

The Need to Move Beyond Explainability Toward Chain-of-Thought Reasoning: A Focus on AI for Mammography

Yalda Zafari¹, Shahd Soliman¹, Essam A. Rahed^{2,3}, Mohamed Mabrok¹

¹Department of Mathematics and Statistics, Qatar University, Doha, Qatar

²Graduate School of Information Science, University of Hyogo, Kobe, 650-0047, Japan

³Advanced Medical Engineering Research Institute, University of Hyogo, Himeji, 670-0836, Japan
yaldazafari5@gmail.com, ss2004689@qu.edu.qa, rashed@gcis.u-hyogo.ac.jp, m.a.mabrok@qu.edu.qa

Abstract

Mammography is one of the primary imaging modalities for breast cancer screening and diagnosis, playing a pivotal role in early detection and mortality reduction. To alleviate the burden on radiologists interpreting mammographic data, artificial intelligence-based models have emerged as decision-making assistants, demonstrating promising results in several studies. However, a critical gap remains between AI capabilities and clinical integration. Current AI systems predominantly employ end-to-end classification approaches that bypass the structured, multi-step reasoning radiologists use in practice. Radiologists systematically detect abnormalities, characterize their features using standardized descriptors, correlate findings across imaging views, perform temporal comparisons, assess information sufficiency, and synthesize evidence into risk-stratified recommendations. In contrast, most AI models map directly from images to diagnoses without transparent intermediate reasoning, limiting their interpretability, clinical utility, and ability to generalize beyond training distributions. This paper examines the radiological chain-of-thought process in mammography interpretation, reviews current AI approaches and their limitations, and proposes a multi-stage reasoning framework that explicitly models each step of clinical decision-making. By decomposing the diagnostic task into sequential reasoning-based stages, this framework aims to create AI systems that not only predict outcomes but also reason transparently in alignment with clinical workflow. Crucially, we explain how contextual information such as breast density serves as a conditioning variable that dynamically optimizes subsequent reasoning stages, mirroring the adaptive decision-making process radiologists employ in clinical practice. We discuss the implications of this approach and identify critical dataset limitations that must be addressed to enable the development of truly reasoning-aware AI in breast imaging.

Introduction

Breast cancer remains the most common cancer among women globally and a leading cause of cancer-related mortality (Sung et al. 2021). Early detection through routine screening is one of the most effective strategies to reduce mortality, and mammography has been established as the primary imaging tool for breast cancer screening and diag-

nosis (Ren et al. 2022). Despite its widespread use, mammographic interpretation remains inherently challenging, heavily reliant on radiologist expertise, and susceptible to inter-reader variability and diagnostic error, particularly in dense breast tissues or in the presence of subtle abnormalities (Lehman et al. 2015).

With the growing complexity of diagnostic imaging and increasing screening volumes, artificial intelligence (AI) has emerged as a promising solution to support radiologists in interpreting mammograms more efficiently and accurately. In prospective and randomized evaluations, AI-supported reading achieved a similar cancer detection rate to standard double reading while substantially reducing screen-reading workload (Lång et al. 2023). In a prospective population-based study, replacing one radiologist with AI yielded a ~4% higher, non-inferior cancer detection rate compared with double reading (Dembrower et al. 2023). Recent work also explores multi-modal systems that integrate 2D and 3D imaging to improve detection and workflow metrics (Park et al. 2025).

However, most existing AI systems focus on end-to-end classification, predicting malignancy directly from the image without incorporating the nuanced chain of reasoning that radiologists follow in clinical practice. In a real-world diagnostic setting, radiologists employ a sequential and multi-faceted reasoning process. This includes identifying and characterizing a variety of abnormalities (e.g., masses, calcifications, architectural distortions, asymmetries), evaluating their features (shape, margins, distribution), assessing breast density, correlating findings across views, and synthesizing clinical context. They may also recommend additional imaging modalities such as ultrasound, tomosynthesis, or MRI based on certain features. Many state-of-the-art AI models bypass these intermediate steps entirely, opting instead for a shortcut from raw pixels to diagnosis. This leads to a major gap between AI performance in experimental settings and its usability in clinical workflows. Furthermore, explainability methods such as saliency maps or attention visualizations, while helpful, often provide post-hoc justifications that do not adequately capture the structured reasoning required in radiological decision-making.

This paper highlights a critical missing component in current AI systems for mammography: the integration of clinical reasoning or “chain-of-thought” processes that mirror

how radiologists arrive at decisions. We argue that effective AI in breast imaging must not just replicate outcomes but also emulate the reasoning steps leading to those outcomes. This includes abnormality detection and classification, understanding lesion subtypes, evaluating imaging features, and proposing appropriate next steps. These are functions that are vital in the diagnostic process but often absent in AI pipelines.

