Original Article

Clinical Validation of AI-Assisted Mammography Analysis: Generalizability and Performance Evaluation in Breast Cancer Detection

Kewalin Rungsinaporn[1] ; Suwara Issaragrisil[2] ; Pramaporn Kimhamanon[3] ;Supphanut Chaiyungyuen[4] ; Nuttawadee Prasawang[4] ;Isarun Chamveha[4]; Sasinun Worakriangkrai[4]



Kewalin rungsinaporn

[1] Assistant Hospital Director, Bangkok Hospital Headquarters, Bangkok, Thailand
[2] Health Design Center, Bangkok Hospital Headquarters, Bangkok, Thailand
[3] Breast Center, Bangkok Hospital Headquarters, Bangkok, Thailand
[4] Perceptra Co., Ltd, Bangkok, Thailand

*Address Correspondence to author:
Kewalin rungsinaporn, MD.
Bangkok Hospital Headquarters,
2 Soi Soonvijai7,New Petchburi Rd.,
Bangkok 10310, Thailand.
email: kewalin.ru@bangkokhospital.com

Received: July 7, 2025
Revision received: September 12, 2025
Accepted after revision: September 15, 2025
BKK Med J 2025;21(2): 145-154.
DOI: 10.31524/bkkmedj.2025.21.007
www.bangkokmedjournal.com

Abstract

OBJECTIVES: To evaluate the generalizability and clinical performance of Inspectra Mammography Model (MMG), an artificial intelligence (AI) system for breast cancer detection in mammography, across different clinical settings and to assess its utility as a second reader in reducing inter-radiologist variability.

MATERIALS AND METHODS: The Inspectra MMG model, developed on a modified EfficientNetV2 architecture, was evaluated using two out-of-domain datasets from Bangkok Dusit Medical Services (BDMS) hospitals: a **Radiologist-validated Set** (172 cases) and a **Biopsy-confirmed Set** (181 cases). Model performance was assessed by area under the receiver operating characteristic curve (AUROC), sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). Clinical usability was further examined through concordance analysis with radiologists' reports, expert acceptance ratings, and system usability scale (SUS) scores. To assess the potential second-reader benefit, the model's impact was analyzed in 119 cases with documented inter-radiologist disagreements.

RESULTS: The model demonstrated robust performance, achieving AUROCs of 0.915 and 0.907 for cancer detection in the Radiologist-validated and Biopsy-confirmed sets, respectively. Lesion localization showed high accuracy with a lesion localization fraction (LLF) of 74.5 and a non-lesion localization fraction (NLF) of 0.53. Clinical usability assessment indicated strong concordance with radiologists' reports (76.2% for classification, 77.9% for localization). Expert radiologists reported high acceptance of AI-generated results (94.5% and 94.0% acceptance rates). The system achieved SUS score of 69.33, reflecting good usability. In the second reader benefit analysis, the AI aligned with final radiological assessments in 71.4% of cases with inter-radiologist disagreement and identified additional potential findings in 16.8% of cases.

CONCLUSION: AI-powered mammography analysis maintained reliable performance across different clinical environments and effectively supported radiologists in breast cancer screening workflows, The system demonstrated potential to reduce inter-reader variability while enhancing detection sensitivity, supporting its role as a clinically valuable second reader.

Keywords: artificial intelligence, AI, mammography, second reader

In 2020, breast cancer was the most frequently diagnosed cancer type globally, with over 2.26 million new cases reported.[1] By 2022, the number of new breast cancer cases had risen to more than 2.31 million, making it the second most common cancer type after lung cancer. Breast mammography is the most common, accessible, and affordable way of screening breast cancer in women around the world. However, with the increase in breast cancer screenings but limited expert medical personnel, it is crucial to optimize the cancer screening workflow.

Advancements in computing and programming technology have improved image interpretation in manifold ways. Computer-aided diagnosis has evolved by leaps and bounds. With advancements in AI technology, particularly deep

learning, which is capable of analyzing big data and learning complex patterns in images, these algorithms have been used to detect cancerous tumors in medical images accurately, even at early stages of the disease.[2]

Recent large-scale prospective trials, including the MASAI trial in Sweden[3] and the German PRAIM study involving over 460,000 women[4], have demonstrated that deep learning-based breast cancer detection systems can improve cancer detection rates by 20-29% while reducing radiologist workload. The AI-STREAM multicenter study further validated AI performance across diverse clinical settings.[5] Studies by Kim et al.[6] and Schaffter et al.[7] have demonstrated that AI can improve diagnostic accuracy and serve as an effective support tool for radiologists. Systematic reviews have confirmed that AI systems achieve comparable or superior performance to radiologists in mammography interpretation[8], with multi-reader studies demonstrating particular benefit when AI serves as a second reader to reduce inter-observer.[9] AI-based mammography systems have also been accepted by the Food and Drug Administration (FDA), USA[10], and CE Mark[11] as products in the category of medical diagnostic tools.

However, a major concern for computer-aided diagnosis of mammograms is clinical validation to ensure generalizability, consistency, and efficiency. Most deep learning models suffer from limited generalizability, i.e., models trained on a dataset from one region or population may not generalize well on other populations due to variability in imaging protocol or demographics.[12] More in-depth investigation and study are indispensable to prove the effectiveness of AI models. Moreover, it is extremely important to validate the effectiveness of the models in a clinical setting for real-world implementation. To the best of our knowledge, no such validation study has been conducted exclusively on the Thai population to date.

