ERIA Research Project Report FY2025, No. 18

Enhancing Breast Cancer Detection with FxMammo: Results from a Multi-Experience AI Evaluation

Economic Research Institute for ASEAN and East Asia

Enhancing Breast Cancer Detection with FxMammo: Results from a Multi-Experience AI Evaluation

Economic Research Institute for ASEAN and East Asia (ERIA)
Sentral Senayan II 6th Floor
Jalan Asia Afrika No. 8, Gelora Bung Karno
Senayan, Jakarta Pusat 12710
Indonesia

ERIA Research Project Report FY2025, No. 18
Published in August 2025

The report is prepared for the Economic Research Institute for ASEAN and East Asia (ERIA) by FathomX and supported by Universitas Gadjah Mada, Department of Radiology, Faculty of Medicine, Public Health, and Nursing.

List of Project Members

FathomX

Stephen Lim

Galvin Lian

Du Hao

Mikael Hartman

Feng Mengling

Abigail Chu

Universitas Gadjah Mada, Faculty of Medicine, Public Health, and Nursing

Lina Choridah (Department Radiology)

Devina Yudistiarta (Department of Radiology)

Vincent Laiman (Department of Radiology)

Didik Setyo Hariyanto (Department of Anatomical Pathology)

Ika Puspitasari (Department of Pharmacy/UGM Academic Hospital)

Sariningsih Hikmawati (Student, Doctoral Program of Medical and Health Sciences)

Rozan Muhammad Irfan (Radiology Research and Training Office)

Zannuba Arifah Noor (Radiology Research and Training Office)

Contents

List of Tables

Glossary

Term	Definition
AI (Artificial Intelligence)	Simulation of human intelligence by machines. In this study, AI assists in interpreting mammograms for breast cancer detection.
AUC (Area Under the Curve)	Performance metric of classification models; higher AUC indicates better diagnostic accuracy.
BI-RADS	Breast Imaging Reporting and Data System used to classify mammogram findings and guide follow-up.
Blinded Reading Trial	Readers interpret images without knowing the diagnosis to prevent bias.
CAD (Computer-Aided Detection)	Software aiding radiologists in identifying abnormalities on medical images. FxMammo is an advanced CAD.
Confidence Interval (CI)	Statistical range indicating the reliability of an estimate, e.g. 95% CI.
Cohen's Kappa Score	Statistic measuring agreement between raters, accounting for chance.
Dense Breast Tissue	Breast tissue with more fibroglandular elements that obscure tumors on mammograms.
Diagnostic Accuracy	Test's ability to correctly identify presence and absence of disease.
Digital Mammography	Electronic capture and storage of breast images for analysis.
False Positive	A result incorrectly indicating presence of disease.
False Negative	A result incorrectly indicating absence of disease.

Term	Definition
FxMammo	FathomX's deep learning AI system for scoring malignancy likelihood in mammograms.
Inter-reader Variability	Differences in diagnostic interpretation amongst readers.
McNemar Test	Statistical test comparing paired proportions (e.g. with and without AI).
MRMC (Multi-Reader Multi-Case) Study	Study with multiple readers and cases evaluating diagnostic tools.
NPV (Negative Predictive Value)	Probability that a negative test correctly indicates absence of disease.
PACS	Picture Archiving and Communication System for storing and sharing medical images.
PPV (Positive Predictive Value)	Probability that a positive test correctly indicates presence of disease.
Sensitivity	Ability of a test to correctly identify individuals with the disease.
Specificity	Ability of a test to correctly identify individuals without the disease.
Triage Tool	Tool to prioritise high-risk cases for review, used in AI-assisted diagnosis.
Malignant	Describes cells or tumors that are cancerous, invasive, and capable of spreading to other parts of the body.
Benign	Describes cells or tumors that are non-cancerous and typically do not spread to surrounding tissue or other parts of the body.