In the following sections, we outline the clinical workflow of mammographic interpretation, including screening and diagnostic imaging, use of additional modalities, and BI-RADS-based reporting. We review current AI models, including advanced vision-language models and explainability methods, and discuss their limitations in replicating clinical reasoning. Finally, we propose a reasoning-aware AI framework that aligns more closely with real-world radiology practice and discuss its implications for future development. By bridging this gap, we aim to move toward AI systems that function as reliable, transparent, and clinically integrated tools in breast cancer detection and diagnosis.

Radiological Chain-of-Thought and Reasoning in Mammogram Interpretation

Mammography image interpretation is a complex, cognitive, and systematic process that requires structured and sequential reasoning, from gathering evidence and forming hypotheses to reaching conclusions that guide patient care. In clinical practice, radiologists employ a chain-of-thought approach that progresses from initial image acquisition to final diagnostic assessment, with each step building upon previous observations to reach clinically actionable conclusions. Figure 1 illustrates the sequential steps and decision-making process undertaken by radiologists to reach the final risk assessment and recommendation.

The reasoning process begins even before examining breast tissue. Factors such as patient age and family history influence workflow decisions for different patients. Screening imaging, aimed at early detection of malignancy signs, typically begins with standard mammography views: craniocaudal (CC) and mediolateral oblique (MLO) views of both breasts. An important initial assessment is breast tissue composition (categorized as almost entirely fatty, scattered fibroglandular tissue, heterogeneously dense, or extremely dense), which affects both cancer detection sensitivity and informs recommendations for supplemental imaging.

The diagnostic phase begins when a patient presents with clinical symptoms or when screening mammography identifies an abnormality requiring further evaluation. Diagnostic studies often include additional specialized mammographic views or other imaging modalities such as ultrasound, MRI, or tomosynthesis (3D mammography) to better characterize identified findings. The choice of additional modalities is based on patient condition or abnormality characteristics. For instance, in patients with dense breasts where mammogram sensitivity is lower, ultrasound serves as a critical adjunct for distinguishing masses from cysts or better characterizing mass margins. Breast MRI provides high sensitivity for cancer detection and can be used for high-risk patients

or evaluating disease extent in newly diagnosed cases. Tomosynthesis reduces tissue superimposition artifacts, a common issue in 2D imaging of 3D tissue, and improves detection and characterization of breast lesions. This decision regarding supplementary modalities is an important clinical reasoning step.

Mammogram analysis starts with systematically examining standard views, performing contralateral and ipsilateral comparisons, and seeking disruptions in expected patterns. When an abnormality is detected, whether from single-view or multi-view analysis, the radiologist not only notes its presence but systematically evaluates its features according to standardized descriptors. Each finding type requires assessment of specific features influencing malignancy probability. For example, a perfectly round mass suggests a benign finding, whereas an irregular mass raises concern for malignancy. Each abnormality has distinct characteristics and descriptors based on established standards that facilitate communication between healthcare professionals (see Figure 2). All abnormalities, along with their characteristics and locations, are documented in the radiologist's report and the patient's electronic health records.

Temporal reasoning represents another critical dimension, where radiologists leverage prior examinations to better analyze current images. Newly developed or growing lesions require more investigation than stable ones. By incorporating prior examinations, static analysis becomes dynamic. For example, with a developing asymmetry abnormality, static evaluation alone may miss the finding if it appears normal in the current image; however, comparison with prior examinations may reveal that the area has become larger or denser. This reasoning incorporates not only what tissue looks like now but also how it has changed over time, enabling identification and modeling of subtle temporal variations.

The Breast Imaging Reporting and Data System (BI-RADS) provides a standardized framework that organizes and guides the radiological reasoning process. It offers consistent terminology for describing mammographic findings and converts imaging observations into clinically meaningful risk categories. BI-RADS classifications range from 0 (incomplete assessment requiring further imaging) to 6 (biopsy-proven malignancy), with intermediate levels representing increasing suspicion: BI-RADS 1 (negative), BI-RADS 2 (benign finding), BI-RADS 3 (probably benign, with less than 2% risk of malignancy), BI-RADS 4 (suspicious abnormality), and BI-RADS 5 (highly suggestive of malignancy, with greater than 95% likelihood). This categorical evaluation reflects the final stage of the radiological interpretive chain-of-thought, integrating all imaging features into a comprehensive risk assessment that informs clinical decision-making.

For suspicious cases or when cancer confirmation is required, patients undergo biopsy, yielding the definitive diagnostic label. This represents the transition from imaging-based reasoning to pathological confirmation, serving as the gold standard for diagnosis. Importantly, based on imaging data alone, radiologists cannot always determine malignancy presence with full confidence. Some cancers, particularly those with subtle or overlapping features, may be

