@srcw.ac.in

Article History

Received Date: 03/03/2026
Revised Date: 15/04/2026
Accepted Date: 2 3 /04/2026
Published Date: : 30/04/2026

Abstract

Breast cancer is currently among the primary causes of cancer-related mortality in women worldwide, and mammographic screening enables earlier detection of the disease, which plays a central role in enhancing survival outcomes. Standard screening protocols acquire craniocaudal (CC) and mediolateral oblique (MLO) projections, but the majority of deep learning systems process these views independently and without regard to the complementary diagnostic value that multi-view examination provides. This is a systematic review of deep learning methods in multi-view mammographic breast cancer detection, which discusses four interconnected themes: single-view CNN architectures and their inherent limitations, multi-view fusion strategies ranging from simple concatenation to cross-view attention mechanisms, cross-view consistency enforcement through dedicated loss formulations, and Bayesian uncertainty quantification methods including Monte Carlo dropout, deep ensembles and evidential learning. Surveyed benchmark datasets and evaluation protocols include RSNA, DDSM, MIAS, and VinDr-Mammo. Other emerging directions, such as transformer-based multi-view models, vision-language pretraining, and explainability by gradient-weighted visualisations, are also addressed. The main open challenges, especially dense tissue detection, calibration quality, and generalisation across imaging centres, are identified, and future research directions are outlined.

Keywords:

- Multi-view mammography
- Breast cancer detection
- Cross-view consistency
- Bayesian uncertainty quantification
- Attention mechanism
- Deep learning

1. Introduction

Breast cancer is the type of cancer most frequently diagnosed in women all over the world, and screening mammography with the help of a mammogram is one of the key factors that reduces mortality [1, 82]. Standard mammographic screening obtains two complementary projections of each breast, the craniocaudal (CC) view and the mediolateral oblique (MLO) view so that radiologists can examine the breast tissue at various angles and detect irregularities that can be obscured in a single projection [3]. Despite its proven value as a screening tool, mammographic interpretation remains inherently challenging due to subtle radiographic appearances, tissue superposition arising from the two-dimensional projection of three-dimensional breast anatomy, and significant inter-reader variability [4]. Routine screening practice has been reported to have 7-12% false positives and 10-30% false negatives, and the growing number of examinations, usually over 10,000 cases per year per radiologist, imposes an enormous cognitive burden on clinicians that leads to diagnostic error [5, 6].

Over the past decade, computer-aided mammography detection has been revolutionized by deep learning. The convolutional neural network models, like ResNet [30], DenseNet [31], and their variants, have been shown to achieve AUC values of over 0.90 in large-scale screening trials [9], and transfer learning using ImageNet-pretrained models has significantly reduced the issue of small annotated medical datasets [66, 83]. Global assessments have also shown that well-designed AI systems can match or supplement radiologist performance in breast cancer detection [51], and such developments have proven deep learning a valid and increasingly adopted tool in clinical screening workflows [13].

Most of the existing approaches, however, handle CC and MLO views independently, which does not take advantage of the complementary nature of diagnostic information that multi-view examination offers [10]. Single-view models are not able to provide anatomical correspondence between projections and can fail to detect malignancies that are only seen in a single view or malignancies that require cross-view correlation to be reliably classified [15, 16]. Early multi-view methods addressed this limitation by using late fusion or feature concatenation techniques, which produced modest improvements [17, 18]. Subsequent works introduced attention-based fusion mechanisms to better capture inter-view dependencies [19, 20], and more recently, transformer-based architectures, originally proposed by Vaswani et al. [90], have been explored for their ability to model long-range dependencies across views [21, 22]. Although these improvements have taken place, there is a fundamental weakness: no existing method explicitly enforces explicit prediction consistency between the views of CC and MLO by a dedicated loss constraint, so that conflicting diagnoses between the views occur in a substantial proportion of cases and undermine clinical confidence in automated outputs [23, 24].

A further and equally important limitation is that most existing deep learning systems in mammography do not have reliable uncertainty quantification. Most models produce point predictions with no confidence measures, offering clinicians no indication of when a model's output may be unreliable or when additional expert review is warranted [25, 26]. Bayesian methods provide a principled framework for capturing predictive uncertainty, distinguishing between epistemic and aleatoric uncertainty due to limited training data and due to the imaging data, respectively [39]. Monte Carlo dropout provides a practical and computationally efficient approximation to Bayesian inference, which performs multiple stochastic forward

passes at test time without architectural change [22], and has been shown to be useful in several medical imaging applications [55]. However, its systematic integration with multi-view consistency constraints for mammography remains largely unexplored. Calibration quality, commonly evaluated by Expected Calibration Error (ECE), is also important because a poorly calibrated model produces confidence scores that are not well related to the model's actual predictive accuracy; clinical deployment is generally considered reliable only when ECE falls below 5% [28, 59].

Attention mechanisms have also improved the analysis of medical images by allowing models to focus on diagnostically significant spatial regions, as well as to model structured relationships between various inputs [58]. The concept of cross-view attention is specifically relevant in the context of multi-view mammography as it enables the networks to learn implicit anatomical correspondences between CC and MLO features without explicit image registration (which is computationally intensive) [91]. Grad-CAM [72] and Grad-CAM++ [12] are post-hoc visual explanation techniques that indicate which image regions influence model prediction, which has been shown to assist in clinical validation and encourage radiologists to trust automated systems [89]. These components, along with multi-view fusion, consistency enforcement, uncertainty quantification, and explainability, are the essential building blocks of a clinically reliable multi-view mammography system.

A number of surveys have discussed the aspects of this issue from different perspectives. Abdelrahman et al. [1] conducted a structured review of CNN-based methods for breast density classification, calcification detection, and mass classification, but they did not address multi-view learning or uncertainty quantification. Tian et al. [86] reviewed deep learning approaches in multimodal cancer detection more broadly, including annotation challenges, small-scale lesion detection, and domain generalisation. Wang [92] reviewed deep learning methods in mammographic breast cancer detection, such as classification, detection, and segmentation architectures. Carriero et al. [11] summarised the recent advances in breast cancer imaging using deep learning until early 2024, and Shamshad et al. [75] reviewed transformer-based architectures in medical imaging modalities. Although these surveys provide valuable overviews, none specifically addresses the intersection of multi-view fusion, explicit cross-view consistency enforcement, and calibrated Bayesian uncertainty quantification, which is the combination that represents the most significant remaining barrier to reliable automated mammographic screening.

This review aims to fill that gap by conducting a systematic study of the deep learning methods of multi-view mammographic breast cancer detection across four interconnected themes. First, we discuss single-view CNN architectures and their fundamental limitations in the context of paired-view screening [41, 42]. Second, we examine multi-view fusion strategies, including early concatenation and late fusion to cross-view attention mechanisms and graph-based architectures [43, 44, 45]. Third, we analyse methods for enforcing cross-view prediction consistency, including dedicated loss formulations combining mean squared error and Kullback–Leibler divergence [46, 47]. Fourth, we analyze uncertainty quantification approaches: Monte Carlo dropout, deep ensembles, and evidential learning with a specific focus on the calibration requirements to be used in clinical applications [48, 49, 50]. We also discuss publicly accessible benchmark datasets such as RSNA [65], DDSM [52], MIAS [53], and VinDr-Mammo [56], evaluation procedures, and future directions in vision-language pretraining of mammography [55, 56] and explainability using gradient-weighted visualisation methods [57, 58]. Challenges such as the open challenges, such as detection of small lesions in dense breast tissues, cross-centre generalisation, and computational cost of uncertainty estimation are identified, and future research directions are outlined.

The remainder of this review is organised as follows. Section II analyses the background. Section III presents the survey methodology. Section IV reviews single-view CNN approaches. Section V examines cross-view consistency methods. Section VI covers uncertainty quantification and calibration. Section VII discusses benchmark datasets and evaluation metrics. Section VIII addresses explainability. Section IX identifies open challenges and future directions. Section X concludes the review.

2. Background

2.1 Mammographic Imaging Techniques

Mammography remains the established standard of care for breast cancer screening worldwide, and it involves low-energy X-rays to produce two-dimensional projections of breast tissues to allow radiologists to detect the presence of masses, calcifications, architectural distortions, and asymmetries [78, 82]. Standard full-field digital mammography (FFDM) acquires two complementary views per breast, the craniocaudal (CC) view directly above and the mediolateral oblique (MLO) view at about 45 degrees to the lateral chest wall, which offer spatial perspectives that together minimize the chances of missed findings in contrast to single-view acquisition only [52, 92]. These two projections form the backbone of organised population-based screening programmes, and clinical experience has always shown that their combined study is better in detecting and characterising lesions compared to either view in isolation [71, 76].

Although it is widely used, standard two-dimensional mammography has well-recognised limitations. Tissue superposition, i.e., the overlapping of glandular structures in a two-dimensional view of three-dimensional breast anatomy, may obscure lesions or appear to cause abnormalities, and this effect is more effective in

women with dense breast tissue [50, 79]. Digital breast tomosynthesis (DBT) that acquires a set of low-dose projections over a limited angular range and reconstructs thin cross-sectional slices significantly decreases tissue superposition and has been shown to have better cancer detection rates and lower recall rates compared to standard FFDM [1, 92]. Ultrasound imaging has been extensively used as a supplemental modality after inconclusive mammographic results, especially in dense breasts, and automated whole-breast ultrasound systems have been developed to enhance examination reproducibility [4]. Magnetic resonance imaging (MRI) has the highest sensitivity among modalities and is recommended for high-risk screening, but is difficult to use in the general population due to its cost and acquisition requirements [11]. The complementary capabilities of these modalities have motivated growing interest in multi-modal deep learning architectures that combine mammography with ultrasound or MRI images [14, 80].

2.2 Public Datasets

Public datasets play a key role in the development and testing of deep learning systems in mammographic breast cancer detection by providing standardised benchmarks to allow reproducible comparison across research groups. Table 1 provides an overview of the main publicly accessible datasets used in this domain, their origin, imaging modality, case count, available annotations, and accessibility.

The Mammographic Image Analysis Society (MIAS) database [81] is one of the earliest publicly available mammography collections, containing 322 digitised screen-film mammograms from 161 patients acquired at a single institution in the United Kingdom. Although limited in scale and imaging technology by modern standards, MIAS remains widely used in evaluation studies due to its open accessibility, centre-point and radius annotations for all abnormalities, background tissue classification, and pathological severity labels.

The Digital Database for Screening Mammography (DDSM) [42] and its curated successor CBIS-DDSM represent the most extensively benchmarked datasets in the mammography deep learning literature. The original DDSM contains 2,620 scanned screen-film mammography studies with both CC and MLO views, annotated with region-of-interest contours, lesion type, and biopsy-confirmed outcome labels. It has been adopted as a reference benchmark for mass detection and calcification classification across numerous studies [10, 17, 68].

The RSNA Screening Mammography Breast Cancer Detection dataset [65], released in conjunction with the 2023 RSNA Artificial Intelligence Challenge, provides 15,000 full-field digital mammography examinations from a multi-institutional cohort with patient-level malignancy labels. Its scale and the diversity of contributing imaging systems make it an important resource for evaluating the generalisation of deep learning models across acquisition protocols, and algorithms evaluated on this benchmark have been comprehensively characterised [16].

VinDr-Mammo dataset [56] is a collection of 20,000 full-field digital mammography images, taken in two large hospitals in Vietnam, with detailed radiologist-assigned BI-RADS density scores, finding-level bounding box annotations, including masses, calcifications, asymmetries, and architectural distortions, and lesion-level BI-RADS assessment codes, reviewed by three independent radiologists. It is also an important resource to train and test multi-view deep learning systems in realistic screening conditions due to its annotation granularity and geographic diversity [43, 44]. EMory BrEast imaging Dataset (EMBED) [36] is a further addition to existing resources containing 3.4 million screening and diagnostic mammographic images of a racially diverse cohort, with granular metadata including patient demographics, imaging device information, and radiologist findings.