Chapter 1
Introduction

1.1. Background

Breast cancer is a leading cause of cancer mortality worldwide, and early detection through screening mammography is critical for improving outcomes (Elhakim et al., 2024; Shi et al., 2025). Mammography's effectiveness, however, depends on its diagnostic accuracy, which can vary by region and population. A meta-analysis of studies from North America and Europe reported that digital mammography achieves a pooled sensitivity of about 76% and specificity of 94%–97% (Shi et al., 2025). In practice, performance ranges widely: for example, an Indonesian hospital observed approximately 90.1% sensitivity and 93.6% specificity for mammography, whereas a study in Pakistan found sensitivity as high as 97% but a specificity of only about 64.5% (Lehman et al., 2015). Such variability arises from multiple factors, including differences in technology, patient populations, and breast composition. Notably, breast density significantly affects mammographic sensitivity – dense breast tissue can mask tumors, reducing sensitivity to around 62%–68% in very dense breasts compared to 86%–89% in predominantly fatty breasts (Carney et al., 2003; Kerlikowske and Phipps, 2011). Dense breasts are more common in younger women and certain ethnic groups (including many Asian populations), partly explaining regional differences in detection rates (del Carmen et al., 2007).

In addition to patient factors, there is marked variability in interpretive performance amongst radiologists. Studies have documented wide disparities in sensitivity even amongst radiologists working under similar conditions (Elmore et al., 2009). This inter-reader variability means that some cancers go undetected (missed cancers) while some patients without cancer undergo unnecessary recall and anxiety due to false-positive readings (Elmore et al., 2009). The issue is exacerbated in countries like Indonesia, which face a shortage of radiologists, especially those specialising in breast imaging. With limited expert manpower, many screening mammograms may be interpreted by general radiologists or trainees, potentially increasing variability and diagnostic errors. These challenges highlight an urgent need for innovative solutions to support radiologists, improve consistency, and maintain high accuracy in breast cancer detection.

Artificial intelligence (AI) has emerged as a promising adjunct in mammographic screening and diagnosis. Modern AI systems, powered by deep learning, can be trained on vast collections of mammograms to recognise patterns of malignancy. In recent studies, AI algorithms have achieved diagnostic performance on par with, or even exceeding that of human radiologists in retrospective settings (Kim et al., 2020). For instance, one AI system evaluated on mammography screenings yielded an area under the curve (AUC) of 0.94 for cancer detection – significantly higher than the 0.81 AUC of

unaided radiologists – and improved radiologists' performance when used as a second reader (Kim et al., 2020). Such results underscore the potential of AI to act as a 'second pair of eyes,' catching subtle cancers that a human might overlook and reducing variability between readers. Notably, variability in human interpretation can be addressed by AI's more consistent analysis: unlike humans, a validated AI algorithm will apply the same criteria to every case, which may help standardise readings across different practitioners.

Recent AI systems using deep learning have shown substantial improvements over legacy computer-aided detection (CAD) tools. Several contemporary studies demonstrate that AI assistance can boost radiologists' diagnostic performance. Kim et al., reported that using an AI system alongside radiologists increased cancer detection sensitivity by about 9.5% and specificity by 2.7% compared to readings without AI (Kim et al., 2020). However, not all trials have found a statistically significant impact on performance; some have observed that seasoned radiologists' metrics (sensitivity, specificity) remain similar with or without AI support (Pacilè et al., 2020; Dang et al., 2022). These mixed findings suggest that AI's benefit may depend on the context – factors such as the difficulty of cases, the experience of the readers, and the specific AI algorithm's capabilities all influence outcomes. Consequently, rigorous evaluation in diverse settings is necessary to ascertain where AI can provide the most value in breast imaging.

Beyond accuracy, AI tools in mammography offer potential workflow and efficiency benefits. Radiologists typically must carefully scrutinise each mammogram for subtle signs of cancer (e.g. tiny calcifications or faint distortions) – a time-consuming task prone to human fatigue. AI can automate the detection of obvious normal cases and flag suspicious regions, thus streamlining the reading process. Studies have noted that AI support can reduce the time required to search for subtle abnormalities, like microcalcifications, by guiding the radiologist's attention (Lehman et al., 2015). By integrating AI, one simulation projected that radiologists' workload could be reduced by over 50% without compromising diagnostic accuracy (Dembrower et al., 2020). In a large retrospective analysis of a national screening cohort, researchers found that replacing one of two human readers with AI in a double-reading programme could cut the reading volume nearly in half while maintaining cancer detection rates (Elha et al., 2024). Furthermore, using AI as an autonomous triage tool – where the AI clears obviously normal exams and forwards only doubtful or high-risk cases to radiologists – achieved almost 50% workload reduction and even slightly improved cancer detection, compared to standard double reading (Elhakim et al., 2024). These efficiencies are especially relevant for healthcare systems facing high screening volumes and workforce shortages. If radiologists can focus their expertise where it's most needed (on complex cases), and let AI handle the straightforward ones, the overall screening programme can run more effectively.