Materials and Methods

Inspectra MMG (Inspectra MMG v1.1; Perceptra, Bangkok, Thailand) is a deep learning model built on EfficientNetV2 architecture[13] and trained on 58,022 mammograms from Siriraj Hospital (2007–2021). The model detects nine lesion types grouped into Calcification, Mass, and Others categories. Built on the EfficientNetV2 backbone, Inspectra MMG employs a multi-task learning approach for simultaneous classification and localization, enhancing interpretability with heatmap-based visualization.

Study Design and Datasets

This generalization study evaluated Inspectra MMG through three complementary analyses:

1. Classification and localization performance assessment,
2. Clinical usability evaluation including concordance analysis and system usability testing.
3. Second reader benefit analysis examining the model's impact on diagnostic disagreements.

Three datasets were utilized:

1. Radiologist-validated Set (n = 172): Real-world clinical cases collected from BDMS Hospital (2022-2024) to evaluate model performance against radiologist reports in routine clinical scenarios.
2. Biopsy-confirmed Set (n = 181): Biopsy-proven cancer cases and 2-year follow-up confirmed benign and normal cases collected from BDMS Hospital (2022-2024), providing robust ground truth for performance measurement.
3. Second Reader Benefit Set (n = 119): Cases with documented inter-radiologist disagreements from BDMS Hospital (2023 - 2024), categorized as Major Disagreement (n = 29) or Minor Disagreement (n = 90) based on clinical significance.

All datasets included standard 4-view mammography with visible breast lesions, exclusion criteria including lesions detectable only on ultrasound without mammographic correlation, breast implants, pacemaker devices, poor image quality, or nonstandard views.

All patient data were de-identified with study approval from Bangkok Hospital Headquarters Institutional Review Board (BHQ-IRB2024-05-21).

Statistical Analysis

Performance assessment utilized AUROC, sensitivity, specificity, PPV, and NPV for classification evaluation. AUROC analysis measured the model's discriminative ability between cancerous and non-cancerous cases using receiver operating characteristic (ROC) curves plotting true-positive rate against false-positive rate. Patient-level abnormality scores were calculated from maximum lesion-level scores with predefined thresholds for classification.

Localization performance was evaluated using LLF and NLF by comparing AI-generated heatmaps against ground truth annotations. LLF measured per-lesion localization sensitivity[14], while NLF quantified the number of false-positive lesion highlights per image.

Clinical Evaluation Methodology

The clinical evaluation comprised four components:

1. Concordance Analysis: AI predictions from the Biopsy-confirmed Set (n=181) were automatically compared against radiologist reports and ground truth lesion locations using a four-tier framework (Agree, Edit, Add, Reject) for both classifications and localizations.
2. Usefulness Assessment: Two expert radiologists evaluated AI results from the Biopsy-confirmed Set and rated them on a 4-point scale (1= not useful, 2 = neutral, 3 = useful, 4 = extremely useful). Scores ≥ 2 were considered acceptable, excluding 82 cases with complete concordance were automatically classified as accepted.

ISSN 2287-0237 (online)/ 2228-9674 (print)
(Free to Processing fees & Free to Publish & Free to Read) No hidden charges

3. System Usability Assessment: Fifteen certified radiologists (7 general, 8 breast imaging specialists) completed standardized SUS questionnaires following hands-on testing with the system.
4. Second Reader Benefit Analysis: Systematic comparison of pre- and post-addendum radiology reports against AI predictions was conducted using the Second Reader Benefit Set (n = 119) to evaluate diagnostic impact of AI in cases with inter-radiologist disagreements.

Concordance Analysis

Following established methodology[15], concordance analysis was performed on the Biopsy-confirmed Set (n = 181) to evaluate agreement between Inspectra MMG predictions and radiologist interpretations. This analysis serves as a key measure of clinical utility. **High concordance rates** indicate that AI findings closely align with expert radiologist assessments, suggesting its effective use as a diagnostic assistant without introducing conflicting or misleading information. In contrast, **low concordance rates** suggest frequent disagreements, which could undermine radiologist confidence and limit adoption of the system in clinical practice.

AI predictions were automatically compared against radiologist reports and ground truth annotations in two categories: classification concordance and localization concordance. Classification concordance evaluated agreement between AI diagnostic categories (benign/malignant/normal) and radiologists' final assessments. Localization concordance assessed spatial overlap between AI-generated heatmaps and radiologist-annotated lesion boundaries using intersection-over-union (IoU) metrics.

Both concordance domains employed a four-tier framework, with levels defined as follows:

- **Agree:** Complete match between AI findings and radiologist assessments.
- **Edit:** AI correctly detected all lesions but required refinement (e.g., removal of over-detected lesions or reduction of excessive highlighting).
- **Add:** AI provided accurate findings, but radiologists identified additional lesions or highlighting areas not detected by AI.
- **Reject:** Fundamental disagreement where AI and radiologist identified entirely different conditions, opposite diagnostic conclusions, or AI highlights showed no spatial overlap with actual lesion locations.

The overall concordance rate was calculated as the sum of Agree, Edit, and Add categories, representing clinically acceptable AI performance.

Usefulness Assessment

The usefulness assessment evaluated radiologist acceptance of Inspectra MMG results in clinical practice, determining whether AI findings functioned as valuable diagnostic aids or potential sources of confusion.[16] **High acceptance rates** indicate that AI outputs are clinically relevant and trustworthy, supporting diagnostic decisions, whereas **low acceptance rates** suggests outputs may conflict with clinical judgment or provide limited added value.