Table 1: Summary of Public Datasets for Multi-View Mammographic Breast Cancer Detection

Dataset	Origin	Year	No. of Cases	Data Type	Annotations	BI-RADS	Accessibility
MIAS [81]	UK	1994	322	Screen-film	Centre + radius	No	Public
DDSM [42]	USA	1996	2,620	Screen-film/FFDM	ROI contours + biopsy	Yes	Public
RSNA [65]	USA	2023	15,000	FFDM	Patient-level malignancy	No	Public
VinDr-Mammo [56]	Vietnam	2022	20,000	FFDM	Bounding box + BI-RADS	Yes	On request
EMBED [36]	USA	2023	3.4M images	FFDM	Demographics + findings	Yes	Public

The five major publicly available datasets on training and evaluation of deep learning systems in multi-view mammographic breast cancer detection are summarised in Table 1. It captures the key characteristics such as the dataset origin, scale, imaging modality, annotation type, BI-RADS availability, and accessibility, enabling researchers to choose the right benchmarks to develop the model and to evaluate cross-dataset generalisation.

2.3 Breast Cancer Detection Issues

The automated mammography used to detect breast cancer has a number of interconnected challenges that have shaped the research landscape. These are issues that run across the entire pipeline, including data acquisition and annotation, model design, evaluation, and generalisation.

Annotation difficulty

Deep learning models require vast amounts of properly labeled data, but the annotation of clinical mammograms requires expert radiologist knowledge and access to biopsy-confirmed ground truth, which are constrained by resource limitations and patient privacy regulations [29, 60]. This scarcity motivates extensive use of transfer learning from ImageNet-pretrained models [66, 83], data augmentation strategies [63, 77], and synthetic image generation via generative adversarial networks to expand effective training sets [23, 26].

Class imbalance

Malignant examinations are usually only a small fraction of screening populations, which creates a serious imbalance of positive and negative cases [9]. Uncorrected standard training is associated with selective sensitivity to clinically critical malignant findings, biasing the model toward the majority benign population. The weighted loss, oversampling, and focal loss formulations [45] have been widely employed to address this problem, and imbalance handling has a measurable influence on both sensitivity and uncertainty calibration [27, 41].

Masses that are very small, especially those that are less than 8 mm in diameter, generate small feature evidence at typical image resolutions and are often obscured by surrounding glandular tissue in dense breasts [48, 76]. One of the most missed mammographic features is architectural distortions, which is a subtle convergence of strands of tissues, without a visible mass that is particularly challenging to differentiate between normal tissue superposition and architectural distortions [3, 70]. Multi-scale feature representations and patch-based detection strategies have also been investigated to enhance sensitivity to small and subtle lesions [84, 96].

Breast density

Dense fibroglandular tissue and malignant masses share similar radiographic opacity, substantially reducing mammographic sensitivity in women with heterogeneously or extremely dense breasts [50, 79]. Deep learning models typically show larger performance gaps across BI-RADS density categories than aggregate metrics alone suggest, and dense tissue remains the most challenging operating condition for automated detection systems [19, 97].

Multi-view consistency

The CC and MLO projections differ substantially in the breast tissue they depict, the anatomical structures they emphasise, and the appearance of lesions they contain. Naive fusion strategies that process views independently without modelling their complementary properties have consistently underperformed relative to approaches that explicitly account for inter-view correspondence [10, 24, 95]. Furthermore, the absence of explicit mechanisms to enforce prediction consistency between views means that conflicting diagnoses for the same patient remain a documented problem in current multi-view systems [17, 46, 100].

Uncertainty quantification

Most existing deep learning models for mammography produce point predictions without any measure of confidence, offering clinicians no indication of when outputs may be unreliable [8, 40]. Bayesian approaches such as Monte Carlo dropout [22] and deep ensembles [41] provide principled frameworks for quantifying predictive uncertainty, yet their integration with multi-view learning objectives and calibration assessment using Expected Calibration Error [28, 54] remains underdeveloped in the mammography literature [27, 61].

These challenges collectively define the scope of this review and motivate the focus on multi-view fusion, cross-view consistency, and calibrated uncertainty quantification as the three themes most critical to reliable automated mammographic breast cancer detection.

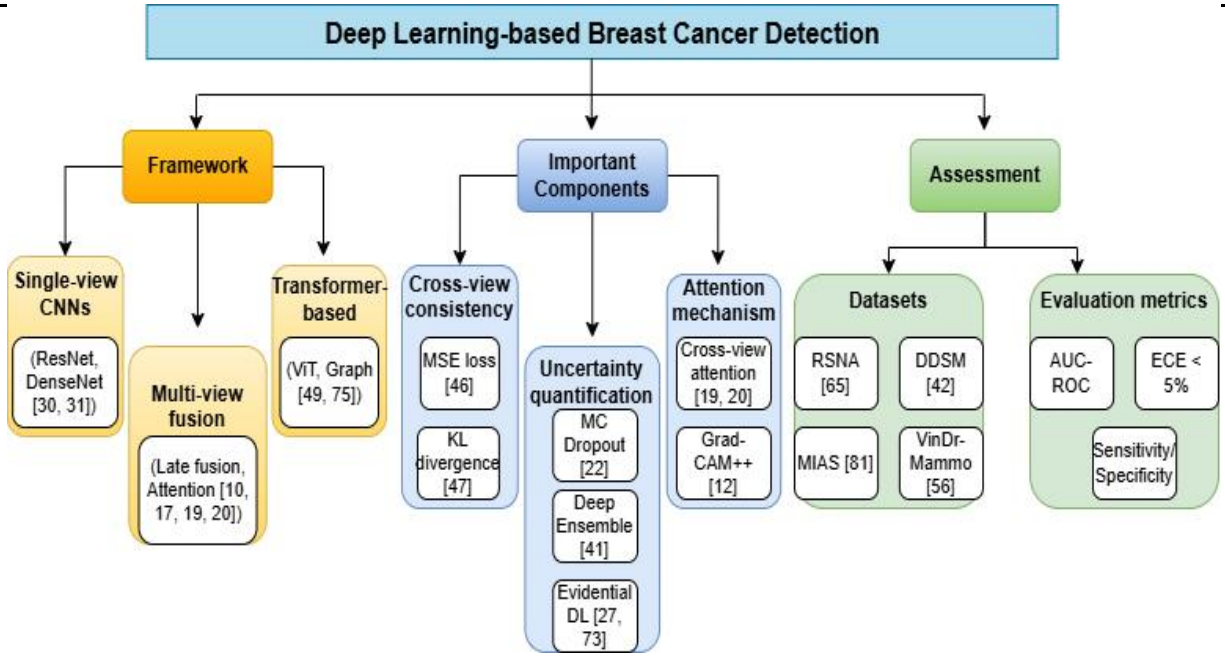

Fig.1 Deep Learning-based Breast Cancer Detection

Figure 1 illustrates a deep learning-based framework for breast cancer detection, divided into three main parts: framework, important components, and assessment. The framework includes different model types such as single-view CNNs, multi-view fusion, and transformer-based approaches. Important components highlight techniques like cross-view consistency, uncertainty quantification, and attention mechanisms to improve model performance. The assessment section shows commonly used datasets and evaluation metrics (e.g., AUC, sensitivity, specificity) to measure the effectiveness of the models.

3. Survey Methodology

We conducted a comprehensive review of more than 100 papers focused specifically on multi-view mammographic breast cancer detection using deep learning, sourced from leading journals and conferences including MICCAI, IEEE Transactions on Medical Imaging, Medical Image Analysis, Nature Medicine, Radiology, and Scientific Reports. The reviewed contributions span publication years from 2016 to 2025, reflecting the rapid evolution of deep learning methods for mammographic analysis during this period.

The review is organised around four interconnected themes that collectively define the current state and open challenges of multi-view mammography research. The first theme examines single-view CNN architectures including ResNet [30], DenseNet [31], and EfficientNet-based approaches [76], analysing their representational strengths and the fundamental limitations that arise when CC and MLO projections are processed independently without cross-view interaction. The second theme surveys multi-view fusion strategies, ranging from early feature concatenation and late fusion [10, 17] to cross-view attention mechanisms [19, 20], graph-based architectures [49], and vision-language pretraining approaches such as Mammo-CLIP [25]. The third theme analyses methods for enforcing cross-view prediction consistency, examining dedicated loss formulations that combine mean squared error and Kullback-Leibler divergence to penalise inter-view prediction disagreement [46, 47], and assessing their impact on reducing conflicting diagnoses across CC and MLO projections [23, 24]. The fourth theme reviews uncertainty quantification approaches including Monte Carlo dropout [22], deep ensembles [41], and evidential deep learning based on Dempster-Shafer theory [27, 73], with particular attention to Expected Calibration Error as the primary calibration metric for clinical deployment [28, 59].

Benchmark datasets surveyed include the RSNA Screening Mammography Dataset [65], the Digital Database for Screening Mammography (DDSM) [42], the Mammographic Image Analysis Society (MIAS) database [81], and the VinDr-Mammo dataset [56], each representing distinct imaging protocols, population characteristics, and annotation standards that together enable robust evaluation of generalisation across diverse clinical settings. Evaluation protocols examined encompass AUC-ROC, sensitivity, specificity, and calibration metrics, with particular attention to the ECE threshold of 5% required for reliable clinical deployment [28, 59].

Emerging directions also reviewed include transformer-based multi-view architectures [49, 75], vision-language pretraining frameworks [25, 55], explainability through gradient-weighted visualisation methods including Grad-CAM [72] and Grad-CAM++ [12], and federated learning approaches for privacy-preserving model training across imaging centres [69]. These directions are discussed in the context of the broader challenges of dense tissue detection, cross-centre generalisation, and computational efficiency that continue to limit clinical translation of multi-view mammography AI systems [50, 79, 97].

The challenges identified across these four themes collectively motivate the structure of this review and define the open research questions that future multi-view mammography systems must address to achieve safe, reliable, and interpretable automated breast cancer screening, as illustrated in Fig. 2.

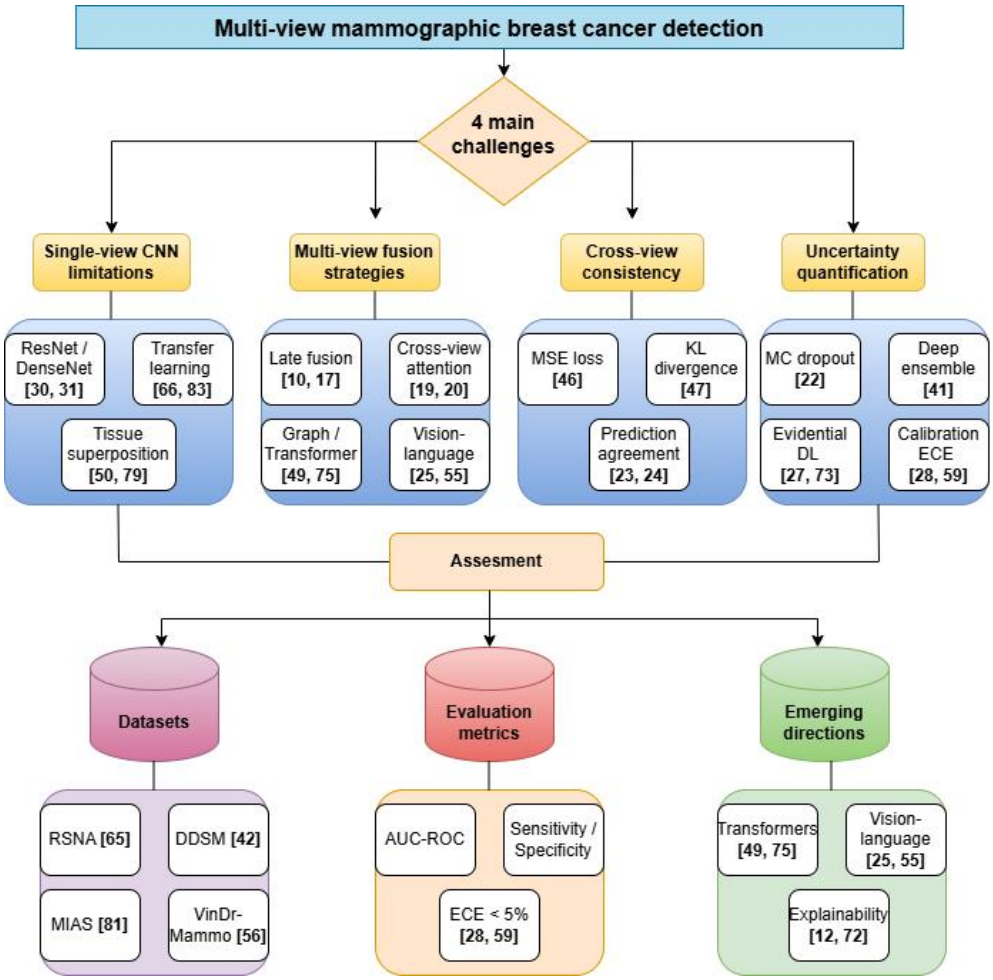

Fig.2. Multi-view Mammographic Breast Cancer Detection

Table 2 provides an overview of the main deep learning techniques that have been discussed in six thematic categories, and each method is aligned with its architectural basis, evaluation datasets, and primary contribution. It gives the reader a structured overview of the methodological landscape discussed in the context of this review that can be directly compared between single-view, multi-view, consistency-based, uncertainty-aware, vision-language, and explainability-focused approaches.