However, integrating AI into clinical practice is not without challenges. One concern is the potential over-reliance on AI or automation bias. Radiologists might trust an AI's judgment too much – for example, if the AI fails to mark a cancerous lesion, a human reader could

be falsely reassured and also miss it. A recent multi-reader study in a different imaging context (chest radiography) found that when an AI system provided incorrect outputs, radiologists were more likely to make errors that they would not have made on their own (Bernstein, Atalay et al., 2023). In other words, an inaccurate AI suggestion can mislead even experienced clinicians, underscoring the need for users to remain vigilant and not defer blindly to AI. Ensuring that radiologists are trained to interpret AI results and retain their critical judgment is crucial. There are also practical considerations: AI algorithms require robust validation on local patient data to ensure their accuracy is generalisable across different populations and imaging equipment. Privacy and data security must be managed when integrating AI software that often relies on large datasets. Despite these challenges, the potential benefits of AI – improved detection, more consistent interpretations, and streamlined workflow – make it a compelling area of research and implementation in breast cancer screening. This study builds on that context by evaluating the effects of the FxMammo AI system (developed by FathomX Pte Ltd) on mammography interpretation in a real-world clinical setting with readers of varying experience. FathomX's solution is designed to reduce false negatives and false positives and to expedite case reading, which could be particularly advantageous in environments with limited expert radiologists. By examining its impact in a controlled study, we aim to provide evidence on how such AI can be best utilised to enhance breast cancer detection in practice.

1.2. Research Objective

The primary objective of this study is to evaluate the diagnostic impact of FxMammo, a deep learning-based decision support system developed by FathomX, on breast cancer detection in Indonesia. Specifically, the research aims to determine whether the integration of FxMammo can enhance radiologists' diagnostic performance in terms of sensitivity, specificity, and overall accuracy when interpreting mammograms. This includes assessing how FxMammo influences inter-reader variability and diagnostic consistency, particularly in challenging cases such as dense breast tissue and amongst less experienced readers. The study further aims to investigate the potential of AI assistance to serve as a second reader, effectively supplementing the limited availability of specialised radiologists in resource-constrained settings. By doing so, the research provides empirical evidence to inform how AI systems can be optimally deployed to improve breast cancer screening outcomes in real-world clinical environments.

1.3. Research Signification

The significance of this study lies in several key perspectives. First, breast cancer remains one of the leading causes of cancer-related deaths in Indonesia, where delayed diagnosis is a persistent problem due to limited screening infrastructure and a shortage of

specialised radiologists. By demonstrating the practical utility of AI tools like FxMammo in a real-world Indonesian setting, this research contributes to the broader conversation on digital health equity and capacity building in low-to-middle-income countries (LMICs).

Second, the findings offer actionable insights for health systems facing similar challenges across Southeast Asia and other developing regions. Unlike many previous studies conducted in high-resource settings, this study focuses on a population with a higher prevalence of dense breast tissue and fewer trained mammographers – factors that compound diagnostic complexity. This enhances the external relevance of the findings for comparable healthcare environments.

Finally, the research adds value to the growing body of literature on the clinical integration of AI in radiology. It not only quantifies the diagnostic gains from AI assistance but also discusses human-AI collaboration dynamics, such as changes in inter-reader agreement and the risk of automation bias. Thus, the significance of this work lies in both its clinical applicability and its contribution to responsible AI adoption in healthcare.