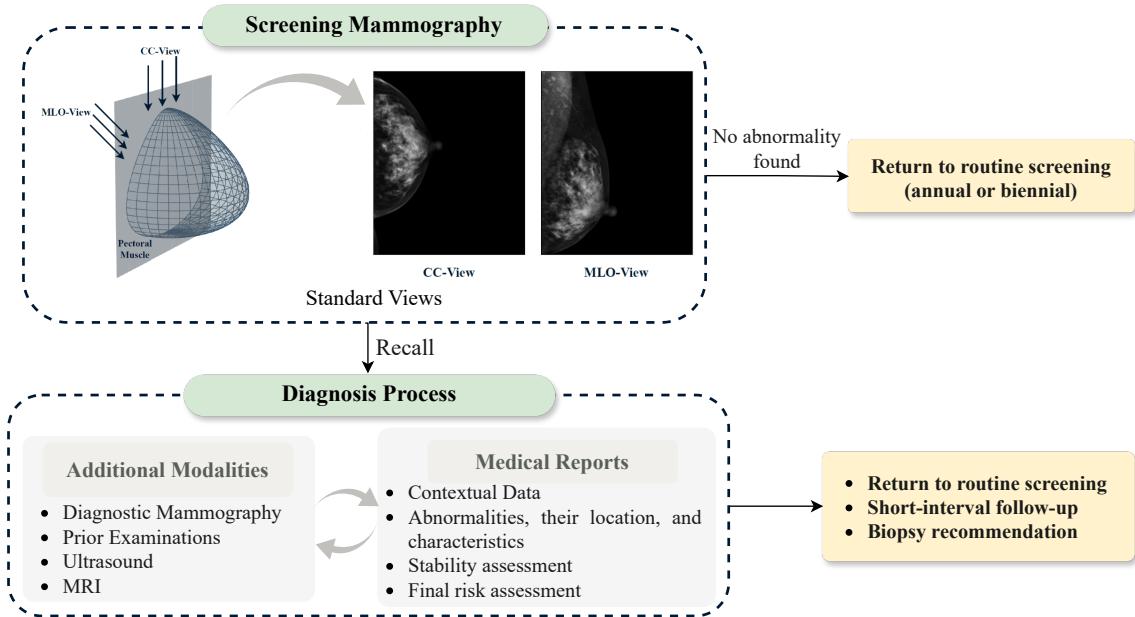


Figure 1: Overview of the transition from screening mammography to the diagnostic phase and the process a radiologist follows to make decisions and final recommendations.

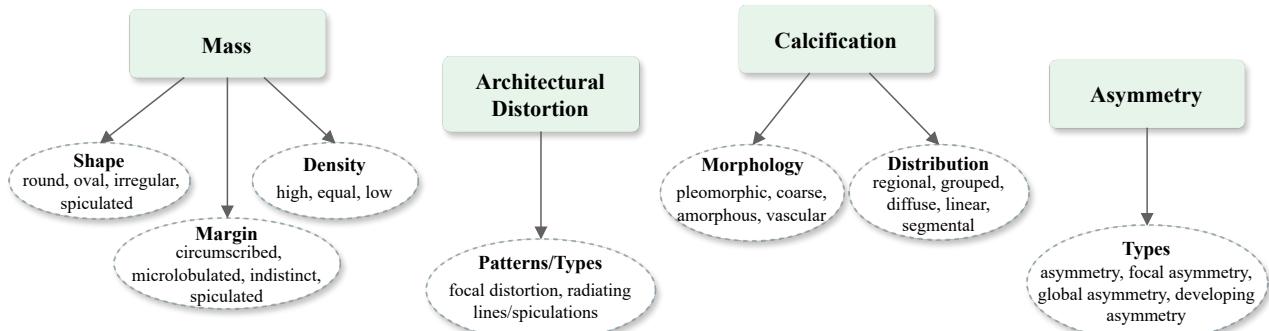


Figure 2: Different categories of abnormalities in mammogram images and their visual attributes.

missed when evaluating only mammography images. Factors such as dense breast tissue, overlapping structures, or non-specific imaging appearances can obscure malignancies and complicate visual interpretation. Therefore, while mammography and other imaging modalities play critical roles in detection and risk stratification, biopsy remains essential for establishing final diagnosis and guiding subsequent clinical management.

This entire reasoning process culminates in the clinical report, a structured document conveying not only what was observed but also the assessment rationale. The report adheres to a standardized format, beginning with clinical indication establishing context (patient gender, age), followed by technical details describing imaging modalities and views analyzed. It then outlines breast composition, details each identified abnormality with features and location, and includes comparison statements when prior examinations are available.

The report concludes with final assessment and corresponding management recommendations. This structured reporting framework ensures that the radiologist's chain-of-thought, from initial observation to final conclusion, is transparent, reproducible, and clinically interpretable, allowing referring physicians to understand both findings and the reasoning informing recommended actions.

Current AI Models for Mammography

This section briefly reviews current AI models, their explainability methods, and the limitations that distinguish these approaches from true reasoning, highlighting the gap between AI systems and clinical workflow.