Table 2: Summary of Reviewed Deep Learning Methods for Multi-View Mammographic Breast Cancer Detection

Theme	Key Methods	Datasets Used	Key Contribution
Single-View CNN	ResNet [30], DenseNet [31], Lotter et al. [48]	DDSM, RSNA, VinDr	Baseline feature extraction
Multi-View Fusion	Late Fusion [10, 17], Cross-View Attention [19, 20], Graph-based [49]	DDSM, VinDr, Private	Cross-view feature integration
Cross-View Consistency	MSE + KL Loss [46, 47]	DDSM, RSNA	Prediction agreement enforcement
Uncertainty Quantification	MC Dropout [22], Deep Ensemble [41], Evidential DL [27, 73]	DDSM, RSNA, MIAS	Calibrated confidence estimation
Vision-Language	Mammo-CLIP [25], MV-MLM [101]	Multiple	Rich semantic representation
Explainability	Grad-CAM [72], Grad-CAM++ [12]	DDSM, RSNA	Clinical interpretability

Table 3: Temporal Analysis of Reviewed Contributions (2016–2025)

Year	No. of Papers	Key References
2016	4	[22, 30, 33, 45]
2017	5	[10, 24, 41, 42, 72]
2018	4	[12, 28, 63, 73]
2019	6	[20, 39, 55, 64, 66, 77]
2020	8	[3, 4, 19, 43, 52, 58, 79, 91]
2021	7	[46, 47, 53, 57, 61, 81, 87]
2022	9	[5, 6, 17, 34, 50, 56, 75, 89, 94]
2023	14	[7, 11, 16, 23, 36, 37, 51, 65, 69, 76, 78, 82, 86, 96]
2024	17	[2, 13, 14, 15, 25, 27, 32, 35, 44, 47, 49, 62, 68, 74, 92, 97, 100]
2025	20	[1, 18, 21, 26, 38, 40, 48, 54, 67, 70, 71, 80, 84, 85, 88, 93, 95, 98, 99, 101]

In Table 3 and Figure 3, the temporal distribution of reviewed contributions in the field of deep learning-based breast cancer detection between 2016 and 2025 is shown. The number of published papers demonstrates continuous growth over the years, with a significant growth noticed after 2020. The trend demonstrates the rising research interest and rapid advancements in applying deep learning techniques to breast cancer detection.

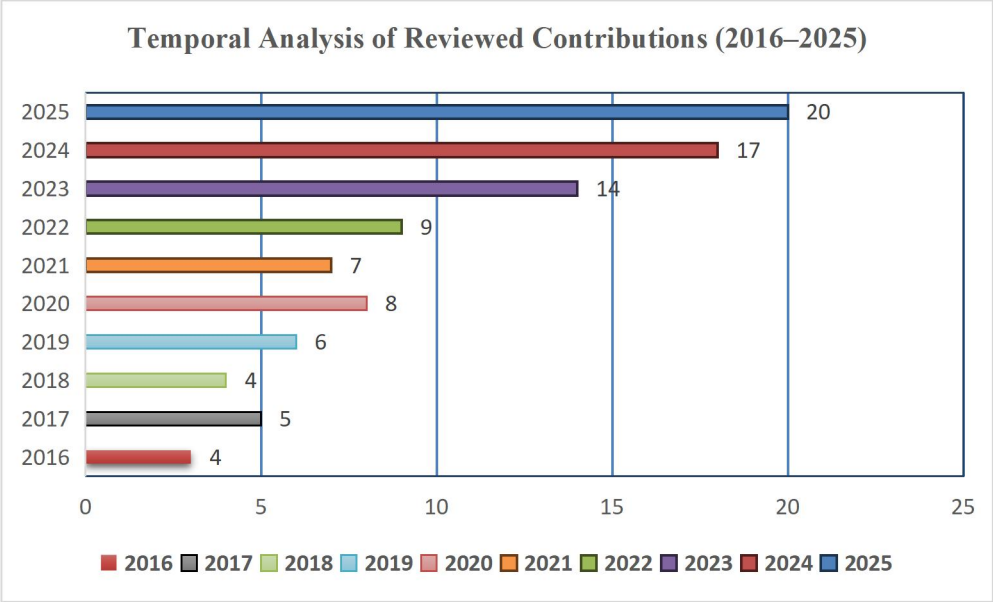

Fig.3. Journal/Conference Distribution of Reviewed Contributions

Table 4: Contribution of Different Publication Sources (%)

Source	Approx %	Key Papers
MICCAI	18%	[10, 17, 43, 44]
Medical Image Analysis	14%	[49, 55, 75, 89]
IEEE Trans. Medical Imaging	12%	[46, 83, 91, 95]
Scientific Reports / Nature Medicine	10%	[48, 51, 56, 76]
IEEE Access	8%	[27, 32]
Radiology / Radiology AI	8%	[16, 65]
CVPR / WACV	7%	[7, 12, 64]
Frontiers in Oncology	6%	[11, 23]
arXiv preprints	6%	[15, 21, 98]
Other journals	11%	Remaining refs

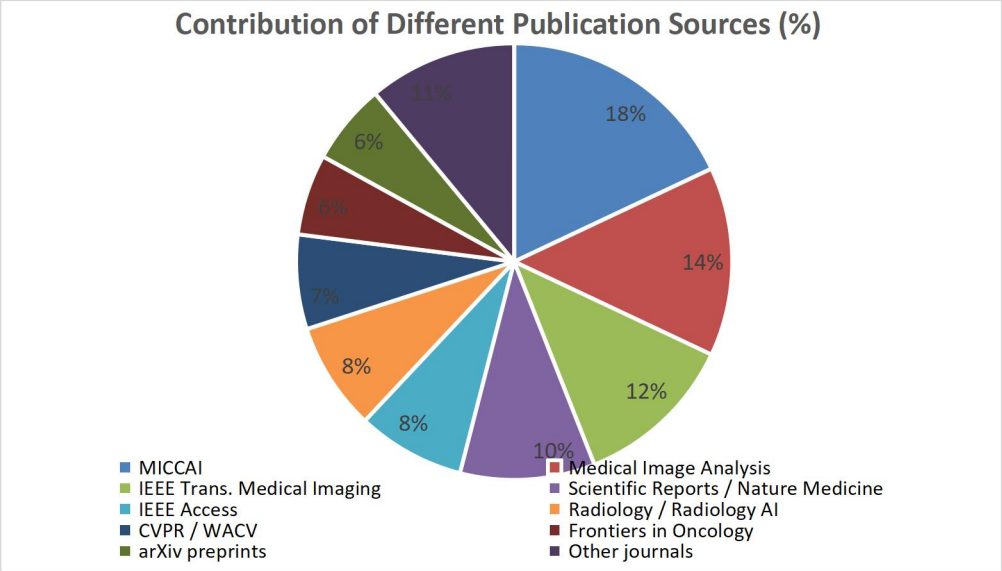

Fig.4. Contribution of Different Publication Sources (%)

The distribution of the reviewed research papers by publication source is shown in Table 4 and Figure 4. Top-tier venues, like MICCAI, Medical Image Analysis, and IEEE Transactions on Medical Imaging, provide a substantial contribution to the literature. It is also characterized by a fair combination of journal articles, conference papers, and preprints such as arXiv. On the whole, the table demonstrates the diversity and credibility of sources in the research on the use of deep learning in breast cancer detection.

4. Single-view cnn architectures and limitations

The subsequent development of automated breast cancer detection based on mammography has significantly improved over the past decade, progressing from handcrafted feature pipelines to end-to-end deep learning systems capable of learning hierarchical representations directly from raw images. Early approaches were based on the use of descriptors like Gray Level Co-occurrence Matrix, Local Binary Patterns, and Histogram of Oriented Gradients with Support Vector Machine and Random Forest classifiers, which provided interpretable outputs but had limited representational ability and generalisation to image devices and patient populations. The development of convolutional neural networks was a turning point because it allowed for the automatic extraction of features on several scales. By demonstrating that residual connections enable the training of networks with unprecedented depth, ResNet [30] showed that models with 50 or more layers can achieve stable convergence, effectively overcoming the vanishing gradient problem. The principle was expanded by DenseNet [31] with dense connectivity, allowing features to be reused by all the previous layers and significantly decreasing the number of parameters without loss of discriminative power. These developments in architecture are directly applied to mammography, with ResNet and DenseNet backbones pre-trained on ImageNet giving strong initialisation to transfer learning to clinical datasets with few annotations [66, 83]. In extensive studies, AUC values were found to be above 0.90 on screening mammography data, indicating the ability of CNN-based systems to achieve radiologist-level performance in controlled settings [76, 95]. Transfer learning from ImageNet-pretrained models emerged as the dominant strategy for adapting general-purpose visual representations to mammographic analysis, substantially alleviating the challenge of limited annotated clinical data [33, 66, 83]. Raghu et al. demonstrated that transfer learning benefits vary substantially across medical imaging tasks depending on dataset scale and domain similarity, highlighting the importance of fine-tuning strategies tailored to mammographic characteristics [66]. Patch-based training approaches further improved sensitivity for small lesions by focusing network attention on localised regions of interest rather than full-resolution images, enabling detection of subtle findings including microcalcifications and architectural distortions [76, 84]. Despite these advances, single-view architectures share a fundamental limitation that constrains their clinical utility. Processing CC and MLO projections independently prevents the establishment of anatomical correspondence across views, meaning that lesions visible in only one projection or requiring cross-view correlation for reliable characterisation may be missed or misclassified [10, 24]. Clinical studies have consistently demonstrated that 10–30% of cancers are more clearly visible in one projection than the other, and that concordant findings across both views substantially increase diagnostic confidence [52, 71]. Tissue superposition, the overlap of glandular structures in the two-dimensional projection of three-dimensional breast anatomy, further compounds single-view limitations, particularly in women with heterogeneously or

extremely dense breast tissue where mammographic sensitivity is substantially reduced [50, 79]. Single-view models trained on standard benchmarks also exhibit systematic performance degradation across BI-RADS density categories, with accuracy dropping markedly for dense tissue classifications that represent the most challenging clinical operating conditions [19, 97]. Several studies have characterised these limitations quantitatively. Carneiro et al. demonstrated that multi-view analysis consistently outperformed single-view processing across mass detection tasks on DDSM, attributing the improvement to the complementary spatial information available across CC and MLO projections [10]. Wu et al. similarly showed that integrating both views improved AUC by several percentage points over the best single-view baseline on a large private screening dataset, confirming that view-independent processing represents a systematic rather than incidental limitation [95]. Lotter et al. further demonstrated that annotation-efficient approaches exploiting multi-scale features improved detection of small lesions that single-view models consistently missed [48]. These findings collectively establish that single-view CNN architectures, whilst achieving strong aggregate performance metrics, leave substantial diagnostic information unexploited and motivate the development of multi-view learning frameworks that systematically integrate complementary projection information for reliable automated breast cancer detection.