1.4. Scope and Limitation

This study focuses on the evaluation of FxMammo's diagnostic utility in a single-center setting at Universitas Gadjah Mada, Indonesia. The study population consists of 500 retrospectively collected digital mammography cases – 250 confirmed malignant and 250 benign or normal – representative of real-world clinical distribution in a tertiary hospital. The readers include both board-certified radiologists and senior radiology residents, offering a range of interpretive expertise. The AI system was assessed under controlled, blinded conditions, allowing for robust comparison of performance with and without AI support.

Several limitations should be acknowledged. First, the study is geographically limited to Indonesia, which may constrain generalisability to other ASEAN or global populations. Although the study highlights the relevance of AI in resource-constrained settings, comparative data on radiologist availability across ASEAN countries could further strengthen the rationale for site selection.

Second, the dataset excludes specific subgroups such as patients with prior breast cancer, those with breast implants, and mammograms of suboptimal quality. These exclusions, while necessary to ensure diagnostic clarity, limit the applicability of findings to more complex or post-treatment cases.

Lastly, while the AI system was evaluated in a simulated screening environment, Indonesia currently lacks a national mammography screening programme. Thus, the performance outcomes may differ in an organised screening setting. The absence of long-term follow-up data also restricts conclusions about interval cancers or long-term impact on patient outcomes.

Chapter 2
Research Methodology

2.1. Research Design

The research team carried out a single-centre, cross-sectional ,multi-reader, multi-case (MRMC) study to assess the effect of AI assistance on mammography interpretation. An MRMC design is well-suited for comparing diagnostic modalities or aids because it involves multiple readers evaluating multiple cases under different conditions (here, with vs. without AI) and allows for robust statistical comparison of performance metrics. In this study, each participating reader interpreted a series of mammographic cases twice: once unaided (relying on their own expertise alone) and once with the support of the FxMammo AI system. All readings were done under blinded conditions with respect to patient outcome; readers did not know the true diagnosis or the proportion of cancer cases, and when reading with AI, they only had the AI's output for that case without feedback about correctness.

2.2. Data Collection

The study was conducted at Universitas Gadjah Mada in Yogyakarta, Indonesia, which is a tertiary referral centre with a Picture Archiving and Communication System (PACS) archiving digital mammograms. We retrospectively collected mammography cases from the PACS database spanning the years 2019 to 2024. From this database, a total of 500 mammographic cases were sampled for the study. We aimed for an even balance of malignant and benign/normal cases to adequately test sensitivity and specificity; thus, the sample included 250 cases with a confirmed breast malignancy and 250 cases that were either normal or benign findings. Each case consisted of the standard four-view mammography series (left and right breast, craniocaudal and mediolateral-oblique views), as is routine for full diagnostic mammographic evaluation. Inclusion criteria required that cases be from women aged 40 and above (the typical screening age range), and that the mammographic study was complete (all four standard views present). Cases with biopsy-proven malignant findings were included as 'malignant' ground truth, and cases deemed normal or benign had to have concordant follow-up evidence (either a negative two-year follow-up or confirmation by ultrasound and/or expert consensus reading as benign). Specifically, for benign/normal cases, we required confirmation by at least two breast radiologists using standard BI-RADS 5th edition criteria and supplemental ultrasound when necessary. This rigorous confirmation was to ensure that our 'non-cancer' cases truly did not harbor malignancy, thereby avoiding false negatives in the study reference standard.

Exclusion criteria were applied to avoid confounding factors that might interfere with either the AI analysis or human reading. We excluded cases from patients with a prior history of breast cancer (before the mammogram date), because their images often contain post-surgical changes or markers that could bias a reader or the AI. Mammograms with any interventional devices visible (e.g. localisation wires or biopsy markers) were excluded, since these could obviously hint at the presence of a lesion and also might trigger AI false markings. We also removed cases with breast implants or other artifacts that substantially alter mammographic appearance. Lastly, very poor-quality mammograms (e.g. underexposed or blurred images) were excluded to ensure that both radiologists and the AI were operating on diagnostically adequate studies. After applying these criteria, we obtained the final set of 500 cases for analysis.