Evolutions of AI models

AI models for mammography analysis have evolved from single-view image interpretation to multi-view and multi-

modality approaches. Early single-view methods included ROI extraction (Yu et al. 2020), sliding window patch detection (Lotter, Sorensen, and Cox 2017), and region-based techniques (Shu et al. 2020). In recent years, multi-view analysis has gained prominence, allowing models to leverage complementary information from different views simultaneously, thereby improving overall performance (Zafari et al. 2025b; Luo et al. 2024). Multi-view approaches have been developed for both ipsilateral pairs (Van Tulder, Tong, and Marchiori 2021) and the complete set of standard four views (Liu et al. 2025), enabling the detection of abnormalities that may only be visible through contralateral comparison. These methods often employ late fusion strategies (Qian et al. 2025), where information from each view is combined after feature extraction, or integrate attention-based mechanisms at intermediate level to capture interview relationships and enhance feature representation (Manigrasso et al. 2025).

Several architectural paradigms have been explored for mammography analysis, including convolutional neural networks (Sun et al. 2025), vision transformers (Manigrasso et al. 2025), and vision state space models (Zafari et al. 2025a; Bayatmakou et al. 2025). Beyond architectural advancements, research has focused on a variety of tasks such as classification, localization, and multi-task learning. Numerous classification models have been developed for diverse objectives, including breast density categorization (Mohamed et al. 2018) and malignancy prediction (Isosalo et al. 2023). Models designed for localization have aimed to identify and delineate abnormalities within mammography images (Liu et al. 2021). Additionally, multi-task frameworks have been widely adopted to perform multiple related tasks simultaneously, such as abnormality localization and cancer classification (Tardy and Mateus 2021), cancer classification combined with radiologist recall prediction (Manigrasso et al. 2025), or cancer classification alongside BI-RADS assessment (Zafari et al. 2025a).

The most recent frontier in mammography AI research involves vision-language models (VLMs) that integrate imaging data with associated clinical reports (Zheng et al. 2025; Ghosh et al. 2024). These models employ architectures such as CLIP-based frameworks and contrastive learning strategies to align visual representations with textual descriptions. By jointly training on paired image-report data, VLMs learn multimodal embeddings that capture both the visual characteristics of breast tissue and the semantic context of radiological interpretations. In cases where clinical reports are unavailable, synthetic reports can be generated from available meta-data to approximate real textual descriptions, thereby enabling continued multimodal training and preserving the semantic richness of the learned representations.

However, despite the integration of textual information, current VLMs in mammography remain primarily oriented toward a single objective: improving final diagnostic accuracy. The incorporation of clinical reports in these models largely functions as a source of weak supervision for representation learning rather than as a mechanism for structured reasoning. Moreover, these models exhibit a notable limitation: they are typically trained to detect only a lim-

ited subset of abnormalities, mainly mass and calcification, while neglecting other critical findings such as asymmetries, architectural distortions, and their respective subtypes. Even within their covered categories, existing models emphasize detection over characterization. For instance, they can detect the presence of a suspicious mass but fail to systematically describe its shape, margin, or density during inference, as the core features that drive radiological reasoning and risk stratification in clinical practice.

Explainability Paradox

Understanding the decision-making process of AI models is a critical prerequisite for their clinical adoption, where trust, accountability, and error analysis depend on transparent reasoning mechanisms. Consequently, numerous explainable AI (XAI) techniques have been developed to elucidate model behavior. These approaches include saliency mapping, gradient-based visualization methods such as Grad-CAM (Selvaraju et al. 2017), and attention-based mechanisms. Most of these techniques aim either to highlight image regions that most strongly influence model predictions or to approximate network behavior through simplified surrogate models, thereby improving interpretability and aiding human understanding of the underlying decision rationale (Shifa et al. 2025).

Although these methods have enhanced our ability to visualize and inspect model behavior, their usefulness in clinical practice is still limited. Explainability alone rarely meets the needs of high-stakes applications such as cancer diagnosis. Most existing XAI methods provide post-hoc explanations, describing after the fact what the model attended to during inference rather than reproducing the reasoning process that clinicians rely on. For instance, a saliency map may show that a region of a mammogram contributed to a model's prediction, directing attention to an area considered suspicious. Yet this visual cue does not clarify what specific features the model examined within that region. It remains uncertain whether the model assessed clinically relevant attributes such as lesion shape, margin irregularity, or calcification pattern and distribution, all of which are central to radiological assessment. Consequently, the logic underlying model decisions often remains unclear.

Another important limitation of using XAI methods as descriptors of model decision-making is that explanations can appear clinically plausible without faithfully reflecting the model's true computational process (Arun et al. 2021). In some cases, a model may generate a heatmap that highlights a suspicious mass, creating the impression that its decision was based on that region, while in reality, the underlying prediction was influenced by subtle background texture patterns or spurious correlations elsewhere in the image. This disconnect between perceived and actual reasoning undermines the faithfulness of explanations and poses a critical barrier to clinical trust. When explanations are merely plausible rather than truthful, the interpretability of the model becomes superficial, risking misplaced confidence in its outputs and limiting its safe integration into diagnostic workflows.