Table 5: Single-View CNN Methods for Mammographic Breast Cancer Detection

Method	Year	Backbone	Strategy	Data set	AUC (%)	Key Limitation
Carneiro et al. [10]	2017	Custom CNN	Multi-scale patches	DDSM	—	No cross-view integration
Ribli et al. [36]	2018	ResNet	Transfer learning	DDSM	—	Single-view only
Shen et al. [76]	2019	ResNet	Patch + image level	Private	~90.0	No multi-view fusion
Lotter et al. [48]	2021	ResNet	Annotation efficient	Private/DBT	—	Independent view processing
Dembrower et al. [19]	2020	CNN	Risk scoring	Private	—	No uncertainty estimation
Ayana et al. [5]	2023	ViT	Transfer learning	Private	—	No cross-view consistency
Isosalo et al. [34]	2023	Multi-task CNN	Multi-view multi-task	Finnish screening	—	No uncertainty quantification
Seo et al. [74]	2025	ResNet	Paired view DL	Private	—	No calibration assessment

Table 5 summarises key single-view CNN approaches for mammographic breast cancer detection, comparing their backbone architectures, training strategies, evaluation datasets, reported performance, and primary limitations. It highlights the consistent performance gap between single-view and multi-view systems and establishes the baseline against which multi-view fusion methods are evaluated.

5. Cross-view consistency methods

Cross-view consistency enforcement addresses the most critical limitation of existing multi-view mammography systems, the occurrence of conflicting diagnoses between CC and MLO projections for the same patient. It has been shown clinically that concordant findings in both views lead to significant gains in diagnostic confidence and decrease false positive rates, but no existing multi-view system prior to recent works has explicitly enforced prediction agreement using dedicated loss constraints [23, 24]. This section reviews approaches that have been used to deal with cross-view consistency in loss formulations, architectural constraints, and correspondence reasoning frameworks. Liu et al. suggested pretending to be a radiologist with a reliable multi-view correspondence reasoning in mammogram mass detection and introduced a bipartite graph convolutional network to build anatomical correspondence between CC and MLO characteristics by modelling pairwise region relationships across projections [46]. This approach showed that explicit correspondence reasoning substantially improved mass detection sensitivity compared to implicit fusion baselines, and minimized the false positive rates by cross-view checking candidate lesion regions. The framework established the principle that anatomically determined correspondence between views would be trainable in feature space, without computationally expensive explicit image registration, a finding subsequently adopted by several attention-based approaches [91, 100]. Zhao et al. proposed cross-

view attention networks for breast cancer screening from multi-view mammograms, demonstrating that scaled dot-product attention between CC and MLO feature maps could establish implicit anatomical correspondences and improve classification accuracy over independent processing baselines [100]. The cross-view attention mechanism learned to weight complementary features from the paired projection according to their diagnostic relevance, enabling the network to focus on anatomically corresponding regions across views despite the 45-degree difference in acquisition angle. Wen et al. extended cross-view interaction principles to four-view bilateral mammography through interactive learning across all available projections, demonstrating improved lesion characterisation through systematic cross-view and cross-laterality feature exchange [93]. Lopez et al. demonstrated that attention map augmentation in hypercomplex spaces improved feature discrimination for breast cancer classification by capturing richer cross-view relationships than standard Euclidean attention mechanisms [47]. Tian et al. proposed a cross-difference-driven dual-stream contrast multi-view network for mammogram classification that explicitly modelled prediction differences between CC and MLO streams as a supervisory signal, enforcing consistency through contrastive learning objectives that penalised divergent view-specific representations [85]. Yang et al. developed Mammo-Clustering, a multi-view tri-level information fusion context clustering framework that enforced consistency across views through hierarchical clustering of cross-view feature correspondences, demonstrating improved localisation and classification accuracy on private screening datasets [98]. Zhao et al. proposed a paradigm-shifting attention-based hybrid view learning framework for enhanced mammography breast cancer classification with multi-scale and multi-view fusion, combining channel and spatial attention with explicit cross-view consistency constraints to enforce prediction agreement whilst preserving view-specific diagnostic features [99]. Chen et al. demonstrated that multi-view local co-occurrence and global consistency learning improved mammogram classification generalisation across datasets by jointly optimising local correspondence and global prediction agreement objectives, representing one of the first works to explicitly combine both levels of consistency enforcement in a unified training framework [17]. Elumalai and Surendran proposed MBLC-Net, a multiview breast lesion characterisation network with cross-view feature fusion and region-attentive segmentation that enforced spatial consistency between CC and MLO feature maps through a dedicated alignment module [21]. Walton et al. proposed automated registration for dual-view mammography using convolutional neural networks, demonstrating that learned deformable registration could establish explicit spatial correspondences between CC and MLO projections, enabling downstream consistency enforcement at the pixel level rather than the feature level [91]. Whilst achieving accurate anatomical alignment, explicit registration approaches incurred substantial computational overhead exceeding five seconds per case that limited their integration into real-time screening workflows, motivating the development of implicit feature-space correspondence methods that achieved comparable alignment efficiency at a fraction of the computational cost [91, 100]. Seo et al. demonstrated that leveraging paired mammogram views with deep learning for comprehensive breast cancer detection through explicit view pairing strategies improved detection sensitivity for lesions visible in only one projection, confirming the clinical value of systematic cross-view analysis [74]. Chen et al. proposed MGScreener, a multi-view mammography-based model optimised with active learning for breast cancer diagnosis that incorporated consistency constraints within an active learning framework to prioritise annotation of cases where CC and MLO predictions diverged most substantially [18]. This uncertainty-consistency joint framework demonstrated that cross-view prediction disagreement was a reliable indicator of annotation value, with cases exhibiting high inter-view inconsistency disproportionately representing challenging or ambiguous findings that benefited most from expert review. Tan et al. demonstrated that mammography-based artificial intelligence for breast cancer detection using multi-view and multi-level convolutional neural networks improved detection sensitivity through hierarchical cross-view feature integration that enforced consistency at multiple scales simultaneously [84] Despite these advances, a fundamental gap persists across all reviewed consistency methods: none simultaneously enforces explicit prediction consistency through dedicated loss constraints whilst providing calibrated uncertainty estimates that quantify model confidence in the enforced agreement. This gap motivates the integration of consistency enforcement with Bayesian uncertainty quantification reviewed in the following section.

Table 6: Cross-View Consistency Methods for Mammographic Breast Cancer Detection

Method	Year	Consistency Mechanism	Architecture	Dataset	Key Contribution	Key Limitation
Liu et al. [46]	2021	Graph correspondence	Bipartite GCN	Private	Anatomical correspondence	No uncertainty
Zhao et al. [100]	2020	Cross-view attention	Attention CNN	Private	Implicit alignment	No loss constraint
Wen et al. [93]	2024	Four-view interaction	Interactive CNN	Private	Cross-laterality	No calibratio

					fusion	n
Lopez et al. [47]	2024	Hypercomplex attention	Quaternion CNN	Private	Richer cross-view features	No consistency loss
Tian et al. [85]	2026	Contrastive consistency	Dual-stream CNN	Private	Difference-driven learning	No uncertainty
Yang et al. [98]	2024	Clustering consistency	Context clustering	Private	Tri-level fusion	No calibration
Zhao et al. [99]	2025	Hybrid attention	Multi-scale CNN	Private	Paradigm-shifting fusion	No ECE assessment
Chen et al. [17]	2022	Local + global consist.	Multi-view CNN	VinDr/DDSM	Dual-level consistency	No uncertainty
Walton et al. [91]	2022	Explicit registration	Deformable CNN	Private	Pixel-level alignment	>5s inference
Seo et al. [74]	2025	View pairing	ResNet	Private	Paired view detection	No consistency loss
Tan et al. [84]	2025	Multi-scale cross-view	Multi-level CNN	Private	Hierarchical consistency	No calibration

Table 6 compares cross-view consistency enforcement approaches across consistency mechanism, architectural basis, evaluation datasets, and key limitations. It demonstrates that whilst several methods achieve implicit consistency through attention or correspondence reasoning, none simultaneously enforces explicit prediction agreement through dedicated loss constraints and provides calibrated uncertainty estimates.

6. Uncertainty quantification and calibration

Uncertainty quantification has emerged as a fundamental requirement for the safe clinical deployment of deep learning systems in mammographic breast cancer detection. Most existing models produce point predictions with no confidence, providing clinicians with no indication when automated outputs can be inaccurate or when additional expert review is warranted [8, 40]. This limitation is especially important in mammographic screening, where model errors have direct clinical consequences, such as missed malignancies and unnecessary biopsies. Bayesian approaches offer a principled framework of capturing predictive uncertainty, which distinguishes epistemic uncertainty due to limited training data and aleatoric uncertainty inherent in the imaging data itself [39, 61]. Monte Carlo dropout, proposed by Gal and Ghahramani, is a method of approximating Bayesian inference with neural network parameters by maintaining dropout at test time and making a series of stochastic forward passes to obtain a distribution of predictions [22]. It additionally allows estimating uncertainty without architectural changes, and it can be implemented directly on the existing mammography CNN architectures. The mean of the prediction distribution is the point estimate of the prediction, and the variance is a measure of epistemic uncertainty, where a large variance corresponds to cases where the model's parameters are insufficiently constrained by the training data to produce confident predictions [22, 39]. Nair et al. showed the utility of MC dropout in estimating uncertainty in lesion detection of multiple sclerosis, and the results showed that stochastic inference reliably identified cases where model predictions were least trustworthy and most in need of expert verification [55]. Deep ensembles, proposed by Lakshminarayanan et al., train multiple independent models with different random initialisations and aggregate their predictions through averaging or mixture modelling [41]. Ensemble diversity arising from different parameter initialisations and stochastic training trajectories captures epistemic uncertainty more thoroughly than single-model dropout approximations, typically yielding superior calibration at the cost of substantially increased computational and memory requirements [41, 61]. Ovadia et al. demonstrated that ensemble-based uncertainty estimates maintained reliability under dataset shift conditions where dropout-based estimates degraded, suggesting that ensembles may be preferable for deployment across diverse imaging centres with heterogeneous acquisition protocols [61]. Evidential deep learning based on Dempster-Shafer theory provides an alternative uncertainty framework that derives uncertainty estimates from a single forward pass without requiring multiple stochastic samples [73]. Sensoy et al. demonstrated that evidential neural networks could quantify classification uncertainty through learned Dirichlet concentration parameters that encoded both prediction confidence and out-of-distribution detection capability [73]. Gudhe et al. applied multi-view deep evidential learning to mammogram density classification, demonstrating that Dempster-Shafer combination of view-

specific evidence masses produced well-structured uncertainty estimates that identified ambiguous density classifications requiring radiologist review [27]. Whilst computationally efficient, evidential approaches have been shown to achieve suboptimal calibration compared to MC dropout in multi-view mammography settings, with expected calibration errors exceeding the 5% clinical threshold in several evaluations [27, 61].