2.3. Research Informants – Radiologist Participants (Readers)

Six readers participated in the interpretation of the mammograms, comprising three early-career radiologists and three senior radiology residents who had completed the breast imaging component of their training. Each radiologist had a minimum of 3 years of post-residency experience in general radiology, with varying degrees of exposure to breast imaging. All readers had experience exclusively with diagnostic mammography, as the hospital does not currently operate a national mammographic screening programme. To avoid bias, none of the participants had substantial prior exposure to the FxMammo AI system.

The residents were in the final stage of their radiology training and had completed a dedicated breast imaging rotation or course, equipping them with baseline competency in mammogram interpretation, though less experience than the attending radiologists.

All readers were blinded to clinical information (e.g. patient history, prior imaging), to simulate a screening-like environment. They were instructed to interpret each case as they would in routine clinical practice. For each mammogram, they recorded a binary assessment – positive (suspicious for malignancy) or negative/benign – as well as an optional BI-RADS category. Sensitivity and specificity calculations were based on the binary assessments.

2.4. AI System (FxMammo)

The AI tool used, FxMammo, is a proprietary deep learning system designed to analyse mammographic images and provide decision support. It is a form of advanced CAD that generates a score indicating the likelihood of malignancy in the study. Before the reading sessions, the AI was installed and integrated with our viewing workstation such that readers could toggle the AI results on and off. In AI-assisted reads, the reader could see the AI's risk score on the mammograms. The FxMammo algorithm has been trained on a large dataset of mammograms and tuned to reduce false positives. For the purpose of our

study, readers were told that the AI was a support tool and that they should use their judgment in conjunction with the AI output. We did not enforce how they were to incorporate the AI suggestion – some might choose to trust AI on subtle findings, others might use it as a second opinion.

2.5. Statistical Analysis Method

We used several statistical approaches tailored to the MRMC study design. The primary endpoints were sensitivity (the proportion of malignant cases correctly identified as suspicious) and specificity (the proportion of benign cases correctly identified as non-suspicious) for each reader under each condition (with AI vs. without AI). These were first calculated for each reader and condition. To compare performance, we employed the McNemar test for paired proportions to test whether sensitivity with AI was significantly different from sensitivity without AI (and similarly for specificity) for the pooled readers. McNemar's test is appropriate for comparing correlated binary outcomes (each case was read twice by the same reader). We also report the average sensitivity and specificity across readers in each condition, with 95% confidence intervals, and the differences in these averages. We set a significance threshold of $p < 0.05$ for all hypothesis tests.

We also examined secondary metrics such as the inter-reader agreement: we calculated Cohen's kappa for each pair of readers in each condition to see if AI assistance led to more concordant readings amongst different readers (the hypothesis being that AI might guide everyone to notice the same lesions, thus increasing agreement). Data analysis was performed using Python and R software for specialised analyses. All statistical tests were two-tailed. The study was approved by the institutional ethics review board, and being a retrospective analysis of anonymised imaging data, informed consent was waived.

Chapter 3
Research Findings

3.1. Study Participants and Case Characteristics:

All six readers completed the reading sessions with no missing data. Each radiologist-resident pair reviewed an identical case set, and the randomised crossover ensured balanced conditions. The 500 selected cases had a mean patient age of 52.09 years (IQR: 45–58 years). By design, 50% of cases (n=250) had malignancies, which included a mix of masses and microcalcification-dominant lesions spanning various sizes and breast density categories. Amongst the 250 malignant cases, the most common histopathological diagnosis was invasive carcinoma (228 out of 250, 91.2%). The 250 benign/normal cases included a variety of findings such as benign calcifications, cysts, fibroadenomas, and a substantial subset (approximately half) that were completely normal studies (BI-RADS 1 on initial assessment). Breast density distribution (by reference standard) in the case set was: 6.4% BI-RADS A (almost entirely fatty), 27.0% BI-RADS B (scattered fibroglandular densities), 58.0% BI-RADS C (heterogeneously dense), and 8.6% BI-RADS D (extremely dense). The younger median age of breast cancer patients (52 years) compared to breast cancer in America and Europe, as well as the high proportion of dense breasts (C: 58%, D: 8.6%) causes an increased level of difficulty in interpreting mammography. This condition also causes AI to be needed to find hidden lesions due to dense breasts and increase the accuracy of mammography readings.