The consequence is that current explainability methods,

while valuable for fostering trust, are insufficient to ensure that AI models engage in correct or clinically aligned reasoning. These approaches cannot reliably distinguish between models that have learned genuine causal relationships and those that have merely exploited dataset-specific correlations or confounding artifacts. This limitation becomes especially critical when models are exposed to distribution shifts, unseen patient populations, or edge cases, where such spurious correlations no longer hold. Under these circumstances, models that appear interpretable in controlled settings can fail unpredictably in real-world clinical environments, revealing the fragility of explainability as a proxy for reasoning fidelity.

Gap between AI Models and Clinical Workflow

One of the most fundamental limitations of current AI models in mammography analysis lies in their misalignment with clinical workflow and absence of explicit reasoning. Radiologists follow a structured chain of thought: they first detect abnormalities, then characterize them using standardized descriptors, correlating findings across different projections, compare with prior examinations, integrate relevant clinical context, and finally synthesize these elements into a risk stratification that guides management. Each stage in this process is deliberate, interpretable, and inherently accountable, qualities that current AI systems have yet to replicate.

In contrast, most mammography AI research frames the task as direct mapping from images to diagnostic outcomes. Models are trained with mammograms as inputs and categorical labels (benign vs. malignant) as outputs, without constraints or incentives to emulate the intermediate reasoning steps radiologists employ. There is no explicit requirement for the model to first localize findings, describe their morphological attributes, then integrate these observations into a diagnostic conclusion. Instead, the model learns whatever internal representations minimize prediction error on training data, pathways that may diverge substantially from clinically meaningful reasoning.

The gap extends beyond raw performance metrics. Measures such as sensitivity capture how often predictions are correct but reveal little about why those predictions were made. Without transparent reasoning chains, concerns regarding bias, fairness, and generalization naturally arise. By bypassing interpretive reasoning and jumping directly to final decisions, models risk learning correlational rather than causal relationships. Prior studies have revealed that models frequently rely on spurious cues or dataset-specific artifacts that fail to generalize to new imaging distributions (DeGrave, Janizek, and Lee 2021; Banerjee et al. 2023). A model depending on such shortcuts may perform confidently within its training domain yet yield clinically unreasonable or unsafe decisions when confronted with real-world variability.

Moreover, the absence of intermediate chain-of-thought deprives current models of an explicit mechanism to manage the compositional nature of radiological findings. Complex mammographic cases often contain mixed evidence, for instance, a mass exhibiting both benign and suspicious features. Radiologists explicitly weigh these competing cues,

integrating visual characteristics, contextual data, and prior examinations to reach balanced assessments. An AI model trained to output a single malignancy score provides no transparent account of how such contradictory evidence is reconciled.

The research community has largely overlooked the characterization step central to clinical practice and radiological reasoning. While most detection models can successfully localize abnormalities such as masses or calcifications, they rarely produce structured feature descriptions. These descriptions are not auxiliary outputs; they constitute the intermediate reasoning steps enabling systematic evaluation, facilitating error analysis, and supporting clinical decision-making. Their absence leaves current models opaque: they reveal where a model is looking but not how it is thinking, providing clinicians with limited means to interpret, verify, or correct the model's diagnostic logic.

This gap between clinical workflow and AI model design represents not only a major limitation but also a critical opportunity for advancement. Closing this gap requires re-thinking model architecture, training objectives, and evaluation metrics to prioritize structured, interpretable, and clinically aligned reasoning alongside predictive accuracy.

Toward Reasoning AI in mammography

This section proposes a framework for developing AI systems that mirror radiological chain-of-thought processes, explicitly reasoning through intermediate steps rather than directly producing final diagnoses.

Proposed Multi-Stage Reasoning Framework

The core concept of reasoning-based AI for mammography is to decompose the diagnostic task into explicit, sequential reasoning stages; a multi-stage pipeline in which each stage corresponds to a distinct reasoning type (see Figure 3). Outputs from earlier stages serve as structured inputs for subsequent ones. This architectural decomposition not only generates interpretable intermediate representations but also allows stage-specific supervision, evaluation, and clinical validation at each reasoning step, thereby bridging the gap between automated prediction and radiological reasoning.

The model should integrate multi-modality information, encompassing multi-view mammography images, complementary imaging modalities, and relevant textual data such as patient risk factors. It should possess robust decision-making mechanisms to handle cases with missing or incomplete information. Furthermore, when previous imaging studies and reports are available, the model should leverage historical data efficiently, enabling longitudinal reasoning and enhancing diagnostic consistency over time.