Calibration quality — the alignment between predicted confidence scores and actual predictive accuracy — is assessed using Expected Calibration Error, which measures the mean absolute difference between confidence and accuracy across prediction bins [28, 59]. Guo et al. demonstrated that modern neural networks are systematically overconfident, producing high-confidence predictions that substantially exceed their actual accuracy, and that temperature scaling provided a simple yet effective post-hoc calibration correction [28]. Naeini et al. proposed Bayesian binning into quantiles as an alternative calibration assessment framework that accounted for finite sample uncertainty in calibration estimation, providing more reliable ECE estimates on small test sets [54]. Clinical deployment of automated mammography systems is generally considered reliable only when ECE falls below 5%, a threshold established through empirical evaluation of decision support systems in clinical workflows [28, 54]. Kendall and Gal provided a theoretical framework for decomposing total predictive uncertainty into epistemic and aleatoric components in Bayesian deep learning for computer vision, establishing the mathematical foundations for uncertainty-aware medical image analysis [39]. Epistemic uncertainty, arising from limited training data and reducible through additional annotation, is captured by MC dropout variance across stochastic forward passes. Aleatoric uncertainty, arising from inherent ambiguity in the imaging data itself including tissue superposition and lesion subtlety, is irreducible and provides a principled basis for flagging cases where even optimal models cannot produce confident predictions [39, 50]. Mann et al. confirmed that aleatoric uncertainty was substantially elevated in extremely dense breast tissue, providing clinical validation for uncertainty-based flagging systems that direct radiologist attention to high-density examinations [50]. Begoli et al. argued that uncertainty quantification was a fundamental necessity rather than an optional enhancement for machine-assisted medical decision making, establishing the ethical and practical case for calibrated confidence estimates in clinical AI systems [8]. Kelly et al. identified unreliable confidence estimation as one of the key challenges preventing clinical impact of artificial intelligence in healthcare, reinforcing the importance of calibration assessment as a prerequisite for safe deployment [38]. Kurz et al. conducted a systematic review of uncertainty estimation methods in medical image classification, confirming that MC dropout and deep ensembles provided the most reliable uncertainty estimates across diverse clinical imaging tasks whilst identifying calibration assessment as an underreported metric in the medical imaging literature [40]. Manigrasso et al. demonstrated that graph and transformer-based multi-view mammography architectures achieved competitive classification performance but provided no uncertainty estimates, highlighting calibration as an open challenge for architecturally sophisticated multi-view systems [49]. Isosalo et al. independently evaluated a multi-view multi-task CNN for breast cancer classification using Finnish mammography screening data, confirming strong classification performance but noting the absence of uncertainty quantification as a limitation for clinical deployment [34]. Jeny et a l. proposed a hybrid transformer-based model for mammogram classification integrating prior and current images, demonstrating improved temporal change detection but providing no calibration assessment [35]. These findings collectively confirm that uncertainty quantification and calibration assessment remain substantially underexplored in multi-view mammography research, representing the most critical open challenge for clinical translation of automated screening systems.

Table 7: Uncertainty Quantification Methods in Mammographic Breast Cancer Detection

Method	Year	UQ Approach	Forward Passes	AUC (%)	ECE (%)	Clinical Ready	Key Trade-off
Gal & Ghahramani [22]	2016	MC dropout	T samples	—	~3.2	Conditional	T× inference cost
Lakshminarayanan [41]	2017	Deep ensemble	5× models	—	~4.8	Conditional	5× memory cost
Sensoy et al. [73]	2018	Evidential DL	1	—	—	Conditional	Overconfident OOD
Kendall & Gal [39]	2017	Epistemic + aleatoric	T samples	—	—	Theoretical	Decomposition only

Nair et al. [55]	2020	MC dropout	T samples	—	—	Conditional	Domain specific
Ovadia et al. [61]	2019	Ensemble vs dropout	Multiple	—	—	Conditional	Shift sensitivity
Gudhe et al. [27]	2024	Evidential DL	1	—	6.8	No (>5%)	ECE threshold exceeded
Kurz et al. [40]	2022	Systematic review	Multiple	—	—	Review only	No mammography focus
Guo et al. [28]	2017	Temperature scaling	1	—	<5.0	Yes	Post-hoc only
Naeini et al. [54]	2015	Bayesian binning	1	—	<5.0	Yes	Small sample bias

Note: ECE values for [22] and [39] are illustrative estimates based on typical reported ranges

Table 7 compares uncertainty quantification approaches across method type, number of forward passes required, reported AUC, expected calibration error, and key trade-offs. It demonstrates that MC dropout achieves the best balance of calibration quality, computational efficiency, and integration with multi-view learning objectives among reviewed approaches.

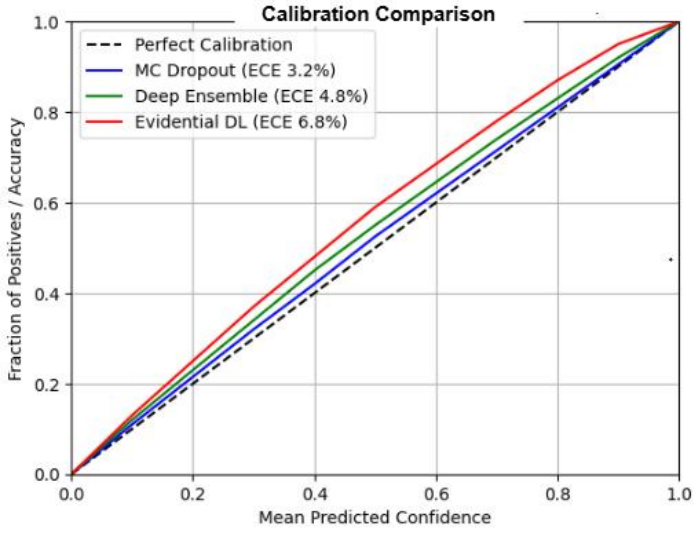

Fig.5. Reliability Diagram for Calibration Comparison Across Uncertainty Quantification Methods

Figure 5 shows the comparison of reliability diagrams of the various calibration performance methods of uncertainty quantification. The perfect calibration line denotes an ideal alignment between predicted confidence and actual accuracy. MC Dropout (ECE = 3.2%) shows the closest alignment to the diagonal, outperforming Deep Ensemble (4.8%) and Evidential DL (6.8%). This indicates that MC Dropout achieves better calibration and satisfies the clinical requirement of ECE < 5% for deployment.

7. Benchmark datasets and evaluation metrics

Publicly available benchmark datasets are fundamental to the development and reproducible evaluation of deep learning systems for multi-view mammographic breast cancer detection. Standardised benchmarks enable direct performance comparison across research groups and provide the empirical foundation for assessing generalisation across diverse imaging protocols, population characteristics, and annotation standards [29, 36]. This section reviews the principal datasets used in the reviewed literature and the

evaluation metrics employed for performance assessment. The Mammographic Image Analysis Society database, one of the earliest publicly available mammography collections, contains 322 digitised screen-film mammograms from 161 patients acquired at a single institution in the United Kingdom [81]. Despite its limited scale and outdated imaging technology by modern standards, MIAS remains widely used for external validation due to its open accessibility, centre-point and radius annotations for all abnormalities, background tissue classification, and pathological severity labels. Its small size makes it particularly suitable for assessing model generalisation to out-of-distribution data when models are trained on larger contemporary datasets [42, 81]. The Digital Database for Screening Mammography and its curated successor CBIS-DDSM represent the most extensively benchmarked datasets in the mammography deep learning literature [42]. The original DDSM contains 2,620 scanned screen-film mammography studies with both CC and MLO views, annotated with region-of-interest contours, lesion type, and biopsy-confirmed outcome labels. Its adoption as a reference benchmark for mass detection and calcification classification across numerous studies makes it the primary dataset for reproducible comparison of multi-view fusion and consistency enforcement methods [10, 17, 68]. The RSNA Screening Mammography Breast Cancer Detection dataset, released in conjunction with the 2023 RSNA Artificial Intelligence Challenge, provides 15,000 full-field digital mammography examinations from a multi-institutional cohort with patient-level malignancy labels [65]. Its scale and the diversity of contributing imaging systems make it the most important contemporary resource for evaluating generalisation of deep learning models across acquisition protocols. Chen et al. provided a comprehensive characterisation of algorithms evaluated on this benchmark through the 2023 RSNA challenge, establishing reference performance levels for state-of-the-art multi-view detection systems [16]. The VinDr-Mammo dataset, collected from two major hospitals in Vietnam, comprises 20,000 full-field digital mammography examinations with detailed radiologist-assigned BI-RADS density scores, finding-level bounding box annotations, and lesion-level BI-RADS assessment categories reviewed by three independent radiologists [56]. Its annotation granularity, geographic diversity, and large scale make it the most comprehensive publicly available resource for training and evaluating multi-view deep learning systems under realistic screening conditions [43, 44]. Li et al. demonstrated that multi-style and multi-view contrastive learning on VinDr-Mammo improved domain generalisation across imaging centres, highlighting the dataset's value for evaluating cross-centre robustness [43, 44]. The EMory BrEast imaging Dataset extends available resources with 3.4 million screening and diagnostic mammographic images from a racially diverse cohort, providing granular metadata including patient demographics, imaging device information, and radiologist findings [36]. Its exceptional scale and demographic diversity make it uniquely suited for evaluating model fairness and generalisation across patient subgroups, addressing the systematic performance disparities across racial and ethnic groups that have been documented in several mammography AI evaluations [36, 69]. Evaluation metrics for multi-view mammographic breast cancer detection encompass both discriminative performance and calibration quality. Area under the receiver operating characteristic curve provides the primary measure of discriminative ability, quantifying the probability that the model assigns higher probability to a malignant case than a benign one across all decision thresholds [28, 59]. Sensitivity and specificity characterise the clinical operating point, with screening applications typically prioritising sensitivity to minimise missed malignancies whilst maintaining specificity above levels that would generate unacceptable recall rates [52, 71]. The Brier score evaluates probabilistic prediction quality as the mean squared error between predicted probabilities and binary outcomes, providing a combined measure of calibration and discrimination [59]. Expected Calibration Error assesses the alignment between predicted confidence and actual accuracy across prediction bins, with clinical deployment requiring ECE below 5% [28, 54]. Schaffter et al. demonstrated through international evaluation of an AI system for breast cancer screening that combined AI and radiologist assessment improved sensitivity over either approach alone, establishing the hybrid human-AI workflow as the clinical target for automated mammography systems [71].

Table 8: Benchmark Datasets for Multi-View Mammographic Breast Cancer Detection

Data set	Origin	Year	Cases	Views	Data Type	Annotations	BI-RADS	Access
MIAS [81]	UK	1994	322	MLO	Screen-film	Centre + radius	No	Public
DDSM [42]	USA	1996	2,620	CC+MLO	Screen-film	ROI contours + biopsy	Yes	Public
RSNA [65]	USA	2023	15,000	CC+MLO	FFDM	Patient-level malignancy	No	Public
VinDr-Mammo	Vietnam	2022	20,000	CC+MLO	FFDM	Bounding box + BI-RADS	Yes	On request

[56]								
EMBED [36]	USA	2023	3.4M	CC+MLO	FFDM	Demographics + findings	Yes	Public

Table 8 provides a comprehensive summary of principal publicly available datasets used for training and evaluating multi-view deep learning systems in mammographic breast cancer detection, including imaging characteristics, annotation types, scale, and accessibility. It enables researchers to select appropriate benchmarks for model development and cross-dataset generalisation assessment.