Two board-certified breast imaging specialists independently evaluated AI results from the Biopsy-confirmed Set using a standardized review protocol. Each radiologist was provided with AI prediction scores, lesion heatmaps, and ground truth annotations for reference. Cases were then rated based on the accuracy and clinical relevance of AI classification and localization results using a 4-point scale:

1 = not useful: AI findings would mislead clinical decision-making.
2 = neutral: AI provides no additional clinical value.
3 = useful: AI supports clinical assessment and aids diagnosis.
4 = extremely useful: AI provides crucial diagnostic insights potentially missed by radiologists.

Scores ≥2 were considered clinically acceptable.

To optimize radiologist workload while maintaining evaluation rigor, 82 cases with complete concordance ('Agree') in both classification and localization were automatically categorized as **accepted**, reflecting clear AI-radiologist alignment. The remaining 99 cases with partial agreement ('Edit'), additional findings ('Add'), or disagreements ('Reject') underwent detailed evaluation to assess AI utility in more challenging diagnostic scenarios. Individual radiologist acceptance scores were calculated as the ratio of accepted cases to total evaluated cases, and the mean across both radiologists was reported as the final acceptance rate.

System Usability Test

The SUS assessment evaluated the practical integration of Inspectra MMG into clinical workflows, recognizing that even highly accurate AI systems may fail if they are difficult to use or disrupt established practices.

Fifteen certified radiologists (7 general, 8 breast imaging specialists) participated in standardized usability testing, which involved hands-on navigation of diverse mammographic cases followed by completion of the validated 10-question SUS questionnaire. SUS scores range from 0 to100, with scores above 68 considered acceptable for clinical software. The scale assesses factors such as ease of learning, user confidence, system complexity, and likelihood of adoption in clinical practice.

This assessment complements technical performance metrics by determining whether the AI system can function effectively as a practical tool in real-world radiology settings.

ISSN 2287-0237 (online)/ 2228-9674 (print)
(Free to Processing fees & Free to Publish & Free to Read) No hidden charges

Second Reader Benefit Analysis

The second reader benefit analysis evaluated Inspectra MMG's potential to reduce inter-radiologist variability, a well-documented challenge in mammographic interpretation that can impact diagnostic consistency and patient care. This assessment determines whether AI could serve as an effective arbitrator in cases of diagnostic disagreement, potentially improving consensus accuracy while reducing the need for additional human consultation.

A total of 119 cases from BDMS Hospital with documented inter-radiologist disagreements (January 2023 - October 2024) were analyzed. Cases were stratified by disagreement severity: **Major Disagreement (n = 29)**, where second reader recommendations resulted in significant report modifications affecting patient management; and **Minor Disagreement (n = 90)**, where discrepancies were clinically insignificant and final reports remained unchanged.

AI predictions were systematically compared against original radiology reports, second reader comments, and final consensus reports. This multi-reference approach allowed assessment of AI's potential as an independent diagnostic arbitrator, measuring its ability to:

1. Detect lesions identified by second readers but initially missed by primary readers.
2. Align with final consensus diagnoses in amended reports.
3. Identify additional findings overlooked by both human readers.

The analysis evaluated whether AI could contribute to improved diagnostic consistency and potentially reduce the burden of resolving inter-radiologist disagreements in clinical practice.

Results

Dataset Composition and Lesion Distribution

The characteristics of the two primary validation datasets—the Radiologist-validated Set (n = 172) and the Biopsy-confirmed Set (n = 181)—are summarized in Table 1. Lesion localization ground truth labels were collected for the Biopsy-confirmed Set to validate detection capabilities. The Second Reader Benefit Set (n = 119), used for inter-radiologist disagreement analysis, included 29 cases with major discrepancies and 90 cases with minor discrepancies; this dataset is described separately and is not included in Tables 1 and 2.

Model Performance

Model predictions were performed on the datasets described in Table 1 to evaluate generalizability across clinical cases. The model demonstrated high classification performance, achieving strong AUROC scores for overall cancer detection as well as across different lesion types. Localization performance was assessed using LLF and NLF metrics, providing insight into the model's accuracy in detecting and marking lesion areas.

Tables 3 and 4 shows the detection performance of the model on both validation datasets. On the **Radiologist-validated Set,** the model demonstrated strong predictive ability, achieving AUROCs of 0.898 for calcifications, 0.933 for masses, 0.850 for other lesions, and 0.915 for overall cancer detection. On the **Biopsy-confirmed Set**, AUROCs remained high at 0.943 for calcifications, 0.897 for masses, 0.776 for other lesions, and 0.907 for cancer detection. Sensitivity for overall cancer detection reached 90.3% on the Radiologist-validated Set and 88% on the Biopsy-confirmed Set, indicating consistent classification performance across datasets.

The observed differences in performance between the Radiologist-validated Set and the Biopsy-confirmed Set can be attributed to their distinct sampling approaches. The Radiologist-validated Set focuses on real-world data distribution, capturing the complete spectrum of clinical presentations, while the Biopsy-confirmed Set specifically targets cancer detection performance with histologically verified cases. Despite these methodological differences resulting in some variation in lesion-specific metrics (calcifications, masses, and other lesions) due to the limited sample sizes in individual classes, the overall cancer detection performance remains remarkably similar across both datasets (AUROC 0.915 vs. 0.907), demonstrating the model's robust generalization capabilities.

Table 1: Distribution of validation datasets used for model evaluation.