Table 3.1. Reader Diagnostic Performance with and without AI Assistance

Reader	Without AI					With AI				
	Sensitivity	Specificity	PPV	NPV	Accuracy	Sensitivity	Specificity	PPV	NPV	Accuracy
JR1	0.640	0.992	0.988	0.734	0.816	0.604	0.992	0.987	0.715	0.798
JR2	0.876	0.740	0.771	0.856	0.808	0.832	0.856	0.852	0.836	0.844
JR3	0.776	0.984	0.980	0.815	0.880	0.740	0.992	0.989	0.792	0.866
SR1	0.596	0.976	0.961	0.707	0.786	0.776	0.900	0.886	0.801	0.838
SR2	0.708	0.992	0.989	0.773	0.850	0.760	0.992	0.99	0.805	0.876
SR3	0.804	0.868	0.859	0.816	0.836	0.876	0.924	0.92	0.882	0.900
JR Group	**0.764**	**0.905**	**0.913**	**0.802**	**0.835**	**0.725**	**0.947**	**0.943**	**0.781**	**0.836**
SR Group	**0.703**	**0.945**	**0.936**	**0.765**	**0.824**	**0.804**	**0.939**	**0.932**	**0.829**	**0.871**

* JR – Junior radiologist; SR: Senior residents.
Source: Authors' data (2025).

3.2. Impact of AI on Cancer Detection

The introduction of AI assistance led to a clear improvement in diagnostic performance for both radiologists and residents. Overall accuracy in classifying cases as malignant or benign increased significantly for all readers with the help of the AI. For the three attending radiologists, the mean accuracy (fraction of cases correctly diagnosed) rose from the unaided readings to the AI-assisted readings by a notable margin. When averaged across the radiologist group, accuracy improved from approximately 82.9% without AI to about 85.4% with AI. For the senior residents, who initially had slightly lower performance, the improvement was also striking: their average accuracy increased from around 82.4% unaided to 87.1% with AI support.

Statistical analysis confirmed that these gains were highly significant. Using the McNemar test across the pooled decisions of each group, we found $p < 0.001$ for the difference in accuracy between unaided and AI-assisted reading for the resident group. Notably, this improvement did not come at the expense of sensitivity or specificity. With AI support, readers were able to either maintain or improve sensitivity and specificity concurrently. The radiologists, for example, achieved a sensitivity very similar to their unaided sensitivity (indeed, if a radiologist missed a cancer on unaided reading, the AI often alerted them to it, boosting sensitivity, and none of the radiologists experienced a drop in cancer detection with AI). Their specificity was improved, as the AI helped reduce some false-positive interpretations without causing over-calling of benign findings.

A similar pattern was observed with the residents: AI assistance enabled them to catch substantially more of the cancers they initially missed, thereby raising sensitivity, while also helping them avoid some incorrect 'cancer' calls on benign cases (maintaining specificity). The net effect for both groups was a significant increase in the proportion of correct diagnoses overall. Importantly, no case of cancer that was correctly identified without AI was missed with AI; if anything, the AI led to additional cancers being detected that would have otherwise been overlooked. The improvement in accuracy underscores the robust benefit of the AI – both experienced radiologists and trainees benefited from the decision support, albeit the residents showed a slightly larger absolute jump in performance due to their lower baseline.

Table 3.2. The Inter-reader Agreement of Readers' Diagnostic Performance with and without AI Assistance

Reader	Cohen's Kappa Score*	
	Without AI	AI Assisted
JR1	0.632	0.596
JR2	0.616	0.688
JR3	0.767	0.732
SR1	0.572	0.676
SR2	0.700	0.752
SR3	0.672	0.800
JR Group	**0.642**	**0.672**
SR Group	**0.648**	**0.773**

Source: Authors' data (2025).

Inter-reader agreement was formally quantified with Cohen's kappa for every possible pair amongst the six readers under each reading condition. When working unaided the radiologist pairs demonstrated an average kappa value of 0.642 (commonly interpreted as substantial agreement), whereas resident readers demonstrated 0.648, at a higher level of substantial agreement. Once FxMammo assistance was introduced, kappa increased across both groups: radiologists rose to 0.672 and residents to 0.773. A permutation-based comparison of the paired kappa coefficients confirmed that these upward shifts were statistically significant. Because kappa corrects for chance agreement, the rise indicates that the readers converged on the same judgments more often because of AI guidance rather than coincidence.