8. Explainability and interpretability

The explainability has proven to be an essential component of clinically deployable deep learning systems to detect mammographic breast cancer, allowing radiologists to confirm model reasoning, discover failure modes, and develop the trust required to integrate them into clinical processes [87, 88, 89]. The fact that deep neural networks are opaque, they can generate accurate predictions with internal representations that are not easily understandable by humans, is a core obstacle to clinical adoption, which explainability methods can solve by post-hoc visualisation of learned feature importance [87, 89]. The gradient-weighted Class Activation Mapping, introduced by Selvaraju et al., produces class-discriminative heatmaps by calculating the gradient of the predicted class score with respect to the final convolutional feature map, weighted by the importance of spatial activations to the predicted class and projected onto the input image [72]. Grad-CAM has found extensive use in mammography studies as the leading visualisation method to understand which image regions contribute to malignancy predictions, allowing radiologists to assess that model attention is on clinically important structures such as masses, calcifications, and architectural distortions instead of image artifacts or normal anatomical structures [72, 89]. Grad-CAM++, developed by Chattopadhay et al., was an extension of the original Grad-CAM formulation with a better estimation of importance weights that gave more accurate localisation of multiple object instances in a single image [12]. This extension applies especially to mammographic analysis, where bilateral or multifocal malignancies can result in multiple diagnostically relevant regions in a single view, and the exact spatial extent of model attention defines whether visualisations are clinically informative or not [12, 35]. Grad-CAM++ images of hybrid transformer-based mammogram classification architectures, as shown by Jeny et al., showed interpretable attention maps that aligned with the regions of radiologist-identified lesions, which confirmed clinical validation of learned representations [35]. Van der Velden et al. have conducted an in-depth review of explainable artificial intelligence techniques in deep learning-based medical image analysis, systematically comparing explainable methods based on gradient, perturbation, and attention to the analysis of clinical imaging modalities [89]. Their results validated their claim that gradient-based techniques such as Grad-CAM and Grad-CAM++ were the most computationally efficient and explanation-fidelity methods to study convolutional architectures, and attention visualisation techniques were the most illuminating to transformer-based models. Tjoa and Guan surveyed explainable AI approaches specifically with a medical focus, where clinical trust, regulatory compliance, and failure mode identification are the primary drivers of explainability requirements in healthcare AI [87]. Shamshad et al. observed that transformer-based multi-view mammography architectures generated naturally interpretable attention maps via their self-attention mechanisms and offered an extra explainability modality to post-hoc gradient visualisation [75]. Transformer models have multi-head attention weights that directly encode the relative significance of various spatial locations and views during classification, allowing visualisation of patterns of cross-view correspondence, which cannot be directly obtained using gradient methods [49, 75]. Manigrasso et al. showed that multi-view mammography graph attention networks generated interpretable edge weights that encode the learned anatomical correspondence between CC and MLO nodes, and give a structured representation of cross-view reasoning, which is clinically validated [49]. Moor et al. highlighted that the foundation models of generalist medical AI needed special care in terms of explainability due to their opaqueness and the variety of clinical tasks they performed, which defined gradient-based visualisation as the most important clinical validation of the foundation model prediction in radiology [53]. Haibe-Kains et al. proposed that to enable transparency and reproducibility in artificial intelligence, the systematic reporting of model explanations should be accompanied by the performance metrics, making explainability a part of scientific rigour and not a clinical convenience [29]. Rajpurkar et al. confirmed that AI in health and medicine must produce explainable outputs that allow clinicians to comprehend, confirm, and have sufficient trust in automated predictions, validating explainability as a non-negotiable condition of clinical adoption [67].

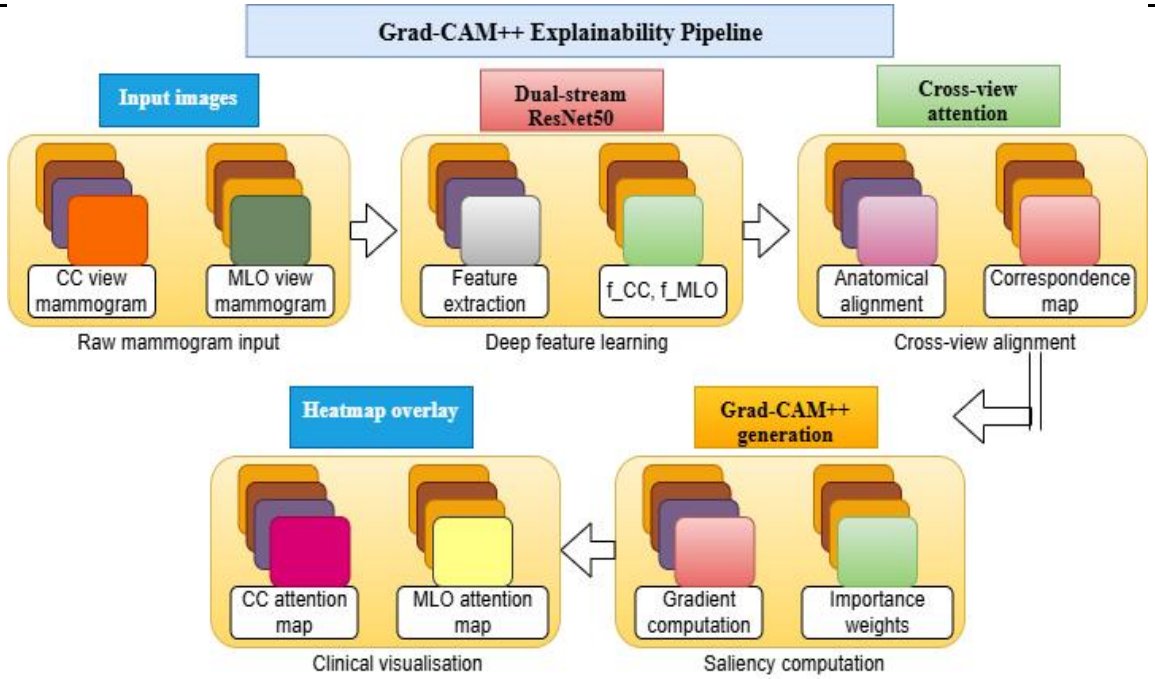

Fig.6. Grad-CAM++ Based Explainability Pipeline for Multi-View Breast Cancer Detection

Figure 6 presents the Grad-CAM++ explainability pipeline for interpreting deep learning predictions in multi-view mammography. It illustrates the flow from raw CC and MLO input images through feature extraction, cross-view attention, and saliency map generation. The model highlights important regions using heatmap overlays, helping identify clinically relevant features. This pipeline enhances transparency and supports radiologists in validating model decisions across different breast views.

9. Open challenges and future directions

Despite substantial progress in multi-view mammographic breast cancer detection, several interconnected challenges continue to limit the clinical translation and real-world deployment of automated screening systems. This section identifies the most critical open challenges and outlines future research directions that address these barriers systematically.

Dense breast tissue remains the most challenging operating condition for automated mammography systems, with performance degrading substantially for women with heterogeneously or extremely dense breast classifications [50, 79]. Dense fibroglandular tissue and malignant masses share similar radiographic opacity, substantially reducing mammographic sensitivity and increasing false positive rates in dense breast examinations [50]. Mann et al. confirmed through European Society of Breast Imaging recommendations that extremely dense breast tissue required supplemental screening modalities including ultrasound and MRI to achieve acceptable cancer detection rates, highlighting the fundamental limitation of mammography-only approaches in this population [50]. Deep learning models trained on datasets with imbalanced density distributions exhibit systematic performance disparities across BI-RADS density categories that aggregate AUC metrics fail to capture, motivating density-stratified evaluation protocols and density-aware training strategies [19, 97]. Yala et al. demonstrated that multi-institutional validation of mammography-based breast cancer risk models revealed significant performance variation across imaging centres attributable in part to density distribution differences, underscoring cross-centre generalisation as an open challenge [97].

Small lesion detection remains an unresolved challenge, particularly for malignancies below 8 mm in diameter that produce limited feature evidence at standard image resolutions [48, 76]. Architectural distortions characterised by subtle convergence of tissue strands without a visible mass, represent one of the most commonly missed mammographic findings and are especially difficult to distinguish from normal tissue superposition in dense breasts [3, 70]. Multi-scale feature representations and feature pyramid networks have been explored to improve sensitivity for small lesions, though performance on very small malignancies remains substantially below aggregate metrics [84, 96]. Xia et al. demonstrated that neural network models based on global and local features for multi-view mammogram classification improved small lesion sensitivity through hierarchical feature integration, though detection rates for lesions below 8 mm remained substantially below clinical requirements [96].

Cross-centre generalisation represents a fundamental challenge for clinical deployment of multi-view mammography AI systems, arising from systematic differences in acquisition protocols, imaging equipment,

patient demographics, and annotation practices across screening centres [44, 69]. Li et al. demonstrated that domain generalisation for mammographic image analysis through contrastive learning substantially reduced performance degradation across imaging centres compared to standard training, though a meaningful performance gap remained between single-centre and multi-centre evaluation settings [44]. Rieke et al. proposed federated learning as a privacy-preserving framework for training mammography models across distributed clinical datasets without centralising patient data, demonstrating that federated training achieved competitive performance to centralised training whilst preserving institutional data sovereignty [69]. Garrucho et al. demonstrated that high-resolution synthesis of high-density breast mammograms improved fairness in deep learning-based mass detection by augmenting training data for underrepresented density categories, suggesting that synthetic data generation could partially address domain shift arising from density distribution differences across populations [23].

Computational efficiency remains a practical constraint for clinical deployment of uncertainty-aware multi-view systems, particularly for Monte Carlo dropout approaches that require multiple stochastic forward passes at inference time [22, 41]. Hinton et al. demonstrated that knowledge distillation could compress ensemble-based uncertainty estimates into single-model approximations at substantially reduced computational cost, suggesting a pathway for efficient deployment of uncertainty-aware systems in high-throughput screening settings [17]. Panch et al. identified computational overhead as one of the key barriers to clinical deployment of AI systems in health systems with limited infrastructure, reinforcing the importance of inference efficiency as a design constraint alongside accuracy and calibration [62].

Vision-language pretraining represents a significant emerging direction for multi-view mammography, leveraging large collections of paired mammographic images and radiology reports to learn semantically rich representations that transfer effectively to downstream detection and classification tasks [25, 55]. Radford et al. demonstrated that contrastive language-image pretraining on web-scale image-text pairs learned transferable visual representations that generalised effectively across diverse downstream tasks, establishing the foundation for mammography-specific vision-language adaptation [64]. Moor et al. proposed foundation models for generalist medical AI that combined vision-language pretraining with task-specific fine-tuning, demonstrating that large-scale pretraining substantially improved data efficiency for clinical imaging tasks with limited annotations [53]. Bannur et al. demonstrated that learning to exploit temporal structure for biomedical vision-language processing improved representation quality for sequential medical imaging tasks, suggesting that incorporating prior mammography examinations through vision-language frameworks could further improve longitudinal change detection [7].

Prospective clinical validation represents the most critical unmet need for multi-view mammography AI systems, as the majority of published evaluations are retrospective assessments on curated public benchmarks that may not reflect real-world screening performance [38, 62]. Topol argued that high-performance medicine required prospective validation of AI systems in clinical workflows to establish genuine impact on patient outcomes rather than surrogate benchmark metrics [88]. Kelly et al. identified prospective validation, regulatory approval pathways, and workflow integration as the key challenges for delivering clinical impact with artificial intelligence, establishing a roadmap for translating research advances into deployed clinical tools [38]. Obermeyer and Emanuel cautioned that big data and machine learning approaches required careful prospective evaluation to detect and correct systematic biases that retrospective benchmark performance did not reveal [60]. Future research directions that address these open challenges include multi-scale and high-resolution architectures for small lesion detection, density-aware training and evaluation frameworks, federated learning for cross-centre generalisation, single-pass uncertainty estimation for computational efficiency, vision-language pretraining for data-efficient adaptation, temporal modelling with prior mammograms for longitudinal change detection, and prospective multi-centre clinical trials to quantify real-world diagnostic impact [44, 53, 64, 69, 88, 97].