Characteristic	Radiologist-validated Set (n = 172)	Biopsy-confirmed Set (n = 181)
Breast composition		
Almost entirely fat	5	5
Scattered areas of fibroglandular densities	30	37
Heterogeneously dense	123	125
Extremely dense	14	14
BI-RADS categories		
1	45	50
2	21	50
3	3	2
4	73	47
5	29	24
6	1	8
Mammographic features		
Mass	40	61
Calcification	57	34
Others	41	14
Histopathologic type		
Invasive Carcinomas	–	57
In Situ Carcinomas	–	24
Histologic grade		
Grade 1 (Low)	–	8
Grade 1–2 (Low–Intermediate)	–	9
Grade 2 (Intermediate)	–	44
Grade 2–3 (Intermediate–High)	–	8
Grade 3 (High)	–	12

ISSN 2287-0237 (online)/ 2228-9674 (print)
(Free to Processing fees & Free to Publish & Free to Read) No hidden charges

Table 2: Number of lesions with localization ground truth labels in the Biopsy-confirmed Set

Lesion type	Number of localization labels
Mass	111
Calcification	63
Others	23

Table 3: Detection performance on the Radiologist-validated Set.

Condition	AUROC	PPV	NPV	Sens	Spec
a. Lesion detection performance					
Calcification	0.898	75.0%	89.3%	79.0%	87.0%
Mass	0.933	57.1%	96.3%	90.0%	79.6%
Others	0.850	49.2%	90.8%	75.6%	75.6%
b. Cancer detection performance					
Cancer	0.915	80.8%	86.76%	90.3%	74.7%

Table 4: Detection performance on Biopsy-confirmed Set

Condition	AUROC	PPV	NPV	Sens	Spec
a. Lesion detection performance					
Calcification	0.943	63.6%	95.6%	82.0%	89.0%
Mass	0.897	68.7%	86.8%	75.0%	83.0%
Others	0.776	16.9%	97.4%	78.6%	67.7%
b. Cancer detection performance					
Cancer	0.907	71.7%	87.8%	87.7%	72.0%

Table 5: Localization performance on Biopsy-confirmed Set.

Condition	#Lesions	LLF	NLF
a. Lesion detection performance			
Calcification	63	71.4	0.49
Mass	111	67.6	0.27
Others	23	30.4	0.48
b. Cancer detection performance			
Cancer	165	74.5	0.53

The model's localization performance is displayed in Table 5, demonstrating strong cancer detection capabilities, with cancer lesions achieving an LLF score of 74.5 and maintaining an NLF of 0.53, confirming reliable detection with minimal over highlights. Among specific lesion types, calcifications showed particularly strong performance with an LLF exceeding 70, while the model maintained favorable metrics across all categories with NLF values consistently below 0.5.

Notably, the "Others" lesion category showed a substantially lower LLF score of 30.4, indicating challenges in precisely localizing these more subtle or complex lesion types. Our analysis reveals that architectural distortions, which belong to this category, are often mislocalized and misclassified as masses, suggesting radiologists should exercise additional caution when evaluating AI recommendations for these findings. However, the high cancer LLF score of 74.5 suggests that despite these challenges with specific lesion subtypes, the model still localizes cancer lesions correctly overall. Improving the differentiation between architectural distortions and masses represents a key direction for future model enhancements.

Overall, the model exhibited promising performance on the BDMS dataset for both classification and localization tasks, indicating its potential to generalize well to new, out-of-domain data. These balanced results between detection accuracy and precision suggest the model's utility in clinical settings.

Figures 1-3 illustrate the model's performance in lesion detection and localization, achieving high agreement with radiologist assessments. Additionally, we included examples where the model failed to detect certain abnormalities, emphasizing areas for potential improvement. These examples provided deeper insight into the model's clinical applicability and areas requiring refinement.

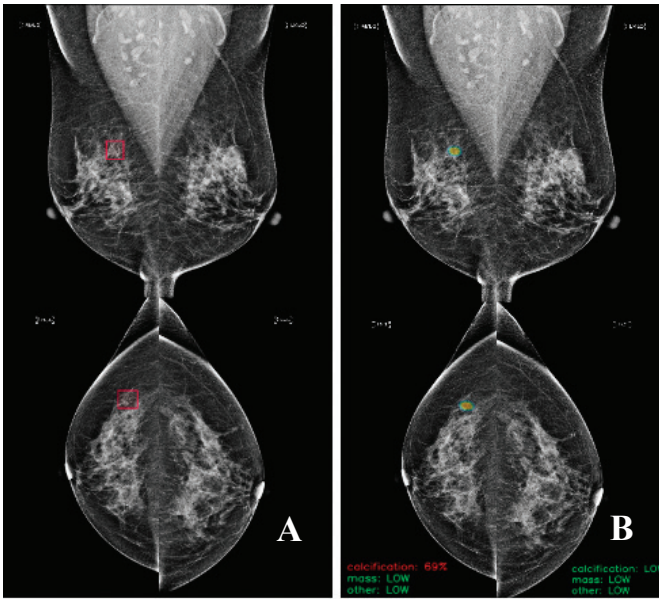


Figure 1: The AI's heatmap highlighting detected cluster of amorphous calcifications in the upper outer right breast (B) were consistent with the radiologist' assessment (A), demonstrating its high agreement in the lesion identification.