Chapter 4

Policy Recommendation

The positive impact of AI assistance observed in this study suggests several policy and practice recommendations for healthcare administrators, professional bodies, and policymakers in the domain of breast cancer screening:

4.1. Integrate AI as a Second Reader in Screening Programmes

Health systems, especially those with high screening volumes or limited specialist workforce, should consider incorporating validated AI tools like FxMammo into their mammography interpretation workflow. As our findings and other studies indicate, AI can effectively serve as a second reader that catches additional cancers and reduces variability (Kim et al., 2020; Darmiati et al., 2023). For instance, national screening programmes could adopt a protocol where every mammogram is analysed by an AI algorithm in parallel with a human radiologist. Any case where the AI disagrees with the radiologist (either the AI finds a potential cancer the radiologist missed or vice versa) could be flagged for a consensus review or a second human opinion.

This model would emulate double-reading, which has known benefits, but with the AI taking on one of the reader roles. By doing so, countries with radiologist shortages could ensure that each mammogram effectively gets two readings (one human, one AI) without doubling the human resource requirement. Pilot implementations in parts of Europe have already tested replacing one human reader in a double-read system with AI, showing maintained accuracy and significantly reduced workload (Elhakim et al., 2024). Policymakers should allocate funding and develop infrastructure to gradually deploy such AI-assisted screening, starting perhaps with high-volume centers.

4.2. Training and Education

Introducing AI into clinical practice should be accompanied by comprehensive training for radiologists and radiology trainees. Residency programmes and continuing medical education courses need to include modules on AI in imaging. Radiologists should learn about the basics of how these algorithms work, their known failure modes, and best practices for using AI outputs in decision-making. This is akin to training pilots when new cockpit automation is introduced – users must know when to trust the system and when to be cautious.

Our findings showed that AI can both increase sensitivity and help reduce some false positives, but improper use could lead to over-reliance. Training should emphasise maintaining one's own interpretative skills and using AI as a tool for confirmation or a second opinion. Radiologists might review past cases with AI to see what they missed,

thereby learning and potentially improving their skills (the AI, in essence, could be a feedback mechanism). Institutions should also foster a culture where radiologists discuss AI findings openly, perhaps in regular meetings (e.g. 'missed-case conference' where AI-detected misses are reviewed, turning them into learning opportunities). This educational approach will help mitigate resistance to AI and improve user acceptance by demonstrating AI's value in a controlled manner.

4.3. Workflow Integration and Resource Planning

Healthcare administrators must plan for the practical aspects of AI integration. This includes ensuring IT infrastructure can handle AI software – high-performance servers or cloud services may be needed to run the algorithms quickly so as not to delay reading times. PACS vendors should be involved to seamlessly integrate AI results into radiologists' workstations (for example, overlaying AI annotations on images). Typically, the integration of AI can be carried out seamlessly without requiring additional processing time, as it leverages the existing CAD architecture. The analysis of mammogram data generally takes approximately 3 to 5 minutes. Workflow needs to be designed so that AI results are available at the right time; an ideal scenario is for AI to pre-reads the exams and be ready with results as the radiologist begins reading, so there is no time lost (Elhakim et al., 2024).

In terms of resource allocation, while AI software can be expensive, the cost might be offset by gains in efficiency. Policymakers could consider funding models or reimbursements for AI-supported readings, incentivising adoption. If AI allows one radiologist to do the work that previously required two readers, that could alleviate staffing bottlenecks. However, it should be noted that initial phases might require running AI and full double reading in parallel to build trust, effectively increasing workload in the short term. Planning should account for this transitional phase where AI is monitored. Furthermore, with AI taking on some workload, radiologists might be able to spend more time on complex cases or other duties, which is a beneficial redistribution of human resources.