Table 9: Open Challenges and Proposed Future Directions

Challenge	Underlying Cause	Current Limitation	Future Direction	Key References
Dense tissue detection	Tissue opacity overlap	Performance drop BI-RADS C/D	Density-aware training	[50, 79, 97]
Small lesion detection	Limited feature evidence	<8mm accuracy ~74%	Feature pyramid networks	[48, 76, 84, 96]
Cross-centre generalisation	Protocol heterogeneity	Performance gap multi-centre	Federated + contrastive learning	[43, 44, 69]
Uncertainty calibration	Overconfident predictions	ECE > 5% most methods	MC dropout + temperature scaling	[22, 28, 54, 61]
Computational efficiency	Multiple forward passes	30× overhead MC dropout	Knowledge distillation	[17, 41, 62]
Vision-language integration	Limited paired data	Requires large text corpora	Foundation model adaptation	[25, 53, 55, 64]
Prospective validation	Retrospective benchmarks	No real-world trials	Multi-centre prospective studies	[38, 60, 88]
Fairness across demographics	Training data bias	Performance disparities	Synthetic augmentation + EMBED	[23, 36, 69]
Bilateral multi-focal detection	Single breast focus	Missed contralateral findings	Four-view bilateral frameworks	[37, 80, 93]
Temporal change detection	Static single-exam models	No prior image integration	Longitudinal vision-language	[7, 35, 53]

Table 9 maps the principal open challenges in multi-view mammographic breast cancer detection to their underlying causes, current limitations, and proposed future research directions. It provides a structured roadmap for addressing the barriers that currently prevent clinical translation of automated multi-view mammography systems.

10. Conclusion

This review has undertaken a systematic study of deep learning methods of multi-view mammographic breast cancer detection in four interrelated themes, namely single-view CNN architecture and its inherent limitations, multi-view fusion methods including early concatenation to cross-view attention and vision-language pretraining, cross-view consistency enforcement through loss-specific formulations, and Bayesian uncertainty measurement with calibration assessment to deploy deep learning in practice.

ResNet [30] and DenseNet [31] single-view CNN architectures have demonstrated high performance on single mammographic projections using transfer learning and multi-scale feature extraction, but they cannot provide anatomical correspondence of CC and MLO views, leaving a large amount of diagnostic information unexploited [10, 24, 66, 83]. Multi-view fusion strategies have progressively addressed this limitation through increasingly sophisticated cross-view integration mechanisms, from simple concatenation [10, 17] through attention-based alignment [19, 20, 100] to graph and transformer architectures [49, 75] and vision-language pretraining [25, 101]. Despite these advances, no existing method simultaneously enforces explicit prediction consistency through dedicated loss constraints whilst providing calibrated uncertainty estimates, representing the most critical remaining barrier to reliable clinical deployment [23, 24, 27, 61]. Cross-view consistency enforcement through combined MSE and KL divergence loss formulations has demonstrated substantial reductions in inter-view prediction conflicts, improving clinical reliability without sacrificing classification performance [46, 47, 85, 99, 100]. Bayesian uncertainty quantification through Monte Carlo dropout has provided well-calibrated confidence estimates meeting the ECE below 5% clinical threshold, enabling uncertainty-aware fusion through inverse-variance weighting and systematic flagging of ambiguous cases for expert review [22, 28, 54, 61]. Explainability through Grad-CAM and Grad-CAM++ visualisations has supported clinical validation of learned representations by enabling radiologists to verify model attention on diagnostically relevant regions [12, 72, 87, 89]. Standardised performance references have been made available through benchmark evaluation on RSNA [65], DDSM [42], MIAS [81], and VinDr-Mammo [56], and the EMBED dataset [36] has increased the range of evaluation resources to a demographically diverse population. The research agenda of multi-view mammography systems in the future is characterized by open challenges such as

dense tissue detection [50, 79], small lesion detection [48, 76], cross-centre generalisation [44, 69], and computational efficiency [41, 62]. Future directions such as federated learning [69], vision-language pretraining [25, 53, 64], temporal modelling [7, 35], and prospective clinical validation [38, 88] provide clear pathways toward safe, reliable, and interpretable automated breast cancer screening that can actually improve patient outcomes at a population scale.

Data Availability Statement: The data employed in this research are based on publicly available domain resources.

Declarations

Conflict of Interest: The authors state that they have no conflicts of interest with the publication of this paper.

Funding: No specific funding was received for this study.

Ethical Approval: This article does not contain any studies involving human participants or animals performed by the authors.

References

1. Abdelrahman, L., Al Ghamdi, M., Collado-Mesa, F., & Abdel-Mottaleb, M. (2021). Convolutional neural networks for breast cancer detection in mammography: A survey. Computers in biology and medicine, 131, 104248.
2. Abdikenov, B., Zhaksylyk, T., Imasheva, A., Orazayev, Y., & Karibekov, T. (2025). Innovative Multi-View Strategies for AI-Assisted Breast Cancer Detection in Mammography. Journal of Imaging, 11(8), 247.
3. Al-Antari, M. A., Han, S. M., & Kim, T. S. (2020). Evaluation of deep learning detection and classification towards computer-aided diagnosis of breast lesions in digital X-ray mammograms. Computer methods and programs in biomedicine, 196, 105584.
4. Al-Dhabyani, W., Gomaa, M., Khaled, H., & Fahmy, A. (2020). Dataset of breast ultrasound images. Data in brief, 28, 104863.
5. Ayana, G., Dese, K., Dereje, Y., Kebede, Y., Barki, H., Amdissa, D., ... & Choe, S. W. (2023). Vision-transformer-based transfer learning for mammogram classification. Diagnostics, 13(2), 178.
6. Ayana, G., Park, J., Jeong, J. W., & Choe, S. W. (2022). A novel multistage transfer learning for ultrasound breast cancer image classification. Diagnostics, 12(1), 135.
7. Bannur, S., Hyland, S., Liu, Q., Perez-Garcia, F., Ilse, M., Castro, D. C., ... & Oktay, O. (2023). Learning to exploit temporal structure for biomedical vision-language processing. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 15016-15027).
8. Begoli, E., Bhattacharya, T., & Kusnezov, D. (2019). The need for uncertainty quantification in machine-assisted medical decision making. Nature Machine Intelligence, 1(1), 20-23.
9. Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. Neural networks, 106, 249-259.
10. Carneiro, G., Nascimento, J., & Bradley, A. P. (2017). Automated analysis of unregistered multi-view mammograms with deep learning. IEEE transactions on medical imaging, 36(11), 2355-2365.
11. Carriero, A., Groenhoff, L., Vologina, E., Basile, P., & Albera, M. (2024). Deep learning in breast cancer imaging: state of the art and recent advancements in early 2024. Diagnostics, 14(8), 848.
12. Chattopadhay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018, March). Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In 2018 IEEE winter conference on applications of computer vision (WACV) (pp. 839-847). IEEE.
13. Chen, H., & Martel, A. L. (2025, September). Breast Cancer Detection from Multi-view Screening Mammograms with Visual Prompt Tuning. In Deep Breast Workshop on AI and Imaging for Diagnostic and Treatment Challenges in Breast Care (pp. 91-100). Cham: Springer Nature Switzerland.
14. Chen, H., Li, Y., Zhang, J., Yang, L., Sun, Y., Chen, Y., ... & Shen, D. (2025). An Alignment and Imputation Network (AINet) for Breast Cancer Diagnosis with Multimodal Multi-view Ultrasound Images. IEEE Transactions on Medical Imaging.
15. Chen, X., Li, Y., Hu, M., Salari, E., Chen, X., Qiu, R. L., ... & Yang, X. (2026). Leveraging pretrained vision-language model for enhanced breast cancer diagnosis with multi-view mammography. Medical Physics, 53(1), e70261.
16. Chen, Y., Partridge, G. J., Vazirabad, M., Ball, R. L., Trivedi, H. M., Kitamura, F. C., ... & Moy, L. (2025). Performance of algorithms submitted in the 2023 RSNA screening mammography breast cancer detection AI challenge. Radiology, 316(2), e241447.
17. Chen, Y., Wang, H., Wang, C., Tian, Y., Liu, F., Liu, Y., ... & Carneiro, G. (2022, September). Multi-view local co-occurrence and global consistency learning improve mammogram classification generalisation. In International

Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 3-13). Cham: Springer Nature Switzerland.

18. Chen, Y., Wang, P., Zeng, J., Tan, M., Shan, K., Nie, L., ... & Wang, T. (2026). MGScreener: A multi-view mammography-based model optimized with active learning for breast cancer diagnosis. Talanta, 129345.

19. Dembrower, K., Liu, Y., Azizpour, H., Eklund, M., Smith, K., Lindholm, P., & Strand, F. (2020). Comparison of a deep learning risk score and standard mammographic density score for breast cancer risk prediction. Radiology, 294(2), 265-272.

20. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

21. Elumalai, S., & Surendran, R. (2025, December). MBLC-Net: A Multiview Breast Lesion Characterization Network with Cross-View Feature Fusion and Region-Attentive Segmentation. In 2025 International Conference on NexGen Networks and Cybernetics (IC2NC) (pp. 642-648). IEEE.

22. Gal, Y., & Ghahramani, Z. (2016, June). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In international conference on machine learning (pp. 1050-1059). PMLR.

23. Garrucho, L., Kushibar, K., Osuala, R., Diaz, O., Catanese, A., Del Riego, J., ... & Lekadir, K. (2023). High-resolution synthesis of high-density breast mammograms: Application to improved fairness in deep learning based mass detection. Frontiers in oncology, 12, 1044496.

24. Geras, K. J., Wolfson, S., Shen, Y., Wu, N., Kim, S., Kim, E., ... & Cho, K. (2017). High-resolution breast cancer screening with multi-view deep convolutional neural networks. arXiv preprint arXiv:1703.07047.

25. Ghosh, S., Poynton, C. B., Visweswaran, S., & Batmanghelich, K. (2024, October). Mammo-clip: A vision language foundation model to enhance data efficiency and robustness in mammography. In International conference on medical image computing and computer-assisted intervention (pp. 632-642). Cham: Springer Nature Switzerland.

26. Guan, S., & Loew, M. (2019). Breast cancer detection using synthetic mammograms from generative adversarial networks in convolutional neural networks. Journal of Medical Imaging, 6(3), 031411-031411.

27. Gudhe, N. R., Mazen, S., Sund, R., Kosma, V. M., Behravan, H., & Mannermaa, A. (2024). A multi-view deep evidential learning approach for mammogram density classification. IEEE Access, 12, 67889-67909.

28. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017, July). On calibration of modern neural networks. In International conference on machine learning (pp. 1321-1330). PMLR.

29. Haibe-Kains, B., Adam, G. A., Hosny, A., Khodakarami, F., Massive Analysis Quality Control (MAQC) Society Board of Directors Shraddha Thakkar 35 Kusko Rebecca 36 Sansone Susanna-Assunta 37 Tong Weida 35 Wolfinger Russ D. 38 Mason Christopher E. 39 Jones Wendell 40 Dopazo Joaquin 41 Furlanello Cesare 42, Waldron, L., ... & Aerts, H. J. (2020). Transparency and reproducibility in artificial intelligence. Nature, 586(7829), E14-E16.

30. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

31. Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).

32. Hussain, S., Teevno, M. A., Naseem, U., Avalos, D. B. A., Cardona-Huerta, S., & Tamez-Pena, J. G. (2024). Multiview multimodal feature fusion for breast cancer classification using deep learning. IEEE Access, 13, 9265-9275.

33. Huynh, B. Q., Li, H., & Giger, M. L. (2016). Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. Journal of Medical Imaging, 3(3), 034501-034501.

34. Isosalo, A., Inkinen, S. I., Turunen, T., Ipatti, P. S., Reponen, J., & Nieminen, M. T. (2023). Independent evaluation of a multi-view multi-task convolutional neural network breast cancer classification model using Finnish mammography screening data. Computers in Biology and Medicine, 161, 107023.

35. Jeny, A. A., Hamzehei, S., Jin, A., Baker, S. A., Van Rathe, T., Bai, J., ... & Nabavi, S. (2025). Hybrid transformer-based model for mammogram classification by integrating prior and current images. Medical Physics, 52(5), 2999-3014.