ISSN 2287-0237 (online)/ 2228-9674 (print)
(Free to Processing fees & Free to Publish & Free to Read) No hidden charges

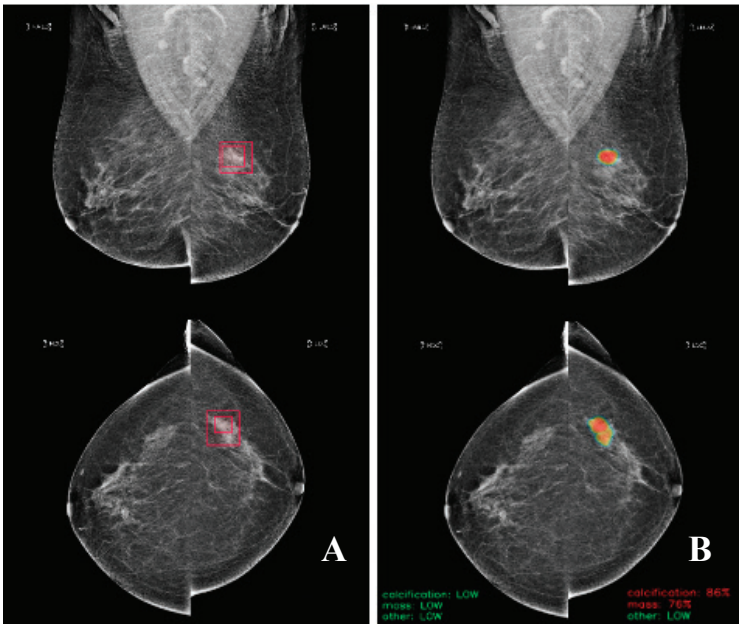


Figure 2: The AI's heatmap highlighting detected cluster of amorphous calcifications ovaerling an ill-defined mass in the upper outer left breast (B), aligned with the radiologist's interpretation (A) and wer future validated by the pathological report confirming invasive ductal carcinoma (IDC).

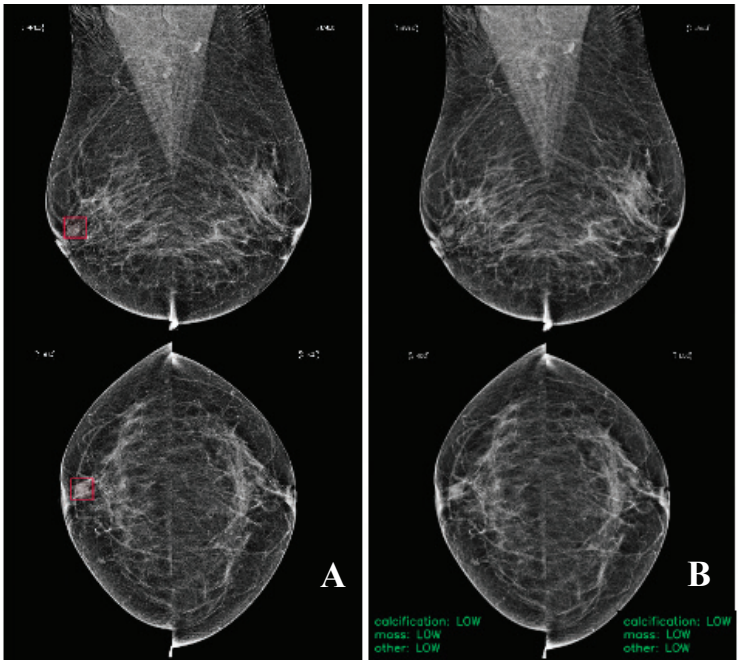


Figure 3: The radiologist's detected a 8mm nodule with internal round microcalcifications in the right suareolar region (A). The AI's heatmap (B) misses small size lesion in a heteregeneously dense breast, this highlighted an area for potential improvement in detecting certain abnormalities, particularly in dense breast tissue

Concordance Analysis Results

Table 6 shows the concordance result on the Biopsy-confirmed Set, which consisted of 181 participants. The 76.2% classification and 77.9% localization concordance rates exceed the >70% threshold typically considered acceptable for AI diagnostic tools, indicating that radiologists can expect AI findings to align with their assessments in approximately 3 out of 4 cases. This level of agreement reduces the likelihood of conflicting interpretations that could disrupt clinical workflow or undermine diagnostic confidence. The relatively low 'Reject' rates (23.8% classification, 22.1% localization) suggest that fundamental disagreements between AI and radiologists are infrequent, supporting the system's potential as a reliable diagnostic assistant. The higher proportion of 'Edit' cases compared to 'Add' cases indicates that AI tends toward over-detection rather than missing lesions, which is generally preferable from a patient safety perspective as it reduces the risk of overlooked abnormalities while allowing radiologists to refine AI recommendations based on their clinical expertise.

ISSN 2287-0237 (online)/ 2228-9674 (print)
(Free to Processing fees & Free to Publish & Free to Read) No hidden charges

Usefulness Assessment Results

As shown in Table 7, Radiologist 1 accepted 89 cases and rejected 10, while Radiologist 2 accepted 88 cases and rejected 11. When including the 82 cases automatically categorized as accepted, the overall acceptance rates were 94.5% (171/181) for Radiologist 1 and 94.0% (170/181) for Radiologist 2.

The high acceptance rates demonstrate that radiologists find AI outputs clinically valuable even in challenging diagnostic scenarios. The detailed scoring breakdown reveals that the majority of accepted cases (54 and 52 cases for each radiologist) received "useful" ratings (score 3), indicating that AI findings provided meaningful diagnostic support that aided clinical assessment. Additionally, 25 and 23 cases respectively received "neutral" ratings (score 2), suggesting that while these AI outputs didn't actively mislead, they provided limited additional clinical value. The presence of "extremely useful" ratings (10 and 13 cases) indicates that AI occasionally identified crucial diagnostic insights that might otherwise be missed, demonstrating potential for significant clinical impact.