The proposed model begins with a vision encoder processing multi-view mammographic images to extract rich, high-dimensional visual representations. Importantly, one of the first reasoning outputs should be breast density assessment, a foundational decision shaping all subsequent analytical steps. This output is not only descriptive; it functions as a conditioning variable for downstream reasoning. Dense breast tissue fundamentally alters diagnostic strategy: it in-

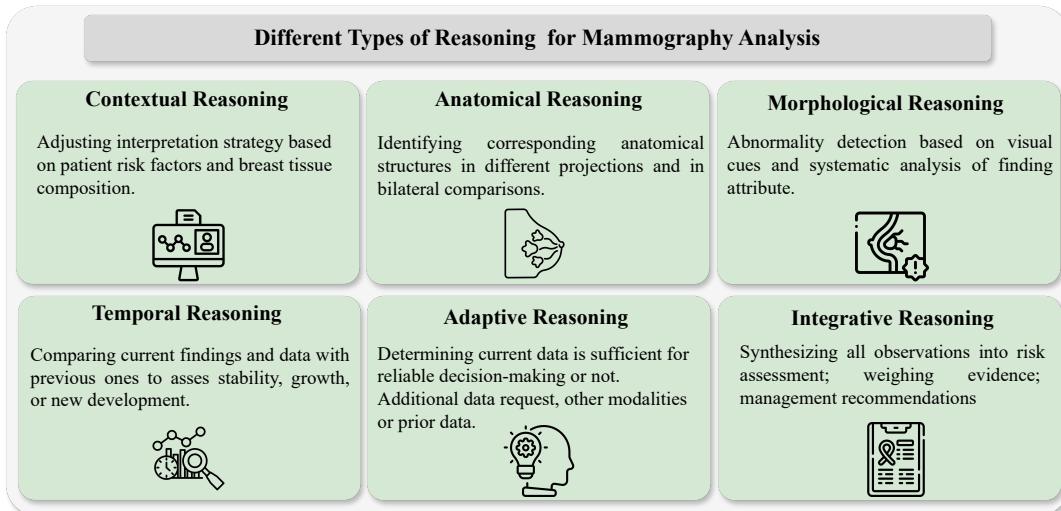


Figure 3: Different types of reasoning employed by radiologists during various stages of the mammogram interpretation and decision-making process.

creases the likelihood that additional imaging will be required, raises suspicion for subtle asymmetries, and influences decision thresholds for recommending supplemental modalities. By explicitly generating this assessment as text, the model introduces an interpretable reasoning branch point mirroring clinical decision-making logic. Subsequent reasoning stages can thus be conditioned or optimized based on this variable, emulating radiologist workflow.

The following stage involves detection, localization, and classification of abnormalities into primary diagnostic categories. Once a category is assigned, the model determines descriptive features, reasoning through visual attributes to generate structured textual characterizations for each detected finding. This stage can be effectively implemented using VLMs to produce outputs resembling clinical reports, describing abnormalities, their features, locations, and extent. Such structured textual reasoning not only enhances interpretability and supervision but also facilitates human-in-the-loop correction and refinement. This stage demonstrates how breast density assessment from the earlier stage can condition subsequent characterization reasoning, leading the model to adjust confidence levels or recommendations according to tissue density.

When prior examinations and reports are available, the model should integrate historical textual descriptions with both current and previous images. This multi-modal temporal reasoning, combining visual comparison with textual historical context, enables more comprehensive and clinically meaningful assessment than image comparison alone. Such reasoning is essential for accurately characterizing developing asymmetries and providing diagnostic reassurance through stability assessment. The model's output should include explicit temporal descriptors (stable, new, increasing, decreasing), which subsequently inform overall risk assessment and guide clinical decision-making.

Moreover, the model should recognize when available in-

formation is insufficient for confident assessment and articulate what additional imaging is required. This reasoning can draw upon multiple contextual factors: breast density, patient risk factors from electronic health records, characteristics of detected findings, or temporal changes observed across prior studies. Through this adaptive reasoning module, the model can generate structured recommendations, allowing it to contribute to imaging workup decisions rather than producing static diagnostic assessments. The underlying architecture should express conditional logic, quantify uncertainty, and communicate decision rationales, capabilities extending far beyond conventional classification-based models. This represents higher-order cognitive reasoning about information sufficiency, a dimension of clinical intelligence current AI systems fundamentally lack.

The final stage integrates all textual reasoning chains and visual information from preceding stages to perform comprehensive evidence synthesis and risk stratification. At this stage, the model leverages multiple information sources to generate unified diagnostic interpretation. The model's output takes the form of a structured clinical assessment mirroring the organization and reasoning style of a radiology report, explicitly articulating how evidence is integrated to reach a conclusion. This structured reasoning allows clinicians to evaluate finding characterization, evidence weighting, and logical validity of the model's decision process. Importantly, breast density serves as a key factor influencing the model's confidence estimation, evidence weighting, and final inference, reflecting its critical role in real-world diagnostic reasoning.

Discussion and Conclusion

This paper has examined the fundamental misalignment between how radiologists reason through mammographic interpretation and how current AI models approach the same task. Although deep learning has achieved impressive

classification performance, most models still follow a direct image-to-diagnosis paradigm, bypassing the structured, multi-step reasoning that underpins clinical practice. In contrast, radiologists engage in a systematic chain-of-thought, from technical assessment and breast density evaluation, through finding detection and BI-RADS-based characterization, to temporal comparison and adaptive decision-making about information sufficiency, ultimately leading to risk synthesis and management recommendations.