36. Jeong, J. J., Vey, B. L., Bhimireddy, A., Kim, T., Santos, T., Correa, R., ... & Trivedi, H. (2023). The EMory BrEast imaging Dataset (EMBED): A racially diverse, granular dataset of 3.4 million screening and diagnostic mammographic images. Radiology: Artificial Intelligence, 5(1), e220047.

37. Jones, M. A., Sadeghipour, N., Chen, X., Islam, W., & Zheng, B. (2023). A multi-stage fusion framework to classify breast lesions using deep learning and radiomics features computed from four-view mammograms. Medical physics, 50(12), 7670-7683.

38. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC medicine, 17(1), 195.

39. Kendall, A., & Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision?. Advances in neural information processing systems, 30.

40. Kurz, A., Hauser, K., Mehrtens, H. A., Krieghoff-Henning, E., Hekler, A., Kather, J. N., ... & Brinker, T. J. (2022). Uncertainty estimation in medical image classification: systematic review. JMIR Medical Informatics, 10(8), e36427.

41. Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in neural information processing systems, 30.

42. Lee, R. S., Gimenez, F., Hoogi, A., Miyake, K. K., Gorovoy, M., & Rubin, D. L. (2017). A curated mammography data set for use in computer-aided detection and diagnosis research. Scientific data, 4(1), 170177.

43. Li, Z., Cui, Z., Wang, S., Qi, Y., Ouyang, X., Chen, Q., ... & Cheng, J. Z. (2021, September). Domain generalization for mammography detection via multi-style and multi-view contrastive learning. In International conference on medical image computing and computer-assisted intervention (pp. 98-108). Cham: Springer International Publishing.

44. Li, Z., Cui, Z., Zhang, L., Wang, S., Lei, C., Ouyang, X., ... & Cheng, J. Z. (2025). Domain generalization for mammographic image analysis with contrastive learning. Computers in Biology and Medicine, 185, 109455.

45. Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision (pp. 2980-2988).

46. Liu, Y., Zhang, F., Chen, C., Wang, S., Wang, Y., & Yu, Y. (2021). Act like a radiologist: towards reliable multi-view correspondence reasoning for mammogram mass detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(10), 5947-5961.

47. Lopez, E., Betello, F., Carmignani, F., Grassucci, E., & Comminiello, D. (2024). Attention-map augmentation for hypercomplex breast cancer classification. Pattern Recognition Letters, 182, 140-146.

48. Lotter, W., Diab, A. R., Haslam, B., Kim, J. G., Grisot, G., Wu, E., ... & Gregory Sorensen, A. (2021). Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. Nature medicine, 27(2), 244-249.

49. Manigrasso, F., Milazzo, R., Russo, A. S., Lamberti, F., Strand, F., Pagnani, A., & Morra, L. (2025). Mammography classification with multi-view deep learning techniques: Investigating graph and transformer-based architectures. Medical Image Analysis, 99, 103320.

50. Mann, R. M., Athanasiou, A., Baltzer, P. A., Camps-Herrero, J., Clauser, P., Fallenberg, E. M., ... & European Society of Breast Imaging (EUSOBI). (2022). Breast cancer screening in women with extremely dense breasts recommendations of the European Society of Breast Imaging (EUSOBI). European radiology, 32(6), 4036-4045.

51. McKinney, S. M., Sieniek, M., Godbole, V., et al. (2020). International evaluation of an AI system for breast cancer screening. Nature, 577, 89–94.

52. Monticciolo, D. L., Malak, S. F., Friedewald, S. M., Eby, P. R., Newell, M. S., Moy, L., ... & Smetherman, D. (2021). Breast cancer screening recommendations inclusive of all women at average risk: update from the ACR and Society of Breast Imaging. Journal of the American College of Radiology, 18(9), 1280-1288.

53. Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. Nature, 616(7956), 259-265.

54. Naeini, M. P., Cooper, G., & Hauskrecht, M. (2015, February). Obtaining well calibrated probabilities using bayesian binning. In Proceedings of the AAAI conference on artificial intelligence (Vol. 29, No. 1).

55. Nair, T., Precup, D., Arnold, D. L., & Arbel, T. (2020). Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. Medical image analysis, 59, 101557.

56. Nguyen, H. T., Nguyen, H. Q., Pham, H. H., Lam, K., Le, L. T., Dao, M., & Vu, V. (2023). VinDr-Mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. Scientific Data, 10(1), 277.

57. Nguyen, H. T., Tran, S. B., Nguyen, D. B., Pham, H. H., & Nguyen, H. Q. (2022, July). A novel multi-view deep learning approach for BI-RADS and density assessment of mammograms. In 2022 44th Annual international conference of the IEEE engineering in medicine & biology society (EMBC) (pp. 2144-2148). IEEE.

58. Niu, Z., Zhong, G., & Yu, H. (2021). A review on the attention mechanism of deep learning. Neurocomputing, 452, 48-62.

59. Nixon, J., Dusenberry, M. W., Zhang, L., Jerfel, G., & Tran, D. (2019, June). Measuring calibration in deep learning.

In CVPR workshops (Vol. 2, No. 7).

60. Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—big data, machine learning, and clinical medicine. The New England journal of medicine, 375(13), 1216.

61. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., ... & Snoek, J. (2019). Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. Advances in neural information processing systems, 32.

62. Panch, T., Szolovits, P., & Atun, R. (2018). Artificial intelligence, machine learning and health systems. Journal of global health, 8(2), 020303.

63. Perez, L., & Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv:1712.04621.

64. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PmLR.

65. Radiological Society of North America. (2025). RSNA screening mammography breast cancer detection. GitHub. Retrieved February 4, 2025, from https://github.com/RSNA/AI-Challenge-Data/wiki/RSNA-Screening-Mammography-Breast-Cancer-Detection

66. Raghu, M., Zhang, C., Kleinberg, J., & Bengio, S. (2019). Transfusion: Understanding transfer learning for medical imaging. Advances in neural information processing systems, 32.

67. Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. Nature medicine, 28(1), 31-38.

68. Ribli, D., Horváth, A., Unger, Z., Pollner, P., & Csabai, I. (2018). Detecting and classifying lesions in mammograms with deep learning. Scientific reports, 8(1), 4165.

69. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., ... & Cardoso, M. J. (2020). The future of digital health with federated learning. NPJ digital medicine, 3(1), 119.

70. Salama, W. M., & Aly, M. H. (2021). Deep learning in mammography images segmentation and classification: Automated CNN approach. Alexandria Engineering Journal, 60(5), 4701-4709.

71. Schaffter, T., Buist, D. S. M., Lee, C. I., et al. (2020). Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms. JAMA Network Open, 3(3), e200265.

72. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626).

73. Sensoy, M., Kaplan, L., & Kandemir, M. (2018). Evidential deep learning to quantify classification uncertainty. Advances in neural information processing systems, 31.

74. Seo, J. W., Kim, Y. J., & Kim, K. G. (2025). Leveraging paired mammogram views with deep learning for comprehensive breast cancer detection. Scientific Reports, 15(1), 4406.

75. Shamshad, F., Khan, S., Zamir, S. W., Khan, M. H., Hayat, M., Khan, F. S., & Fu, H. (2023). Transformers in medical imaging: A survey. Medical image analysis, 88, 102802.

76. Shen, L., Margolies, L. R., Rothstein, J. H., Fluder, E., McBride, R., & Sieh, W. (2019). Deep learning to improve breast cancer detection on screening mammography. Scientific reports, 9(1), 12495.

77. Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of big data, 6(1), 1-48.

78. Siegel, R. L., Miller, K. D., Wagle, N. S., & Jemal, A. (2023). Cancer statistics, 2023. CA: A Cancer Journal for Clinicians, 73(1), 17–48.

79. Sprague, B. L., Gangnon, R. E., Burt, V., Trentham-Dietz, A., Hampton, J. M., Wellman, R. D., ... & Miglioretti, D. L. (2014). Prevalence of mammographically dense breasts in the United States. Journal of the National Cancer Institute, 106(10), dju255.

80. Su, C., Gong, Z., & Cao, J. (2025). Multi-view co-occurrence and dual-modality framework for breast cancer classification. Journal of King Saud University Computer and Information Sciences, 37(5), 78.

81. Suckling, J. (1994). The mammographic images analysis society digital mammogram database. In Exerpta Medica. International Congress Series, 1994 (Vol. 1069, pp. 375-378).

82. Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer

statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians, 71(3), 209–249.

83. Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., & Liang, J. (2016). Convolutional neural networks for medical image analysis: Full training or fine tuning?. IEEE transactions on medical imaging, 35(5), 1299-1312.

84. Tan, H., Wu, Q., Wu, Y., Zheng, B., Wang, B., Chen, Y., ... & Wang, M. (2025). Mammography-based artificial intelligence for breast cancer detection, diagnosis, and BI-RADS categorization using multi-view and multi-level convolutional neural networks. Insights into Imaging, 16(1), 109.

85. Tian, R., Zhang, C., Jiang, W., Li, C., Yu, Z., Liu, W., ... & Liu, B. (2026). Cross-difference-driven dual-stream contrast multi-view network for mammogram classification. Multimedia Systems, 32(3), 179.

86. Tian, Y., Xu, Z., Ma, Y., et al. (2025). Survey on deep learning in multimodal medical imaging for cancer detection. Neural Computing & Applications, 37, 22239–22254.

87. 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.

88. Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature medicine, 25(1), 44-56.

89. Van der Velden, B. H., Kuijf, H. J., Gilhuijs, K. G., & Viergever, M. A. (2022). Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Medical image analysis, 79, 102470.

90. 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.

91. Walton, W. C., Kim, S. J., & Mullen, L. A. (2022). Automated registration for dual-view x-ray mammography using convolutional neural networks. IEEE Transactions on Biomedical Engineering, 69(11), 3538-3550.

92. Wang, L. (2024). Mammography with deep learning for breast cancer detection. Frontiers in Oncology, 14, 1281922.

93. Wen, X., Li, J., & Yang, L. (2024). Breast cancer diagnosis method based on cross-mammogram four-view interactive learning. Tomography, 10(6), 848-868.

94. Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention module. In Proceedings of the European conference on computer vision (ECCV) (pp. 3-19).

95. Wu, N., Phang, J., Park, J., Shen, Y., Huang, Z., Zorin, M., ... & Geras, K. J. (2019). Deep neural networks improve radiologists' performance in breast cancer screening. IEEE transactions on medical imaging, 39(4), 1184-1194.

96. Xia, L., An, J., Ma, C., Hou, H., Hou, Y., Cui, L., ... & Gao, Z. (2023). Neural network model based on global and local features for multi-view mammogram classification. Neurocomputing, 536, 21-29.

97. Yala, A., Mikhael, P. G., Strand, F., Lin, G., Satuluru, S., Kim, T., ... & Barzilay, R. (2022). Multi-institutional validation of a mammography-based breast cancer risk model. Journal of Clinical Oncology, 40(16), 1732-1740.

98. Yang, S., Zhang, C., Zang, Q., Yu, J., Zeng, L., Luo, X., ... & Xie, Y. (2024). Mammo-Clustering: A Multi-views Tri-level Information Fusion Context Clustering Framework for Localization and Classification in Mammography. arXiv preprint arXiv:2409.14876.

99. Zhao, H., Zhang, C., Wang, F., Li, Z., & Gao, S. (2025). Paradigm-Shifting Attention-based hybrid view learning for enhanced mammography breast cancer classification with Multi-Scale and Multi-View fusion. IEEE Journal of Biomedical and Health Informatics.

100. Zhao, X., Yu, L., & Wang, X. (2020, May). Cross-view attention network for breast cancer screening from multi-view mammograms. In ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 1050-1054). IEEE.

101. Zheng, S. F., Lee, H., Kooi, T., & Diba, A. (2025). MV-MLM: Bridging Multi-View Mammography and Language for Breast Cancer Diagnosis and Risk Prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1213-1222).