4.4. Continuous Monitoring and Feedback

Once AI is integrated, there should be a system for ongoing monitoring of its performance and its interaction with radiologist performance. This could be a registry or regular audit where outcomes (cancers detected, interval cancers, false positives) are tracked and compared against historical benchmarks. If any drop in performance is noted, one should investigate whether the AI is causing issues or whether radiologists are perhaps over-relying on it.

Also, user feedback loops to AI developers are important; for example, our study identified a couple of scenarios where the AI made false predictions. Reporting such cases back to the developers can help them improve the algorithm in future versions. Policymakers

could encourage a framework where data from real-world use of AI (while protecting patient privacy) are aggregated to further refine AI tools – essentially creating health systems that make AI better over time.

Chapter 5
Conclusion

This study examined the impact of the AI system FxMammo on mammographic breast cancer detection and found that it significantly enhances diagnostic performance. In a controlled multi-reader, multi-case setting, AI use led to improved cancer detection sensitivity across all readers, with particularly notable gains in challenging cases such as dense breast tissue and among less-experienced readers. Crucially, these improvements were achieved without a substantial rise in false positives, suggesting that FxMammo improved the diagnostic signal without adding excessive noise.

The consistent performance gains across multiple readers also indicate that AI can help reduce inter-reader variability – a longstanding challenge in breast imaging. By providing analytical support and triaging capabilities, FxMammo addresses key issues in mammography, including interpretive inconsistency and heavy clinical workloads.

In conclusion, integrating AI systems like FxMammo into breast imaging workflows offers a tangible step forward in cancer diagnostics. It enables earlier and more accurate detection while maintaining specificity, helping overstretched healthcare systems extend the capabilities of existing radiology resources. For patients, this translates to a greater chance of timely cancer detection and fewer unnecessary follow-ups.

References

Bernstein, M.H. et al. (2023), 'Can incorrect artificial intelligence (AI) results impact radiologists, and if so, what can we do about it? A multi-reader pilot study of lung cancer detection with chest radiography', *European radiology*, 33(11), pp.8263–69.

Carney, P. A. et al. (2003), 'Individual and combined effects of age, breast density, and hormone replacement therapy use on the accuracy of screening mammography,' *Annals of Internal Medicine,* 138(3), pp.168–75.

Dang, L.-A. et al. (2022), 'Impact of artificial intelligence in breast cancer screening with mammography,' *Breast Cancer,* 29(6), pp.967–77.

Darmiati, S. et al. (2023), 'Impact of Artificial Intelligence on Mammography Interpretation by Breast Radiologists, Non-Breast Radiologists, and Senior Residents', *Indonesian Journal of Cancer*, 17(4), pp.327–37.

del Carmen, M.G. et al. (2007), 'Mammographic breast density and race', *American Journal of Roentgenology*, 188(4), pp.1147–50.

Dembrower, K. et al. (2020), 'Comparison of a deep learning risk score and standard mammographic density score for breast cancer risk prediction', *Radiology*, 294(2), pp.265–72.

Elhakim, M.T. et al. (2024), 'AI-integrated screening to replace double reading of mammograms: a population-wide accuracy and feasibility study', *Radiology: Artificial Intelligence,* 6(6), e230529.

Elmore, J.G. et al. (2009), 'Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy', *Radiology*, 253(3), pp.641–51.

Kerlikowske, K. and A.I. Phipps (2011), 'Breast density influences tumor subtypes and tumor aggressiveness', *Journal of the National Cancer Institute*, 103, pp.1143–45.

Kim, H.-E. et al. (2020), 'Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study', *The Lancet Digital Health,* 2(3), e138–e148.

Lehman, C.D. et al. (2015), 'Diagnostic accuracy of digital screening mammography with and without computer-aided detection', *JAMA Internal Medicine,* 175(11), pp.1828–37.

Pacilè, S. et al. (2020), 'Improving breast cancer detection accuracy of mammography with the concurrent use of an artificial intelligence tool', *Radiology: Artificial Intelligence,* 2(6), e190208.

Shi, J. et al. (2025), 'The screening value of mammography for breast cancer: an overview of 28 systematic reviews with evidence mapping', Journal of Cancer Research and Clinical Oncology, 151(3), pp.1–20.