Current AI systems, including advanced vision–language architectures, fail to replicate this reasoning process. They often emphasize abnormality detection while neglecting characterization, and they produce final diagnoses without transparent intermediate reasoning. This makes them prone to learning correlations rather than causal relationships, thereby limiting both generalization and clinical interpretability. Furthermore, existing explainability methods, while effective at highlighting salient image regions, offer only post-hoc justifications rather than modeling genuine reasoning. They cannot reliably distinguish between models truly capturing clinical concepts and those simply exploiting dataset-specific biases.

To address these limitations, we propose a reasoning framework that explicitly decomposes mammographic interpretation into sequential stages mirroring the radiological workflow. The framework must be capable of efficiently handling multi-modality inputs, including both images and text, while adapting to the specific needs of each reasoning stage. Each stage produces structured textual outputs such as breast density assessments, localized finding descriptions, BI-RADS characterizations, temporal comparisons, adaptive imaging recommendations, and synthesized risk assessments, thereby constructing transparent reasoning chains that clinicians can verify and interact with. Within this framework, breast density assessment serves as a foundational decision, conditioning subsequent reasoning steps by influencing characterization confidence, guiding supplemental imaging recommendations, and modulating overall risk thresholds. Furthermore, the framework integrates multiple types of reasoning, including anatomical, contextual, morphological, temporal, adaptive, and integrative reasoning, that collectively emulate the interpretive process of expert radiologists. By embedding these structured reasoning stages, the system moves beyond prediction toward clinically grounded inference, fostering interpretability, accountability, and clinical trust.

However, implementing such reasoning framework faces substantial barriers, most critically the limitations of existing public datasets. Current publicly available mammography datasets suffer from several fundamental constraints that prevent the development of reasoning-based AI systems (Zafari et al. 2025c). First, they are not sufficiently large or diverse to support the training of robust, multi-stage models capable of learning complex reasoning chains. While data volumes may be adequate for binary or multi-class classification tasks, they are insufficient when models must learn to capture fine-grained diagnostic features and context-dependent relationships. Second, available annotations are limited, inconsistent, and vary across datasets. Some collec-

tions provide only image-level malignancy labels, while a few include bounding boxes for lesions. Even then, bounding boxes are often incomplete, provided for only a subset of patients or specific abnormality types. Third, critical clinical information is missing. Elements such as recommendations for additional imaging, notations of data insufficiency, and reasoning statements indicating when current information is insufficient are rarely documented. Yet, these components are essential to the radiological decision-making process and for modeling clinical reasoning. Finally, longitudinal data with temporal comparisons remain scarce, limiting opportunities to train AI systems that reason across multiple exams over time. These dataset limitations represent technical inconveniences and fundamental barriers for developing AI systems that reason as radiologists do.

Acknowledgment

This work is supported under the International Research Collaboration Co-Fund (IRCC) between Qatar University and University of Hyogo. Grant number IRCC-2025-633.

References

- Arun, N.; Gaw, N.; Singh, P.; Chang, K.; Aggarwal, M.; Chen, B.; Hoebel, K.; Gupta, S.; Patel, J.; Gidwani, M.; et al. 2021. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiology: Artificial Intelligence*, 3(6): e200267.
- Banerjee, I.; Bhattacharjee, K.; Burns, J. L.; Trivedi, H.; Purkayastha, S.; Seyyed-Kalantari, L.; Patel, B. N.; Shiradkar, R.; and Gichoya, J. 2023. “Shortcuts” causing bias in radiology artificial intelligence: causes, evaluation, and mitigation. *Journal of the American College of Radiology*, 20(9): 842–851.
- Bayatmakou, F.; Taleei, R.; Simone, N.; and Mohammadi, A. 2025. Mammo-Mamba: A Hybrid State-Space and Transformer Architecture with Sequential Mixture of Experts for Multi-View Mammography. *arXiv preprint arXiv:2507.17662*.
- DeGrave, A. J.; Janizek, J. D.; and Lee, S.-I. 2021. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7): 610–619.
- Dembrower, K.; Crippa, A.; Colón, E.; Eklund, M.; and Strand, F. 2023. Artificial intelligence for breast cancer detection in screening mammography in Sweden: a prospective, population-based, paired-reader, non-inferiority study. *The Lancet Digital Health*, 5(10): e703–e711.
- Ghosh, S.; Poynton, C. B.; Visweswaran, S.; and Barmaghelich, K. 2024. 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*, 632–642. Springer.
- Isosalo, A.; Inkkinen, S. I.; Turunen, T.; Ipatti, P. S.; Reponen, J.; and 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.