The remarkably low rejection rates (5.5-6%) with only 10-11 cases receiving "not useful" ratings indicate that fundamental disagreements where AI results would mislead clinical decision-making are rare. This distribution suggests that when radiologists do find AI outputs unhelpful, it represents genuine diagnostic conflicts rather than systematic bias against AI assistance. The consistency between both radiologists' scoring patterns validates these findings, with both showing similar distributions across all rating categories. These results indicate that Inspectra MMG successfully generates outputs that experienced radiologists consider clinically relevant and trustworthy, with the majority providing active diagnostic support rather than mere confirmation of obvious findings.

System Usability Test Results

The resulting average SUS score was 69.33, suggesting that Inspectra MMG demonstrates acceptable usability and potential for integration into clinical workflows. This score falls just above the 68-point threshold typically considered acceptable for clinical software systems, indicating that radiologists found the interface learnable and efficient enough for practical use. However, the score also suggests substantial room for user experience improvements, particularly in areas such as interface intuitiveness, workflow streamlining, and reducing the learning curve for new users. Individual SUS responses revealed that while most radiologists felt confident using the system independently after training, some noted concerns about system complexity during initial adoption phases.

Second Reader Benefit Analysis Results

By comparing the predictions of Inspectra MMG with the radiologists' reports and comments, our study revealed

Table 6. Concordance between radiologists' reports and Inspectra MMG predictions on the Biopsy-confirmed Set (n = 181). *Overall concordance rate is defined as the sum of Agree, Edit, and Add categories.*

Concordance category	Classification		Localization	
	# cases	%	# cases	%
AGREE	92	50.8	96	53.0
EDIT	39	21.6	36	19.9
ADD	7	3.9	9	5.0
REJECT	43	23.8	40	22.1
Overall concordance	138	76.2	141	77.9

Table 7. Radiologists' usefulness scores for the Biopsy-confirmed Set (n = 181). *'Automatic accept' cases represent those with complete concordance in both detection and localization.*

Score	Accept / Not Accept	Radiologist 1	Radiologist 2
1 (Not useful)	Not Accept	10	11
2 (Neutral)	Accept	25	23
3 (Useful)	Accept	54	52
4 (Extremely useful)	Accept	10	13
Automatic accept	Accept	82	82
Acceptance rate	–	94.5%	94.0%

significant findings that in The Major Disagreement, the second radiologist identified lesions missed by the first in 11 cases; notably, the AI accurately detected 7 of these lesions (63.6%) and identified additional lesions in 13.8% of cases. In the minor group, the AI's findings concorded with the final reports in 71.1% of instances while uncovering extra lesions in 17.8% of cases. Overall, the AI demonstrated agreement with final radiological assessments in 71.4% of cases and revealed additional potential lesions in 16.8%. These results underscore the AI's potential as an effective tertiary reader, enhancing diagnostic consistency and mitigating inter-radiologist discrepancies in mammographic interpretation.

To further illustrate the impact of Inspectra MMG in reducing radiologist disagreements, we conducted a detailed case analysis, highlighting instances where the AI successfully aligned with the second reader's findings or identified additional lesions. Figure 4 provides visual representations of Inspectra MMG's role in bridging discrepancies. The model accurately flagged a calcification that was identified by the second reader and confirmed in the final radiology report.

In addition to confirming findings already present in the final radiology reports, Inspectra MMG also detected certain abnormalities that were not noted by both readers. Figure 5 shows a case in which the model detected a benign mass in the right breast, which was not reported by either radiologist in the mammographic evaluation. These findings suggest a potential for AI-assisted mammography to expand its role in broader breast abnormality detection, which may warrant further study.

ISSN 2287-0237 (online)/ 2228-9674 (print)
(Free to Processing fees & Free to Publish & Free to Read) No hidden charges

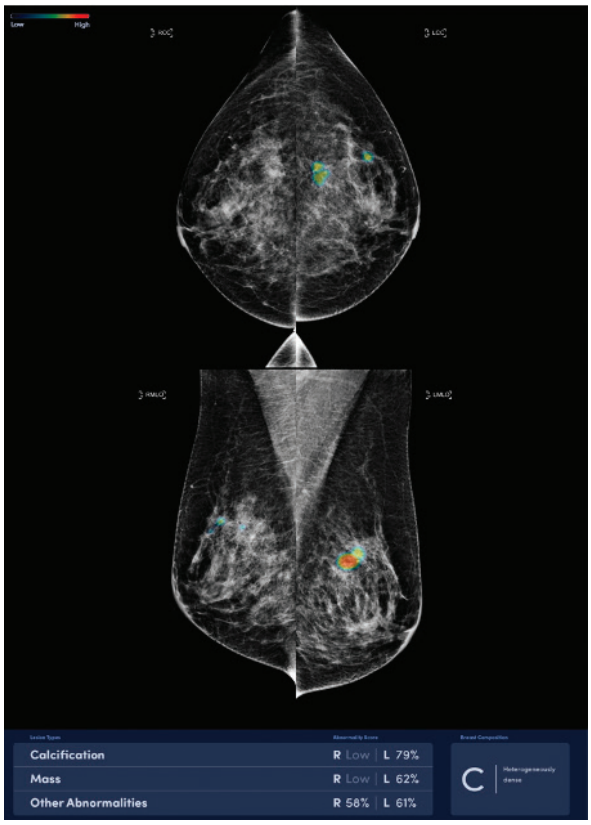


Figure 4: In this case from the Major Disagreement group, the model accurately flagged a calcification that was identified by the second reader and confirmed in the final radiology report.

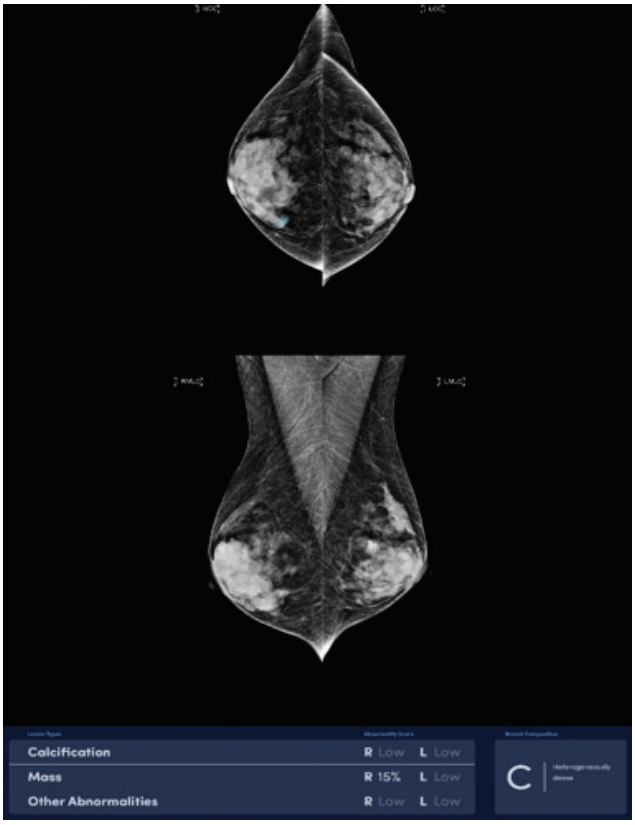


Figure 5: A case in which the model detected a benign mass in the right breast, which was not reported by either radiologist in the mammographic evaluation.

The findings of this study highlighted the potential of Inspectra MMG as an AI-aided second reader in mammographic interpretation, particularly in cases where radiologists have discrepant assessments. In the Major Disagreement group, the AI demonstrated a strong ability to detect significant lesions, correctly flagging masses, calcifications, adenopathy, and architectural distortions that were confirmed by the second radiologist. In 63.6% of cases, it successfully enhanced diagnostic consistency by potentially reducing radiologist disagreements. Moreover, in a subset of cases, the AI identified additional findings that were not initially reported, suggesting its capacity to improve lesion detection sensitivity. In the Minor Disagreement group, Inspectra MMG aligned with final reports in 71.1% of cases, demonstrating its reliability in standard interpretations while also identifying subtle additional findings in 17.8% of cases. Notably, the AI also flagged abnormalities outside its primary scope, such as lesions later confirmed via ultrasound or asymmetry that may warrant further evaluation, raising questions about its potential broader clinical utility. However, Inspectra MMG has demonstrated notable accuracy in detecting significant lesions, but it is occasionally unable to identify subtle or stable (benign) abnormalities. These cases highlight the AI's tendency to prioritize clinically significant lesions while overlooking findings that are either stable or deemed non-urgent by radiologists. This selective detection approach may be advantageous in reducing unnecessary recalls but also underscores the need for further AI refinement.

These results underscore the value of AI in supporting radiologists by improving consistency in mammographic assessments, reducing inter-reader variability, and potentially catching overlooked abnormalities. Despite its promising performance, Inspectra MMG should remain a supportive tool rather than a replacement for human expertise, serving as an adjunct to refine decision-making and improve patient outcomes.

Discussion

Our findings of consistent cancer detection performance (AUROCs 0.915 and 0.907) and effective localization capabilities (cancer LLF of 74.5, NLF of 0.53) across independent datasets align with previous studies demonstrating that deep learning models can maintain diagnostic accuracy across different clinical environments.[6] This consistency in both classification and localization performance suggests that AI-assisted mammography could be reliably deployed across diverse healthcare settings without requiring extensive retraining or recalibration, addressing practical implementation concerns for healthcare systems.[12] The balance between accurate lesion

ISSN 2287-0237 (online)/ 2228-9674 (print)
(Free to Processing fees & Free to Publish & Free to Read) No hidden charges

detection and minimal false-positive highlights is crucial for clinical utility, as it helps radiologists focus on truly suspicious areas without being overwhelmed by excessive false alarms.

The concordance rates achieved in this study (76.2% for classification, 77.9% for localization) are comparable to those reported in AI-assisted diagnostic imaging studies[15], suggesting that AI-radiologist concordance is approaching the level of human expert agreement. This finding supports AI's potential role as a reliable second opinion rather than merely a screening tool.

The high radiologist acceptance rates (94.5% and 94.0%) demonstrate strong clinical confidence in the AI system's outputs, with most accepted cases rated as "useful" for clinical assessment and a subset achieving "extremely useful" scores for providing critical diagnostic insights. This scoring pattern indicates that Inspectra MMG consistently generates clinically relevant outputs that actively aid diagnostic decision-making rather than simply confirming obvious findings.

The SUS score of 69.33 indicates good usability, exceeding the threshold of 68 that suggests acceptable system usability.[17] This score reflects the system's potential for successful integration into clinical workflows, though continued refinement of the user interface could further enhance user experience.