- Lång, K.; Josefsson, V.; Larsson, A.-M.; Larsson, S.; Höglberg, C.; Sartor, H.; Hofvind, S.; Andersson, I.; and Rosso, A. 2023. Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. *The Lancet Oncology*, 24(8): 936–944.
- Lehman, C. D.; Wellman, R. D.; Buist, D. S.; Kerlikowske, K.; Tosteson, A. N.; Miglioretti, D. L.; Consortium, B. C. S.; et al. 2015. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA internal medicine*, 175(11): 1828–1837.
- Liu, X.; Sun, L.; Li, C.; Han, B.; Jiang, W.; Yuan, T.; Liu, W.; Liu, Z.; Yu, Z.; and Liu, B. 2025. Lesion Asymmetry Screening Assisted Global Awareness Multi-view Network for Mammogram Classification. *IEEE Transactions on Medical Imaging*.
- Liu, Y.; Zhang, F.; Chen, C.; Wang, S.; Wang, Y.; and 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.
- Lotter, W.; Sorensen, G.; and Cox, D. 2017. A multi-scale CNN and curriculum learning strategy for mammogram classification. In *International Workshop on Deep Learning in Medical Image Analysis*, 169–177. Springer.
- Luo, L.; Wang, X.; Lin, Y.; Ma, X.; Tan, A.; Chan, R.; Vardhanabuti, V.; Chu, W. C.; Cheng, K.-T.; and Chen, H. 2024. Deep learning in breast cancer imaging: A decade of progress and future directions. *IEEE Reviews in Biomedical Engineering*.
- Manigrasso, F.; Milazzo, R.; Russo, A. S.; Lamberti, F.; Strand, F.; Pagnani, A.; and Morra, L. 2025. Mammography classification with multi-view deep learning techniques: Investigating graph and transformer-based architectures. *Medical Image Analysis*, 99: 103320.
- Mohamed, A. A.; Berg, W. A.; Peng, H.; Luo, Y.; Jankowitz, R. C.; and Wu, S. 2018. A deep learning method for classifying mammographic breast density categories. *Medical physics*, 45(1): 314–321.
- Park, J.; Witowski, J.; Xu, Y.; Trivedi, H.; Gichoya, J.; Brown-Mulry, B.; Westerhoff, M.; Moy, L.; Heacock, L.; Lewin, A.; et al. 2025. A Multi-Modal AI System for Screening Mammography: Integrating 2D and 3D Imaging to Improve Breast Cancer Detection in a Prospective Clinical Study. *arXiv preprint arXiv:2504.05636*.
- Qian, X.; Pei, J.; Han, C.; Liang, Z.; Zhang, G.; Chen, N.; Zheng, W.; Meng, F.; Yu, D.; Chen, Y.; et al. 2025. A multi-modal machine learning model for the stratification of breast cancer risk. *Nature Biomedical Engineering*, 9(3): 356–370.
- Ren, W.; Chen, M.; Qiao, Y.; and Zhao, F. 2022. Global guidelines for breast cancer screening: a systematic review. *The Breast*, 64: 85–99.
- Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.
- Shifa, N.; Saleh, M.; Akbari, Y.; and Al Maadeed, S. 2025. A review of explainable AI techniques and their evaluation in mammography for breast cancer screening. *Clinical Imaging*, 110492.
- Shu, X.; Zhang, L.; Wang, Z.; Lv, Q.; and Yi, Z. 2020. Deep neural networks with region-based pooling structures for mammographic image classification. *IEEE transactions on medical imaging*, 39(6): 2246–2255.
- Sun, L.; Han, B.; Jiang, W.; Liu, W.; Liu, B.; Tao, D.; Yu, Z.; and Li, C. 2025. Multi-scale region selection network in deep features for full-field mammogram classification. *Medical Image Analysis*, 100: 103399.
- Sung, H.; Ferlay, J.; Siegel, R. L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; and 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.
- Tardy, M.; and Mateus, D. 2021. Looking for abnormalities in mammograms with self-and weakly supervised reconstruction. *IEEE Transactions on Medical Imaging*, 40(10): 2711–2722.
- Van Tulder, G.; Tong, Y.; and Marchiori, E. 2021. Multi-view analysis of unregistered medical images using cross-view transformers. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III* 24, 104–113. Springer.
- Yu, X.; Pang, W.; Xu, Q.; and Liang, M. 2020. Mammographic image classification with deep fusion learning. *Scientific Reports*, 10(1): 14361.
- Zafari, Y.; Elalfy, R.; Mabrok, M.; Al-Maadeed, S.; Khattab, T.; and Rashed, E. A. 2025a. A Hybrid CNN-VSSM model for Multi-View, Multi-Task Mammography Analysis: Robust Diagnosis with Attention-Based Fusion. *arXiv preprint arXiv:2507.16955*.
- Zafari, Y.; Elalfy, R.; Nouman, M.; Al-Maadeed, S.; Khattab, T.; Rashed, E. A.; and Mabrok, M. 2025b. Multi-Modal Deep Learning in Breast Cancer Diagnosis: A Review of Recent Advances. In *2025 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA)*, 1–6. IEEE.
- Zafari, Y.; Pan, H.; Durak, G.; Bagci, U.; Rashed, E. A.; and Mabrok, M. 2025c. MammoClean: Toward Reproducible and Bias-Aware AI in Mammography through Dataset Harmonization. *arXiv preprint arXiv:2511.02400*.
- Zheng, S.-F.; Lee, H.; Kooi, T.; and 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*, 1213–1222.