The second reader benefit analysis demonstrates significant potential for reducing inter-radiologist variability, addressing a well-documented challenge in mammographic interpretation. Previous studies have shown that AI can serve as an effective diagnostic support tool, improving radiologists' performance and serving as an arbitrator in diagnostic disagreements[6,7], which aligns with our findings that the AI aligned with final radiological assessments in 71.4% of disagreement cases.
Several limitations should be acknowledged in this study. The validation datasets were derived from a single medical institution (BDMS), which may introduce selection bias and limit generalizability to other healthcare systems or populations. The sample sizes, while adequate for initial validation, may not be sufficient to detect subtle performance differences across all lesion subtypes, particularly for the 'Others' category with only 14 samples. The study population was exclusively Thai, and further validation across diverse ethnic groups and imaging protocols is needed. Additionally, the retrospective design of the second reader benefit analysis may not fully capture the dynamic decision-making process in real-time clinical practice. The observed challenges in localizing architectural distortions reflect a known limitation in current deep learning approaches for mammography, underscoring the ongoing need for architectural improvements in AI models to better differentiate subtle lesion types, particularly in dense breast tissue environments common in Asian populations.

Conclusion

The results demonstrate Inspectra MMG's robust cancer detection performance (AUROCs 0.915 and 0.907) and effective lesion localization (LLF 74.5) across independent clinical datasets. High concordance rates with radiologist assessments (76.2% classification, 77.9% localization) and strong expert acceptance (94.0-94.5%) indicate reliable clinical utility. The system achieved acceptable usability (SUS score 69.33) and demonstrated value as a second reader by aligning with final radiological assessments in 71.4% of disagreement cases while identifying additional findings in 16.8% of cases. These findings support Inspectra MMG's potential to enhance mammographic interpretation by reducing inter-reader variability and improving detection sensitivity in clinical breast cancer screening workflows.

Ethics approval

This study complied with the Declaration of Helsinki and was approved by the institutional review board from the Bangkok Hospital Headquarters Institutional Review Board (BHQ-IRB2024-05-21). Informed consent was waived due to the retrospective nature of the study.

References

1. The International Agency for Research on Cancer (IARC), World Health Organization (WHO). Breast Cancer. (Accessed on July 15, 2024, at https://www.iarc.who.int/cancer-type/breast-cancer/).
2. Martinez RG, van Dongen DM. (2023). Deep learning algorithms for the early detection of breast cancer: A comparative study with traditional machine learning. *Inf Med Unlock* 2023;41:101317. doi: 10.1016/j.imu.2023.101317.
3. Lång K, Dustler M, Dahlblom V, et al. 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. *Lancet Oncol* 2023;24(8):936-44. doi:10.1016/S1470-2045(23)00298-X.
4. Schaffter T, Buist DSM, Lee CI, et al. Nationwide real-world implementation of AI for cancer detection in population-based mammography screening. *Nat Med* 2024;30(12):3501-10. doi:10.1038/s41591-024-03408-6.
5. Kim HE, Yoon JH, Kim EK, et al. Artificial intelligence for breast cancer screening in mammography (AI-STREAM): preliminary analysis of a prospective multicenter cohort study. *Nat Commun* 2024;15:3224. doi:10.1038/s41467-025-57469-3.
6. Kim HE, Kim HH, Han BK, et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *Lancet Digit Health* 2020;2(3):e138-48. doi: 10.1016/S2589-7500(20)30003-0.
7. Schaffter T, Buist DSM, Lee CI, et al. Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms. *JAMA Netw Open* 2020;3(3):e200265. doi:10.1001/jamanetworkopen.2020.0265

8. Freeman K, Geppert J, Stinton C, et al. Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy. *BMJ*. 2021;374:n1872. doi:10.1136/bmj.n1872.
9. Rodriguez-Ruiz A, Krupinski E, Mordang JJ, et al. Detection of breast cancer with mammography: effect of an artificial intelligence support system. *Radiology* 2019;290(2):305-14. doi:10.1148/radiol.2018181371
10. BusinessWire. Zebra Medical Vision Secures its first FDA clearance in oncology, boosting early detection of breast cancer in mammograms [Online], Jul 27, 2020.(Accessed on July 15, 2024, at https://www.businesswire.com/news/home/20200727005442/en/Zebra-Medical-Vision-Secures-its-First-FDA-Clearance-in-Oncology-Boosting-Early-Detection-of-Breast-Cancer-in-Mammograms)
11. Lunit. Lunit INSIGHT MMG, an AI Solution for Breast Cancer Detection, Now CE Certified [Online], Jun 2,2020. (Accessed on July 15, 2024, at https://www.lunit.io/en/company/news/lunit-insight-mmg-an-ai-solution-for-breast-cancer-detection-now-ce-certified).
12. Guan S, Loew M. Analysis of generalizability of deep neural networks based on the complexity of decision boundary. In 2020 19th IEEE international conference on machine learning and applications (ICMLA). 2020:101-6. doi: 10.1109/ICMLA51294.2020.00025.
13. Tan M, Le QV. Efficientnetv2: Smaller models and faster training. 2021. doi: 10.48550/arXiv.2104.00298.
14. Chakraborty DP, Yoon HJ. Operating characteristics predicted by models for diagnostic tasks involving lesion localization. *Med Phys* 2008;35(2):435-45. doi: 10.1118/1.2820902.
15. Kim EY, Kim YJ, Choi WJ, et al. Concordance rate of radiologists and a commercialized deep-learning solution for chest X-ray: Real-world experience with a multicenter health screening cohort. *PLoS One* 2022;17(2):e0264383. doi: 10.1371/journal.pone.0264383.
16. Gaube S, Suresh H, Raue M, et al. Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ Digit Med* 2021;4(1):31. doi: 10.1038/s41746-021-00385-9.
17. Brooke J. SUS: a retrospective. *J Usabi Stud* 2013;8:29-40

ISSN 2287-0237 (online)/ 2228-9674 (print)
(Free to Processing fees & Free to Publish & Free to Read) No hidden charges